SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 1–25

RETRIEVING LÉVY PROCESSES FROM OPTION PRICES: REGULARIZATION OF AN ILL-POSED INVERSE PROBLEM*

RAMA CONT† AND PETER TANKOV‡

Abstract. We propose a stable nonparametric method for constructing an option pricing model of exponential Lévy type, consistent with a given data set of option prices. After demonstrating the ill-posedness of the usual and least squares version of this inverse problem, we suggest to regularize the calibration problem by reformulating it as the problem of finding an exponential Lévy model that minimizes the sum of the pricing error and the relative entropy with respect to a prior exponential Lévy model. We prove the existence of solutions for the regularized problem and show that it yields solutions which are continuous with respect to the data, stable with respect to the choice of prior, and which converge to the minimum entropy least squares solution of the initial problem when the noise level in the data vanishes.

Key words. inverse problem, entropy, Lévy process, jump process, model calibration, option pricing, regularization, ill-posed problem

AMS subject classifications. 49N45, 60G51, 60J75, 91B70

DOI. 10.1137/040616267

1. Introduction.

The specification of an arbitrage-free option pricing model on a time horizon T_∞ involves the choice of a *risk-neutral* measure [25]: a probability measure Q on the set Ω of possible trajectories $\{S_t\}_{t\in[0,T_\infty]}$ of the underlying asset such that the discounted asset price $e^{-rt}S_t$ is a martingale (where r is the discount rate). Such a probability measure Q then specifies a pricing rule which attributes to an option with terminal payoff H_T at T the value $C(H_T) = e^{-rT}E^Q[H_T]$. For example, the value under the pricing rule Q of a call option with strike K and maturity T is given by $e^{-rT}E^Q[(S_T - K)^+]$. Given that data sets of option prices have become increasingly available, a common approach for selecting the pricing model is to choose, given option prices $(C(H^i))_{i\in I}$ with maturities T_i payoffs H^i, a risk-neutral measure Q compatible with the observed market prices, i.e., such that $C(H^i) = e^{-rT_i}E^Q[H^i]$. This *inverse problem* of determining a pricing model Q verifying these constraints is known as the "model calibration" problem. The number of observed options can be large ($\simeq$ 100–200 for index options) and the Black–Scholes model has to be replaced with models with richer structure such as nonlinear diffusion models [18] or models with jumps [13]. The inverse problem is ill-posed in these settings [14, 33] and various methods have been proposed for solving it in a stable manner, mostly in the framework of diffusion models [1, 4, 5, 6, 9, 16, 18, 26, 28, 32, 33].

Given the ill-posed nature of the inverse problem, an extra criterion must be used to select a model compatible with observed option prices. The use of *relative entropy* as a model selection criterion has solid theoretical foundations [17] and has been investigated by many authors in the context of option pricing.

*Received by the editors October 4, 2004; accepted for publication (in revised form) August 4, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sicon/45-1/61626.html
†CMAP–Ecole Polytechnique, F 91128 Palaiseau, France (Rama.Cont@polytechnique.org).
‡Laboratoire de Probabilites et Modeles Aleatoires, Université de Paris VII, 2, Place Jussieu, 75251 Paris, France (Peter.tankov@polytechnique.org). This author was supported by an AMX fellowship from the French Ministry of Education.

The notion of minimal entropy martingale measure (MEMM)—the pricing measure Q that minimizes the relative entropy with respect to a reference probability P—has been investigated by many authors [22, 19, 29]. However, option prices computed using the MEMM are in general not consistent with the market-quoted prices of traded European options and can lead to arbitrage opportunities with respect to market-traded options.

The notion of minimal entropy distribution consistent with observed market prices was introduced in a static framework in [4, 3, 35]; given prices of call options $\{C_M(T_i, K_i)\}_{i\in I}$ and a prior distribution P on scenarios, it is obtained by minimizing relative entropy over all probability measures $Q \sim P$ such that

$$C_M(T_i, K_i) = E^Q[e^{-rT_i}(S_{T_i} - K_i)^+] \quad \text{for } i \in I. \tag{1.1}$$

This approach is based on relative entropy minimization under constraints [17] and yields a computable result. It was extended to the case of stochastic processes by the weighted Monte Carlo method of Avellaneda et al. [5], but the martingale property is not taken into account since it would yield an infinite number of constraints [30]. As a result, derivative prices computed with the weighted Monte Carlo algorithm may contain arbitrage opportunities, especially when applied to forward start contracts.

Goll and Rüschendorf [24] consider the notion of consistent (or calibrated) MEMM (CMEMM), defined as the solution of

$$I(Q^*|P) = \min_{Q\in\mathcal{M}^*} I(Q|P),$$

where the minimum is taken over all martingale measures $Q \sim P$ verifying (1.1). While this notion seems to conciliate the advantages of the MEMM and Avellaneda's entropy minimization under constraints, no algorithm is proposed in [24] to compute the CMEMM. In fact, the notion of CMEMM does not in general preserve the structure of the prior—e.g., the Markov property—and it may be difficult to represent.[1]

We also note that such model selection methods based on relative entropy are not convenient when dealing with one-dimensional diffusion models since as soon as the model has a diffusion coefficient different from the prior their measures become singular and the relative entropy is infinite.

In this paper we show that the shortcomings of the above approaches can be overcome by enlarging the class of models to include processes with jumps and using relative entropy as a regularization criterion rather than a selection criterion. On one hand, introducing jumps in the prior model allows us to obtain a large class of equivalent martingale measures which also have finite relative entropy with respect to the prior, avoiding the singularity which arises in diffusion models. On the other hand, by restricting the class of pricing models to exponential Lévy models—where the risk-neutral dynamics of the logarithm of the stock price is given by a Lévy process—we are able to go beyond a simple existence result and obtain a computable alternative to the CMEMM. Also, unlike the weighted Monte Carlo approach, our approach yields as a solution a continuous-time price process whose discounted value is a martingale. Finally, the use of regularization yields a stable solution to the inverse problem for which a computational approach is possible [14].

The relation between the option prices and the parameters of the process (its Lévy measure) being nonlinear, we face a nonlinear, infinite-dimensional inverse problem.

[1] In particular, if X is a Lévy process under the prior P, it will in general no longer be a Lévy process under a consistent MEMM.

After demonstrating the ill-posedness of the usual and least squares version of this inverse problem, we show that it can be regularized by using as a penalization term the relative entropy with respect to a prior exponential Lévy model. We show that our approach yields solutions which are continuous with respect to the data, stable with respect to the choice of prior, and which converge to the minimum entropy least squares solution of the initial problem.

Unlike linear inverse problems for which general results on regularization methods and their convergence properties are available [20], nonlinear inverse problems have been explored less systematically. Our study is an example of rigorous analysis of regularization using entropy for a nonlinear, infinite-dimensional inverse problem. Previous results on regularization using entropy have been obtained in a Banach space setting [21] by mapping the problem to a Tikhonov regularization problem. Using probabilistic methods, we are able to use a direct approach and extend these results to the spaces of probability measures considered here.

The paper is structured as follows. Section 2 recalls basic facts about Lévy processes and exponential Lévy models. In section 3 we formulate the calibration problem as that of finding a martingale measure Q, consistent with market-quoted prices of traded options, under which the logarithm of the stock price process remains a Lévy process. We show that both this problem and its least squares version are ill-posed: a solution need not exist and, when it exists, may be unstable with respect to perturbations in the data. Section 4 discusses relative entropy in the case of Lévy processes and its use as a criterion for selecting solutions, and introduces the notion of minimum entropy least squares solution. In section 5 we formulate the regularized version of the calibration problem, show that it always admits a solution depending continuously on market data, discuss conditions for the solutions to be equivalent martingale measures, and formulate conditions under which they converge to the minimum entropy least squares solutions as the noise level in the data goes to zero.

In section 6 we show that the solutions of the regularized calibration problem are stable with respect to small perturbations of the prior measure. The solutions of the regularized calibration problem with any prior measure can thus be approximated (in the weak sense) by the solutions of regularized problems with discretized priors, which has implications for the discretization and the numerical solution of the regularized calibration problem, further discussed in [14]. In the appendix we discuss some properties of relative entropy in the case of Lévy processes.

2. Definitions and notation. Consider a time horizon $T_\infty < \infty$ and denote by Ω the space of real-valued cadlag functions on $[0, T_\infty]$, equipped with the Skorokhod topology [27]. The time horizon T_∞ must be chosen finite since we will work with the class of Lévy processes absolutely continuous with respect to a given Lévy process, and on an infinite time interval this class is trivial since in this case absolute continuity of Lévy processes is equivalent to identity in law [27, Theorem IV.4.39]. Unless otherwise mentioned, X is the coordinate process: for every $\omega \in \Omega$, $X_t(\omega) := \omega(t)$. $\mathcal{F}$ is the smallest σ-field, for which the mappings $\omega \in \Omega \mapsto \omega(s)$ are measurable for all $s \in [0, T_\infty]$ and for any $t \in [0, T_\infty]$, and $(\mathcal{F}_t)$ is the natural filtration of $(X_t)_{t\in[0,T_\infty]}$. Weak convergence of measures will be denoted by $\Rightarrow$.

Lévy processes. A Lévy process $\{X_t\}_{t\geq 0}$ on $(\Omega, \mathcal{F}, P)$ is a stochastic process with stationary independent increments satisfying $X_0 = 0$. The characteristic function of X_t has the following form, called the Lévy–Khinchin representation [34]:

$$E[e^{izX_t}] = e^{t\psi(z)} \text{ with } \psi(z) = -\frac{1}{2}Az^2 + i\gamma z + \int_{-\infty}^{\infty} (e^{izx} - 1 - izh(x))\nu(dx), \tag{2.1}$$

where $A \geq 0$ is the unit variance of the Brownian motion part of the Lévy process, $\gamma \in \mathbb{R}$, ν is a positive measure on $\mathbb{R}$ verifying $\nu(\{0\}) = 0$ and

$$\int_{-\infty}^{\infty} (x^2 \wedge 1)\nu(dx) < \infty,$$

and h is the truncation function: any bounded measurable function $\mathbb{R} \to \mathbb{R}$ such that $h(x) \equiv x$ on a neighborhood of zero. The most common choice of truncation function is $h(x) = x1_{|x|\leq 1}$ but sometimes in this paper we will need h to be continuous. The triplet (A, ν, γ) is called the characteristic triplet of X with respect to the truncation function h.

Model setup. We consider exponential Lévy models, where the stock price S_t is modeled, under a *risk-neutral measure* Q [25], as the exponential of a Lévy process:

$$S_t = S_0 e^{rt+X_t}, \tag{2.2}$$

where r is the interest rate. Since Q is a risk-neutral probability measure, e^{X_t} must be a martingale. It follows from (2.1) that this is the case if and only if

$$\frac{A}{2} + \gamma + \int_{-\infty}^{\infty} (e^x - 1 - h(x))\nu(dx) = 0. \tag{2.3}$$

Under Q call option prices can be evaluated as discounted expectations of terminal payoffs:

$$C^Q(T, K) = e^{-rT} E^Q[(S_T - K)^+] = e^{-rT} E^Q[(S_0 e^{rT+X_T} - K)^+]. \tag{2.4}$$

Notation. In what follows $\mathcal{P}(\Omega)$ denotes the set of probability measures (stochastic processes) on $(\Omega, \mathcal{F})$, $\mathcal{L}$ denotes the set of all probability measures $P \in \mathcal{P}(\Omega)$ under which the coordinate process X is a Lévy process, and $\mathcal{M}$ stands for the set of all probability measures $P \in \mathcal{P}(\Omega)$, under which $\exp(X_t)$ is a martingale. $\mathcal{L}_{NA}$ is the set of all probability measures $P \in \mathcal{L}$ corresponding to arbitrage-free exponential Lévy models, that is, to Lévy processes that are not almost surely increasing nor almost surely decreasing. Furthermore, for $B > 0$ we define

$$\mathcal{L}_B^+ = \{P \in \mathcal{L},\ P[\Delta X_t \leq B\ \forall t \in [0, T_\infty]\ ] = 1\ \},$$

the set of Lévy processes with jumps bounded from above by B.

The following lemma shows the usefulness of the above definitions.

LEMMA 2.1. *The set $\mathcal{M} \cap \mathcal{L}_B^+$ is weakly closed for every $B > 0$.*

Proof. Let $\{Q_n\}_{n=1}^{\infty} \subset \mathcal{M} \cap \mathcal{L}_B^+$ with characteristic triplets (A_n, ν_n, γ_n) with respect to a continuous truncation function h and let Q be a Lévy process with characteristic triplet (A, ν, γ) with respect to h, such that $Q_n \Rightarrow Q$. Note that the limit in distribution of a sequence of Lévy processes is necessarily a Lévy process; due to convergence of characteristic functions, the limiting process must have stationary and independent increments. Define a function f by

$$f(x) := \begin{cases} 0, & x \leq B, \\ 1, & x \geq 2B, \\ \frac{x-B}{B} & B < x < 2B. \end{cases}$$

By Corollary VII.3.6 in [27], $\int_{-\infty}^{\infty} f(x)\nu(dx) = \lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)\nu_n(dx) = 0$, which implies that the jumps of Q are bounded by B. Define

$$g(x) := \begin{cases} e^x - 1 - h(x) - \frac{1}{2}h^2(x), & x \leq B, \\ e^B - 1 - h(B) - \frac{1}{2}h^2(B), & x > B. \end{cases}$$

Then, once again by Corollary VII.3.6 in [27] and because Q_n satisfies the martingale condition (2.3) for every n,

$$\gamma + \frac{A}{2} + \int_{-\infty}^{\infty} (e^x - 1 - h(x))\nu(dx) = \gamma + \frac{A + \int_{-\infty}^{\infty} h^2(x)\nu(dx)}{2} + \int_{-\infty}^{\infty} g(x)\nu(dx)$$
$$= \lim_{n\to\infty} \left\{ \gamma_n + \frac{A_n + \int_{-\infty}^{\infty} h^2(x)\nu_n(dx)}{2} + \int_{-\infty}^{\infty} g(x)\nu_n(dx) \right\} = 0,$$

which shows that Q also satisfies the condition (2.3). □

3. The calibration problem and its least squares formulation. Suppose first that the market data C_M are consistent with the class of exponential Lévy models. This is, for example, the case when the market pricing rule is an exponential Lévy model but can hold more generally since many models may give the same prices for a given set of European options. For instance one can construct, using Dupire's formula [18], a diffusion model that gives the same prices, for a set of European options, as a given exp-Lévy model [12]. Using the notation defined in the preceding section, the calibration problem assumes the following form.

PROBLEM 1 (calibration with equality constraints). *Given prices of call options* $\{C_M(T_i, K_i)\}_{i\in I}$*, find an arbitrage-free exponential Lévy model* $Q^* \in \mathcal{M} \cap \mathcal{L}$ *such that*

$$\forall i \in I, \quad C^{Q^*}(T_i, K_i) = C_M(T_i, K_i). \tag{3.1}$$

When the market data are not consistent with the class of exponential Lévy models, the exact calibration problem may not have a solution. In this case one may consider an approximate solution: instead of reproducing the market option prices exactly, one may look for a Lévy triplet which reproduces them in a least squares sense. Let w be a probability measure on $[0, T_\infty] \times [0, \infty)$ (the weighting measure, determining the relative importance of different data points). An option data set is defined as a mapping $C : [0, T_\infty] \times [0, \infty) \to [0, \infty)$ and the data sets that coincide w-almost everywhere are considered identical. One can introduce a norm on option data sets via

$$\|C\|_w^2 := \int_{[0,T_\infty]\times[0,\infty)} C(T, K)^2 w(dT \times dK). \tag{3.2}$$

The quadratic pricing error in model Q is then given by $\|C_M - C^Q\|_w^2$. If the number of constraints is finite, then $w = \sum_{i=1}^N w_i \delta_{(T_i,K_i)}(dT \times dK)$ (with, e.g., N constraints), where $\{w_i\}_{1\le i\le N}$ are positive weights that sum up to one. Therefore, in this case

$$\|C_M - C^Q\|_w^2 = \sum_{i=1}^N w_i (C_M(T_i, K_i) - C^Q(T_i, K_i))^2. \tag{3.3}$$

The following lemma establishes some useful properties of the pricing error functional.

LEMMA 3.1. *The pricing error functional* $Q \mapsto \|C_M - C^Q\|_w^2$ *is uniformly bounded and weakly continuous on* $\mathcal{M} \cap \mathcal{L}$.

Proof. From (2.4), $C^Q(T, K) \le S_0$. Absence of arbitrage in the market implies that the market option prices satisfy the same condition. Therefore, $(C_M(T, K) - C^Q(T, K))^2 \le S_0^2$ and, since w is a probability measure, $\|C_M - C^Q\|_w^2 \le S_0^2$.

Let $\{Q_n\}_{n\geq 1} \subset \mathcal{M}\cap\mathcal{L}$ and $Q \in \mathcal{M}\cap\mathcal{L}$ be such that $Q_n \Rightarrow Q$. For all T, K,

$$\begin{aligned}\lim_n C^{Q_n}(T,K) &= e^{-rT}\lim_n E^{Q_n}[(S_0e^{rT+X_T}-K)^+]\\ &= e^{-rT}\lim_n E^{Q_n}[S_0e^{rT+X_T}-K] + e^{-rT}\lim_n E^{Q_n}[(K-S_0e^{rT+X_T})^+]\\ &= S_0 - Ke^{-rT} + e^{-rT}E^Q[(K-S_0e^{rT+X_T})^+] = C^Q(T,K).\end{aligned}$$

Therefore, by the dominated convergence theorem, $\|C_M - C^{Q_n}\|_w^2 \to \|C_M - C^Q\|_w^2$. □

The calibration problem now takes the following form.

PROBLEM 2 (least squares calibration problem). *Given prices C_M of call options, find $Q^* \in \mathcal{M}\cap\mathcal{L}$ such that*

$$\|C_M - C^{Q^*}\|_w^2 = \inf_{Q\in\mathcal{M}\cap\mathcal{L}} \|C_M - C^Q\|_w^2. \tag{3.4}$$

In what follows, any such Q^* will be called a least squares solution and the set of all least squares solutions will be denoted by $\mathrm{LSS}(C_M)$.

Several authors [2, 8] have used least squares formulations similar to (3.4) for calibrating parametric models without taking into account that the least squares calibration problem is ill-posed in several ways:

Lack of identification. Option prices are only available for a finite number of strikes (typically between 10 and 100) and knowing the prices of a finite number of options is not sufficient to reconstruct the Lévy process. This problem is discussed in detail in [14, 36].

Absence of solution. In some cases even the least squares problem may not admit a solution, as shown by the following (artificial) example.

Example 3.1. Suppose that $S_0 = 1$, there are no interest rates or dividends, and the (equally weighted) market data consist of the following two observations:

$$C_M(T=1, K=1) = 1 - e^{-\lambda} \quad \text{and} \quad C_M(T=1, K=e^{\lambda}) = 0, \tag{3.5}$$

with some $\lambda > 0$. It is easy to see that these prices are, for example, compatible with the (martingale) asset price process $S_t = e^{\lambda t}1_{t\leq\tau_1}$, where τ_1 is the time of the first jump of a Poisson process with intensity λ. We will show that if the market data are given by (3.5), the calibration problem (3.4) does not admit a solution.

Equation (2.4) implies that in every risk-neutral model Q, for fixed T, $C^Q(T,K)$ is a convex function of K and that $C^Q(T, K=0) = 1$. The only convex function which satisfies this equality and passes through the market data points (3.5) is given by $C(T=1,K) = (1-Ke^{-\lambda})^+$. Therefore, in every arbitrage-free model that is an exact solution of the calibration problem with market data (3.5), for every $K \geq 0$, $P[S_1 \leq K] = e^{-\lambda}1_{K\leq e^{\lambda}}$. Since in an exponential Lévy model $P[S_1 > 0] = 1$, there is no risk-neutral exponential Lévy model for which $\|C_M - C^Q\|_w = 0$.

On the other hand, $\inf_{Q\in\mathcal{M}\cap\mathcal{L}} \|C_M - C^Q\|_w^2 = 0$. Indeed, let $\{N_t\}_{t\geq 0}$ be a Poisson process with intensity λ. Then for every n, the process

$$X_t^n := -nN_t + \lambda t(1-e^{-n}) \tag{3.6}$$

belongs to $\mathcal{M}\cap\mathcal{L}$ and

$$\lim_{n\to\infty} E[(e^{X_t^n}-K)^+] = \lim_{n\to\infty}\sum_{k=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\left(e^{-nk+\lambda t(1-e^{-n})}-K\right)^+ = (1-Ke^{-\lambda t})^+.$$

We have shown that $\inf_{Q\in\mathcal{M}\cap\mathcal{L}} \|C_M - C^Q\|^2 = 0$ and that for no Lévy process $Q \in \mathcal{M}\cap\mathcal{L}$, $\|C_M - C^Q\|^2 = 0$. Thus the calibration problem (3.4) does not admit a solution.

Lack of continuity of solutions with respect to market data. Market option prices are typically defined up to a bid-ask spread and the prices used for calibration may therefore be subject to perturbations of this order. If the solution of the calibration problem is not continuous with respect to market data, these small errors may dramatically alter the result of calibration, rendering it useless. In addition, in absence of continuity small daily changes in prices could lead to large variations of calibrated parameters and of quantities computed using these parameters, such as prices of exotic options.

When the calibration problem has more than one solution, care should be taken in defining continuity. In what follows, we will use the following definition [7, 20].

DEFINITION 3.2 (continuity with respect to data). *The solutions of a calibration problem are said to depend continuously on input data at the point C_M if for every sequence of data sets $\{C_M^n\}_{n\geq 0}$ such that $\|C_M^n - C_M\|_w \xrightarrow[n\to\infty]{} 0$; if, for every n, Q_n is a solution of the calibration problem with data C_M^n, then*

1. *$\{Q_n\}_{n\geq 1}$ has a weakly convergent subsequence $\{Q_{n_m}\}_{m\geq 1}$.*
2. *The limit Q of every weakly convergent subsequence of $\{Q_n\}_{n\geq 1}$ is a solution of the calibration problem with data C_M.*

If the solution of the calibration problem with the limiting data C_M is unique, this definition reduces to the standard definition of continuity, because in this case every subsequence of $\{Q_n\}$ has a further subsequence converging towards Q, which implies $Q_n \Rightarrow Q$.

Remark 3.1. Note that the above definition can accommodate the presence of random errors ("noise") in the data. In this case the observational error can be described by a separate probability space $(\Omega_0, \mathcal{E}, p_0)$. The continuity property must then be interpreted as almost-sure continuity with respect to the law p_0 of the observational errors: for every (random) sequence $\{C_M^n\}_{n\geq 0}$ such that $\|C_M^n - C_M\|_w \xrightarrow[n\to\infty]{} 0$ almost surely, any sequence of solutions with data $\{C_M^n\}_{n\geq 0}$ must verify the properties of Definition 3.2 p_0-almost surely.

It is easy to construct an example of market data leading to a least squares calibration problem (3.4) that does not satisfy the above definition.

Example 3.2. Assume $S_0 = 1$, no interest rates or dividends, and observations given by a single option price:

$$C_M^n(T=1, K=1) = E[(e^{X_1^n} - 1)^+] \quad \text{for } n \geq 1 \quad \text{and} \quad C_M(T=1, K=1) = 1 - e^{-\lambda},$$

where X_t^n is defined by (3.6) and $\lambda > 0$. Then $\|C_M^n - C_M\|_w \xrightarrow[n\to\infty]{} 0$ and X_t^n is a solution for data C_M^n, but the sequence $\{X_t^n\}$ has no convergent subsequence (cf. Corollary VII.3.6 in [27]).

In addition to these theoretical obstacles, even if a solution exists, it may be difficult to compute numerically since, as shown in [14, 36], the pricing error $\|C_M - C^Q\|^2$ is typically nonconvex and can have many local minima, preventing a gradient-based minimization algorithm from finding the solution.

4. Relative entropy as a selection criterion. When constraints given by option prices do not determine the exponential Lévy model completely, additional information may be introduced into the problem by specifying a *prior model*: we start from a reference Lévy process P and look for the solution of the problem (3.4) that has the smallest relative entropy with respect to P. For two probabilities P and Q on the same measurable space $(\Omega, \mathcal{F})$, the relative entropy of Q with respect to P

is defined by

$$I(Q|P) = \begin{cases} E^P\left[\frac{dQ}{dP}\log\frac{dQ}{dP}\right] & \text{if } Q \ll P, \\ \infty & \text{otherwise,} \end{cases} \tag{4.1}$$

where by convention $x \log x = 0$ when $x = 0$. We define $I_T(Q|P) := I(Q|_{\mathcal{F}_T}|P|_{\mathcal{F}_T})$.

Problem 3 (minimum entropy least squares calibration problem). *Given prices C_M of call options and a prior Lévy process P, find a least squares solution $Q^* \in \mathrm{LSS}(C_M)$, such that*

$$I(Q^*|P) = \inf_{Q\in\mathrm{LSS}(C_M)} I(Q|P). \tag{4.2}$$

In what follows, any such Q^* will be called a minimum entropy least squares solution (MELSS) and the set of all such solutions will be denoted by $\mathrm{MELSS}(C_M)$.

P reflects a priori knowledge about the nature of possible trajectories of the underlying asset and their probabilities of occurrence. A natural choice of prior, ensuring absence of arbitrage in the calibrated model, is an exponential Lévy model estimated from the time series of returns. Whether this choice is adopted or not does not affect our discussion below. Other possible ways to choose the prior model in practice are discussed in [14], which also gives an empirical analysis of the effect of the choice of prior on the solution of the calibration problem.

The choice of relative entropy as a method for selection of solutions of the calibration problem is driven by the following considerations:

- Relative entropy can be interpreted as a (pseudo-)distance to the prior P: it is convex, nonnegative functional of Q for fixed P, and equal to zero if and only if $\frac{dQ}{dP} = 1$ P-a.s. To see this, observe that

$$E^P\left[\frac{dQ}{dP}\log\frac{dQ}{dP}\right] = E^P\left[\frac{dQ}{dP}\log\frac{dQ}{dP} - \frac{dQ}{dP} + 1\right],$$

  and that $z \log z - z + 1$ is a convex nonnegative function of z, equal to zero if and only if $z = 1$.
- Relative entropy for Lévy processes is easily expressed in terms of their characteristic triplets (see Theorem A.1).
- Relative entropy has an information-theoretic interpretation and has been repeatedly used for model selection in finance (see section 1).

Using relative entropy for the selection of solutions removes, to some extent, the identification problem of least squares calibration. Whereas in the least squares case this was an important nuisance, now, if two measures reproduce market option prices with the same precision and have the same entropy relative to the prior, this means that both measures are compatible with all the available information. Knowledge of many such probability measures instead of one may be seen as an advantage, because it allows us to estimate model risk and provide confidence intervals for the prices of exotic options [12]. However, the calibration problem (4.2) remains ill-posed: since the minimization of entropy is done over the results of least squares calibration, problem (4.2) may only admit a solution if problem (3.4) does. Also, $\mathrm{LSS}(C_M)$ is not necessarily a compact set, so even if it is nonempty, (4.2) may not have a solution. Other undesirable properties such as absence of continuity and numerical instability are also inherited from the least squares approach.

The MELSS does not always exist, but if the prior is chosen correctly such that (3.4) admits a solution with finite relative entropy with respect to the prior, then MELSSs will also exist.

LEMMA 4.1. *Let $P \in \mathcal{L}_{NA} \cap \mathcal{L}_B^+$ for some $B > 0$ and assume problem* (3.4) *admits a solution Q^+ with $I(Q^+|P) = C < \infty$. Then problem* (4.2) *admits a solution.*

Proof. Under the condition of the lemma, it is clear that the solution Q^* of problem (4.2), if it exists, satisfies $I(Q^*|P) \leq C$. This entails that $Q^* \ll P$, which means by Theorem IV.4.39 in [27] that $Q^* \in \mathcal{L}_B^+$. Therefore, Q^* belongs to the set

$$\mathcal{L}_B^+ \cap \{Q \in \mathcal{M} \cap \mathcal{L} : \|C^Q - C_M\| = \|C^{Q^+} - C_M\|\} \cap \{Q \in \mathcal{L} : I(Q|P) \leq C\}. \tag{4.3}$$

Lemma A.2 and the Prohorov's theorem entail that the level set $\{Q \in \mathcal{L} : I(Q|P) \leq C\}$ is relatively weakly compact. On the other hand, by Corollary A.4, $I(Q|P)$ is weakly lower semicontinuous with respect to Q for fixed P. Therefore, the set $\{Q \in \mathcal{P}(\Omega) : I(Q|P) \leq C\}$ is weakly closed, and since by Lemma 2.1 $\mathcal{M} \cap \mathcal{L}_B^+$ is also weakly closed, the set $\mathcal{M} \cap \mathcal{L}_B^+ \cap \{Q \in \mathcal{L} : I(Q|P) \leq C\}$ is weakly compact. Lemma 3.1 then implies that the set (4.3) is also weakly compact. Since $I(Q|P)$ is weakly lower semicontinuous, it reaches its minimum on this set. □

Remark 4.1. It is essential for our analysis that the model has discontinuous trajectories, i.e., the prior P corresponds to a process with jumps, not a diffusion process. If P corresponds to the law of a Markovian diffusion model, then the set of processes which have both the martingale property and finite entropy with respect to P is reduced to a single element and the solution to 4.2 is trivial (this follows, e.g., from Theorem IV.4.39 in [27]).

5. Regularization using relative entropy. As observed in [14] and in section 4, problem (4.2) is ill-posed and hard to solve numerically. In particular its solutions, when they exist, may not be stable with respect to perturbations of market data. If we do not know the prices C_M exactly but only dispose of observations C_M^δ with $||C_M^\delta - C_M||_w \leq \delta$ and want to construct an approximation to MELSS(C_M), it is not a good idea to solve problem (4.2) with the noisy data C_M^δ because MELSS(C_M^δ) may be very different from MELSS(C_M). We therefore need to regularize the problem (4.2), that is, construct a family of continuous "regularization operators" $\{R_\alpha\}_{\alpha>0}$, where α is the regularization parameter such that $R_\alpha(C_M^\delta)$ converges to an MELSS as the noise level δ tends to zero if an appropriate parameter choice rule $\delta \mapsto \alpha(\delta)$ is used [20]. The approximation to MELSS(C_M) using the noisy data C_M^δ is then given by $R_\alpha(C_M^\delta)$ with an appropriate choice of α.

Following a classical approach to regularization of ill-posed problems [20, 4], we regularize (4.2) by using the relative entropy as a penalization term:

$$J_\alpha(Q) = \|C_M^\delta - C^Q\|_w^2 + \alpha I(Q|P), \tag{5.1}$$

where α is the regularization parameter. We then solve the following optimization problem.

PROBLEM 4 (regularized calibration problem). *Given prices C_M of call options, a prior Lévy process P, and a regularization parameter $\alpha > 0$, find $Q^* \in \mathcal{M} \cap \mathcal{L}$ such that*

$$J_\alpha(Q^*) = \inf_{Q \in \mathcal{M} \cap \mathcal{L}} J_\alpha(Q). \tag{5.2}$$

Problem (5.2) can be thought of in two ways:

- If the MELSS with the true data C_M exists, (5.2) allows us to construct a stable approximation of this solution using the noisy data.
- If the MELSS$(C_M) = \emptyset$, either because the set of least squares solutions is empty or because the least squares solutions are incompatible with the prior, the regularized problem (5.2) allows us to achieve, in a stable manner, a trade-off between matching the constraints and the prior information.

In the rest of this section we study the regularized calibration problem. Under our standing hypothesis that the prior Lévy process has jumps bounded from above and corresponds to an arbitrage-free market ($P \in \mathcal{L}_{NA} \cap \mathcal{L}_B^+$), we show that the regularized calibration problem always admits a solution that depends continuously on the market data. In addition, we give a sufficient condition on the prior P for the solution to be an *equivalent* martingale measure and show how the regularization parameter α must be chosen depending on the noise level δ if the regularized solutions are to converge to the solutions of the minimum entropy least squares calibration problem (4.2).

5.1. Existence of solutions. The following result shows that, unlike the exact or the least squares formulations, the regularized inverse problem always admits a solution.

THEOREM 5.1. *Let $P \in \mathcal{L}_{NA} \cap \mathcal{L}_B^+$ for some $B > 0$. Then the calibration problem* (5.2) *has a solution $Q^* \in \mathcal{M} \cap \mathcal{L}_B^+$.*

Proof. By Lemma A.5, there exists $Q^0 \in \mathcal{M} \cap \mathcal{L}$ with $I(Q^0|P) < \infty$. The solution, if it exists, must belong to the level set $L_{J_\alpha(Q^0)} := \{Q \in \mathcal{L} : I(Q|P) \leq J_\alpha(Q^0)\}$. Since $J_\alpha(Q^0) = \|C_M - C^{Q^0}\|_w^2 + I(Q^0|P) < \infty$, by Lemma A.2 $L_{J_\alpha(Q^0)}$ is tight and, by Prohorov's theorem, weakly relatively compact. Corollary A.4 entails that $I(Q|P)$ is weakly lower semicontinuous with respect to Q. Therefore, $\{Q \in \mathcal{P}(\Omega) : I(Q|P) \leq J_\alpha(Q^0)\}$ is weakly closed, and since by Lemma 2.1 $\mathcal{M} \cap \mathcal{L}_B^+$ is weakly closed, $\mathcal{M} \cap \mathcal{L}_B^+ \cap L_{J_\alpha(Q^0)}$ is weakly compact. Moreover, by Lemma 3.1 the squared pricing error is weakly continuous, which entails that $J_\alpha(Q)$ is weakly lower semicontinuous. Therefore, $J_\alpha(Q)$ achieves its minimum value on $\mathcal{M} \cap \mathcal{L}_B^+ \cap L_{J_\alpha(Q^0)}$, which proves the theorem. □

Since $P \in \mathcal{L}_B^+$ (i.e., with jumps of X bounded from above P-a.s.), solutions Q are also in $\mathcal{L}_B^+$. This may seem a limitation if the data is generated by a Lévy process with jumps unbounded from above. This case is unlikely in financial applications; the form of Lévy densities found empirically in [14] gives little evidence for large upward jumps. Even in the theoretical case where the observed option prices are generated by an exponential-Lévy model with jumps unbounded from above, the localization estimates in [15, Proposition 4.2.] show that we can reproduce such prices with a Lévy process in $\mathcal{L}_B^+$ by choosing B large enough.

Since every solution Q^* of the regularized calibration problem (5.2) has finite relative entropy with respect to the prior P, necessarily $Q^* \ll P$. However, Q^* need not in general be equivalent to the prior. When the prior corresponds to the "objective" probability measure, absence of arbitrage is guaranteed if options are priced using an *equivalent* martingale measure [25]. The following theorem gives a sufficient condition for this equivalence.

THEOREM 5.2. *Let $P \in \mathcal{L}_{NA} \cap \mathcal{L}_B^+$ and assume the characteristic function Φ_T^P of P satisfies*

$$\int_{-\infty}^{\infty} |\Phi_T^P(u)| du < \infty \tag{5.3}$$

for some $T < T_0$, where T_0 is the shortest maturity, present in the market data. Then every solution Q^* of the calibration problem (5.2) *satisfies $Q^* \sim P$.*

Remark 5.1. Condition (5.3) implies that the prior Lévy process has a continuous density at time T and all subsequent times. Two important examples of processes satisfying the condition (5.3) for all T are

- processes with nonzero Gaussian component ($A > 0$).
- processes with stablelike behavior of small jumps, that is, processes whose Lévy measure satisfies

$$\exists \beta \in (0,2), \quad \liminf_{\varepsilon \downarrow 0} \varepsilon^{-\beta} \int_{-\varepsilon}^{\varepsilon} |x|^2 \nu(dx) > 0. \tag{5.4}$$

  For a proof, see [34, Proposition 28.3]. This class includes tempered stable processes [13] with $\alpha_+ > 0$ and/or $\alpha_- > 0$.

To prove Theorem 5.2 we will use the following lemma.

LEMMA 5.3. *Let $P \in \mathcal{M} \cap \mathcal{L}_B^+$ with characteristic triplet (A, ν, γ) and characteristic exponent ψ. There exists $C < \infty$ such that*

$$\left| \frac{\psi(v-i)}{(v-i)v} \right| \le C \quad \forall v \in \mathbb{R}.$$

Proof. From the Lévy–Khinchin formula and (2.3),

$$\psi(v-i) = -\frac{1}{2} A v (v-i) + \int_{-\infty}^{\infty} (e^{i(v-i)x} + iv - e^x - ive^x) \nu(dx). \tag{5.5}$$

Observe first that

$$e^{i(v-i)x} + iv - e^x - ive^x = iv(xe^x + 1 - e^x) + \frac{\theta v^2 x^2 e^x}{2} \quad \text{for some } \theta \text{ with } |\theta| \le 1.$$

Therefore, for all v with $|v| \ge 2$,

$$\left| \frac{e^{i(v-i)x} + iv - e^x - ive^x}{(v-i)v} \right| \le xe^x + 1 - e^x + x^2 e^x. \tag{5.6}$$

On the other hand,

$$\begin{aligned} \frac{e^{i(v-i)x} + iv - e^x - ive^x}{(v-i)v} &= \frac{ie^x(e^{ivx} - 1)}{v} - \frac{i(e^{i(v-i)x} - 1)}{v-i} \\ &= -xe^x - \frac{ivx^2}{2} e^{\theta_1 ivx} + x + \frac{i(v-i)x^2}{2} e^{\theta_2 i(v-i)x} \end{aligned}$$

with some $\theta_1, \theta_2 \in [0,1]$. Therefore, for all v with $|v| \le 2$,

$$\left| \frac{e^{i(v-i)x} + iv - e^x - ive^x}{(v-i)v} \right| \le x(1-e^x) + \frac{x^2}{2}(v + \sqrt{1+v^2} e^x) \le x(1-e^x) + x^2(1+2e^x). \tag{5.7}$$

Since the support of ν is bounded from above, the right-hand sides of (5.6) and (5.7) are ν-integrable and the proof of the lemma is completed. □

Proof of Theorem 5.2. Let Q^* be a solution of (5.2) with prior P. By Lemma A.5, there exists $Q^0 \in \mathcal{M} \cap \mathcal{L}$ such that $Q^0 \sim P$. Denote the characteristic triplet of Q^* by (A, ν^*, γ^*) and that of Q^0 by (A, ν^0, γ^0).

Let Q_x be a Lévy process with characteristic triplet $(A, x\nu^0 + (1-x)\nu^*, x\gamma^0 + (1-x)\gamma^*)$. From the linearity of the martingale condition (2.3), it follows that for all $x \in [0,1]$, $Q_x \in \mathcal{M} \cap \mathcal{L}$. Since Q^* realizes the minimum of $J_\alpha(Q)$, necessarily $J_\alpha(Q_x) - J_\alpha(Q^*) \geq 0$ for all $x \in [0,1]$. Our strategy for proving the theorem is first to show that $\frac{\|C_M - C^{Q_x}\|^2 - \|C_M - C^{Q^*}\|^2}{x}$ is bounded as $x \to 0$ and then to show that if $\frac{I(Q_x|P) - I(Q^*|P)}{x}$ is bounded from below as $x \to 0$, necessarily $Q^* \sim P$.

The first step is to prove that the characteristic function Φ^* of Q^* satisfies the condition (5.3) for some $T < T_0$. If $A > 0$, this is trivial. Assume that $A = 0$. In this case, $|\Phi^*_T(u)| = \exp(T\int_{-\infty}^{\infty}(\cos(ux)-1)\nu^*(dx))$. Denote $\frac{d\nu^*}{d\nu^P} := \phi^*$. Since $Q^* \ll P$, by Theorem IV.4.39 in [27], $\int_{-\infty}^{\infty}(\sqrt{\phi^*(x)}-1)^2\nu^P(dx) \leq K < \infty$ for some constant K. Therefore, there exists another constant $C > 0$ such that

$$\int_{\{\phi^*(x)>C\}} (1-\cos(ux))|\phi^* - 1|\nu^P(dx) < C$$

uniformly on u. For all $r > 0$,

$$\begin{aligned}\int_{-\infty}^{\infty}(1-\cos(ux))|\phi^*-1|\nu^P(dx) &\leq C + \int_{\{\phi^*(x)\leq C\}}(1-\cos(ux))|\phi^*-1|\nu^P(dx)\\ &\leq C + \frac{r}{2}\int_{\{\phi^*(x)\leq C\}}(1-\cos(ux))^2\nu^P(dx) + \frac{1}{2r}\int_{\{\phi^*(x)\leq C\}}(\phi^*-1)^2\nu^P(dx)\\ &\leq C + r\int_{-\infty}^{\infty}(1-\cos(ux))\nu^P(dx) + \frac{K(\sqrt{C}+1)^2}{2r}.\end{aligned}$$

This implies

$$\int_{-\infty}^{\infty}(\cos(ux)-1)\nu^*(dx) \leq (1+r)\int_{-\infty}^{\infty}(\cos(ux)-1)\nu^P(dx) + C + \frac{K(\sqrt{C}+1)^2}{2r}$$

for all $r > 0$. Therefore, if the characteristic function of P satisfies the condition (5.3) for some T, the characteristic function of Q^* will satisfy it for every $T' > T$.

Since $P \in \mathcal{L}_{NA} \cap \mathcal{L}^+_B$, $Q_x \in \mathcal{M} \cap \mathcal{L}^+_B$ for all $x \in [0,1]$. Therefore, condition (11.15) in [13] is satisfied and option prices can be computed using equation (11.20) of [13]:[2]

$$\begin{aligned}C^{Q_x}(T,K) &= (1 - Ke^{-rT})^+\\ &\quad + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iv\log K + ivrT}\frac{\exp(T(1-x)\psi^*(v-i) + Tx\psi^0(v-i)) - 1}{iv(1+iv)}dv,\end{aligned}$$

where ψ^0 and ψ^* denote the characteristic exponents of Q_0 and Q^*. It follows that

$$\begin{aligned}&\frac{C^{Q_x}(T,K) - C^{Q^*}(T,K)}{x}\\ &\qquad = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iv\log K + ivrT}\frac{e^{T(1-x)\psi^*(v-i) + Tx\psi^0(v-i)} - e^{T\psi^*(v-i)}}{iv(1+iv)x}dv.\end{aligned}$$

[2]This method for option pricing by Fourier transform is originally due to Carr and Madan [10].

Since $\Re\psi^0(v-i) \le 0$ and $\Re\psi^*(v-i) \le 0$ for all $v \in \mathbb{R}$, Lemma 5.3 implies

$$\left| e^{-iv\log K + ivrT} \frac{e^{T(1-x)\psi^*(v-i)+Tx\psi^0(v-i)} - e^{T\psi^*(v-i)}}{iv(1+iv)x} \right| \le T\frac{|e^{T(1-x)\psi^*(v-i)}||\psi^0(v-i)-\psi^*(v-i)|}{|v(1+iv)|} \le T|e^{T(1-x)\psi^*(v-i)}|C'$$

for some constant C'. From the dominated convergence theorem and since Q^* satisfies (5.3), $\frac{\partial C^{Q_x}(T,K)}{\partial x}|_{x=0}$ exists and is bounded uniformly on T and K in the market data set. This in turn means that $\frac{\|C_M - C^{Q_x}\|^2 - \|C_M - C^{Q^*}\|^2}{x}$ is bounded as $x \to 0$. To complete the proof, it remains to show that if $\frac{I(Q_x|P)-I(Q^*|P)}{x}$ is bounded from below as $x \to 0$, necessarily $Q^* \sim P$. Using the convexity (with respect to ν^Q and γ^Q) of the two terms in the expression (A.1) for relative entropy, we have

$$\begin{aligned}
&\frac{I(Q_x|P) - I(Q^*|P)}{x} \\
&= \frac{T_\infty}{2Ax}\left\{x\gamma^0 + (1-x)\gamma^* - \gamma^P - \int_{|z|\le 1} z(x\nu^0 + (1-x)\nu^* - \nu^P)(dz)\right\}^2 1_{A\neq 0} \\
&\quad - \frac{T_\infty}{2Ax}\left\{\gamma^* - \gamma^P - \int_{|z|\le 1} z(\nu^* - \nu^P)(dz)\right\}^2 1_{A\neq 0} \\
&\quad + \frac{T_\infty}{x}\int_{-\infty}^{\infty}\{(x\phi^0 + (1-x)\phi^*)\log(x\phi^0 + (1-x)\phi^*) - x\phi^0 - (1-x)\phi^* + 1\}\nu^P(dz) \\
&\quad - \frac{T_\infty}{x}\int_{-\infty}^{\infty}\{\phi^*\log(\phi^*) - \phi^* + 1\}\nu^P(dz) \\
&\le \frac{T_\infty}{2A}\left\{\gamma^0 - \gamma^P - \int_{|z|\le 1} z(\nu^0 - \nu^P)(dz)\right\}^2 1_{A\neq 0} \\
&\quad - \frac{T_\infty}{2A}\left\{\gamma^* - \gamma^P - \int_{|z|\le 1} z(\nu^* - \nu^P)(dz)\right\}^2 1_{A\neq 0} \\
&\quad + T_\infty\int_{\{\phi^*>0\}}\{\phi^0\log(\phi^0) - \phi^0 + 1\}\nu^P(dz) - T_\infty\int_{\{\phi^*>0\}}\{\phi^*\log(\phi^*) - \phi^* + 1\}\nu^P(dz) \\
&\quad + T_\infty\int_{\{\phi^*=0\}}\{\phi^0\log(x\phi^0) - \phi^0\}\nu^P(dz) \le I(Q_0|P) + T_\infty\int_{\{\phi^*=0\}}(\phi_0\log x - 1)\nu^P(dz).
\end{aligned}$$

Since $J_\alpha(Q_x) - J_\alpha(Q^*) \ge 0$, this expression must be bounded from below. Therefore, because $\phi_0 > 0$, necessarily $\nu^P(\{\phi^* = 0\}) = 0$ and Theorem IV.4.39 in [27] entails that $P \ll Q^*$. □

5.2. Continuity of solutions with respect to data.

THEOREM 5.4 (continuity of solutions with respect to data). *Let $\{C_M^n\}_{n\ge 1}$ and C_M be data sets of option prices such that*

$$\|C_M^n - C_M\|_w \underset{n\to\infty}{\to} 0.$$

Let $P \in \mathcal{L}_{NA} \cap \mathcal{L}_B^+$, $\alpha > 0$, and for each n let Q_n be a solution of the calibration problem (5.2) with data C_M^n, prior Lévy process P, and regularization parameter α.

Then $\{Q_n\}_{n\geq 1}$ has a subsequence converging weakly to $Q^* \in \mathcal{M} \cap \mathcal{L}_B^+$, and the limit of every converging subsequence of $\{Q_n\}_{n\geq 1}$ is a solution of calibration problem (5.2) *with data C_M, prior P, and regularization parameter α.*

Proof. By Lemma A.5, there exists $Q^0 \in \mathcal{M} \cap \mathcal{L}$ with $I(Q^0|P) < \infty$. Since, by Lemma 3.1, $\|C^{Q^0} - C_M^n\|^2 \leq S_0^2$ for all n, $\alpha I(Q_n|P) \leq S_0^2 + \alpha I(Q^0|P)$ for all n. Therefore, by Lemmas 2.1 and A.2 and Prohorov's theorem, $\{Q_n\}_{n\geq 1}$ is weakly relatively compact. Together with Lemma 2.1 this proves the first part of the theorem. □

Choose any subsequence of $\{Q_n\}_{n\geq 1}$ converging weakly to $Q^* \in \mathcal{M} \cap \mathcal{L}_B^+$. To simplify notation, this subsequence is denoted again by $\{Q_n\}_{n\geq 1}$. The triangle inequality and Lemma 3.1 imply that

$$\|C^{Q_n} - C_M^n\|^2 \xrightarrow[n\to\infty]{} \|C^{Q^*} - C_M\|^2. \tag{5.8}$$

Since, by Lemma A.3, the relative entropy functional is weakly lower semicontinuous with respect to Q, for every $Q \in \mathcal{M} \cap \mathcal{L}_B^+$,

$$\begin{aligned}
\|C^{Q^*} - C_M\|^2 + \alpha I(Q^*|P) &\leq \liminf_n \{\|C^{Q_n} - C_M^n\|^2 + \alpha I(Q_n|P)\} \\
&\leq \liminf_n \{\|C^{Q} - C_M^n\|^2 + \alpha I(Q|P)\} \\
&= \lim_n \|C^{Q} - C_M^n\|^2 + \alpha I(Q|P) \\
&= \|C^{Q} - C_M\|^2 + \alpha I(Q|P),
\end{aligned}$$

where the second inequality follows from the fact that Q_n is the solution of the calibration problem with data C_M^n and the last line follows from the triangle inequality.

5.3. Convergence to MELSSs. The convergence analysis of regularization methods for inverse problems usually involves the study of the solution of the regularized problem as the noise level δ vanishes, the regularization parameter being chosen as a function $\alpha(\delta)$ of the noise level using some parameter choice rule. The following result gives conditions on the parameter choice rule $\delta \mapsto \alpha(\delta)$ under which the solutions of the regularized problem (5.2) converge to MELSSs defined by (4.2).

THEOREM 5.5. *Let $\{C_M^\delta\}$ be a family of data sets of option prices such that $\|C_M - C_M^\delta\| \leq \delta$, let $P \in \mathcal{L}_{NA} \cap \mathcal{L}_B^+$, and assume there exists a solution Q of problem* (3.4) *with data C_M (a least squares solution) such that $I(Q|P) < \infty$.*

In the case where the constraints are attainable, i.e., $\|C^Q - C_M\| = 0$, let $\alpha(\delta)$ be such that $\alpha(\delta) \to 0$ and $\frac{\delta^2}{\alpha(\delta)} \to 0$ as $\delta \to 0$. Otherwise, let $\alpha(\delta)$ be such that $\alpha(\delta) \to 0$ and $\frac{\delta}{\alpha(\delta)} \to 0$ as $\delta \to 0$.

Then every sequence $\{Q^{\delta_k}\}$, where $\delta_k \to 0$ and Q^{δ_k} is a solution of problem (5.2) *with data $C_M^{\delta_k}$, prior P, and regularization parameter $\alpha(\delta_k)$, has a weakly convergent subsequence. The limit of every convergent subsequence is a solution of problem* (4.2) *with data C_M and prior P. If the MELSS is unique (MELSS$(C_M) = \{Q^+\}$), then*

$$Q^\delta \underset{\delta\to 0}{\Rightarrow} Q^+.$$

Proof. By Lemma 4.1, there exists at least one MELSS with data C_M and prior P, with finite relative entropy with respect to the prior. Let $Q^+ \in \text{MELSS}(C_M)$. Since Q^{δ_k} is the solution of the regularized problem, for every k,

$$\|C^{Q^{\delta_k}} - C_M^{\delta_k}\|^2 + \alpha(\delta_k) I(Q^{\delta_k}|P) \leq \|C^{Q^+} - C_M^{\delta_k}\|^2 + \alpha(\delta_k) I(Q^+|P).$$

Using the fact that for every $r > 0$ and for every $x, y \in \mathbb{R}$,

$$(1-r)x^2 + (1-1/r)y^2 \le (x+y)^2 \le (1+r)x^2 + (1+1/r)y^2,$$

we obtain that

$$(1-r)\|C^{Q^{\delta_k}} - C_M\|^2 + \alpha(\delta_k) I(Q^{\delta_k}|P) \le (1+r)\|C^{Q^+} - C_M\|^2 + \frac{2\delta_k^2}{r} + \alpha(\delta_k) I(Q^+|P), \tag{5.9}$$

and since $Q^+ \in \mathrm{LSS}(C_M)$, this implies for all $r \in (0,1)$ that

$$\alpha(\delta_k) I(Q^{\delta_k}|P) \le 2r\|C^{Q^+} - C_M\|^2 + \frac{2\delta_k^2}{r} + \alpha(\delta_k) I(Q^+|P). \tag{5.10}$$

If the constraints are met exactly, $\|C^{Q^+} - C_M\| = 0$, and with the choice $r = 1/2$, the above expression yields

$$I(Q^{\delta_k}|P) \le \frac{4\delta_k^2}{\alpha(\delta_k)} + I(Q^+|P).$$

Since, by the theorem's statement, in the case of exact constraints $\frac{\delta_k^2}{\alpha(\delta_k)} \to 0$, this implies

$$\limsup_k \{I(Q^{\delta_k}|P)\} \le I(Q^+|P). \tag{5.11}$$

If $\|C^{Q^+} - C_M\| > 0$ (misspecified model), then the right-hand side of (5.10) achieves its maximum when $r = \delta_k \|C^{Q^+} - C_M\|^{-1}$, in which case we obtain

$$I(Q^{\delta_k}|P) \le \frac{4\delta_k}{\alpha(\delta_k)}\|C^{Q^+} - C_M\| + I(Q^+|P).$$

Since in the case of approximate constraints $\frac{\delta_k}{\alpha(\delta_k)} \to 0$, we obtain (5.11) once again.

Inequality (5.11) implies in particular that $I(Q^{\delta_k}|P)$ is uniformly bounded, which proves, by Lemmas A.2 and 2.1, that $\{Q^{\delta_k}\}$ is relatively weakly compact in $\mathcal{M} \cap \mathcal{L}_B^+$.

Choose a subsequence of $\{Q^{\delta_k}\}$ converging weakly to $Q^* \in \mathcal{M} \cap \mathcal{L}_B^+$. To simplify notation, this subsequence is denoted again by $\{Q^{\delta_k}\}_{k \ge 1}$. Substituting $r = \delta$ into (5.9) and making k tend to infinity shows that

$$\limsup_k \|C^{Q^{\delta_k}} - C_M\|^2 \le \|C^{Q^+} - C_M\|^2.$$

Together with Lemma 3.1 this implies that

$$\|C^{Q^*} - C_M\|^2 \le \|C^{Q^+} - C_M\|^2;$$

hence Q^* is a least squares solution. By weak lower semicontinuity of I (cf. Lemma A.3) and using (5.11),

$$I(Q^*|P) \le \liminf_k I(Q^{\delta_k}|P) \le \limsup_k I(Q^{\delta_k}|P) \le I(Q^+|P),$$

which means that $Q^* \in \text{MELSS}(C_M)$. The last assertion of the theorem follows from the fact that in this case every subsequence of $\{Q^{\delta_k}\}$ has a further subsequence converging toward Q^+. □

Remark 5.2 (random errors). In line with Remark 3.1, it is irrelevant whether the noise in the data is "deterministic" or "random," as long the error level δ is interpreted as a worst-case error level, i.e., an almost-sure bound on the error:

$$\mathbb{P}_0(||C_M^\delta - C_M|| \le \delta) = 1. \tag{5.12}$$

In this case, Theorem 5.5 holds for random errors, convergence being interpreted as almost-sure convergence with respect to the law $\mathbb{P}_0$ of the errors.

6. Stability with respect to the prior. If we choose a prior Lévy process P with a finite number of jump sizes (sometimes called *simple* Lévy processes),

$$\nu^P = \sum_{k=0}^{M-1} p_k \delta_{\{x_k\}}(dx), \tag{6.1}$$

then the solution Q satisfies $Q \ll P$ by Theorem IV.4.39 in [27]. Thus the corresponding Lévy measure ν^Q satisfies $\nu^Q \ll \nu^P$ and is of the form

$$\nu^Q = \sum_{k=0}^{M-1} q_k \delta_{\{x_k\}}(dx). \tag{6.2}$$

The calibration problem (5.2) is then a finite-dimensional optimization problem and can be solved using a numerical optimization algorithm [14]. The advantage of this method is that we are simply solving (5.2) with a specific choice of prior, so all results of section 5 hold. Numerical methods for solving this problem are discussed in the companion paper [14]. Here we will complement these results by a theorem showing that the solution of a calibration problem with any prior can be approximated (in the weak sense) by a sequence of solutions of calibration problems with simple Lévy processes as priors. We start by showing that every Lévy process can be approximated by simple Lévy processes of the form (6.1).

LEMMA 6.1. *Let P be a Lévy process with characteristic triplet (A, ν, γ) with respect to a continuous truncation function h; for every n, let P_n be a Lévy process with characteristic triplet (A, ν_n, γ) (with respect to h), where*

$$\nu_n := \sum_{k=1}^{2n} \delta_{\{x_k\}}(dx) \frac{\mu([x_k - 1/\sqrt{n}, x_k + 1/\sqrt{n}))}{1 \wedge x_k^2},$$

$x_k := (2(k-n)-1)/\sqrt{n}$, and μ is a (positive and finite) measure on $\mathbb{R}$, defined by $\mu(B) := \int_B (1 \wedge x^2)\nu(dx)$ for all $B \in \mathcal{B}(\mathbb{R})$. Then $P_n \Rightarrow P$.

Proof. For a function $f \in C_b(\mathbb{R})$, define

$$f_n(x) := \begin{cases} 0, & x \ge 2\sqrt{n}, \\ 0, & x < -2\sqrt{n}, \\ f(x_i), & x \in [x_i - 1/\sqrt{n}, x_i + 1/\sqrt{n}) \quad \text{with } 1 \le i \le 2n. \end{cases}$$

Then $\int (1 \wedge x^2) f(x) \nu_n(dx) = \int f_n(x)\mu(dx)$. Since $f(x)$ is continuous, $f_n(x) \xrightarrow{n\to\infty} f(x)$ for all x and since f is bounded, the dominated convergence theorem implies

$$\lim_n \int (1 \wedge x^2) f(x) \nu_n(dx) = \lim_n \int f_n(x)\mu(dx) = \int (1 \wedge x^2) f(x)\nu(dx). \tag{6.3}$$

With $f(x) \equiv \frac{h^2(x)}{1 \wedge x^2}$ the above yields

$$\int h^2(x)\nu_n(dx) \overset{n\to\infty}{\to} \int h^2(x)\nu(dx).$$

On the other hand, for every $g \in C_b(\mathbb{R})$ such that $g(x) \equiv 0$ on a neighborhood of 0, $f(x) := \frac{g(x)}{1\wedge x^2}$ belongs to $C_b(\mathbb{R})$. Therefore, from (6.3), $\lim_n \int g(x)\nu_n(dx) = \int g(x)\nu(dx)$ and by Corollary VII.3.6 in [27], $P_n \Rightarrow P$. □

To compute numerically the solution of the calibration problem (5.2) with a given prior P, we can construct, using Lemma 6.1, an approximating sequence $\{P_n\}$ of simple Lévy processes such that $P_n \Rightarrow P$. Problem (5.2) with P replaced by P_n is then a finite-dimensional optimization problem that can be solved. The resulting sequence $\{Q_n\}$ of solutions will converge, as shown in the following theorem, to a solution of the calibration problem with prior P.

THEOREM 6.2. *Let $P, \{P_n\}_{n\geq 1} \subset \mathcal{L}_{NA} \cap \mathcal{L}_B^+$ such that $P_n \Rightarrow P$. Let $\alpha > 0$, let C_M be a data set of option prices, and for each $n \geq 1$ let Q_n be a solution of the calibration problem* (5.2) *with prior P_n, regularization parameter α, and data C_M. Denote the characteristic triplet (with respect to a continuous truncation function h) of P_n by $(A_n, \nu_n^P, \gamma_n^P)$ and that of P by (A, ν^P, γ^P). If $A_n \to A > 0$, then the sequence $\{Q_n\}_{n\geq 1}$ has a weakly convergent subsequence and the limit of every weakly convergent subsequence of $\{Q_n\}_{n\geq 1}$ is a solution of the calibration problem* (5.2) *with prior P.*

Proof. By Lemma A.5, there exists $C < \infty$ such that for every n one can find $\tilde{Q}_n \in \mathcal{M} \cap \mathcal{L}$ with $I(\tilde{Q}_n|P_n) \leq C$. Since $\|C^{\tilde{Q}_n} - C_M\|_w^2 \leq S_0^2$ for every n and Q_n is the solution of the calibration problem, $I(Q_n|P_n) \leq S_0^2/\alpha + C < \infty$ for every n. Therefore, by Lemma A.2, $\{Q_n\}$ is tight and, by Prohorov's theorem and Lemma 2.1, weakly relatively compact in $\mathcal{M} \cap \mathcal{L}_B^+$. Choose a subsequence of $\{Q_n\}$ converging weakly to $Q \in \mathcal{M} \cap \mathcal{L}_B^+$. To simplify notation, this subsequence is also denoted by $\{Q_n\}_{n\geq 1}$. It remains to show that Q is indeed a solution of (5.2). We can parameterize the characteristic triplet of any $Q^\phi \in \mathcal{M} \cap \mathcal{L}$ with $I(Q|P) < \infty$ as

$$\left(A, \phi\nu^P, \gamma^\phi := -\frac{A}{2} - \int_{-\infty}^{\infty} (e^x - 1 - h(x))\phi\nu^P(dx)\right),$$

where $\phi \in L^1(\ (|x|^2 \wedge 1)\nu^P(dx)\ )$, $\phi \geq 0$. To prove that Q is a solution of (5.2), we need to establish that

$$\|C^Q - C_M\|_w^2 + \alpha I(Q, P) \leq \|C^{Q^\phi} - C_M\|_w^2 + \alpha I(Q^\phi|P). \tag{6.4}$$

This will be shown in three steps.

Step 1. Let $C_b^{(k)}(\mathbb{R})$ denote the set of continuous bounded functions $\phi : \mathbb{R} \to \mathbb{R}$ equal to k on some neighborhood of 0: for each $\phi \in C_b^{(k)}(\mathbb{R})$ there exists $\delta > 0$ with $\phi(x) = k$ for all $x : |x| < \delta$. The first step is to prove (6.4) for every $\phi \in C_b^{(1)}$. Choose one such ϕ and let Q_n^ϕ denote an element of $\mathcal{M} \cap \mathcal{L}$ with triplet

$$\left(A_n, \phi\nu_n^P, \gamma_n^\phi := -\frac{A_n}{2} - \int_{-\infty}^{\infty} (e^x - 1 - h(x))\phi\nu_n^P(dx)\right).$$

Corollary VII.3.6 in [27] and the fact that $P_n \Rightarrow P$ imply that $Q_n^\phi \Rightarrow Q^\phi$, and therefore by Lemma 3.1,

$$\lim_n \|C^{Q_n^\phi} - C_M\|_w^2 = \|C^{Q^\phi} - C_M\|_w^2.$$

Moreover, $\phi \log \phi + 1 - \phi \in C_b^{(0)}(\mathbb{R})$ and $h(\phi - 1) \in C_b^{(0)}(\mathbb{R})$. Therefore, using once again Corollary VII.3.6 in [27], we obtain (here, we use the hypothesis $\lim A_n = A > 0$)

$$\lim_n I(Q_n^\phi | P_n) = \lim_n \frac{T_\infty}{2A_n} \left\{ \gamma_n^\phi - \gamma_n^P - \int_{-\infty}^{\infty} h(x)(\phi - 1)\nu_n^P(dx) \right\}^2 1_{A_n \neq 0} + \lim_n T_\infty \int_{-\infty}^{\infty} (\phi \log \phi + 1 - \phi)\nu_n^P(dx) = I(Q^\phi | P).$$

Lemma A.3 entails that

$$I(Q, P) \leq \liminf_n I(Q_n, P_n), \tag{6.5}$$

and since, by Lemma 3.1, the pricing error is weakly continuous, we have, using the optimality of Q_n,

$$\begin{aligned} \|C^Q - C_M\|_w^2 + \alpha I(Q, P) &\leq \liminf_n \{\|C^{Q_n} - C_M\|_w^2 + \alpha I(Q_n, P_n)\} \\ &\leq \liminf_n \{\|C^{Q_n^\phi} - C_M\|_w^2 + \alpha I(Q_n^\phi, P_n)\} = \|C^{Q^\phi} - C_M\|_w^2 + \alpha I(Q^\phi, P). \end{aligned} \tag{6.6}$$

This proves (6.4) for all $\phi \in C_b^{(1)}(\mathbb{R})$.

Step 2. Let $\phi \in L^1((|x|^2 \wedge 1)\nu^P)$ such that $\phi \geq 0$ and $|\phi(x) - 1| \leq C(|x| \wedge 1)$ for every $x \in \mathbb{R}$. Then there exists a sequence $\{\phi_n\} \subset C_b^{(1)}(\mathbb{R})$ such that $\phi_n \to \phi$ ν^P-a.e. and $|\phi_n(x) - 1| \leq C(|x| \wedge 1)$ for every n and every $x \in \mathbb{R}$. Then by Step 1,

$$\|C^Q - C_M\|_w^2 + \alpha I(Q, P) \leq \|C^{Q^{\phi_n}} - C_M\|_w^2 + \alpha I(Q^{\phi_n}, P) \quad \forall n. \tag{6.7}$$

Using the dominated convergence theorem and Corollary VII.3.6 in [27] yields that $Q^{\phi_n} \Rightarrow Q^\phi$. Since $|h(x)(\phi_n - 1)| \leq Ch(x)(|x| \wedge 1)$ and

$$\phi_n \log \phi_n + 1 - \phi_n \leq (\phi_n - 1)^2 \leq C^2(|x|^2 \wedge 1),$$

the dominated convergence theorem yields

$$\lim_n I(Q^{\phi_n} | P_n) = \lim_n \frac{T_\infty}{2A} \left\{ \gamma^{\phi_n} - \gamma^P - \int_{-\infty}^{\infty} h(x)(\phi_n - 1)\nu^P(dx) \right\}^2 1_{A \neq 0} + \lim_n T_\infty \int_{-\infty}^{\infty} (\phi_n \log \phi_n + 1 - \phi_n)\nu^P(dx) = I(Q^\phi | P).$$

Therefore, by passing to the limit $n \to \infty$ in (6.7), we obtain that (6.4) holds for every $\phi \in L^1((|x|^2 \wedge 1)\nu^P)$ such that $\phi \geq 0$ and $|\phi(x) - 1| \leq C(|x| \wedge 1)$.

Step 3. Let us now choose a nonnegative $\phi \in L^1((|x|^2 \wedge 1)\nu^P)$. Without loss of generality, we can assume $I(Q^\phi | P) < \infty$. Let

$$\phi_n(x) = \begin{cases} 1 - n(|x| \wedge 1), & \phi(x) < 1 - n(|x| \wedge 1), \\ 1 + n(|x| \wedge 1), & \phi(x) > 1 + n(|x| \wedge 1), \\ \phi(x) & \text{otherwise.} \end{cases}$$

Then $\phi_n \leq \phi \vee 1$ and, once again, the dominated convergence theorem and Corollary VII.3.6 in [27] entail that $Q^{\phi_n} \Rightarrow Q^\phi$. Since $|h(x)(\phi_n - 1)| \leq |h(x)(\phi - 1)|$ and $\phi_n \log \phi_n + 1 - \phi_n \leq \phi \log \phi + 1 - \phi$, again by dominated convergence we obtain that

$\lim_n I(Q^{\phi_n}|P_n) = I(Q^\phi|P)$ and by passage to the limit $n \to \infty$ in (6.7), (6.4) holds for all $\phi \in L^1((|x|^2 \wedge 1)\nu^P)$ with $\phi \geq 0$, which completes the proof of the theorem. □

Another implication of the above theorem is that small changes in the prior Lévy process lead to small changes in the solution: the solution is not very sensitive to minor errors in the determination of the prior measure. This result confirms the empirical observations made in [14].

7. Conclusion. We have proposed here a stable method for constructing an option pricing model of exponential Lévy type, consistent with a given data set of option prices. Our approach is based on the regularization of the calibration problem using the relative entropy with respect to a prior exp-Lévy model as penalization term. The regularization restores existence and stability of solutions; the use of relative entropy links our approach to previous work using relative entropy as a criterion for selection of pricing rules. This technique is readily amenable to numerical implementation, as shown in [14], where empirical applications to financial data are also discussed.

The problem studied here is an example of regularization of a nonlinear, infinite-dimensional inverse problem with noisy data. The above results may also be useful for other nonlinear inverse problems where positivity constraints on the unknown parameter make regularization by relative entropy suitable.

Finally, although we have considered the setting of Lévy processes, this approach can also be adapted to other models with jumps—such as stochastic volatility models with jumps (see [13, Chapter 15] for a review)—where the jump structure is described by a Lévy measure to be retrieved from observations.

Appendix A. Relative entropy for Lévy processes. In this appendix we explicitly compute the relative entropy of two Lévy processes in terms of their characteristic triplets and establish some properties of the relative entropy viewed as a functional on Lévy processes. Under additional assumptions the relative entropy of two Lévy processes was computed in [11] in the case where Q is equivalent to P and the Lévy process has finite exponential moments under P and in [30] in the case where $\log \frac{d\nu^Q}{d\nu^P}$ is bounded. We give here an elementary proof valid for all Lévy processes.

THEOREM A.1 (relative entropy of Lévy processes). *Let $\{X_t\}_{t\geq 0}$ be a real-valued Lévy process on $(\Omega, \mathcal{F}, Q)$ and on $(\Omega, \mathcal{F}, P)$ with respective characteristic triplets (A_Q, ν_Q, γ_Q) and (A_P, ν_P, γ_P). Suppose that $Q \ll P$ (by Theorem* IV.4.39 *in* [27], *this implies that $A^Q = A^P$ and $\nu^Q \ll \nu^P$) and denote $A := A_Q = A_P$. Then for every time horizon $T \leq T_\infty$ the relative entropy of $Q|_{\mathcal{F}_T}$ with respect to $P|_{\mathcal{F}_T}$ can be computed as follows:*

$$I_T(Q|P) = I(Q|_{\mathcal{F}_T}|P|_{\mathcal{F}_T}) = \frac{T}{2A}\left\{\gamma^Q - \gamma^P - \int_{-1}^{1} x(\nu^Q - \nu^P)(dx)\right\}^2 1_{A\neq 0} + T\int_{-\infty}^{\infty}\left(\frac{d\nu^Q}{d\nu^P}\log\frac{d\nu^Q}{d\nu^P} + 1 - \frac{d\nu^Q}{d\nu^P}\right)\nu^P(dx). \tag{A.1}$$

Proof. Let $\{X_t^c\}_{t\geq 0}$ be the continuous martingale part of X under P (a Brownian motion), μ be the jump measure of X, and $\phi := \frac{d\nu^Q}{d\nu^P}$. From [27, Theorem III.5.19], the density process $Z_t := \frac{dQ|_{\mathcal{F}_t}}{dP|_{\mathcal{F}_t}}$ is the Doléans-Dade exponential of the Lévy process $\{N_t\}_{t\geq 0}$ defined by

$$N_t := \beta X_t^c + \int_{[0,t]\times\mathbb{R}} (\phi(x) - 1)\{\mu(ds \times dx) - ds\, \nu^P(dx)\},$$

where β is given by

$$\beta = \begin{cases} \frac{1}{A}\{\gamma^Q - \gamma^P - \int_{|x|\le 1} x(\phi(x)-1)\nu^P(dx)\} & \text{if } A > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Choose $0 < \varepsilon < 1$ and let $I := \{x : \varepsilon \le \phi(x) \le \varepsilon^{-1}\}$. We split N_t into two independent martingales:

$$N'_t := \beta X^c_t + \int_{[0,t]\times I} (\phi(x)-1)\{\mu(ds\times dx) - ds\,\nu^P(dx)\} \quad \text{and}$$

$$N''_t := \int_{[0,t]\times(\mathbb{R}\setminus I)} (\phi(x)-1)\{\mu(ds\times dx) - ds\,\nu^P(dx)\}.$$

Since N' and N'' never jump together, $[N', N'']_t = 0$ and $\mathcal{E}(N'+N'')_t = \mathcal{E}(N')_t\mathcal{E}(N'')_t$ (cf. equation II.8.19 in [27]). Moreover, since N' and N'' are Lévy processes and martingales, their stochastic exponentials are also martingales (Proposition 8.23 in [13]). Therefore,

$$\begin{aligned} I_T(Q|P) &= E^P[Z_T \log Z_T] \\ &= E^P[\mathcal{E}(N')_T\mathcal{E}(N'')_T \log \mathcal{E}(N')_T] + E^P[\mathcal{E}(N')_T\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T] \\ &= E^P[\mathcal{E}(N')_T \log \mathcal{E}(N')_T] + E^P[\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T] \end{aligned} \tag{A.2}$$

if these expectations exist.

Since $\Delta N'_t > -1$ a.s., $\mathcal{E}(N')_t$ is almost surely positive. Therefore, from Lemma 5.8 in [23], $U_t := \log \mathcal{E}(N')_t$ is a Lévy process with the following characteristic triplet:

$$\begin{aligned} A^U &= \beta^2 A, \\ \nu^U(B) &= \nu^P(I \cap \{x : \log\phi(x) \in B\})\ \forall B \in \mathcal{B}(\mathbb{R}), \\ \gamma^U &= -\frac{\beta^2 A}{2} - \int_{-\infty}^{\infty} (e^x - 1 - x1_{|x|\le 1})\nu^U(dx). \end{aligned}$$

This implies that U has bounded jumps and all exponential moments. Therefore, $E[U_T e^{U_T}] < \infty$ and can be computed as follows:

$$\begin{aligned} E^P[U_T e^{U_T}] &= -i\frac{d}{dz}E^P[e^{izU_T}]|_{z=-i} = -iT\psi'(-i)E^P[e^{U_T}] = -iT\psi'(-i) \\ &= T\left(A^U + \gamma^U + \int_{-\infty}^{\infty}(xe^x - x1_{|x|\le 1})\nu^U(dx)\right) \\ &= \frac{\beta^2 AT}{2} + T\int_I (\phi(x)\log\phi(x) + 1 - \phi(x))\nu^P(dx). \end{aligned} \tag{A.3}$$

It remains to compute $E^P[\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T]$. Since N'' is a compound Poisson process, $\mathcal{E}(N'')_t = e^{bt}\prod_{s\le t}(1+\Delta N''_s)$, where $b = \int_{\mathbb{R}\setminus I}(1-\phi(x))\nu^P(dx)$. Let ν'' be the Lévy measure of N'' and λ its jump intensity. Then

$$\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T = bT\mathcal{E}(N'')_T + e^{bT}\prod_{s\le T}(1+\Delta N''_s)\sum_{s\le T}\log(1+\Delta N''_s)$$

and

$$\begin{aligned} &E^P[\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T] \\ &\quad = bT + e^{bT}\sum_{k=0}^{\infty} e^{-\lambda T}\frac{(\lambda T)^k}{k!} E\left[\prod_{s\le T}(1+\Delta N''_s)\sum_{s\le T}\log(1+\Delta N''_s) | k \text{ jumps}\right]. \end{aligned}$$

Since, under the condition that N'' jumps exactly k times in the interval $[0,T]$, the jump sizes are independent and identically distributed, we find, denoting the generic jump size by $\Delta N''$,

$$\begin{aligned}
&E^P[\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T] \\
&\quad = bT + e^{bT} \sum_{k=0}^{\infty} e^{-\lambda T} \frac{(\lambda T)^k}{k!} k E[1+\Delta N'']^{k-1} E[(1+\Delta N'')\log(1+\Delta N'')] \\
&\quad = bT + \lambda T E[(1+\Delta N'')\log(1+\Delta N'')] \\
&\quad = bT + T\int_{-\infty}^{\infty} (1+x)\log(1+x)\nu''(dx) \\
&\quad = T\int_{\mathbb{R}\setminus I} (\phi(x)\log\phi(x) + 1 - \phi(x))\nu^P(dx).
\end{aligned}$$

In particular, $E^P[\mathcal{E}(N'')_T \log \mathcal{E}(N'')_T]$ is finite if and only if the integral in the last line is finite. Combining the above expression with (A.3) and (A.2) finishes the proof. □

LEMMA A.2. *Let $P, \{P_n\}_{n\geq 1} \subset \mathcal{L}_B^+$ for some $B > 0$, such that $P_n \Rightarrow P$. Then for every $r > 0$, the level set $L_r := \{Q \in \mathcal{L} : I(Q|P_n) \leq r \text{ for some } n\}$ is tight.*

Proof. For every $Q \in L_r$, choose any element of $\{P_n\}_{n\geq 1}$ for which $I(Q|P_Q) \leq r$ and denote it by P_Q. The characteristic triplet of Q is denoted by (A^Q, ν^Q, γ^Q) and that of P_Q by $(A^{P_Q}, \nu^{P_Q}, \gamma^{P_Q})$. In addition, we define $\phi^Q := \frac{d\nu^Q}{d\nu^{P_Q}}$. From Theorem A.1,

$$\int_{-\infty}^{\infty} (\phi^Q(x)\log\phi^Q(x) + 1 - \phi^Q(x))\nu^{P_Q}(dx) \leq r/T_\infty.$$

Therefore, for u sufficiently large,

$$\int_{\{\phi^Q > u\}} \phi^Q \nu^{P_Q}(dx) \leq \int_{\{\phi^Q > u\}} \frac{2\phi^Q[\phi^Q\log\phi^Q + 1 - \phi^Q]\nu^{P_Q}(dx)}{\phi^Q\log\phi^Q} \leq \frac{2r}{T_\infty \log u},$$

which entails that for u sufficiently large,

$$\int_{\{\phi^Q > u\}} \nu^Q(dx) \leq \frac{2r}{T_\infty \log u}$$

uniformly with respect to $Q \in L_r$. Let $\varepsilon > 0$ and choose u such that $\int_{\{\phi^Q>u\}} \nu^Q(dx) \leq \varepsilon/2$ for every $Q \in L_r$. By Corollary VII.3.6 in [27],

$$\int_{-\infty}^{\infty} f(x)\nu^{P_n}(dx) \to \int_{-\infty}^{\infty} f(x)\nu^P(dx)$$

for every continuous bounded function f that is identically zero on a neighborhood of zero. Since the measures ν^P and ν^{P_n} for all $n \geq 1$ are finite outside a neighborhood of zero and $P_n \Rightarrow P$, we can choose a compact K such that $\nu^{P_n}(\mathbb{R}\setminus K) \leq \varepsilon/2u$ for every n. Then

$$\nu^Q(\mathbb{R}\setminus K) = \int_{(\mathbb{R}\setminus K)\cap\{\phi^Q\leq u\}} \phi^Q\nu^{P_Q}(dx) + \int_{(\mathbb{R}\setminus K)\cap\{\phi^Q>u\}} \nu^Q(dx) \leq \varepsilon. \tag{A.4}$$

It is easy to check by computing derivatives that for every $u > 0$, on the set $\{x : \phi^Q(x) \leq u\}$,

$$(\phi^Q - 1)^2 \leq 2u(\phi^Q\log\phi^Q + 1 - \phi^Q).$$

Therefore, for u sufficiently large and for all $Q \in L_r$,

$$\begin{aligned}
&\left|\int_{|x|\le 1} x(\phi^Q - 1)\nu^{P_Q}(dx)\right| \\
&\quad\le \left|\int_{|x|\le 1,\ \phi^Q \le u} x(\phi^Q - 1)\nu^{P_Q}(dx)\right| + \left|\int_{|x|\le 1,\ \phi^Q > u} x(\phi^Q - 1)\nu^{P_Q}(dx)\right| \\
&\quad\le \int_{|x|\le 1} x^2\nu^{P_Q}(dx) + \int_{|x|\le 1,\ \phi^Q \le u} (\phi^Q - 1)^2\nu^{P_Q}(dx) + 2\int_{\phi^Q > u} \phi^Q\nu^{P_Q}(dx) \\
&\quad\le \int_{|x|\le 1} x^2\nu^{P_Q}(dx) + 2u\int_{-\infty}^{\infty} (\phi^Q \log\phi^Q + 1 - \phi^Q)\nu^{P_Q}(dx) + \frac{4r}{T_\infty \log u} \\
&\quad\le \int_{|x|\le 1} x^2\nu^{P_Q}(dx) + \frac{3ru}{T_\infty}. \qquad \text{(A.5)}
\end{aligned}$$

By Proposition VI.4.18 in [27], the tightness of $\{P_n\}_{n\ge 1}$ implies that

$$A^{P_n} + \int_{|x|\le 1} x^2\nu^{P_n}(dx) \qquad \text{(A.6)}$$

is bounded uniformly on n, which means that the right-hand side of (A.5) is bounded uniformly with respect to $Q \in L_r$. From Theorem IV.4.39 in [27], $A^Q = A^{P_Q}$ for all $Q \in L_r$ because for the relative entropy to be finite, necessarily $Q \ll P_Q$. Theorem A.1 then implies that

$$\left\{\gamma^Q - \gamma^P - \int_{-1}^{1} x(\nu^Q - \nu^P)(dx)\right\}^2 \le \frac{2A^{P_Q} r}{T_\infty}.$$

From (A.6), A^{P_n} is bounded uniformly on n. Therefore, inequality (A.5) shows that $|\gamma^Q|$ is bounded uniformly with respect to Q. For u sufficiently large,

$$\begin{aligned}
A^Q + \int_{-\infty}^{\infty} (x^2 \wedge 1)\phi^Q\nu^{P_Q}(dx) &\le A^Q + u\int_{\phi^Q \le u} (x^2 \wedge 1)\nu^{P_Q}(dx) \\
&\quad + \int_{\phi^Q > u} \phi^Q\nu^{P_Q}(dx) \le A^{P_Q} + u\int_{-\infty}^{\infty} (x^2 \wedge 1)\nu^{P_Q}(dx) + \frac{2r}{T_\infty \log u}
\end{aligned} \qquad \text{(A.7)}$$

and (A.6) implies that the right-hand side is bounded uniformly with respect to $Q \in L_r$. By Proposition VI.4.18 in [27], (A.4), (A.7), and the fact that $|\gamma^Q|$ is bounded uniformly with respect to Q entail that the set L_r is tight. □

LEMMA A.3. *Let Q and P be two probability measures on $(\Omega, \mathcal{F})$. Then*

$$I(Q|P) = \sup_{f \in C_b(\Omega)} \left\{\int_\Omega f dQ - \int_\Omega (e^f - 1)dP\right\}, \qquad \text{(A.8)}$$

where $C_b(\Omega)$ is space of bounded continuous functions on Ω.

Proof. Observe that

$$\phi(x) = \begin{cases} x\log x + 1 - x, & x > 0, \\ \infty, & x \le 0, \end{cases}$$

and $\phi^*(y) = e^y - 1$ are proper convex functions on $\mathbb{R}$, conjugate to each other, and apply Corollary 2 in [31, p. 538]. □

COROLLARY A.4. *The relative entropy functional $I(Q|P)$ is weakly lower semi-continuous with respect to Q for fixed P.*

LEMMA A.5. *Let $P, \{P_n\}_{n\geq 1} \subset \mathcal{L}_{NA} \cap \mathcal{L}_B^+$ for some $B > 0$ such that $P_n \Rightarrow P$. There exists a sequence $\{Q_n\}_{n\geq 1} \subset \mathcal{M} \cap \mathcal{L}_B^+$ with $Q_n \sim P_n$ for every n and a constant $C < \infty$ such that $I(Q_n|P_n) \leq C$ for every n.*

Proof. Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous truncation function. For $n \geq 1$, let (A_n, ν_n, γ_n) be the characteristic triplet of P_n with respect to h and let

$$f(\beta, P_n) := \gamma_n + \left(\frac{1}{2} + \beta\right) A_n + \int_{-\infty}^{\infty} \left\{(e^x - 1)e^{\beta(e^x-1)} - h(x)\right\} \nu_n(dx).$$

The first step is to show that for every n, there exists a unique β_n such that $f(\beta_n, P_n) = 0$ and that the sequence $\{\beta_n\}_{n\geq 1}$ is bounded.

Since for every n, $P_n \in \mathcal{L}_B^+$, the dominated convergence theorem yields

$$f'_\beta(\beta, P_n) = A_n + \int_{-\infty}^{\infty} (e^x - 1)^2 e^{\beta(e^x-1)} \nu_n(dx) > 0,$$

and since $P_n \in \mathcal{L}_{NA}$, the Lévy process (X, P_n) is not a.s. increasing or a.s. decreasing, which means that at least one of the following conditions holds:

1. $A_n > 0$,
2. $\nu_n((-\infty, 0)) > 0$ and $\nu_n(0, \infty) > 0$,
3. $A_n = 0$, $\nu_n((-\infty, 0)) = 0$ and $\gamma_n - \int_{-\infty}^{\infty} h(x)\nu_n(dx) < 0$,
4. $A_n = 0$, $\nu_n((0, \infty)) = 0$ and $\gamma_n - \int_{-\infty}^{\infty} h(x)\nu_n(dx) > 0$.

Since $f'_\beta(\beta, P_n) \geq A_n + \min(\int_{-\infty}^{0} (e^x - 1)^2 \nu_n(dx), \int_0^{\infty} (e^x - 1)^2 \nu_n(dx))$, if conditions 1 or 2 above hold, $f'_\beta(\beta, P_n)$ is bounded from below by a positive constant and therefore

$$\exists! \beta_n : f(\beta_n, P_n) = 0. \tag{A.9}$$

If condition 3 above holds, $\lim_{\beta\to-\infty} f(\beta, P_n) = \gamma_n - \int_{-\infty}^{\infty} h(x)\nu_n(dx) < 0$ and $\lim_{\beta\to\infty} f(\beta, P_n) = \infty$, which means that (A.9) also holds. The case when condition 4 above is satisfied may be treated similarly.

Let us now show that the sequence $\{\beta_n\}_{n\geq 1}$ is bounded. Rewrite $f(\beta, P_n)$ as

$$\begin{aligned} f(\beta, P_n) := \gamma_n + \left(\frac{1}{2} + \beta\right) \left(A_n + \int_{-\infty}^{\infty} h^2(x)\nu_n(dx)\right) \\ + \int_{-\infty}^{\infty} \left\{(e^x - 1)e^{\beta(e^x-1)} - h(x) - \left(\frac{1}{2} + \beta\right) h^2(x)\right\} \nu_n(dx). \end{aligned} \tag{A.10}$$

Since $(e^x - 1)e^{\beta(e^x-1)} - x - \left(\frac{1}{2} + \beta\right) x^2 = o(|x|^2)$ and the integrand in the last term of (A.10) is bounded on $(-\infty, B]$, by Corollary VII.3.6 in [27], for every β, $\lim_n f(\beta, P_n) = f(\beta, P)$.

Since P also belongs to $\mathcal{L}_B^+ \cap \mathcal{L}_{NA}$, by the same argument as above, there exists a unique β^* such that $f(\beta, P) = 0$ and $f'_\beta(\beta^*, P) > 0$. This means that there exist $\varepsilon > 0$ and finite constants $\beta_- < \beta^*$ and $\beta^+ > \beta^*$ such that $f(\beta_-, P) < -\varepsilon$ and $f(\beta_+, P) > \varepsilon$. One can then find N such that for all $n \geq N$, $f(\beta_-, P_n) < -\varepsilon/2$ and $f(\beta_+, P_n) > \varepsilon/2$, which means that $\beta_n \in [\beta_-, \beta_+]$ and the sequence $\{\beta_n\}$ is bounded. Let (X, Q_n) be the Lévy process with characteristic triplet (with respect to h)

$$\begin{aligned} A_n^Q &= A_n, \qquad \nu_n^Q = e^{\beta_n(e^x-1)}\nu_n, \\ \gamma_n^Q &= \gamma_n + A_n\beta_n + \int_{-\infty}^{\infty} h(x)(e^{\beta(e^x-1)} - 1)\nu_n(dx). \end{aligned}$$

The measure Q_n is in fact the minimal entropy martingale measure for P_n [29], but this result is not used here. From Theorem A.1,

(A.11)
$$I(Q_n|P_n) = -T\left\{\frac{\beta_n}{2}(1+\beta_n)A_n + \beta_n\gamma_n + \int_{-\infty}^{\infty}\{e^{\beta_n(e^x-1)} - 1 - \beta_n h(x)\}\nu_n(dx)\right\}.$$

To show that the sequence $\{I(Q_n|P_n)\}_{n\geq 1}$ is bounded, observe that for

$$\forall x \in [-1,1], \quad \left|e^{\beta(e^x-1)} - 1 - \beta x\right| \leq \beta e^{\beta(e-1)+1}(1+\beta e)|x|^2,$$
$$\forall x \leq B, \quad \left|e^{\beta(e^x-1)} - 1 - \beta x 1_{|x|\leq 1}\right| \leq \beta e^{\beta(e^B+1)} + 1 + \beta B.$$

The uniform boundedness of the sequence of relative entropies now follows from (A.11) and Theorem VI.4.18 in [27]. □

REFERENCES

[1] Y. Achdou, *An inverse problem for a parabolic variational inequality arising in volatility calibration with American options*, SIAM J. Control Optim., 43 (2005), pp. 1583–1615.

[2] L. Andersen and J. Andreasen, *Jump-diffusion models: Volatility smile fitting and numerical methods for pricing*, Rev. Derivatives Research, 4 (2000), pp. 231–262.

[3] M. Avellaneda, *The minimum-entropy algorithm and related methods for calibrating asset-pricing models*, in Proceedings of the International Congress of Mathematicians, (Berlin, 1998), Vol. III, Documenta Mathematica, 1998, pp. 545–563.

[4] M. Avellaneda, *Minimum entropy calibration of asset pricing models*, Internat. J. Theoret. Appl. Finance, 1 (1998), pp. 447–472.

[5] M. Avellaneda, R. Buff, C. Friedman, N. Grandchamp, L. Kruk, and J. Newman, *Weighted Monte Carlo: A new technique for calibrating asset-pricing models*, Internat. J. Theoret. Appl. Finance, 4 (2001), pp. 91–119.

[6] M. Avellaneda, C. Friedman, R. Holmes, and D. Samperi, *Calibrating volatility surfaces via relative entropy minimization*, Appl. Math. Finance, 4 (1997), pp. 37–64.

[7] H. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser Boston, Cambridge, MA, 1989.

[8] D. Bates, *Jumps and stochastic volatility: The exchange rate processes implicit in Deutschemark options*, Rev. Fin. Studies, 9 (1996), pp. 69–107.

[9] S. Ben Hamida and R. Cont, *Recovering volatility from option prices by evolutionary optimization*, J. Comput. Finance, 8 (2005), pp. 43–76.

[10] P. Carr and D. Madan, *Option valuation using the fast Fourier transform*, J. Comput. Finance, 2 (1998), pp. 61–73.

[11] T. Chan, *Pricing contingent claims on stocks driven by Lévy processes*, Ann. Appl. Probab., 9 (1999), pp. 504–528.

[12] R. Cont, *Model uncertainty and its impact on the pricing of derivative instruments*, Math. Finance, to appear.

[13] R. Cont and P. Tankov, *Financial Modelling with Jump Processes*, Chapman and Hall/CRC Press, Boca Raton, FL, 2004.

[14] R. Cont and P. Tankov, *Non-parametric calibration of jump-diffusion option pricing models*, J. Comput. Finance, 7 (2004), pp. 1–49.

[15] R. Cont and E. Voltchkova, *A finite difference scheme for option pricing in jump diffusion and exponential Lévy models*, SIAM J. Numer. Anal., 43 (2005), pp. 1596–1626.

[16] S. Crépey, *Calibration of the local volatility in a trinomial tree using Tikhonov regularization*, Inverse Problems, 19 (2003), pp. 91–127.

[17] I. Csiszar, *I-divergence geometry of probability distributions and minimization problems*, Ann. Probab., 3 (1975), pp. 146–158.

[18] B. Dupire, *Pricing with a smile*, RISK, 7 (1994), pp. 18–20.

[19] N. El Karoui and R. Rouge, *Pricing via utility maximization and entropy*, Math. Finance, 10 (2000), pp. 259–276.

[20] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[21] H. W. Engl and G. Landl, *Convergence rates for maximum entropy regularization*, SIAM J. Numer. Anal., 30 (1993), pp. 1509–1536.
[22] M. Frittelli, *The minimal entropy Martingale measure and the valuation problem in incomplete markets*, Math. Finance, 10 (2000), pp. 39–52.
[23] T. Goll and J. Kallsen, *Optimal portfolios for logarithmic utility*, Stochastic Process. Appl., 89 (2000), pp. 31–48.
[24] T. Goll and L. Rüschendorf, *Minimal distance Martingale measures and optimal portfolios consistent with observed market prices*, in Stochastic Processes and Related Topics (Siegmundsburg, 2000), Stochastics Monogr. 12, Taylor and Francis, London, 2002, pp. 141–154.
[25] J. M. Harrison and S. R. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11 (1981), pp. 215–260.
[26] N. Jackson, E. Süli, and S. Howison, *Computation of deterministic volatility surfaces*, J. Comput. Finance, 2 (1999), pp. 5–32.
[27] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, 2nd ed., Springer, Berlin, 2003.
[28] R. Lagnado and S. Osher, *A technique for calibrating derivative security pricing models: Numerical solution of the inverse problem*, J. Comput. Finance, 1 (1997), pp. 13–25.
[29] Y. Miyahara and T. Fujiwara, *The minimal entropy martingale measures for geometric Lévy processes*, Finance Stoch., 7 (2003), pp. 509–531.
[30] L. Nguyen, *Calibration de Modèles Financiers par Minimisation d'Entropie Relative et Modèles avec Sauts*, Ph.D. thesis, Ecole Nationale de Ponts et Chaussées, Paris, France, 2003.
[31] R. T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[32] M. Rubinstein, *Implied binomial trees*, J. Finance, 49 (1994), pp. 771–819.
[33] D. Samperi, *Calibrating a diffusion pricing model with uncertain volatility: Regularization and stability*, Math. Finance, 12 (2002), pp. 71–87.
[34] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, UK, 1999.
[35] M. Stutzer, *A simple nonparametric approach to derivative security valuation*, J. Finance, 51 (1996), pp. 1633–1652.
[36] P. Tankov, *Lévy Processes in Finance: Inverse Problems and Dependence Modelling*, Ph.D. thesis, Ecole Polytechnique, Palaiseau, France, 2004.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 26–50

DISTANCE TO UNCONTROLLABILITY FOR CONVEX PROCESSES*

ADRIAN LEWIS†, RENÉ HENRION‡, AND ALBERTO SEEGER§

Abstract. The classical study of controllability of linear systems assumes unconstrained control inputs. The "distance to uncontrollability" measures the size of the smallest perturbation to the matrix description of the system rendering it uncontrollable and is a key measure of system robustness. We extend the standard theory of this measure of controllability to the case where the control input must satisfy given linear inequalities. Specifically, we consider the control of differential inclusions, concentrating on the particular case where the control input takes values in a given convex cone.

Key words. convex processes, controllability, rank condition, uncontrollable modes, adjoint processes, cone-constrained controls, distance to uncontrollability

AMS subject classifications. 15A18, 93B05, 47H04, 34A60

DOI. 10.1137/050628982

1. Introduction.

Classical linear control theory concerns a system of the form

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1.1}$$

where, at each time t, the state vector $x(t)$ lies in the space $\mathbb{R}^n$, the input control $u(t)$ lies in the space $\mathbb{R}^m$, and the given matrices A and B are real and of appropriate dimensions. A key question is controllability—whether x can be steered from the origin to an arbitrary point in the state space. To fix the ideas, suppose the input function $u(\cdot)$ is taken from

$$\mathcal{U} = \left\{ u : [0,T] \to \mathbb{R}^m \mid \int_0^T |u(t)|dt < \infty \right\},$$

the space of integrable functions over a prescribed time interval $[0,T]$. The associated trajectory

$$t \mapsto x_{u,A,B}(t) = \int_0^t e^{(t-s)A} Bu(s)ds$$

is then an element of the function space

$$\mathcal{X} = \{x : [0,T] \to \mathbb{R}^n \mid x \text{ is absolutely continuous}\}.$$

Controllability of the linear system (1.1), or equivalently of the pair (A,B), simply means that

$$\left\{ x_{u,A,B}(T) \mid u \in \mathcal{U} \right\} = \mathbb{R}^n.$$

*Received by the editors April 11, 2005; accepted for publication (in revised form) September 27, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sicon/45-1/62898.html

†Cornell University, School of Operations Research and Industrial Engineering, 234 Rhodes Hall, Ithaca, NY 14853 (aslewis@orie.cornell.edu).

‡Weierstrass Institute for Applied Analysis and Stochastics, 10117 Berlin, Germany (henrion@wias-berlin.de)

§University of Avignon, Department of Mathematics, 33, rue Louis Pasteur, 84000 Avignon, France (alberto.seeger@univ-avignon.fr).

For convenience, we see (A, B) not just as a pair of matrices but also as an element of $\mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)$, the space of linear maps from $\mathbb{R}^{n+m}$ to $\mathbb{R}^n$. Spaces of this type are equipped with the operator (or spectral) norm, which we denote by $\|\cdot\|$. Norms in standard Euclidean spaces are denoted simply by $|\cdot|$. To avoid a possible misunderstanding, let us be more explicit:

$$\|(A, B)\| = \sup_{|(s,w)|=1} |As + Bw|.$$

As pointed out by Lee and Markus [25], the set

$$\Xi = \{(A, B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \mid \text{the system (1.1) is uncontrollable}\}$$

is closed. This fact prompted Paige [27] to introduce the number

$$\mu(A, B) = \inf_{(C,D)\in\Xi} \|(A, B) - (C, D)\| \tag{1.2}$$

as measure for the "degree of controllability" of a given (A, B). The number (1.2) indicates how much we need to perturb the system (1.1) in order to destroy its controllability.

The problem of estimating (1.2) is of importance for control theorists and engineers alike. In section 3 we review what has been done already in connection with the evaluation of Paige's distance function $\mu : \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \to \mathbb{R}$. We also clarify a point that remained a bit obscure until now, namely, the difference between real and complex controllability.

The purpose of our work is to go beyond the traditional context of the unconstrained linear model (1.1). As shown in section 5, the discussion becomes more involved when the input function $u(\cdot)$ is subject to constraints. New concepts and tools are needed to handle this more general situation. Sections 7 and 8 are devoted to the controllability analysis of dynamical systems described by convex processes.

The notation that we employ is for the most part standard; however, a partial list is provided for the reader's convenience:

$\mathrm{Im}L = \{Ls \mid s \in \mathbb{R}^n\}$ (range of an operator L defined on $\mathbb{R}^n$),

$\mathrm{Ker}L = \{s \in \mathbb{R}^n \mid Ls = 0\}$ (nullspace of an operator L defined on $\mathbb{R}^n$),

$\mathrm{dist}[z, \Gamma] = \inf_{\gamma\in\Gamma}|z-\gamma|$ (distance from z to the set Γ),

$\mathrm{span}K = K-K$ (space spanned by the cone $K \subset \mathbb{R}^n$),

$\mathrm{lin}K = K\cap -K$ (lineality space of the cone $K \subset \mathbb{R}^n$),

$K^+ = \{q \in \mathbb{R}^n \mid q^T s \geq 0 \ \forall s \in K\}$ (dual cone of $K \subset \mathbb{R}^n$),

$S^\perp = \{q \in \mathbb{R}^n \mid q^T s = 0 \ \forall s \in S\}$ (orthogonal space of $S \subset \mathbb{R}^n$),

$\mathrm{gr}\mathcal{F} = \{(s, v) \in \mathbb{R}^n \times \mathbb{R}^n \mid v \in \mathcal{F}(s)\}$ (graph of a process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$),

$\mathrm{dom}\mathcal{F} = \{s \in \mathbb{R}^n \mid \mathcal{F}(s) \neq \emptyset\}$ (domain of a process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$),

$\mathrm{Im}\mathcal{F} = \cup_{s\in\mathbb{R}^n} \mathcal{F}(s)$ (image of a process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$).

2. The reduction lemma. Controllability is a linear-algebraic property of the matrix pair (A, B), and in this framework, the problem of computing the distance to uncontrollability is a matrix distance problem. As is often the case for such problems, rank-one perturbations are important. We capture the essential idea in the following abstract linear algebra result that plays a ubiquitous role throughout this work. The notation x^T indicates the transpose of the column vector x.

LEMMA 2.1 (reduction lemma). *Let $\Gamma \subset \mathbb{R}^p$ be a nonempty set, $x \in \mathbb{R}^n$ a nonzero vector, $y \in \mathbb{R}^p$, and $F \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$. Then,*

$$\inf_{\substack{E \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p) \\ Ex - y \in \Gamma}} \|E - F\| \;=\; \frac{1}{|x|}\, \mathrm{dist}[Fx - y, \Gamma]. \tag{2.1}$$

Furthermore, if γ is a point in Γ at minimal distance from $Fx - y$, then

$$E = F + \frac{1}{|x|^2}\,(y - Fx + \gamma)x^T \tag{2.2}$$

achieves the infimum on the left-hand side of (2.1).

Proof. Denote by α the term on the left-hand side of (2.1). Then,

$$\begin{aligned}
\alpha = \inf_{\substack{E \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p) \\ Ex - y \in \Gamma}} \sup_{|s|=1} |(E - F)s| \;&\geq\; \inf_{\substack{E \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p) \\ Ex - y \in \Gamma}} \left| (E - F)\left(\frac{x}{|x|}\right)\right| \\
\geq \frac{1}{|x|} \inf_{\substack{E \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p) \\ Ex - y \in \Gamma}} |(Ex - y) - (Fx - y)| \;&\geq\; \frac{1}{|x|}\, \mathrm{dist}[Fx - y, \Gamma].
\end{aligned}$$

To prove the reverse inequality, we find a sequence $\{\gamma_\nu\}_{\nu \in \mathbb{N}}$ in Γ such that

$$|Fx - y - \gamma_\nu| \leq \mathrm{dist}[Fx - y, \Gamma] + \nu^{-1} \quad \forall \nu \in \mathbb{N}.$$

(Recall that Γ is not assumed to be closed.) The corresponding linear map

$$E_\nu = F + \frac{1}{|x|^2}\,(y - Fx + \gamma_\nu)x^T$$

satisfies $E_\nu x - y \in \Gamma$, and therefore

$$\begin{aligned}
\alpha \leq \|E_\nu - F\| = \frac{1}{|x|^2} \sup_{|s|=1} |(y - Fx + \gamma_\nu)x^T s| \\
\leq \frac{|y - Fx + \gamma_\nu|}{|x|} \;\leq\; \frac{\mathrm{dist}[Fx - y, \Gamma] + \nu^{-1}}{|x|}.
\end{aligned}$$

We now let $\nu \to \infty$ and arrive at the desired conclusion. The second part of the lemma is obtained by working with γ instead of the minimizing sequence $\{\gamma_\nu\}_{\nu \in \mathbb{N}}$. ∎

What formula (2.1) says is that our complicated approximation problem in the space $(\mathcal{L}(\mathbb{R}^n, \mathbb{R}^p), \|\cdot\|)$ can be reduced to a simpler approximation problem over the Euclidean space $(\mathbb{R}^p, |\cdot|)$.

3. The unconstrained linear model. Paige's measure of controllability (1.2), while rather natural, is not the most amenable to analysis. We therefore begin our exposition by discussing the easier case first analyzed by Paige, allowing the pair (A, B) to have *complex* entries. Most of the material presented in this section is well known,

but we take the opportunity to clarify some common points of confusion. The original Paige measure of controllability is the distance function $\mu_{\text{complex}} : \mathcal{L}(\mathbb{C}^{n+m}, \mathbb{C}^n) \to \mathbb{R}$ defined by

$$\mu_{\text{complex}}(A,B) = \inf_{\substack{(C,D)\ \text{uncontrollable} \\ (C,D)\in\mathcal{L}(\mathbb{C}^{n+m},\mathbb{C}^n)}} \|(A,B)-(C,D)\|. \tag{3.1}$$

A celebrated result due to Eising [9] asserts that

$$\mu_{\text{complex}}(A,B) = \inf_{z\in\mathbb{C}} \sigma_{\min}[A - zI, B], \tag{3.2}$$

where the term on the right-hand side concerns the minimization of the smallest singular value of the rectangular matrix $[A - zI, B]$ with respect to the complex scalar z. This minimization problem has been extensively studied in the last years (cf. [6, 10, 11, 14, 17]), so we don't indulge in this matter. Suffice it to say that the Eising formula has its root in the Hautus [16] characterization of controllability:

$$(A,B) \text{ is controllable} \quad\Longleftrightarrow\quad \operatorname{rank}[A - zI, B] = n \quad \forall z \in \mathbb{C}.$$

We would like to stress the fact that in this paper we are going to work with control systems described only in terms of real entries. The field of complex numbers is ill adjusted when it comes to conically constrained control systems or, more generally, with convex processes.

As shown by Gracia and de Hoyos [15], even if (A, B) has real entries, the uncontrollable (C, D) achieving the infimum in (3.1) may well have complex entries. The "real" Paige function (1.2) is not just the restriction of μ_{complex} to the real field. The question of estimating the real Paige function can be answered in at least two different ways.

3.1. The approach of DeCarlo and Wicks. In what follows, we identify the set

$$\mathcal{O}(r,n) = \{Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^n) \mid Q^TQ = I\}$$

with the collection of orthonormal matrices of size $n \times r$. The following variational formula involves a minimization over the collection of orthonormal matrices having at most two columns.

Proposition 3.1 (see DeCarlo and Wicks [8]). *Consider a controllable operator $(A, B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)$. Then, one has*

$$\mu(A,B) = \inf_{Q\in\mathcal{O}(1,n)\cup\mathcal{O}(2,n)} \|(Q^TA(I-QQ^T), Q^TB)\|. \tag{3.3}$$

From a computational point of view, formula (3.3) is not very satisfactory because it involves a minimization problem over a complicated set of matrices. Notice that (3.3) can be written in the form

$$\mu(A,B) = \min\{\mu_1(A,B),\ \mu_2(A,B)\},$$

where the term

$$\begin{aligned}\mu_1(A,B) = \inf_{Q\in\mathcal{O}(1,n)} \|(Q^TA(I-QQ^T), Q^TB)\| &= \inf_{|q|=1} \left[|(I-qq^T)A^Tq|^2 + |B^Tq|^2\right]^{1/2} \\ = \inf_{|q|=1}\inf_{\lambda\in\mathbb{R}} \left[|A^Tq-\lambda q|^2 + |B^Tq|^2\right]^{1/2} &= \inf_{\lambda\in\mathbb{R}} \sigma_{\min}[A-\lambda I, B]\end{aligned}$$

is rather easy to evaluate, but the computation of

$$\mu_2(A,B) = \inf_{Q\in\mathcal{O}(2,n)} \|(Q^TA(I-QQ^T), Q^TB)\|$$

remains a difficult task. As observed in [15], the term $\mu_1(A,B)$ is not necessarily equal to $\mu(A,B)$. As a general rule, it is only an upper bound.

3.2. The approach of Hu and Davison. An alternative formula for estimating the real Paige function has been suggested by Hu and Davison [19, 20]. In the proposition stated below, the symbols $\mathcal{R}W$ and $\mathcal{I}W$ refer, respectively, to the real part and the imaginary part of a complex linear map $W \in \mathcal{L}(\mathbb{C}^{n+m}, \mathbb{C}^n)$. The notation $\mathrm{sssv}(E)$ stands for the second-smallest singular value of the matrix E.

PROPOSITION 3.2 (see Hu and Davison [19, 20]). *Consider a controllable operator* $(A,B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)$. *Then,*

$$\mu(A,B) = \inf_{z\in\mathbb{C}} \sup_{\gamma\in]0,1]} \mathrm{sssv}\left(\begin{bmatrix} \mathcal{R}W_z & -\gamma\mathcal{I}W_z \\ \gamma^{-1}\mathcal{I}W_z & \mathcal{R}W_z \end{bmatrix}\right) \tag{3.4}$$

with $W_z = [A - zI, B]$.

Paradoxically, the evaluation of the real Paige function is much more involved than the evaluation of the complex counterpart. This should not be very surprising, however, for readers who have encountered a similar phenomenon while comparing the real stability radius of a matrix to the complex one. (See the survey paper of Hinrichsen and Pritchard [18].)

3.3. Partial perturbations. The case of perturbations in the pair (A,B) is the most popular one, but other situations could be considered as well. It may happen, for instance, that only the component A is subject to perturbations. The partial index

$$\partial_A\mu(A,B) = \inf_{\substack{C\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n) \\ (C,B) \text{ uncontrollable}}} \|A-C\|$$

indicates how much one needs to perturb the first component of (A,B) in order to produce a pair which is uncontrollable. A similar interpretation must be given to the number

$$\partial_B\mu(A,B) = \inf_{\substack{D\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n) \\ (A,D) \text{ uncontrollable}}} \|B-D\|.$$

Later on, these indices are used in the more general context of cone-constrained linear systems (section 5) and control systems governed by convex processes (section 7).

4. Incorporating linear constraints on the input function. Our aim in this work is to extend the classical theory of the distance to uncontrollability to the case where the control u is constrained. As a first, easy but illuminating, step, let us consider the case of linear equality constraints. The works of DeCarlo and Wicks [8] and Hu and Davison [19, 20] can both be extended to the case of a linear system with linear constraints on the input function:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ u(t) \in S. \end{cases} \tag{4.1}$$

Controllability for the model (4.1) simply means that $\{x_{u,A,B}(T) \mid u \in \mathcal{U}_S\} = \mathbb{R}^n$, with

$$\mathcal{U}_S = \{u \in \mathcal{U} \mid u(t) \in S \text{ a.e. on } [0,T]\}.$$

For convenience, we introduce the notation

$$\Xi(S) = \{(A,B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \mid \text{the system (4.1) is uncontrollable}\}.$$

THEOREM 4.1 (transfer theorem). *Let S be an r-dimensional subspace of $\mathbb{R}^m$. Then, the index of controllability*

$$\mu_S(A,B) = \inf_{(C,D)\in\Xi(S)} \|(A,B) - (C,D)\| \tag{4.2}$$

for the model (4.1) *is given simply by*

$$\mu_S(A,B) = \mu(A,BQ), \tag{4.3}$$

where $Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^m)$ is any orthonormal map having S as range.

Proof. The subspace S can be represented as the range of a certain orthonormal map $Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^m)$. By writing the input u in the form $u(t) = Qw(t)$, we arrive at a linear control problem

$$\dot{x}(t) = Ax(t) + BQw(t), \tag{4.4}$$

where the input function w is chosen without restrictions. It is not difficult to see that (4.1) is controllable if and only if the pair (A, BQ) is controllable. This simple but important fact is at the origin of formula (4.3). First, one can write

$$\|(A,B) - (C,D)\| \geq \|(A,BQ) - (C,DQ)\| \qquad \forall (C,D) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)$$

because Q is orthonormal. Thus,

$$\begin{aligned} \mu_S(A,B) &\geq \inf_{\substack{(C,D)\in\mathcal{L}(\mathbb{R}^{n+m},\mathbb{R}^n) \\ (C,DQ) \text{ uncontrollable}}} \|(A,BQ) - (C,DQ)\| \\ &\geq \inf_{\substack{(C,Y)\in\mathcal{L}(\mathbb{R}^{n+r},\mathbb{R}^n) \\ (C,Y) \text{ uncontrollable}}} \|(A,BQ) - (C,Y)\| \;=\; \mu(A,BQ). \end{aligned}$$

For the proof of the reverse inequality $\mu_S(A,B) \leq \mu(A,BQ)$, pick up any solution (C^*, Y^*) to the minimization problem

$$\begin{cases} \text{minimize } \|(A,BQ) - (C,Y)\|, \\ (C,Y) \in \mathcal{L}(\mathbb{R}^{n+r}, \mathbb{R}^n) \text{ uncontrollable.} \end{cases} \tag{4.5}$$

Since the map $D \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n) \mapsto DQ \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^n)$ is surjective, one can write

$$\mu(A,BQ) = \inf_{\substack{D\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n) \\ DQ=Y^*}} \|(A,BQ) - (C^*,DQ)\|. \tag{4.6}$$

We now construct a $D^* \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$D^*Q = Y^* \text{ and } \|(A,BQ) - (C^*, D^*Q)\| = \|(A,B) - (C^*,D^*)\|. \tag{4.7}$$

To see that this is possible, take an orthonormal map $V \in \mathcal{L}(\mathbb{R}^{m-r}, \mathbb{R}^m)$ such that $\mathrm{Im}V = S^\perp$, and define

$$D^* = Y^*Q^T + BVV^T.$$

With this particular choice, one has

$$\begin{aligned} D^*Q &= Y^*Q^TQ + BVV^TQ = Y^*, \\ D^*V &= Y^*Q^TV + BVV^TV = BV. \end{aligned}$$

Hence,

$$\begin{aligned} \|(A,B) - (C^*,D^*)\| &= \sup_{|(s,w)|\le 1} |(A-C^*)s + (B-D^*)w| \\ &= \sup_{|(s,\gamma_1,\gamma_2)|\le 1} |(A-C^*)s + (B-D^*)(Q\gamma_1 + V\gamma_2)| \\ &= \sup_{|(s,\gamma_1)|\le 1} |(A-C^*)s + (B-D^*)Q\gamma_1| \ = \|(A,BQ) - (C^*,D^*Q)\|. \end{aligned}$$

Notice that $(C^*, D^*) \in \Xi(S)$. The combination of (4.6) and (4.7) produces then the desired inequality, completing the proof in this way. □

Remark. The proof technique of the transfer theorem tells us, in fact, how to construct an operator (C^*, D^*) achieving the infimum (4.2) in the definition of $\mu_S(A,B)$. Everything boils down to solving the easier and well-understood minimization problem (4.5).

We end this section with a proposition concerning the partial indices

$$\partial_A\mu_S(A,B) = \inf_{\substack{C\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\(C,B)\in\Xi(S)}} \|A-C\|,$$

$$\partial_B\mu_S(A,B) = \inf_{\substack{D\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)\\(A,D)\in\Xi(S)}} \|B-D\|.$$

As was done in the transfer theorem, it is possible to get rid again of the linear contraint set S.

PROPOSITION 4.2. *Suppose that S is an r-dimensional subspace of $\mathbb{R}^m$ and that $Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^m)$ is an orthonormal map having S as range. Then,*

$$\partial_A\mu_S(A,B) = \partial_A\mu(A,BQ) \quad \textit{and} \quad \partial_B\mu_S(A,B) = \partial_B\mu(A,BQ). \tag{4.8}$$

Proof. We take into account the transformation $u(t) = Qw(t)$ that leads to the unconstrained control system (4.4). One can show straightforwardly the first equality in (4.8), as well as

$$\partial_B\mu_S(A,B) \ge \inf_{\substack{D\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)\\(A,DQ)\text{ uncontrollable}}} \|BQ-DQ\| \ \ge \ \partial_B\mu(A,BQ).$$

For the proof of the reverse inequality $\partial_B\mu_S(A,B) \le \partial_B\mu(A,BQ)$, pick up any solution Y^* to the problem

$$\begin{cases} \text{minimize } \|BQ-Y\| \text{ with respect to} \\ Y \in \mathcal{L}(\mathbb{R}^r,\mathbb{R}^n) \text{ such that } (A,Y) \text{ is uncontrollable,} \end{cases}$$

and observe that

$$\partial_B \mu(A, BQ) = \inf_{\substack{D \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n) \\ DQ = Y^*}} \|BQ - DQ\|.$$

It suffices then to construct a $D^* \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ such that

$$D^*Q = Y^* \quad \text{and} \quad \|BQ - D^*Q\| = \|B - D^*\|.$$

The construction of D^* and the remaining part of the proof is as in Theorem 4.1. □

5. The cone-constrained linear model. In the previous section we saw that restricting controls to take values in a subspace presents no substantial technical difficulties to the classical theory of controllability. In this section we take the next natural step: conical constraints. The problem of controlling a linear system by using positive inputs has been recognized as an important one since the pioneering works of Brammer [5] and Korobov [22] (see also Son [32]).

5.1. Preliminaries. The model under consideration in this section is

$$\left\{ \begin{array}{l} \dot{x}(t) = Ax(t) + Bu(t), \\ u(t) \in P, \end{array} \right. \tag{5.1}$$

where the closed convex cone P is regarded as the set of "positive" elements in $\mathbb{R}^m$. (Typically, P is the positive orthant of $\mathbb{R}^m$.)

Controllability for the model (5.1) is defined in a similar way as before, except that now the contraint set is not the subspace S but the cone P. Controllability of (5.1) implies, of course, controllability of the relaxed control problem

$$\left\{ \begin{array}{l} \dot{x}(t) = Ax(t) + Bu(t), \\ u(t) \in \text{span}P. \end{array} \right. \tag{5.2}$$

Relaxation is a convenient device to be back in a linear setting, where simple and nice controllability tests are available. In what follows, we use the notation

$$\ll A, B, P \gg = B(P) + AB(P) + \cdots + A^{n-1}B(P),$$

where addition of sets is understood in the usual Minkowski sense, and powers of $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ correspond to iterated compositions. Since P is a convex cone, the set $\ll A, B, P \gg$ is also a convex cone and

$$\text{span} \ll A, B, P \gg \; = \; \ll A, B, \text{span}P \gg \; = \; \{\mathcal{C}^{A,B}\gamma : \gamma \in [\text{span}P]^n\}$$

with $\mathcal{C}^{A,B} = [B, AB, \ldots, A^{n-1}B]$ denoting the controllability matrix associated to the pair (A, B). If one represents the space span$P$ as the range of a linear map $Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^m)$, with $r = \dim[\text{span}P]$, then

$$\text{span} \ll A, B, P \gg \; = \; \text{Im}\, \mathcal{C}^{A,BQ}.$$

PROPOSITION 5.1. *The following three conditions are equivalent:*

(i) *the relaxed system* (5.2) *is controllable,*
(ii) $\ll A, B, P \gg$ *spans the whole space* $\mathbb{R}^n$*,*
(iii) $\mathcal{C}^{A,BQ}$ *has full rank.*

Proof. This result is surely well known since it is an obvious extension of Kalman's controllability theorem [21]. □

Unfortunately, the relaxation (or linearization) mechanism $P \mapsto \text{span}P$ destroys part of the information contained in the original model (5.1). For recovering the information that is lost, we introduce the concept of "unilateral uncontrollable mode."

DEFINITION 5.2. *One says that $\lambda \in \mathbb{R}$ is an uncontrollable mode of (A, B) relative to P if*

$$\text{Im}(A - \lambda I) + B(P) \neq \mathbb{R}^n. \tag{5.3}$$

Such an uncontrollable mode λ is declared unilateral if $\text{Im}(A-\lambda I)+B(P)$ *has nonempty interior; otherwise it is declared bilateral.*

If the relaxed system (5.2) is controllable, then we should not worry about the existence of uncontrollable modes of the bilateral type. In fact, one has the next lemma.

LEMMA 5.3. *Suppose that $\ll A, B, P \gg$ spans $\mathbb{R}^n$. Then, (A, B) doesn't have bilateral uncontrollable modes relative to P.*

Proof. This corresponds to a particular case of a more general result stated in section 7, namely, Proposition 7.9. □

That $\ll A, B, P \gg$ spans $\mathbb{R}^n$ doesn't rule out, however, the existence of uncontrollable modes of the unilateral type. This is an important point that deserves to be stressed.

THEOREM 5.4. *Controllability of the cone-constrained linear model* (5.1) *is equivalent to the combination of the following two conditions:*

(i) $\ll A, B, P \gg$ *spans* $\mathbb{R}^n$,

(ii) (A, B) *has no unilateral uncontrollable mode relative to* P.

Proof. According to Brammer [5], controllability of (5.1) is equivalent to the combination of (i) and

$$\left\{ \begin{array}{l} \text{the matrix } A^T \text{ has no (real) eigenvalue with} \\ \text{associated eigenvector in the cone } [B(P)]^+. \end{array} \right. \tag{5.4}$$

Since

$$[B(P)]^+ = \{q \in \mathbb{R}^n \mid B^T q \in P^+\},$$

Brammer's condition (5.4) is just another way of saying that (A, B) has no uncontrollable mode relative to P. Due to Lemma 5.3, bilateral uncontrollable modes can be taken out of the discussion. Indeed, these modes are excluded by the property (i). □

5.2. Divide and conquer. As shown in the above theorem, controllability of a cone-constrained linear model is a concept that can be broken into two different pieces. The first piece is a sort of generalized Kalman's rank condition. It takes into account the span of the cone P, but not the cone itself. This condition is purely linear in the sense that it doesn't recognize the "conic" part of P. The second piece takes care of the possible gap between the cone P and its span. In line with this observation, we split the set

$$\Xi(P) = \{(A, B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \mid \text{the system (5.1) is uncontrollable}\}$$

in two different components:

$$\Xi^{\text{rank}}(P) = \{(A,B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \mid \text{span} \ll A, B, P \gg \neq \mathbb{R}^n\},$$
$$\Xi^{\text{uni}}(P) = \{(A,B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \mid \sigma_P^{\text{uni}}(A,B) \neq \emptyset\}.$$

The notation $\sigma_P^{\text{uni}}(A,B)$ refers, of course, to the set of all unilateral uncontrollable modes of (A,B) relative to P. Since

$$\Xi(P) = \Xi^{\text{rank}}(P) \cup \Xi^{\text{uni}}(P),$$

the index of controllability

$$\mu_P(A,B) = \inf_{(C,D)\in\Xi(P)} \|(A,B) - (C,D)\|$$

for the cone-constrained model (5.1) can be computed by using the rule

$$\mu_P(A,B) = \min\{\mu_P^{\text{rank}}(A,B), \mu_P^{\text{uni}}(A,B)\},$$

where the component indices $\mu_P^{\text{rank}}(A,B)$ and $\mu_P^{\text{uni}}(A,B)$ are defined in an obvious manner.

The evaluation of μ_P^{rank} is the "easy" part of the job. What we have to do is to adjust Hu–Davison's formula to the linearly constrained control system (5.2).

PROPOSITION 5.5. *Let P be an r-dimensional closed convex cone in $\mathbb{R}^m$. Let $Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^m)$ be any orthonormal map such that* span $P =$ Im Q. *Consider an operator $(A,B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)$ such that $\ll A, B, P \gg$ spans $\mathbb{R}^n$. Then,*

$$\mu_P^{\text{rank}}(A,B) = \inf_{z\in\mathbb{C}} \sup_{\gamma\in]0,1]} \text{sssv}\left(\begin{bmatrix} \mathcal{R}W_z & -\gamma\mathcal{I}W_z \\ \gamma^{-1}\mathcal{I}W_z & \mathcal{R}W_z \end{bmatrix}\right) \tag{5.5}$$

with $W_z = [A - zI, BQ]$.

Proof. By definition, μ_P^{rank} is the distance function to the set $\Xi^{\text{rank}}(P)$. Since

$$\mu_P^{\text{rank}}(A,B) = \mu_{\text{span}P}(A,B),$$

it suffices to combine Theorem 4.1 and Proposition 3.2. □

The evaluation of μ_P^{uni} falls beyond the context of Hu–Davison's formula. We no longer seem able to use arguments in the realm of standard linear algebra. The number

$$\mu_P^{\text{uni}}(A,B) = \inf_{(C,D)\in\Xi^{\text{uni}}(P)} \|(A,B) - (C,D)\| \tag{5.6}$$

indicates how much we need to perturb the pair (A,B) if we wish to produce a unilateral uncontrollable mode.

Before trying to compute this number, let us say a few additional words on the set $\Xi^{\text{uni}}(P)$. In the very definition of this set, we use implicitly the expression

$$P^{\oplus} = P^+ \backslash \text{lin}(P^+).$$

We don't know if there is already a name for $P^{\oplus}$, so we call it the pseudo-dual cone of P. Without loss of generality we may suppose that P is not a subspace. If P were a subspace, then $P^{\oplus}$ would be empty, and $\Xi^{\text{uni}}(P)$ would be empty as well. Observe that the cone $P^{\oplus}$ is convex but not necessarily closed.

LEMMA 5.6. *Consider a closed convex cone K in some arbitrary Euclidean space. If K is not a subspace, then one has* $\mathrm{cl}[K\backslash \mathrm{lin}\, K] = K$ *and also* $\mathrm{cl}[K^{\oplus}] = K^{+}$.

Proof. We prove only $K \subset \mathrm{cl}[K\backslash \mathrm{lin} K]$, the reverse inclusion being trivial. Take $c \in K$. Suppose that $c \in \mathrm{lin}(K)$; otherwise we are done. Since K is not a subspace, we can pick up some $c^* \in K\backslash \mathrm{lin} K$ and form

$$c_\alpha = (1-\alpha)c + \alpha c^* \quad \text{with} \quad \alpha \in]0,1[.$$

Since c and c^* are in K, so is the convex combination c_α. The equality

$$c^* = \alpha^{-1} c_\alpha - \alpha^{-1}(1-\alpha)c$$

implies that c_α doesn't belong to $\mathrm{lin}(K)$. Hence, $c_\alpha \in K\backslash \mathrm{lin} K$. The desired conclusion is obtained by letting $\alpha \to 0^+$. □

LEMMA 5.7. *Suppose that P is a closed convex cone but not a subspace. Then,*

$$(C,D) \in \Xi^{\mathrm{uni}}(P) \iff \left\{ \begin{array}{l} \text{one can find } \lambda \in \mathbb{R} \text{ and a unit vector } q \\ \text{such that } C^T q = \lambda q \ \text{ and } D^T q \in P^{\oplus}. \end{array} \right.$$

Proof. The proof is not difficult, and therefore it is omitted. □

PROPOSITION 5.8. *Suppose that P is a closed convex cone but not a subspace. Then, the index* (5.6) *admits the characterization*

$$\mu_P^{\mathrm{uni}}(A,B) = \inf_{\substack{\lambda \in \mathbb{R} \\ |q|=1}} \Big[|A^T q - \lambda q|^2 + \mathrm{dist}^2[B^T q, P^+] \Big]^{1/2}. \tag{5.7}$$

Proof. By using Lemma 5.7, one gets

$$\mu_P^{\mathrm{uni}}(A,B) = \inf_{(C,D)} \ \inf_{\substack{\lambda \in \mathbb{R}, |q|=1 \\ C^T q = \lambda q,\ D^T q \in P^{\oplus}}} \|(A,B)-(C,D)\| \ = \ \inf_{\lambda \in \mathbb{R}, |q|=1} \Psi_{A,B}(\lambda, q)$$

with

$$\Psi_{A,B}(\lambda,q) \ = \ \inf_{\substack{(C,D) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \\ C^T q = \lambda q,\ D^T q \in P^{\oplus}}} \|(A,B)-(C,D)\|.$$

A simple matter of computation shows that

$$\begin{aligned} \Psi_{A,B}(\lambda,q) \ &= \ \inf_{\substack{(C,D) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n) \\ (C^T - \lambda I)q = 0,\ D^T q \in P^{\oplus}}} \left\| \begin{bmatrix} C^T - \lambda I \\ D^T \end{bmatrix} - \begin{bmatrix} A^T - \lambda I \\ B^T \end{bmatrix} \right\| \\ &= \ \inf_{X,Y} \left\| \begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} A^T - \lambda I \\ B^T \end{bmatrix} \right\|, \end{aligned}$$

where the last infimum is taken with respect to

$$\begin{bmatrix} X \\ Y \end{bmatrix} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^{n+m}) \quad \text{such that} \quad \begin{bmatrix} X \\ Y \end{bmatrix} q \in \begin{bmatrix} 0 \\ P^{\oplus} \end{bmatrix}.$$

The reduction lemma yields

$$\Psi_{A,B}(\lambda,q) = \mathrm{dist}\left[\begin{bmatrix} A^T - \lambda I \\ B^T \end{bmatrix} q, \begin{bmatrix} 0 \\ P^{\oplus} \end{bmatrix} \right],$$

and therefore

$$\mu_P^{\mathrm{uni}}(A,B) = \inf_{\substack{\lambda \in \mathbb{R} \\ |q|=1}} \Big[|A^T q - \lambda q|^2 + \mathrm{dist}^2[B^T q, P^{\oplus}] \Big]^{1/2}.$$

But, due to Lemma 5.6, one can change $P^{\oplus}$ by P^+. □

6. Partial perturbations of cone-constrained linear models. Analogously to our earlier discussion of partial perturbations, we might wish to consider perturbing only the matrix A in measuring the distance to uncontrollability of the cone-contrained linear model (5.1). The techniques of the previous section extend in a straightforward manner. The nonnegative real number

$$\partial_A \mu_P^{\text{uni}}(A,B) = \inf_{\substack{C\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ \sigma_P^{uni}(C,B)\neq\emptyset}} \|A-C\|$$

indicates how much one needs to perturb the first component of (A,B) in order to produce a unilateral uncontrollable mode relative to P. A similar interpretation must be given to the number

$$\partial_B \mu_P^{\text{uni}}(A,B) = \inf_{\substack{D\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)\\ \sigma_P^{uni}(A,D)\neq\emptyset}} \|B-D\|.$$

In the next proposition we provide the reader with a recipe for computing these partial indices.

PROPOSITION 6.1. *Suppose that P is a closed convex cone but not a subspace. Then,*

$$\partial_A \mu_P^{\text{uni}}(A,B) = \inf_{\substack{\lambda\in\mathbb{R},|q|=1\\ B^Tq\in P^+}} |A^Tq-\lambda q| \tag{6.1}$$

and

$$\partial_B \mu_P^{\text{uni}}(A,B) = \inf_{\substack{\lambda\in\mathbb{R},|q|=1\\ A^Tq=\lambda q}} \operatorname{dist}[B^Tq,P^+]. \tag{6.2}$$

Proof. Both formulas are obtained by employing a similar proof technique as in Proposition 5.8. By way of example, let us write

$$\partial_A \mu_P^{\text{uni}}(A,B) = \inf_{C\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)} \inf_{\substack{\lambda\in\mathbb{R},|q|=1\\ C^Tq=\lambda q,\ q\in[B(P)]^\oplus}} \|A-C\| \ = \inf_{\substack{\lambda\in\mathbb{R},|q|=1\\ q\in[B(P)]^\oplus}} \Psi_A(\lambda,q) \tag{6.3}$$

with

$$\Psi_A(\lambda,q) \ = \inf_{\substack{C\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ C^Tq=\lambda q,}} \|A-C\| \ = \inf_{\substack{X\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ Xq=0}} \|X-(A^T-\lambda I)\| \ = |A^Tq-\lambda q|.$$

The last equality is obtained, of course, by applying the reduction lemma. Notice that due to Lemma 5.6 and a continuity argument, the last infimum in (6.3) can be written with $[B(P)]^+$ instead of $[B(P)]^\oplus$. □

The partial indices $\partial_A \mu_P^{\text{rank}}$ and $\partial_B \mu_P^{\text{rank}}$ are defined in an obvious manner:

$$\partial_A \mu_P^{\text{rank}}(A,B) \ = \ \partial_A \mu_{\text{span}P}(A,B) = \inf_{\substack{C\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ \ll C,B,\text{span}P\gg\neq\mathbb{R}^n}} \|A-C\|,$$

$$\partial_B \mu_P^{\text{rank}}(A,B) \ = \ \partial_B \mu_{\text{span}P}(A,B) = \inf_{\substack{D\in\mathcal{L}(\mathbb{R}^m,\mathbb{R}^n)\\ \ll A,D,\text{span}P\gg\neq\mathbb{R}^n}} \|B-D\|.$$

The computation of these indices can be carried out with the help of the transfer formulas established in Proposition 4.2.

7. Controllability of convex processes. We can consider the control models we have studied so far in a slightly different light, as controlling differential inclusions of the form $\dot{x} \in Ax + K$ for convex cones K. In the model (5.1), for example, $K = BP$. In this section we broaden this perspective, considering the controllability of a differential inclusion

$$\dot{x}(t) \in \mathcal{F}(x(t)) \tag{7.1}$$

whose right-hand side is a strict closed convex processes $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. That $\mathcal{F}$ is a closed convex process simply means that

$$\mathrm{gr}\mathcal{F} = \{(s, v) \in \mathbb{R}^n \times \mathbb{R}^n \mid v \in \mathcal{F}(s)\}$$

is a closed convex cone. Saying that $\mathcal{F}$ is strict is a short way of indicating that $\mathcal{F}$ is nonempty-valued everywhere, that is to say, $\mathcal{F}(s) \neq \emptyset$ for any $s \in \mathbb{R}^n$.

DEFINITION 7.1. *A strict convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is said to be controllable if the corresponding reachable set*

$$\mathit{Reach}(\mathcal{F}) = \{x(T) \mid x \in \mathcal{X} \text{ solves } (7.1) \text{ and } x(0) = 0\}$$

is the whole space $\mathbb{R}^n$.

7.1. Characterizing controllability. We know exactly what controllability of $\mathcal{F}$ means in terms of the trajectories of its associated differential inclusion, but it would be helpful to have at our disposal some simple algebraic criteria for checking this property. This topic has been handled in a brilliant manner by Aubin, Frankowska and Olech in their 1986 paper [3]. Their contribution admits, however, a certain number of improvements. To put everything in the right perspective, let us start by recalling two algebraic concepts for an arbitrary convex process. The first concept emerges as an extension of the classical rank condition of Kalman.

DEFINITION 7.2. *A convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is said to be reproducing if*

$$\text{there is an integer } k \geq 1 \text{ such that } \mathcal{F}^k(0) \text{ spans } \mathbb{R}^n, \tag{7.2}$$

where the kth power $\mathcal{F}^k = \mathcal{F} \circ \cdots \circ \mathcal{F}$ (k-fold) is understood as an iterated composition in the multivalued sense.

We shall say some extra words on the reproducibility or rank condition (7.2) in a moment. The second concept is an extension of Definition 5.2.

DEFINITION 7.3. *The number $\lambda \in \mathbb{R}$ is called an uncontrollable mode of the convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ if $\mathcal{F} - \lambda I$ is not surjective, that is, if $\mathrm{Im}[\mathcal{F} - \lambda I] \neq \mathbb{R}^n$. The set of uncontrollable modes of $\mathcal{F}$ is denoted by $\sigma(\mathcal{F})$.*

These are the basic ingredients to state the following theorem.

THEOREM 7.4 (see Aubin, Frankowska, and Olech, [3]). *Let $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a strict closed convex process. Then,*

$$\mathcal{F} \text{ is controllable} \iff \mathcal{F} \text{ is reproducing and has no uncontrollable modes.}$$

We mention two ways of rendering this beautiful result even more attractive. First, there is a simple way to characterize the reproducibility condition (7.2). The proposition stated below seems to be new, so we prove it in detail. We rely on two auxiliary lemmas.

LEMMA 7.5. *Consider a strict convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and an integer $k \geq 1$. If the spans of the cones $\mathcal{F}^{k-1}(0)$ and $\mathcal{F}^k(0)$ coincide, then so do the spans of the cones $\mathcal{F}^j(0)$ for exponents $j = k-1, k, k+1, \ldots$.*

Proof. By induction, it suffices to prove the case $j = k+1$. In this case, if the result fails, there is a vector v in the cone $\mathcal{F}^{k+1}(0)$ outside the span of the cone $\mathcal{F}^k(0)$. Choose a vector s in $\mathcal{F}^k(0)$ with $v \in \mathcal{F}(s)$ and a vector w in the relative interior of the cone $\mathcal{F}^{k-1}(0)$. Since s lies in the span of the cone $\mathcal{F}^{k-1}(0)$, the vector $bw + s$ lies in $\mathcal{F}^{k-1}(0)$ for some real $b > 0$ sufficiently large. Since $\mathcal{F}$ is strict, there is a vector z in $\mathcal{F}(w)$ (and hence in $\mathcal{F}^k(0)$). Since the graph of $\mathcal{F}$ is a convex cone, the vector $bz + v$ lies in $\mathcal{F}(bw + s)$, and hence in $\mathcal{F}^k(0)$, contradicting the fact that v lies outside the span of $\mathcal{F}^k(0)$. □

LEMMA 7.6. *Suppose that $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a strict convex process. Then the interiors of the cones $\mathcal{F}^k(0)$ (for exponents $k = n, n+1, \ldots$) are either all empty or all nonempty.*

Proof. The spans of the cones $\mathcal{F}^k(0)$ (for exponents $k = 1, 2, \ldots$) are an increasing sequence of linear subspaces. The previous result implies that equality of two successive elements of the sequence entails constancy thereafter. Hence, by counting dimension, the sequence is constant after at most n elements. The result now follows, since a convex cone has nonempty interior if and only if it spans the whole space. □

Remark. One can construct an easy example showing that $\mathcal{F}^{n+1}(0)$ need not be equal to $\mathcal{F}^n(0)$. Consider, for instance, $n = 2$ and a convex process $\mathcal{F} : \mathbb{R}^2 \to \mathbb{R}^2$ of the form $\mathcal{F}(s) = As + K$, with

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad K = \mathbb{R}_+ \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The angle $\theta > 0$ is chosen small enough. Since K is a ray and A is a rotation matrix, the set

$$\mathcal{F}^k(0) = K + A(K) + \cdots + A^{k-1}(K)$$

reduces to the convex cone generated by the vectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \cos((k-1)\theta) \\ \sin((k-1)\theta) \end{bmatrix}.$$

This happens as long as $(k-1)\theta \leq \pi$, that is, $k \leq 1 + \pi/\theta$. It is only after $k > 1 + \pi/\theta$ that $\mathcal{F}^k(0) = \mathbb{R}^2$ becomes constant. Observe that $1 + \pi/\theta$ goes to infinity as $\theta \to 0^+$, so one can adjust this example to cover the case of an arbitrary power n.

PROPOSITION 7.7. *A strict convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is reproducing if and only if $\mathcal{F}^n(0)$ spans $\mathbb{R}^n$.*

Proof. This follows from the last lemma. □

The second improvement in the presentation of Theorem 7.4 has to do with the nature of uncontrollable modes. The elements of $\sigma(\mathcal{F})$ can be partioned into two different categories. One says that $\lambda \in \sigma(\mathcal{F})$ is of the unilateral type if $\mathrm{Im}[\mathcal{F} - \lambda I]$ has nonempty interior; otherwise, it is declared of the bilateral type. In short, one has a partition

$$\sigma(\mathcal{F}) = \sigma^{\mathrm{uni}}(\mathcal{F}) \cup \sigma^{\mathrm{bi}}(\mathcal{F}),$$

where the notation is self-explanatory.

The different types of uncontrollability modes are perhaps better understood if we characterize them in terms of the adjoint process of F. Recall that the *adjoint* (or transpose) of the convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the convex process $\mathcal{F}^* : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by

$$\mathrm{gr}\mathcal{F}^* = \{(q,p) \in \mathbb{R}^n \times \mathbb{R}^n \mid (-p,q) \in [\mathrm{gr}\mathcal{F}]^+\},$$

that is,

$$(q,p) \in \mathrm{gr}\mathcal{F}^* \iff \langle p,s\rangle \leq \langle q,v\rangle \quad \forall (s,v) \in \mathrm{gr}\mathcal{F}.$$

We assume that the reader is familiar with this transposition mechanism [2, 4, 29]. As observed already in Proposition 2.4 of [30], the convex cone $\mathrm{Im}[\mathcal{F} - \lambda I]$ is related to

$$(\mathcal{F}^* - \lambda I)^{-1}(0) = \{q \in \mathbb{R}^n \mid \lambda q \in \mathcal{F}^*(q)\}$$

by means of the duality formula

$$(\mathcal{F}^* - \lambda I)^{-1}(0) = [\mathrm{Im}(\mathcal{F} - \lambda I)]^+ . \tag{7.3}$$

As a consequence of (7.3), it is clear that an uncontrollable mode of $\mathcal{F}$ is exactly the same thing as an eigenvalue of $\mathcal{F}^*$. In short,

$$\sigma(\mathcal{F}) = \Lambda(\mathcal{F}^*) \tag{7.4}$$

with

$$\Lambda(\mathcal{F}^*) = \{\lambda \in \mathbb{R} \mid \lambda q \in \mathcal{F}^*(q) \ \text{ for some } \ q \neq 0\}$$

denoting the (point) spectrum of $\mathcal{F}^*$. General information on point spectra of convex processes can be found, for instance, in [1, 23, 24]. For bilateral uncontrollable modes, one has the next lemma.

LEMMA 7.8. *Consider a convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. For $\lambda \in \mathbb{R}$, the following three conditions are equivalent:*

(i) *λ is a bilateral uncontrollable mode of $\mathcal{F}$,*

(ii) *the convex cone $(\mathcal{F}^* - \lambda I)^{-1}(0)$ contains a line,*

(iii) *there is a unit vector $q \in \mathbb{R}^n$ such that $\lambda q \in \mathcal{F}^*(q)$ and $-\lambda q \in \mathcal{F}^*(-q)$.*

Proof. The equivalence between (ii) and (iii) is straightforward. The equivalence between (i) and (ii) is again a consequence of the duality formula (7.3). □

Remark. A vector q as in Lemma 7.8 (iii) is called a bilateral eigenvector of $\mathcal{F}^*$. The concept of bilateral eigenvector is used by Gajardo and Seeger [13] in connection with the asymptotic stability analysis of discrete-time evolution systems governed by convex processes.

PROPOSITION 7.9. *For a strict convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, one has the implication*

$$\sigma^{\mathrm{bi}}(\mathcal{F}) \neq \emptyset \quad \Longrightarrow \quad \mathrm{int}[\mathcal{F}^k(0)] = \emptyset \quad \forall k \geq 1.$$

In particular, if a strict convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is reproducing, then it has no bilateral uncontrollable modes.

Proof. Take $\lambda \in \sigma^{\mathrm{bi}}(\mathcal{F})$. By Lemma 7.8, there is a unit vector $q \in \mathbb{R}^n$ such that

$$\lambda q \in \mathcal{F}^*(q), \quad -\lambda q \in \mathcal{F}^*(-q). \tag{7.5}$$

We claim that for every $k \in \mathbb{N}$, one has

$$\lambda^k q \in (\mathcal{F}^*)^k(q), \quad -\lambda^k q \in (\mathcal{F}^*)^k(-q). \tag{7.6}$$

The proof is carried out by using an induction argument. The case $k = 1$ corresponds to (7.5). Suppose that (7.6) is true for a given k, and let us examine the situation for $k + 1$. One has

$$(\mathcal{F}^*)^{k+1}(q) = \mathcal{F}^*[(\mathcal{F}^*)^k(q)] = \cup_{z\in(\mathcal{F}^*)^k(q)}\mathcal{F}^*(z) \supset \mathcal{F}^*(\lambda^k q),$$

and similarly

$$(\mathcal{F}^*)^{k+1}(-q) = \mathcal{F}^*[(\mathcal{F}^*)^k(-q)] = \cup_{z\in(\mathcal{F}^*)^k(-q)}\mathcal{F}^*(z) \supset \mathcal{F}^*(-\lambda^k q).$$

We now use the fact that $\mathcal{F}^*$ is positively homogeneous. If $\lambda^k \geq 0$, then one can write

$$\lambda^k \mathcal{F}^*(q) \subset (\mathcal{F}^*)^{k+1}(q), \quad \lambda^k \mathcal{F}^*(-q) \subset (\mathcal{F}^*)^{k+1}(-q).$$

If $\lambda^k < 0$, then one gets

$$-\lambda^k \mathcal{F}^*(-q) \subset (\mathcal{F}^*)^{k+1}(q), \quad -\lambda^k \mathcal{F}^*(q) \subset (\mathcal{F}^*)^{k+1}(-q).$$

In either case, one obtains

$$\lambda^{k+1} q \in (\mathcal{F}^*)^{k+1}(q), \quad -\lambda^{k+1} q \in (\mathcal{F}^*)^{k+1}(-q),$$

proving in this way our claim. In fact, we don't use the full power of (7.6). We just observe that

$$q \in \mathrm{dom}(\mathcal{F}^*)^k \cap -\mathrm{dom}(\mathcal{F}^*)^k,$$

that is, $\mathrm{dom}(\mathcal{F}^*)^k$ is a convex cone containing a line. By invoking the duality formula

$$[\mathcal{F}^k(0)]^+ = \mathrm{dom}(\mathcal{F}^*)^k$$

of Phat [28, Prop. 2.5], we conclude that $\mathcal{F}^k(0)$ has empty interior. □

In view of Propositions 7.7 and 7.9, the Aubin–Frankowska–Olech controllability theorem can be reformulated in the following form.

COROLLARY 7.10. *Suppose that $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a strict closed convex process. Then,*

$$\mathcal{F} \textit{ is controllable } \iff \mathcal{F}^n(0) \textit{ spans } \mathbb{R}^n \textit{ and } \sigma^{\mathrm{uni}}(\mathcal{F}) \textit{ is empty}$$

7.2. Checking reproducibility. Deviating momentarily from the main stream of the discussion, we make some comments concerning the concept of reproducibility.

The relaxation mechanism (5.2) introduced in section 5.1 can be extended to the framework of a differential inclusion whose right-hand side is a general convex process.

DEFINITION 7.11. *The linear relaxation of a convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is defined as the multivalued operator $\mathcal{F}^{rel} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ whose graph is given by*

$$\mathrm{gr}\mathcal{F}^{rel} = \mathrm{gr}\mathcal{F} - \mathrm{gr}\mathcal{F}.$$

Said in another way, the graph of $\mathcal{F}^{rel}$ is the linear subspace spanned by the convex cone $\mathrm{gr}\mathcal{F}$.

A more explicit formula for $\mathcal{F}^{\rm rel}$ is given by

$$\mathcal{F}^{\rm rel}(s) = \bigcup_{s_2 - s_1 = s} \mathcal{F}(s_2) - \mathcal{F}(s_1) \qquad \forall s \in \mathbb{R}^n. \tag{7.7}$$

By construction, the multivalued operator $\mathcal{F}^{\rm rel}$ is linear in the sense that

$$\mathcal{F}^{\rm rel}(\alpha_1 s_1 + \alpha_2 s_2) = \alpha_1 \mathcal{F}^{\rm rel}(s_1) + \alpha_2 \mathcal{F}^{\rm rel}(s_2)$$

$\forall s_1, s_2 \in {\rm dom}\mathcal{F}^{\rm rel}$ and nonzero $\alpha_1, \alpha_2 \in \mathbb{R}$. (That $\alpha_1, \alpha_2 \in \mathbb{R}$ are nonzero scalars is of importance and should not be neglected.) General information on the theory of linear multivalued operators can be found in the book by Cross [7].

We declare the differential inclusion

$$\dot{x}(t) \in \mathcal{F}^{\rm rel}(x(t)) \tag{7.8}$$

as being the relaxed version of the control model (7.1). As we shall see in the next theorem, reproducibility of $\mathcal{F}$ is equivalent to controllability of (7.8). First, we state the next lemma.

LEMMA 7.12. *For a strict closed convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, one has*

(i) $\mathcal{F}(s_2) - \mathcal{F}(s_1) \subset \mathcal{F}(s_2 - s_1) - \mathcal{F}(0) \quad \forall s_1, s_2 \in \mathbb{R}^n$,
(ii) $\mathcal{F}^{\rm rel}(s) \subset \mathcal{F}(s) - \mathcal{F}(0) \quad \forall s \in \mathbb{R}^n$,
(iii) $\mathcal{F}(s) \subset \mathcal{F}^{\rm rel}(s) \quad \forall s \in \mathbb{R}^n$,
(iv) $(\mathcal{F}^{\rm rel})^k(0) = {\rm span}\mathcal{F}^k(0) \quad \forall k \in \mathbb{N}$.

Proof. For proving part (i), take $s_1, s_2 \in \mathbb{R}^n$ and $v \in \mathcal{F}(s_2) - \mathcal{F}(s_1)$. Write

$$v = v_2 - v_1 \quad \text{with} \quad v_1 \in \mathcal{F}(s_1), v_2 \in \mathcal{F}(s_2).$$

By strictness of $\mathcal{F}$, we can find some element w in $\mathcal{F}(-s_1)$. Since the graph of $\mathcal{F}$ is a convex cone, it follows that

$$v_1 + w \in \mathcal{F}(0), \quad v_2 + w \in \mathcal{F}(s_2 - s_1).$$

Hence,

$$v = (v_2 + w) - (v_1 + w) \ \in \mathcal{F}(s_2 - s_1) - \mathcal{F}(0).$$

Part (ii) follows immediately from (i) and formula (7.7). Part (iii) is trivial because ${\rm gr}\mathcal{F} \subset {\rm gr}\mathcal{F}^{\rm rel}$. The proof of (iv) is more subtle and is based on an induction argument. For $k = 1$, the result is true because the equality

$$\mathcal{F}^{\rm rel}(0) = {\rm span}\mathcal{F}(0)$$

is obtained by combining (ii) and (iii). Suppose the announced formula is true for a given k. For $k+1$, one gets

$$(\mathcal{F}^{\rm rel})^{k+1}(0) = \bigcup_{v \in (\mathcal{F}^{\rm rel})^k(0)} \mathcal{F}^{\rm rel}(v) = \bigcup_{v \in {\rm span}\mathcal{F}^k(0)} \mathcal{F}^{\rm rel}(v) = \bigcup_{v_1, v_2 \in \mathcal{F}^k(0)} \mathcal{F}^{\rm rel}(v_2 - v_1)$$
$$= \bigcup_{v_1, v_2 \in \mathcal{F}^k(0)} \Big\{ \mathcal{F}^{\rm rel}(v_2) - \mathcal{F}^{\rm rel}(v_1) \Big\} \subset \bigcup_{v_1, v_2 \in \mathcal{F}^k(0)} \Big\{ [\mathcal{F}(v_2) - \mathcal{F}(0)] - [\mathcal{F}(v_1) - \mathcal{F}(0)] \Big\}.$$

Therefore,

$$(\mathcal{F}^{\rm rel})^{k+1}(0) \subset [\mathcal{F}(\mathcal{F}^k(0)) - \mathcal{F}(0)] - [\mathcal{F}(\mathcal{F}^k(0)) - \mathcal{F}(0)] = {\rm span}\mathcal{F}^{k+1}(0) + {\rm span}\mathcal{F}(0).$$

The subspace $\mathrm{span}\mathcal{F}(0)$ can be dropped from the last sum because it is contained in $\mathrm{span}\mathcal{F}^{k+1}(0)$. We have proved in this way the inclusion $(\mathcal{F}^{\mathrm{rel}})^{k+1}(0) \subset \mathrm{span}\mathcal{F}^{k+1}(0)$. The reverse inclusion is trivial because the convex cone $\mathcal{F}^{k+1}(0)$ is smaller than the linear space $(\mathcal{F}^{\mathrm{rel}})^{k+1}(0)$. □

LEMMA 7.13. *Let $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a strict closed convex process. Then, $\mathcal{F}^{\mathrm{rel}}$ admits the representation*

$$\mathcal{F}^{\mathrm{rel}}(s) = A^{\mathcal{F}}s + S^{\mathcal{F}} \qquad \forall s \in \mathbb{R}^n, \tag{7.9}$$

where $S^{\mathcal{F}} = \mathrm{span}\mathcal{F}(0)$, and $A^{\mathcal{F}} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is defined by

$$A^{\mathcal{F}}s = \pi^{\mathcal{F}}[\mathcal{F}^{\mathrm{rel}}(s)] \tag{7.10}$$

with $\pi^{\mathcal{F}} : \mathbb{R}^n \to \mathbb{R}^n$ denoting the orthogonal projection onto $[\mathrm{span}\mathcal{F}(0)]^{\perp}$.

Proof. It must be observed that $\pi^{\mathcal{F}}[\mathcal{F}^{\mathrm{rel}}(s)]$ is indeed a singleton. To see this, take $y_1, y_2 \in \pi^{\mathcal{F}}[\mathcal{F}^{\mathrm{rel}}(s)]$ and write

$$y_1 = \pi^{\mathcal{F}}(v_1), \quad y_2 = \pi^{\mathcal{F}}(v_2) \quad \text{with } v_1, v_2 \in \mathcal{F}^{\mathrm{rel}}(s).$$

Hence,

$$y_1 - y_2 = \pi^{\mathcal{F}}(v_1) - \pi^{\mathcal{F}}(v_2) = \pi^{\mathcal{F}}(v_1 - v_2) = 0,$$

the last equality being due to the fact that

$$v_1 - v_2 \in \mathcal{F}^{\mathrm{rel}}(s) - \mathcal{F}^{\mathrm{rel}}(s) = \mathcal{F}^{\mathrm{rel}}(0) = \mathrm{span}\mathcal{F}(0).$$

Checking the linearity of the single-valued operator $A^{\mathcal{F}}$ is essentially a matter of exploiting the linearity of the multivalued operator $\mathcal{F}^{\mathrm{rel}}$. The details are omitted. Finally, we check the representation formula (7.9). Take $s \in \mathbb{R}^n$ and $y \in A^{\mathcal{F}}s + S^{\mathcal{F}}$. Thus,

$$y = q_1 + \pi^{\mathcal{F}}(v) \quad \text{with } q_1 \in S^{\mathcal{F}},\ v \in \mathcal{F}^{\mathrm{rel}}(s).$$

Since $q_2 = v - \pi^{\mathcal{F}}(v) \in S^{\mathcal{F}}$, it follows that

$$y = q_1 - q_2 + v \in S^{\mathcal{F}} + \mathcal{F}^{\mathrm{rel}}(s) = \mathcal{F}^{\mathrm{rel}}(0) + \mathcal{F}^{\mathrm{rel}}(s) \subset \mathcal{F}^{\mathrm{rel}}(s).$$

Conversely, take $y \in \mathcal{F}^{\mathrm{rel}}(s)$. Since $y - \pi^{\mathcal{F}}(y) \in S^{\mathcal{F}}$, it follows that

$$y = \pi^{\mathcal{F}}(y) + [y - \pi^{\mathcal{F}}(y)] \in A^{\mathcal{F}}s + S^{\mathcal{F}}. \qquad \square$$

Remark. The operator $A^{\mathcal{F}}$ used to represent $\mathcal{F}^{\mathrm{rel}}$ is not unique. In fact, one has $\mathcal{F}^{\mathrm{rel}}(\cdot) = A(\cdot) + S^{\mathcal{F}}$ for any $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ with $\mathrm{Im}(A - A^{\mathcal{F}}) \subset S^{\mathcal{F}}$. Such A is called a linear selector of $\mathcal{F}^{\mathrm{rel}}$. We declare $A^{\mathcal{F}}$ to be the standard linear selector of $\mathcal{F}^{\mathrm{rel}}$.

In view of Lemma 7.13, the relaxed version of the differential inclusion (7.1) can be written in the form

$$\begin{cases} \dot{x}(t) = A^{\mathcal{F}}x(t) + u(t), \\ u(t) \in S^{\mathcal{F}}, \end{cases} \tag{7.11}$$

a model that is well understood by now. Such linearly constrained control problem can also be written in the unconstrained form

$$\begin{cases} \dot{x}(t) = A^{\mathcal{F}}x(t) + Qw(t), \\ w(t) \in \mathbb{R}^r, \end{cases} \tag{7.12}$$

where r is the dimension of $S^{\mathcal{F}}$, and $Q \in \mathcal{L}(\mathbb{R}^r, \mathbb{R}^n)$ is any orthonormal map such that $\mathrm{Im}Q = S^{\mathcal{F}}$.

THEOREM 7.14. *For a strict closed convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, the following five conditions are equivalent:*

(i) *$\mathcal{F}$ is reproducing,*
(ii) *$\mathcal{F}^{\mathrm{rel}}$ is reproducing,*
(iii) *the system (7.11) is controllable,*
(iv) *$(A^{\mathcal{F}}, Q)$ is controllable for some $Q \in \mathcal{O}(r,n)$ such that $\mathrm{Im}\, Q = S^{\mathcal{F}}$,*
(v) *$(A^{\mathcal{F}}, Q)$ is controllable for every $Q \in \mathcal{O}(r,n)$ such that $\mathrm{Im}\, Q = S^{\mathcal{F}}$.*

Proof. The equivalence between (i) and (ii) is a consequence of Lemma 7.12(iv). From the controllability theory of linear systems, we know that the conditions (iii), (iv), and (v) are all equivalent to (ii). □

8. Additive versus hybrid perturbations. Equipped with the characterizations of controllability of process differential inclusions we explored in the previous section, we can now return to our central topic of measuring the degree of controllability. Robustness of controllability for a system like (7.1) is a topic that has been studied by Naselli-Ricceri [26], Tuan [33], and Lavilledieu and Seeger [24]. Here we go beyond the qualitative analysis carried out by these authors and focus attention on the quantitative aspect. We want to measure how much we need to perturb the system (7.1) in order to destroy its controllability.

8.1. Additive perturbations. The simplest way to perturb the differential inclusion (7.1) is to add a linear map $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ to the reference or nominal operator $\mathcal{F}$. The perturbed system

$$\dot{x}(t) \in (\mathcal{F} + L)(x(t)) \tag{8.1}$$

may no longer be controllable if the perturbation L is too severe. The index

$$\mu_{\mathrm{add}}(\mathcal{F}) = \inf_{\substack{L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \\ \mathcal{F}+L \text{ uncontrollable}}} \|L\| \tag{8.2}$$

speaks by itself and doesn't need further explanation. In line with the "divide and conquer" strategy adopted in this work, we write

$$\mu_{\mathrm{add}}(\mathcal{F}) = \min\{\mu^{\mathrm{rank}}_{\mathrm{add}}(\mathcal{F}), \mu^{\mathrm{uni}}_{\mathrm{add}}(\mathcal{F})\},$$

where

$$\mu^{\mathrm{rank}}_{\mathrm{add}}(\mathcal{F}) = \inf_{\substack{L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \\ \mathcal{F}+L \text{ irreproducing}}} \|L\| \tag{8.3}$$

measures the distance to irreproducibility, and

$$\mu^{\mathrm{uni}}_{\mathrm{add}}(\mathcal{F}) = \inf_{\substack{L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \\ \sigma^{\mathrm{uni}}(\mathcal{F}+L) \neq \emptyset}} \|L\| \tag{8.4}$$

indicates how much we need to perturb $\mathcal{F}$ in order to produce a unilateral uncontrollable mode.

In the proof of the theorem stated below, we use the notation $\mathcal{F}^{\odot}$ to indicate the *pseudo-adjoint* of the convex process $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. By definition, $\mathcal{F}^{\odot} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the convex process given by $\mathrm{gr}\mathcal{F}^{\odot} = \mathrm{gr}\mathcal{F}^* \backslash \mathrm{lin}[\mathrm{gr}\mathcal{F}^*]$, or, more explicitly,

$$\mathcal{F}^{\odot}(q) = \mathcal{F}^*(q) \backslash -\mathcal{F}^*(-q) \quad \forall q \in \mathbb{R}^n.$$

THEOREM 8.1. *Suppose that $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a strict closed convex process. Then,*

$$\mu^{\rm uni}_{\rm add}(\mathcal{F}) = \begin{cases} \inf\limits_{\substack{\lambda\in\mathbb{R}\\ |q|=1}} {\rm dist}[\lambda q, \mathcal{F}^*(q)] & \text{if } {\rm gr}\mathcal{F} \text{ is not a subspace},\\ \infty & \text{otherwise.}\end{cases} \tag{8.5}$$

On the other hand,

$$\mu^{\rm rank}_{\rm add}(\mathcal{F}) = \partial_A \mu(A^{\mathcal{F}}, Q) \tag{8.6}$$

with $Q \in \mathcal{O}(r,n)$ such that Im $Q = {\rm span}\mathcal{F}(0)$ *and $A^{\mathcal{F}}$ denoting the standard linear selector of $\mathcal{F}^{\rm rel}$.*

Proof. The formula (8.5) is based on the fact that $\sigma^{\rm uni}(\mathcal{F}+L) \neq \emptyset$ if and only if there exist a scalar $\lambda \in \mathbb{R}$ and a unit vector $q \in \mathbb{R}^n$ such that

$$\begin{cases} \lambda q \in \mathcal{F}^*(q) + L^T q,\\ -\lambda q \notin \mathcal{F}^*(-q) - L^T q.\end{cases}$$

The above condition can be written in the more compact form $L^T q - \lambda q \in -\mathcal{F}^{\odot}(q)$. Hence,

$$\mu^{\rm uni}_{\rm add}(\mathcal{F}) = \inf_{\substack{\lambda\in\mathbb{R}\\ |q|=1}} \inf_{\substack{L\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ L^T q-\lambda q\in-\mathcal{F}^{\odot}(q)}} \|L\|.$$

By applying the reduction lemma, one obtains

$$\mu^{\rm uni}_{\rm add}(\mathcal{F}) = \inf_{\substack{\lambda\in\mathbb{R}\\ |q|=1}} {\rm dist}[\lambda q, \mathcal{F}^{\odot}(q)]. \tag{8.7}$$

Both terms in (8.7) are equal to ∞ if ${\rm gr}\mathcal{F}$ is a subspace. Suppose then that ${\rm gr}\mathcal{F}$ is not a subspace. Since $\mathcal{F}$ is a strict closed convex process, it follows that $\mathcal{F}^*(0) = \{0\}$. Hence,

$$\mathcal{F}^*(q) + \mathcal{F}^*(-q) \subset \{0\}.$$

From this relation, one can see that

$${\rm dom}\mathcal{F}^{\odot} = {\rm dom}\mathcal{F}^* \backslash {\rm lin}[{\rm dom}\mathcal{F}^*] \quad \text{and} \quad \mathcal{F}^{\odot}(q) = \mathcal{F}^*(q) \ \ \forall q \in {\rm dom}\mathcal{F}^{\odot}.$$

It has to be shown that, for arbitrary $\lambda \in \mathbb{R}$ and unit vector $q \in {\rm dom}\mathcal{F}^*$, one has

$${\rm dist}[\lambda q, \mathcal{F}^*(q)] \geq \mu^{\rm uni}_{\rm add}(\mathcal{F}).$$

To do this, we take into account (8.7) and the following two facts. First, due to Lemma 5.6, every unit vector in ${\rm dom}\mathcal{F}^*$ can be obtained as limit of a sequence of unit vectors taken from ${\rm dom}\mathcal{F}^* \backslash {\rm lin}[{\rm dom}\mathcal{F}^*]$. Second, since $\mathcal{F}$ is strict, $\mathcal{F}^*$ maps bounded sets to bounded sets and $\mathcal{F}^*$ is single-valued over ${\rm lin}[{\rm dom}\mathcal{F}^*]$ (cf. Corollary 2.5.8 in [2]). As far as (8.6) is concerned, one follows a similar proof technique as in the more general situation discussed in Theorem 8.3. □

8.2. Hybrid perturbations. Perturbing a differential inclusion by adding a linear map to the right-hand side is not the most general perturbational scheme that one may consider. In fact, a perturbational scheme of the additive type is poorly suited to deal with a large number of important situations occurring in practice. To

see this, just think of the particular case of the unconstrained linear control problem (1.1). This model can be represented in terms of the convex process $\mathcal{F}_{A,B} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ given by

$$\mathcal{F}_{A,B}(s) = As + \mathrm{Im}B.$$

By adding a linear map $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ to the convex process $\mathcal{F}_{A,B}$ one recovers a perturbed system

$$\dot{x}(t) = (A + L)x(t) + Bu(t)$$

for which the B component remains unaffected. A more sophisticated operation must be carried out on $\mathcal{F}_{A,B}$ if one wishes to incorporate perturbations in the B component as well.

The concept of hybrid perturbation is based on the simultaneous use of two linear maps, say, $M \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ and $L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, the first acting in an multiplicative way and the second in a additive way. The new convex process

$$s \in \mathbb{R}^n \mapsto [M \circ \mathcal{F} + L](s) = M(\mathcal{F}(s)) + Ls$$

is viewed as a perturbed version of $\mathcal{F}$. Of course, perturbation doesn't occur if one takes $(M, L) = (I, 0)$. All this is for saying that

$$\mu_{\mathrm{hyb}}(\mathcal{F}) = \inf_{\substack{M,L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \\ M \circ \mathcal{F} + L \text{ uncontrollable}}} \|(M, L) - (I, 0)\| \tag{8.8}$$

is a reasonable candidate for measuring the degree of controllability of $\mathcal{F}$. Observe, incidentally, that hybrid perturbations preserve the strictness of $\mathcal{F}$.

We follow once more our old habit of thought and decompose (8.8) in the form

$$\mu_{\mathrm{hyb}}(\mathcal{F}) = \min\{\mu^{\mathrm{rank}}_{\mathrm{hyb}}(\mathcal{F}), \mu^{\mathrm{uni}}_{\mathrm{hyb}}(\mathcal{F})\}$$

with $\mu^{\mathrm{rank}}_{\mathrm{hyb}}(\mathcal{F})$ and $\mu^{\mathrm{uni}}_{\mathrm{hyb}}(\mathcal{F})$ being defined in an obvious way.

THEOREM 8.2. *Suppose that $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a strict closed convex process. Then, one has*

$$\mu^{\mathrm{uni}}_{\mathrm{hyb}}(\mathcal{F}) = \begin{cases} \inf_{\substack{\lambda \in \mathbb{R}, \\ |q|=1}} \mathrm{dist}[(q, \lambda q), \mathrm{gr}\mathcal{F}^*] & \text{if } \mathrm{gr}\mathcal{F} \text{ is not a subspace,} \\ \infty & \text{otherwise.} \end{cases} \tag{8.9}$$

Proof. We consider only the case when $\mathrm{gr}\mathcal{F}$ is not a subspace, the other case being trivial. For any $M, L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, one has $(M \circ \mathcal{F} + L)^* = \mathcal{F}^* \circ M^T + L^T$. Hence, $\sigma^{\mathrm{uni}}(M \circ \mathcal{F} + L) \neq \emptyset$ if and only if there exist a scalar $\lambda \in \mathbb{R}$ and a unit vector $q \in \mathbb{R}^n$ such that

$$\begin{cases} \lambda q \in \mathcal{F}^*(M^T q) + L^T q, \\ -\lambda q \notin \mathcal{F}^*(-M^T q) - L^T q. \end{cases}$$

This can be written in the form

$$(L^T q - \lambda q, M^T q) \in (\mathrm{gr}\mathcal{F})^{\oplus}$$

with $(\mathrm{gr}\mathcal{F})^{\oplus}$ denoting the pseudo-dual of the convex cone $\mathrm{gr}\mathcal{F}$. Hence,

$$\mu^{\mathrm{uni}}_{\mathrm{hyb}}(\mathcal{F}) = \inf_{\substack{\lambda \in \mathbb{R}, \\ |q|=1}} \inf_{\substack{M,L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \\ (L^T q - \lambda q, M^T q) \in (\mathrm{gr}\mathcal{F})^{\oplus}}} \|(M, L) - (I, 0)\| = \inf_{\substack{\lambda \in \mathbb{R}, \\ |q|=1}} \inf_{X,Y} \left\| \begin{bmatrix} X \\ Y \end{bmatrix} - \begin{bmatrix} 0 \\ I \end{bmatrix} \right\|,$$

where the last infimum is taken with respect to

$$\begin{bmatrix} X \\ Y \end{bmatrix} \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^{n+m}) \quad \text{such that} \quad \begin{bmatrix} X \\ Y \end{bmatrix} q - \begin{bmatrix} \lambda q \\ 0 \end{bmatrix} \in (\mathrm{gr}\mathcal{F})^{\oplus}.$$

The reduction lemma yields

$$\mu^{\mathrm{uni}}_{\mathrm{hyb}}(\mathcal{F}) = \inf_{\substack{\lambda\in\mathbb{R},\\ |q|=1}} \mathrm{dist}[(-\lambda q, q), (\mathrm{gr}\mathcal{F})^{\oplus}] = \inf_{\substack{\lambda\in\mathbb{R},\\ |q|=1}} \mathrm{dist}[(-\lambda q, q), (\mathrm{gr}\mathcal{F})^{+}],$$

from which one gets the announced result. □

THEOREM 8.3. *Suppose that $\mathcal{F} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a strict closed convex process. Then,*

$$\mu^{\mathrm{rank}}_{\mathrm{hyb}}(\mathcal{F}) = \inf_{\substack{M,L\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ (MA+L,MQ)\ \text{uncontrollable}}} \|(M, L) - (I, 0)\| \tag{8.10}$$

with $Q \in \mathcal{O}(r, n)$ such that $\mathrm{Im}\, Q = \mathrm{span}\mathcal{F}(0)$ *and A denoting any linear selector of $\mathcal{F}^{\mathrm{rel}}$.*

Proof. One can show that for any $M, L \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$, one has the identity

$$\mathrm{gr}(M \circ \mathcal{F} + L) - \mathrm{gr}(M \circ \mathcal{F} + L) = \mathrm{gr}(M \circ \mathcal{F}^{\mathrm{rel}} + L),$$

and therefore $(M \circ \mathcal{F} + L)^{\mathrm{rel}} = M \circ \mathcal{F}^{\mathrm{rel}} + L$. By combining this fact and Theorem 7.14, one sees that

$$\begin{aligned} M \circ \mathcal{F} + L \text{ is irreproducing} &\iff (M \circ \mathcal{F} + L)^{\mathrm{rel}} \text{ is irreproducing} \\ &\iff M \circ \mathcal{F}^{\mathrm{rel}} + L \text{ is irreproducing} \\ &\iff (MA + L, MQ) \text{ is uncontrollable.} \end{aligned}$$

This proves, of course, the announced formula. □

We end this section by showing how to evaluate the hybrid indices μ^{uni}_{hyb} and $\mu^{\mathrm{rank}}_{hyb}$ in the particular case of a convex process $\mathcal{F}^P_{A,B} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ given by

$$\mathcal{F}^P_{A,B}(s) = As + B(P).$$

This choice may seem very peculiar, but, in fact, it is one of the most prominent examples in the general theory of convex processes. Observe that the cone-constrained model (5.1) can be written in the form of a differential inclusion whose right-hand side is $\mathcal{F}^P_{A,B}$. For the sake of completeness, we mention that the class

$$\mathcal{H}_P = \{\mathcal{F}^P_{A,B} \mid (A, B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)\}$$

is stable with respect to hybrid perturbations. Indeed, one can write the identity

$$M \circ \mathcal{F}^P_{A,B} + L = \mathcal{F}^P_{C,D},$$

where the pairs (A, B) and (C, D) are related through the transformation formulas

$$C = MA + L, \quad D = MB.$$

Observe that the perturbation (M, L) that brings (A, B) to (C, D) is given by

$$M = D(B^TB)^{-1}B^T, \quad L = C - D(B^TB)^{-1}B^TA. \tag{8.11}$$

For writing (8.11) we are implicitly assuming that the transpose of $B \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^n)$ is surjective because otherwise $B^T B$ is not invertible. In the context of our cone-constrained control problem (5.1), surjectivity of B^T can be assumed without loss of generality. It is interesting to note that seemingly more general types of perturbations like pointwise or graphical addition of convex processes do not even allow to recover the class $\mathcal{H}_p$ introduced above.

COROLLARY 8.4. *Let $P \subset \mathbb{R}^m$ be a closed convex cone but not a subspace. Then,*

$$\mu^{\rm uni}_{\rm hyb}(\mathcal{F}^P_{A,B}) = \inf_{\substack{\lambda\in\mathbb{R}\\ |q|=1}} \inf_{B^Th\in P^+} \Big[|A^Th-\lambda q|^2 + |h-q|^2\Big]^{1/2}. \tag{8.12}$$

Proof. An easy calculation shows that

$$\mathrm{gr}(\mathcal{F}^P_{A,B})^* = \{(h,p) \mid B^Th \in P^+,\ p = A^Th\}.$$

It suffices now to apply the general formula (8.9). □

We mention in passing that (8.12) can also be obtained by writing

$$\begin{aligned}\mu^{\rm uni}_{\rm hyb}(\mathcal{F}^P_{A,B}) &= \inf_{\substack{M,L\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ \sigma^{uni}(M\circ\mathcal{F}^P_{A,B}+L)\neq\emptyset}} \|(M,L)-(I,0)\| \\ &= \inf_{\substack{M,L\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n),\lambda\in\mathbb{R},|q|=1\\ A^TM^Tq+L^Tq=\lambda q,\ B^TM^Tq\in P^\oplus}} \|(M,L)-(I,0)\|\end{aligned}$$

and then applying the reduction lemma. This alternative method, however, requires some additional simplificatory work.

COROLLARY 8.5. *Let $(A,B) \in \mathcal{L}(\mathbb{R}^{n+m}, \mathbb{R}^n)$ and $P \subset \mathbb{R}^m$ be a closed convex cone. Take any $Q \in \mathcal{O}(r,n)$ such that* $\mathrm{Im}\, Q = \mathrm{span}\, B(P)$. *Then,*

$$\mu^{\rm rank}_{\rm hyb}(\mathcal{F}^P_{A,B}) = \inf_{\substack{M,L\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)\\ (MA+L,MQ)\ \text{uncontrollable}}} \|(M,L)-(I,0)\|. \tag{8.13}$$

Proof. It suffices to apply Theorem 8.3, keeping in mind that the relaxed version of $\mathcal{F}^P_{A,B}$ is given by $[\mathcal{F}^P_{A,B}]^{\rm rel}(\cdot) = A(\cdot) + \mathrm{span}B(P)$. □

9. By way of conclusion. This paper is about measuring the distance to uncontrollability in cone-constrained linear control problems or, more generally, in control problems described by convex processes. We have adopted the strategy of splitting the analysis into two separate components. One part of our study consists in measuring the distance to irreproducibility. The term *reproducibility* refers to a suitable generalization of Kalman's rank condition. The second part of our study consists in measuring the distance to unilateral modality (i.e., existence of unilateral uncontrollable modes). It is in this part of our study that the conic aspect of the data (convex cones, convex processes, etc.) comes into the picture. Bilateral uncontrollable modes belong to the realm of classical linear algebra and therefore they are left aside (in fact, they are implicitly incorporated in the analysis of reproducibility).

The formulas for measuring the distance to unilateral modality were obtained by exploiting the reduction lemma. There is a different approach which consists in exploiting the concept of ε-eigenvalue for multivalued operators. Following Gajardo and Seeger [12], we denote by

$$\Lambda_\varepsilon(\mathcal{F}^*) = \{\lambda \in \mathbb{R} \mid \exists (q,p) \in \mathrm{gr}\mathcal{F}^* \text{with } q \neq 0, \text{ such that } |p-\lambda q| \leq \varepsilon|q|\}$$

the set of ε-eigenvalues of $\mathcal{F}^*$. For practical purposes, it is important to estimate the smallest value of $\varepsilon \in \mathbb{R}_+$ that guarantees the nonvacuity of $\Lambda_\varepsilon(\mathcal{F}^*)$. This smallest value is called the spectral threshold of $\mathcal{F}^*$. As established in [12], the equality

$$\inf\{\varepsilon \in \mathbb{R}_+ \mid \Lambda_\varepsilon(\mathcal{F}^*) \neq \emptyset\} = \inf_{\substack{\lambda \in \mathbb{R} \\ |q|=1}} \operatorname{dist}[\lambda q, \mathcal{F}^*(q)] \tag{9.1}$$

holds, in particular, when $\mathcal{F}$ is a strict closed convex process. Formula (9.1) gives us an alternative interpretation of the index $\mu^{\text{uni}}_{\text{add}}(\mathcal{F})$ when $\text{gr}\mathcal{F}$ is not a subspace (cf. Theorem 8.1).

Remark. When $\text{gr}\mathcal{F}$ is not a subspace, both expressions in (9.1) serve not only to measure the distance to unilateral modality but also to modality in general (i.e., existence of uncontrollable modes without specification of their nature). This observation is quite subtle because, in general, the sets $\sigma^{\text{uni}}(\mathcal{F})$ and $\sigma(\mathcal{F})$ don't coincide.

As far as the hybrid index $\mu^{\text{uni}}_{\text{hyb}}(\mathcal{F})$ is concerned, we see now appearing an expression of the form

$$\Psi_{\mathcal{F}^*}(\lambda) = \inf_{|q|=1} \operatorname{dist}[(q, \lambda q), gr\mathcal{F}^*] \tag{9.2}$$

which has to be minimized with respect to $\lambda \in \mathbb{R}$. The function $\Psi_{\mathcal{F}^*}$ is used by Seeger [31] in connection with the upper stabilization of the point-spectral set-valued mapping Λ. Observe that in a finite dimensional setting, one has

$$\Lambda(\mathcal{F}^*) = \{\lambda \in \mathbb{R} \mid \Psi_{\mathcal{F}^*}(\lambda) = 0\}.$$

In an infinite dimensional setting, the above equality is no longer true. As shown in [31], the roots of $\Psi_{\mathcal{F}^*}$ produce a set which may be much larger than $\Lambda(\mathcal{F}^*)$ (one gets the so-called approximate or stabilized spectrum of $\mathcal{F}^*$). This observation is just to warn the reader that some of our results (for instance, Theorem 8.2) do not extend to an infinite dimensional setting, unless important modifications are incorporated. Infinite dimensionality introduces various complications that are not addressed in the present work.

REFERENCES

[1] F. Alvarez, R. Correa, and P. Gajardo, *Inner estimation of the eigenvalue set and exponential series solutions to differential inclusions*, J. Convex Anal., 12 (2005), pp. 1–11.

[2] J.P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[3] J.P. Aubin, H. Frankowska and C. Olech, *Controllability of convex processes*, SIAM J. Control Optim., 24 (1986), pp. 1192–1211.

[4] J. Borwein, *Adjoint process duality*, Math. Oper. Res., 8, (1983), pp. 403–434.

[5] R.F. Brammer, *Controllability in linear autonomous systems with positive controllers*, SIAM J. Control Optim., 10 (1972), pp. 339–353.

[6] J.V. Burke, A.S. Lewis, and M.L. Overton, *Pseudospectral components and the distance to uncontrollability*. SIAM J. Matrix Anal. Appl., 26 (2004/05), pp. 350–361.

[7] R. Cross, *Multivalued Linear Operators*, Marcel Dekker, New York, 1988.

[8] R.A. DeCarlo and M. Wicks, *Computing the distance to an uncontrollable system*, IEEE Trans. Automat. Control, 36, (1991), pp. 39–49.

[9] R. Eising, *Between controllable and uncontrollable*, Systems Control Lett., 4 (1984), pp. 263–264.

[10] L. Elsner and C. He, *An algorithm for computing the distance to uncontrollability*, Systems Control Lett., 17 (1991), pp. 453–464.

[11] P. M. Gahinet, *Distance to the nearest uncontrollable pair and algebraic Riccati equation*, in Proceedings of the 29th Conference on Decision and Control, Honolulu, 1990, pp. 273–278.

[12] P. Gajardo and A. Seeger, *Epsilon-eigenvalues of multivalued operators*, Set-Valued Anal., 11 (2003), pp. 273–296.
[13] P. Gajardo and A. Seeger, *Higher-Order Spectral Analysis and Weak Asymptotic Stability of Convex Processes*, J. Math. Anal. Appl., 2005, in press.
[14] M. Gao and M. Neumann, *A global minimum search algorithm for estimating the distance to uncontrollability*, Linear Algebra Appl., 188/189 (1993), pp. 305–350.
[15] J.M. Gracia and I. de Hoyos, *Nearest pair with more nonconstant invariant factors and pseudospectrum*, Linear Algebra Appl., 298 (1999), pp. 143–158.
[16] M.L.J. Hautus, *Controllability and observability conditions of linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A 72, 31 (1969), pp. 443–448.
[17] C. He, *Estimating the distance to uncontrollability: A fast method and a slow one*, Systems Control Lett., 26 (1995), pp. 275–281.
[18] D. Hinrichsen and A.J. Pritchard, *Real and Complex Stability Radii: A Survey*, Progr. Systems Control Theory 6, Birkhäuser, Boston, 1990, pp. 119–162.
[19] G. Hu and E.J. Davison, *A real radius measure for controllability*, in Proceedings American Control Conference, Arlington, VA, June 2001, pp. 3144–3148.
[20] G. Hu and E.J. Davison, *Real controllability/stabilizability radius of LTI systems*, IEEE Trans. Automat. Control, 49 (2004) pp. 254–257.
[21] R. Kalman, Y.C. Ho, and K.S. Narendra, *Controllability of linear dynamical systems*, Contrib. Differential Equations, 1 (1962), pp. 189–213.
[22] V.I. Korobov, *A geometric criterion of local controllability of dynamical systems in the presence of contraints on the control*, Differential Equations, 15 (1980), pp. 1136–1142.
[23] P. Lavilledieu and A. Seeger, *Existence de valeurs propres pour les systèmes multivoques: résultats anciens et nouveaux*, Ann. Sci. Math. Québec, 25 (2001), pp. 47–70.
[24] P. Lavilledieu and A. Seeger, *Rank condition and controllability of parametric convex processes*, J. Convex Anal., 9 (2002), pp. 535–542.
[25] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[26] O. Naselli-Ricceri, *On the controllability of a class of differential inclusions depending on a parameter*, J. Optim. Theory Appl., 65 (1990), pp. 281–288.
[27] C.C. Paige, *Properties of numerical algorithms related to computing controllability*, IEEE Automat. Control, 26 (1981), pp. 130–138.
[28] V.N. Phat, *Weak asymptotic stabilizability of discrete-time systems given by set-valued operators*, J. Math. Anal. Appl., 202 (1996), pp. 363–378.
[29] R.T. Rockafellar, *Convex Analysis*, Princeton Academic Press, Princeton, NJ, 1970.
[30] A. Seeger, *Spectral analysis of set-valued mappings*, Acta Math. Vietnamica, 25 (1998), pp. 49–63.
[31] A. Seeger, *Stabilization of the point-spectral mapping*, J. Nonlinear Convex Anal., 4 (2003), pp. 277–285.
[32] N.K. Son, *Approximate controllability with positive controls*, Acta Math. Vietnamica, 22 (1997), pp. 589–620.
[33] H.D. Tuan, *On controllability of convex differential inclusions in Banach space*, Optimization, 30 (1994), pp. 151–162.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 51–73

BIAS OPTIMALITY FOR CONTINUOUS-TIME CONTROLLED MARKOV CHAINS*

TOMÁS PRIETO-RUMEAU† AND ONÉSIMO HERNÁNDEZ-LERMA‡

Abstract. In this paper we give conditions for the existence of bias optimal policies in a class of continuous-time controlled Markov chains with unbounded reward and transition rates. Several characterizations of bias optimality are proposed. We also introduce new sets of conditions ensuring uniform exponential ergodicity of continuous-time controlled Markov chains.

Key words. continuous-time controlled Markov chain (or Markov decision process), average reward (or gain) criterion, Poisson equation, bias optimality, exponential ergodicity

AMS subject classifications. 93E20, 90C40, 60J27

DOI. 10.1137/S036301290343432

1. Introduction.

The long-run average reward or *gain* is one of the most widely used optimality criteria for stochastic control processes, in particular in applications to manufacturing processes, computer systems, and telecommunication networks, to name just a few [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 26, 28, 32]. However, as it is only concerned with the asymptotic behavior of the process it turns out to be extremely underselective. For instance, to be a little more precise, suppose that $J_T(f)$ denotes the expected total reward during the time interval $[0, T]$, when using a control policy f, and let

$$g(f) := \liminf_{T\to\infty} \frac{1}{T} J_T(f)$$

be the corresponding gain. Then if f and f' are two policies such that

$$J_T(f) = J_T(f') + T^\theta \tag{1.1}$$

for all $T > 0$ and some $\theta \in (0, 1)$, we have two policies that yield the *same gain* although their finite-horizon rewards are quite different. Thus the average reward criterion is "underselective" because it does not distinguish between policies such as f and f'.

To avoid such a behavior we have to restrict ourselves to "better" classes of policies. In this paper we first impose conditions under which the finite-horizon rewards of stationary policies are necessarily of the form

$$J_T(f) = T \cdot g(f) + h_f(\cdot) + e(f, T), \tag{1.2}$$

where $h_f(\cdot)$ is the so-called *bias* of f and $e(f, T)$ is a residual term converging to 0 as $T \to \infty$. Therefore, if f and f' are two stationary policies with the same gain, then

*Received by the editors September 5, 2003; accepted for publication (in revised form) September 19, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sicon/45-1/43432.html

†Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Educación a Distancia, Calle Senda del Rey 9, 28040 Madrid, Spain (tprieto@ccia.uned.es). The research of this author was supported by a grant from the Spanish Secretaría de Estado de Educación y Universidades in cooperation with the European Social Funds.

‡Departamento de Matemáticas, CINVESTAV-IPN, Apartado Postal 14-740, México D.F. 07000, Mexico (ohernand@math.cinvestav.mx). The research of this author was partially supported by CONACyT grant 45693.

instead of (1.1) we now get

$$J_T(f) - J_T(f') = h_f(\cdot) - h_{f'}(\cdot) + [e(f,T) - e(f',T)].$$

If in addition $h_f(\cdot) \geq h_{f'}(\cdot)$, then the policy f, which has the larger bias, will eventually "overtake" f' in the sense that for any given $\varepsilon > 0$,

$$J_T(f) \geq J_T(f') - \varepsilon$$

for all T sufficiently large. In other words, maximizing the bias functions, within the class of gain optimal policies, allows us to take the policy with the larger growth.

The main objective for this paper is to make precise the ideas expressed in the previous paragraph for a class of continuous-time controlled Markov chains (CMCs), also known as Markov decision processes. Hence, starting with the set $\mathbb{F}$ of all stationary policies, we give conditions for the existence of subclasses

$$\mathbb{F}_{ao} \supseteq \mathbb{F}_{ca} \supseteq \mathbb{F}_{bias}, \tag{1.3}$$

which are, respectively, the class of stationary *average optimal* policies, the class of *canonical* policies, and the class of *bias optimal* policies. Our main goal is to show that $\mathbb{F}_{bias}$ is nonempty and to characterize the policies in it. It should be noted that the inclusions in (1.3) can all be *proper*. For instance, Guo and Liu [11] show an average optimal policy which is *not* canonical, whereas our Example 3.7 gives a canonical policy which is *not* bias optimal. Similar examples for discrete-time CMCs can be seen in [12, 15, 18, 19, 26], for instance.

The concept of bias optimality, which was introduced by Veinott [30], has been extensively studied for *discrete-time* CMCs—see, for instance, [2, 15, 16, 17, 18, 19, 26], and their references. However, to the best of our knowledge, this is the first paper dealing with the *continuous-time* case—although it is implicit in Puterman's work [25] for controlled diffusions in compact intervals. It should also be noted that in our control model a control action can be taken at *any* time $t \geq 0$ (as opposed to at certain times) and, therefore, our CMC cannot be reduced to a discrete-time problem.

Our point of departure to study bias optimality is the average reward optimality equation for continuous-time CMCs. This optimality equation was established by Guo and Hernández-Lerma in [10] under a uniform exponential ergodic property of the continuous-time CMC. The authors propose irreducibility and monotonicity conditions ensuring that the model is uniformly exponentially ergodic; see [10, Assumption C]. In this paper, we propose an alternative sufficient condition for uniform ergodicity, namely, we replace Assumption C in [10] with a uniform integrability condition. This result is of valuable interest by itself and not just in the context of average or bias optimality.

The remainder of our paper is organized as follows. In section 2 we introduce the control model we will be dealing with, paying particular attention to its ergodic properties. Our main results are stated in section 3. For expositional clarity, the proofs of the results in sections 2 and 3 are postponed to sections 4 and 5, respectively. Section 6 states the final conclusions and mentions some interesting open issues.

2. Model definition and ergodic properties.

The control model. The control model we are concerned with is

$$\{S, (A(i), i \in S), q_{ij}(a), r(i,a)\}, \tag{2.1}$$

where S is the *state space*, a denumerable set. Without loss of generality we will assume that $S = \{0, 1, 2, \ldots\}$, the set of nonnegative integers. Furthermore, $A(i)$ is a Borel space which stands for the set of *admissible control actions* in the state $i \in S$. Let $\mathcal{B}(A(i))$ be the Borel σ-algebra of subsets of $A(i)$, and $K := \{(i, a) : i \in S, a \in A(i)\}$. The numbers $q_{ij}(a)$, for $(i, a) \in K$ and $j \in S$ are the system's *transition rates* and so they satisfy that $q_{ij}(a) \geq 0$ if $j \neq i$. Moreover, the matrix $[q_{ij}(a)]$ is supposed to be *conservative*, i.e.,

$$\sum_{j \in S} q_{ij}(a) = 0 \quad \text{for all } (i, a) \in K, \tag{2.2}$$

and *stable*, which means that

$$q(i) := \sup_{a \in A(i)} [-q_{ii}(a)] < \infty \quad \text{for } i \in S. \tag{2.3}$$

In addition, $q_{ij}(a)$ is measurable in $a \in A(i)$ for each fixed $i, j \in S$. Finally, $r : K \to \mathbb{R}$ denotes the *reward rate* and is assumed to be measurable in $a \in A(i)$ for each fixed $i \in S$. (As $r(i, a)$ is allowed to take positive and negative values, it can be interpreted as a *cost rate* rather than a "reward.")

Control policies. Let $A := \cup_{i \in S} A(i)$. A function $f : S \to A$ is called a *decision function* if $f(i)$ is in $A(i)$ for all $i \in S$. The set of all such functions is denoted by $\mathbb{F}$. Now let Φ_m be the family of functions $\varphi_t(B|i)$ such that

- for each $i \in S$ and $t \geq 0$, $B \mapsto \varphi_t(B|i)$ is a probability measure on $\mathcal{B}(A(i))$, and
- for each $i \in S$ and $B \in \mathcal{B}(A(i))$, $t \mapsto \varphi_t(B|i)$ is a Lebesgue measurable function on $[0, \infty)$.

A family $\varphi = \{\varphi_t, t \geq 0\}$ in Φ_m is called a *randomized Markov policy*. If in addition there is a function $f \in \mathbb{F}$ such that $\varphi_t(\cdot|i)$ is concentrated at $f(i)$ for all $i \in S$ and $t \geq 0$, then φ is said to be a (deterministic) *stationary policy* and it is identified with f. In other words, we will identify the family $\mathbb{F}$ of decision functions with the family of stationary policies.

Notation. For each $\varphi = \{\varphi_t, t \geq 0\} \in \Phi_m$, i and j in S, and $t \geq 0$, we write

$$r(t, i, \varphi) := \int_{A(i)} r(i, a) \varphi_t(da|i), \tag{2.4}$$

$$q_{ij}(t, \varphi) := \int_{A(i)} q_{ij}(a) \varphi_t(da|i) \tag{2.5}$$

whenever these integrals are well defined. For a stationary policy $\varphi = f \in \mathbb{F}$, (2.4) and (2.5) reduce to $r(i, f) := r(i, f(i))$ and $q_{ij}(f) := q_{ij}(f(i))$, respectively.

DEFINITION 2.1. *A randomized Markov policy $\varphi \in \Phi_m$ is said to be* admissible *if $q_{ij}(t, \varphi)$ is continuous in t for each fixed $i, j \in S$. We shall denote by Φ_{ad} the family of all such policies.*

Observe that Φ_{ad} contains $\mathbb{F}$. The continuity of the transition rates is a standard assumption, e.g., [7, 9, 10]. It allows us to derive the transition function of the system from its transition rates. More precisely, for each $\varphi \in \Phi_{ad}$, the matrix $Q(t, \varphi) := [q_{ij}(t, \varphi)]$ is easily seen to be conservative and stable and, therefore, there exists a *Q-process*—that is, a possibly substochastic and nonhomogeneous transition function—with transition rates $q_{ij}(t, \varphi)$. To ensure the existence of a regular (stochastic) transition function $p^{\varphi}(s, i, t, j)$ and an associated Markov process $\{x^{\varphi}(t)\}$, we will borrow the following "drift" condition from [10].

Assumption A. There exists a sequence $\{S_m, m \geq 1\}$ of subsets of S, a non-decreasing function $w \geq 1$ on S, and constants $b \geq 0$, $c > 0$, and $M > 0$ such that

(a) $S_m \uparrow S$ and $\sup_{i \in S_m} q(i) < \infty$ for each $m \geq 1$;
(b) $\inf_{j \notin S_m} w(j) \to +\infty$ as $m \to \infty$;
(c) $\sum_{j \in S} q_{ij}(a) w(j) \leq -cw(i) + b\delta_{i0}$ for all $(i, a) \in K$, where δ_{ij} denotes the Kronecker delta;
(d) $|r(i, a)| \leq Mw(i)$ for all $(i, a) \in K$.

See Remark 2.1 in [10] for a discussion of Assumption A, in particular, its connection to similar conditions in the previous literature, e.g., Anderson [1] and Lund, Meyn, and Tweedie [20]. On the other hand, Assumption A guarantees that the long-run average reward in (2.7) is finite-valued. To state this we need some notation.

For each initial state $i \in S$ at time $s \geq 0$ and $\varphi \in \Phi_{ad}$, we denote by $P^{\varphi}_{s,i}$ and $E^{\varphi}_{s,i}$ the probability measure determined by $p^{\varphi}(s, i, t, j)$ and the corresponding expectation operator. If $s = 0$, we write $E^{\varphi}_{0,i}$ as E^{φ}_{i} and $p^{\varphi}(0, i, t, j)$ as $p^{\varphi}(i, t, j)$.

DEFINITION 2.2. *For each $\varphi \in \Phi_{ad}$, $i \in S$ and $T \geq 0$, let*

$$J_T(i, \varphi) := E^{\varphi}_{i} \left[\int_0^T r(t, x(t), \varphi) dt \right]. \tag{2.6}$$

The long-run average reward—*also known as the* gain—*of φ is given by*

$$J(i, \varphi) := \liminf_{T \to \infty} \frac{1}{T} J_T(i, \varphi). \tag{2.7}$$

The function

$$J^*(i) := \sup_{\varphi \in \Phi_{ad}} J(i, \varphi) \quad \textit{for } i \in S$$

is referred to as the optimal gain *or* optimal average reward. *A policy $\varphi^* \in \Phi_{ad}$ is said to be* average optimal *if $J(i, \varphi^*) = J^*(i)$ for all $i \in S$.*

Remark 2.3. Observe that in (2.6) the Markov process is not denoted $\{x^{\varphi}(t)\}$. The reason for this is that $P^{\varphi}_{s,i}$ is a probability measure over the set of all sample paths $\{x(t)\} \in S^{[0,\infty)}$ and thus writing, for instance, $E^{\varphi}_{s,i}[x^{\varphi}(t)]$ is redundant.

Under Assumption A, for each $f \in \mathbb{F}$ the corresponding Markov process $\{x^f(t)\}$ is nonexplosive [21, Theorem 2.1] and positive Harris recurrent and, further, it has a unique invariant probability measure μ_f for which

$$\mu_f(w) := \sum_{j \in S} w(j) \mu_f(j) < \infty; \tag{2.8}$$

see [21, Theorem 4.2]. Moreover, the gain or reward in (2.7) turns out to be a constant $J(\cdot, f) \equiv g(f)$ independent of the initial state $i \in S$ given by

$$g(f) = \lim_{T \to \infty} \frac{1}{T} J_T(i, f) = \sum_{j \in S} r(j, f) \mu_f(j) \quad \text{for } i \in S. \tag{2.9}$$

Our next assumption states the standard continuity-compactness conditions (see Assumption B in [10]).

Assumption B. For each $i, j \in S$

(a) $A(i)$ is compact, and

(b) $r(i,a)$, $q_{ij}(a)$, and $\sum_{j\in S} q_{ij}(a)w(j)$ are upper semicontinuous in $a \in A(i)$, where w is the function in Assumption A.

(c) There exists a nonnegative function w' on S and constants $b' \geq 0$, $c' > 0$, $M' > 0$ such that

$$q(i)w(i) \leq M'w'(i) \quad \text{and} \quad \sum_{j\in S} q_{ij}(a)w'(j) \leq c'w'(i) + b'$$

for all $(i,a) \in K$.

To state our last assumption, we need some notation. Let $w(\cdot) \geq 1$ be as in Assumption A. The *w-norm* of a function $u : S \to \mathbb{R}$ is defined as

$$||u||_w := \sup_{i\in S} \frac{|u(i)|}{w(i)}.$$

We shall denote by $B_w(S)$ the Banach space of functions on S with finite w-norm.

Assumption C.

(a) For each $f \in \mathbb{F}$, the Markov process $\{x^f(t)\}$ is irreducible; that is, given a pair of states $i \neq j$ there exists a set of distinct states $i = i_1, \ldots, i_m = j$ such that

$$\prod_{k=1}^{m-1} q_{i_k i_{k+1}}(f) > 0.$$

(b) The CMC model is *w-exponentially ergodic*; that is, there exist constants $\delta > 0$ and $R \geq 0$ such that

$$\sup_{f\in\mathbb{F}} |E_i^f u(x(t)) - \mu_f(u)| \leq Re^{-\delta t}||u||_w w(i) \tag{2.10}$$

for all $i \in S$, $u \in B_w(S)$, and $t \geq 0$, where $\mu_f(u) := \sum_{j\in S} u(j)\mu_f(j)$.

Assumption C(b) seems to be difficult to verify in practice. For noncontrolled Markov processes, several approaches yield exponential ergodicity; see, e.g., [3, 6, 20, 21, 29]. Also, there exist conditions ensuring uniform ergodicity of discrete-time controlled Markov chains; see [5], for instance. However, except for the paper [10], very little has been done for controlled continuous-time Markov chains.

Next, we propose two different sets of sufficient conditions for Assumption C(b). The first, stated in Theorem 2.4, is taken from [10]. (See, in particular, its connection to the monotonicity conditions in [20].)

THEOREM 2.4. *Suppose that Assumptions A(a)–(c) and C(a) are satisfied and, in addition, that*

(i) *for every $i, k \in S$ (where $k \neq i+1$) and for each $f \in \mathbb{F}$,*

$$\sum_{j\geq k} q_{ij}(f) \leq \sum_{j\geq k} q_{i+1,j}(f);$$

(ii) *for every $f \in \mathbb{F}$ and $j > i > 0$, there exist nonzero distinct states $i = i_1, \ldots, i_m \geq j$ such that*

$$\prod_{k=1}^{m-1} q_{i_k i_{k+1}}(f) > 0.$$

Then, Assumption C(b) holds with $\delta = c$, the constant in Assumption A(c).

For a proof, we refer to [10, 20]. Condition (i) in Theorem 2.4 imposes that for each $f \in \mathbb{F}$, the Markov process $\{x^f(t)\}$ is stochastically monotone, whereas condition (ii) means that for every $f \in \mathbb{F}$, we can "travel" with positive probability from state i to the set $\{j, j+1, \ldots\}$ (with $j > i > 0$) without passing through $\{0\}$.

Our next result, rather than the monotonicity conditions in Theorem 2.4, imposes a uniform integrability assumption.

THEOREM 2.5. *Suppose that Assumptions A(a)–(c), B(a), and C(a) are satisfied. Suppose also that for every $i, j \in S$, the function $a \mapsto q_{ij}(a)$ is continuous on $A(i)$ (cf. Assumption B(b)) and that*

$$\lim_{k\to\infty} \sup_{f\in\mathbb{F}} \left\{ \sum_{j=k}^{\infty} p^f(i,t,j) w(j) \right\} = 0 \quad \text{for every } i \in S \text{ and } t \geq 0. \tag{2.11}$$

Then, Assumption C(b) holds.

Observe that by [10, Lemma 3.2], the condition (2.11) holds for every single $f \in \mathbb{F}$. The proof of Theorem 2.5, which is based on the discrete-time ergodicity results in [5], is given in section 4. See Remark 4.7 for an equivalent formulation of (2.11).

Assumptions A, B, and C are supposed to hold throughout the following.

3. Main results. For expositional convenience, in this section we state Theorems 3.3, 3.6, 3.8 and 3.9 but their proofs are postponed to section 5.

Canonical policies. Let $\mathbb{F}_{ao} \subseteq \mathbb{F}$ be the family of stationary average optimal policies. Under Assumptions A, B, and C, the set $\mathbb{F}_{ao}$ is nonempty; see [10]. Actually, its subset $\mathbb{F}_{ca}$ defined next is also nonempty.

DEFINITION 3.1.

(a) *A pair (g^*, h^*) consisting of a constant $g^* \in \mathbb{R}$ and a function $h^* \in B_w(S)$ is said to be a* solution of the average reward optimality equation (*AROE*) *if*

$$g^* = \max_{a\in A(i)} \left[r(i,a) + \sum_{j\in S} q_{ij}(a) h^*(j) \right] \quad \text{for all } i \in S. \tag{3.1}$$

(b) *If $f^* \in \mathbb{F}$ is such that $f^*(i) \in A(i)$ attains the maximum in* (3.1) *for all $i \in S$, that is,*

$$g^* = r(i, f^*) + \sum_{j\in S} q_{ij}(f^*) h^*(j) \quad \text{for } i \in S, \tag{3.2}$$

then f^* is called a canonical policy*, and, furthermore, (g^*, h^*, f^*) is called a* canonical triplet*. The set of canonical policies is denoted $\mathbb{F}_{ca}$.*

Remark 3.2. If (g^*, h^*) is a solution of the AROE, it is easily seen that g^* is unique (see (3.3) below). However, as a consequence of (2.2), h^* can be replaced with $h^* + z$ for any constant z. In fact, as in the proof of Proposition 3.5, it can be shown that h^* is unique up to additive constants.

As in [10, Theorem 4.1], our standing assumptions guarantee the existence of a canonical triplet (g^*, h^*, f^*). Moreover,

$$g^* = J^*(i) = J(i, f^*) \quad \text{for } i \in S, \tag{3.3}$$

which means that g^* is the optimal gain and that f^* is average optimal; hence

$$\mathbb{F}_{ca} \subseteq \mathbb{F}_{ao}. \tag{3.4}$$

In general, the inclusion in (3.4) can be proper; in other words, there can be average optimal policies which are *not* canonical—see Guo and Liu [11], for instance. However, under our Assumptions A, B, and C, the following Theorem 3.3 states that in fact the *equality* holds in (3.4).

THEOREM 3.3. *A stationary policy is canonical if and only if it is average optimal, that is,* $\mathbb{F}_{ca} = \mathbb{F}_{ao}$.

The Poisson equation. For each stationary policy $f \in \mathbb{F}$, we define the *bias* of f as

$$h_f(i) := \int_0^\infty [E_i^f r(x(t), f) - g(f)]dt \quad \text{for } i \in S. \tag{3.5}$$

By (2.10) and Assumptions A(d) and C(b),

$$|E_i^f r(x(t), f) - g(f)| \le RMw(i)e^{-\delta t} \quad \text{for all } t \ge 0$$

and so

$$|h_f(i)| \le \delta^{-1} RMw(i) \quad \text{for } i \in S. \tag{3.6}$$

It follows that h_f is in $B_w(S)$ and also that its w-norm is *uniformly bounded* in $f \in \mathbb{F}$ because

$$||h_f||_w \le \delta^{-1} RM \quad \text{for every } f \in \mathbb{F}.$$

We also get the following.

PROPOSITION 3.4. *For each* $f \in \mathbb{F}$, *the pair* $(g(f), h_f)$ *is the unique solution of the so-called* Poisson equation

$$g(f) = r(i, f) + \sum_{j \in S} q_{ij}(f) h_f(j) \quad \textit{for } i \in S \tag{3.7}$$

for which the μ_f*-expectation of the bias* h_f *is zero:*

$$\sum_{j \in S} h_f(j)\mu_f(j) = 0. \tag{3.8}$$

Proof. For a proof of (3.7) see, for instance, Theorem 5.2(ii) in [4] or Lemma 5.1 in [10]. To obtain (3.8) first note that, by (3.6) and (2.8), h_f is indeed μ_f-integrable. Then in (3.5) choose the distribution of the inital state to be μ_f, and so (3.8) follows from a direct calculation.

To prove uniqueness we will use the fact that, by (2.10),

$$\lim_{t \to \infty} \frac{1}{t} E_i^f [u(x(t))] = 0 \tag{3.9}$$

for all $f \in \mathbb{F}$, $i \in S$, and $u \in B_w(S)$. We will also use the notation

$$(Q^f h)(i) := \sum_{j \in S} q_{ij}(f) h(j). \tag{3.10}$$

Now, for $k = 1, 2$, let $(g_k, h_k) \in \mathbb{R} \times B_w(S)$ be a solution of (3.7), (3.8). Then for each $k = 1, 2$, Dynkin's formula

$$E_i^f h_k(x(t)) = h_k(i) + E_i^f \left[\int_0^t (Q^f h_k)(x(s))ds \right] \tag{3.11}$$

together with (3.7) and (2.6) yields

$$E_i^f h_k(x(t)) = h_k(i) + t \cdot g_k - J_t(i, f). \tag{3.12}$$

Multiplying both sides by $1/t$ and letting $t \to \infty$, it follows from (3.9) and (2.9) that

$$g_1 = g_2 = g(f).$$

It only remains to show that $h_1 = h_2$. Let $u := h_1 - h_2$. Substracting the Poisson equation (3.7) for (g_2, h_2) from the corresponding equation for (g_1, h_1) we get that

$$(Q^f u)(i) = 0 \quad \text{for } i \in S. \tag{3.13}$$

Therefore, using Dynkin's formula (3.11) with u in lieu of h_k we see that u is *harmonic*, that is,

$$u(i) = E_i^f u(x(t)) = \sum_{j \in S} p^f(i, t, j) u(j) \quad \text{for } t \geq 0.$$

Thus, letting $t \to \infty$, (2.10) and (3.8) give

$$u(i) = \mu_f(u) = \mu_f(h_1) - \mu_f(h_2) = 0 \quad \text{for all } i \in S, \tag{3.14}$$

that is, $h_1(\cdot) = h_2(\cdot)$. □

Now let $f \in \mathbb{F}_{ao}$ be an average optimal policy. Then, by Theorem 3.3, f is canonical and, therefore, it satisfies (3.2), i.e.,

$$g^* = r(i, f) + \sum_{j \in S} q_{ij}(f) h^*(j).$$

On the other hand, since $g(f) = g^*$, the Poisson equation (3.7) for f becomes

$$g^* = r(i, f) + \sum_{j \in S} q_{ij}(f) h_f(j),$$

and so substracting the last two equalities we conclude that $u := h^* - h_f$ satisfies (3.13). This implies that u is harmonic and, as in (3.14), we conclude that $u = h^* - h_f = \mu_f(h^*) - \mu_f(h_f) = \mu_f(h^*)$. We can summarize this as follows.

PROPOSITION 3.5. *If $f \in \mathbb{F}$ is average optimal, then its bias h_f and any function h^* in the AROE* (3.1) *are equal up to an additive constant; in fact*

$$h_f(i) = h^*(i) - \mu_f(h^*) \quad \textit{for } i \in S. \tag{3.15}$$

The next theorem provides an interesting characterization of a canonical triplet. We shall need a little notation first. In addition to the finite-horizon reward in (2.6), we consider the reward functional

$$J_T(i, s, \varphi, h) := E_{s,i}^{\varphi} \left[\int_s^T r(t, x(t), \varphi) dt + h(x(T)) \right]$$

for each $i \in S$, $0 \leq s \leq T$, $\varphi \in \Phi_{ad}$, and $h \in B_w(S)$. The corresponding optimal reward function is

$$J_T^*(i, s, h) := \sup_{\varphi \in \Phi_{ad}} J_T(i, s, \varphi, h). \tag{3.16}$$

A policy attaining the supremum in (3.16) for all $i \in S$ and a given $s \in [0, T]$ will be called *optimal for the finite-horizon control problem over* $[s, T]$ *with terminal payoff* h.

THEOREM 3.6. *$(g^*, h^*, f^*) \in \mathbb{R} \times B_w(S) \times \mathbb{F}$ is a canonical triplet if and only if for all $i \in S$ and $0 \le s \le T$,*

(i) $J_T^*(i, s, h^*) = (T - s)g^* + h^*(i)$, *and*

(ii) $J_T^*(i, s, h^*) = J_T(i, s, f^*, h^*)$, *i.e., f^* is optimal for the finite-horizon control problem over $[s, T]$ with terminal payoff h^*.*

The characterization of a canonical triplet in Theorem 3.6 is essentially the same as the one proposed by Yushkevich [33] (see also [14] or [15]) for *discrete-time* CMCs.

Bias optimality. The function $\hat{h} \in B_w(S)$ defined as

$$\hat{h}(i) := \sup_{f \in \mathbb{F}_{ao}} h_f(i) \quad \text{for } i \in S \tag{3.17}$$

is called the *optimal bias function*. (By (3.6), $\hat{h}$ is indeed in $B_w(S)$.) A stationary average optimal policy $\hat{f}$ is said to be *bias optimal* if it attains the maximum in (3.17), i.e.,

$$h_{\hat{f}}(i) = \hat{h}(i) \quad \text{for } i \in S.$$

We shall denote by $\mathbb{F}_{bias}$ the family of bias optimal policies. By definition of bias optimality, $\mathbb{F}_{bias}$ is a subset of $\mathbb{F}_{ao} = \mathbb{F}_{ca}$, and in general the inclusion is *strict*: the following simple example shows an average optimal policy which is *not* bias optimal. Other examples are given by Lewis and Puterman [18, 19] and Haviv and Puterman [12].

Example 3.7. Consider a control model as in (2.1) with components

$$S = \{1, 2\},\ A(1) = \{1\},\ A(2) = \{1, 2\},$$

$$q_{11}(1) = -2,\ q_{22}(1) = -1,\ q_{22}(2) = -2,$$

$$r(1, 1) = 4,\ r(2, 1) = 1,\ r(2, 2) = 0.$$

There are two stationary policies f_1, f_2, with $f_1(2) = 1$, $f_2(2) = 2$. Moreover, the corresponding gain and bias functions are

$$g(f_1) = g(f_2) = 2 \quad \text{and} \quad h_{f_1}(\cdot) = \begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix},\ h_{f_2}(\cdot) = \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}.$$

Hence both policies have the same gain and they are average optimal [10, Theorem 4.1] but only f_1 is bias optimal.

The following result states the existence of bias optimal policies as well as a useful "growth sensitive" characterization (b_3) of bias optimality.

THEOREM 3.8. *Under Assumptions A, B, and C,*

(a) *There exists a bias optimal policy—that is, $\mathbb{F}_{bias}$ is nonempty.*

(b) *The following statements are equivalent:*

(b_1) *$\hat{f} \in \mathbb{F}$ is bias optimal.*

(b_2) *$\hat{f} \in \mathbb{F}$ is a canonical policy and $\mu_{\hat{f}}(\hat{h}) = 0$.*

(b_3) *$\hat{f} \in \mathbb{F}$ is* weakly overtaking optimal *in $\mathbb{F}$, that is,*

$$\liminf_{T \to \infty} [J_T(i, \hat{f}) - J_T(i, f)] \ge 0 \quad \textit{for every } i \in S \textit{ and } f \in \mathbb{F}. \tag{3.18}$$

More explicitly, (3.18) states that for every $\varepsilon > 0$ there exists $T_\varepsilon = T_\varepsilon(i, \hat{f}, f)$ such that

$$J_T(i, \hat{f}) \geq J_T(i, f) - \varepsilon \quad \text{for } T \geq T_\varepsilon. \tag{3.19}$$

(If (3.19) holds for $\varepsilon = 0$, it is said that $\hat{f}$ *overtakes* f, a concept that can be traced back to Ramsey [27]. The weaker form in (3.18) is due to Gale [8] and von Weizsäcker [31].)

To conclude this section we shall introduce another characterization of bias optimality related to (b_2) in Theorem 3.8 and the AROE (3.1).

A pair $(g^*, u) \in \mathbb{R} \times B_w(S)$ is said to verify the *bias optimality equations* if (g^*, u) is a solution of the AROE, i.e.,

$$g^* = \max_{a \in A(i)} \left[r(i,a) + \sum_{j \in S} q_{ij}(a) u(j) \right] \quad \text{for } i \in S, \tag{3.20}$$

and in addition there exists $v \in B_w(S)$ such that

$$u(i) = \max_{a \in A^*(i)} \left[\sum_{j \in S} q_{ij}(a) v(j) \right] \quad \text{for } i \in S, \tag{3.21}$$

where $A^*(i)$ is the set of actions $a \in A(i)$ attaining the maximum in (3.20).

THEOREM 3.9. *$(g^*, \hat{h})$ is the unique pair that verifies the bias optimality equations. Moreover, a policy $f \in \mathbb{F}$ is bias optimal if and only if for all $i \in S$, $f(i)$ attains the maximum in the two equations* (3.20), (3.21).

Remark 3.10. As in (3.12), from Dynkin's formula and the Poisson equation (3.7) we have

$$J_T(i, f) = T \cdot g(f) + h_f(i) - E_i^f h_f(x(T)) \quad \text{for } f \in \mathbb{F},\ i \in S, \text{ and } T \geq 0.$$

This fact and (2.10) yield

$$\sup_{f \in \mathbb{F}} ||J_T(\cdot, f) - T \cdot g(f) - h_f(\cdot)||_w = \mathrm{O}(e^{-\delta T}).$$

This shows that the rate of convergence to zero of the residual term in (1.2) is uniformly exponential.

4. Proof of Theorem 2.5. Throughout this section, we assume that the hypotheses of Theorem 2.5 are verified, namely, Assumptions A(a)–(c), B(a), and C(a), the continuity of the transition rates, and the uniform integrability condition (2.11). To prove Theorem 2.5, we will consider a discretization of the Markov control process; in other words, we will deal with a "skeleton" CMC. We suppose that the system is observed at times $0, \Delta, 2\Delta, \ldots$, where, for the moment, $\Delta > 0$ is left unspecified.

In the next paragraphs, we make important remarks with regard to the discretization of the continuous-time CMC.

Topology of the action spaces. If the state of the system is i at time $k\Delta$, then the action chosen at i at time $k\Delta$ *does not* suffice to determine the state of the system at time $(k+1)\Delta$. The reason for this is that the process can move to any state $j \in S$ between times $k\Delta$ and $(k+1)\Delta$; as a consequence, we must specify the action chosen not only at i, but at *every* state $j \in S$.

This means that when dealing with the discretized CMC, the set of available actions at the state $i \in S$ is $\mathbb{F}$ instead of $A(i)$. (We are focusing on stationary strategies.) This is not a real problem because $\mathbb{F}$ is metrizable and compact with the topology of componentwise convergence, i.e., $f_n \to f$ if $f_n(i) \to f(i)$ for every $i \in S$; see the paragraph after Assumption 2.1 in [5, p. 541].

Definition of the skeleton CMC. The above paragraph implies that the corresponding skeleton CMC is defined by the following elements:

- the denumerable state space $S = \{0, 1, 2, \ldots\}$,
- the action space $\mathbb{F}$, a compact metric space, which is the same for every state $i \in S$, and
- the transition probabilities $p^f(i, \Delta, j)$, for $\Delta > 0$, $i, j \in S$, and $f \in \mathbb{F}$.

As we are mainly concerned with ergodic properties of Markov processes, the definition of the corresponding reward function is not relevant.

Observe that a deterministic stationary strategy for the skeleton CMC does not correspond to a deterministic stationary strategy for the continuous-time CMC; indeed, the chosen action (in $\mathbb{F}$) would depend on the state of the system at the times $k\Delta$ and, therefore, it would not be stationary (in the continuous-time sense). Hence, in the discrete-time model, we will deal with a subclass of deterministic stationary strategies, namely, we will consider *constant* deterministic stationary strategies, that is, the *same* action $f \in \mathbb{F}$ is chosen at *every* state $i \in S$.

Our plan is as follows. Uniform ergodicity has been proved for discrete-time CMCs in Key Theorem II [5, p. 544]. (For ease of reference, we state this result here in Theorem 4.1.) In Propositions 4.5, 4.6, and 4.9, we propose conditions under which the hypotheses of Theorem 4.1 are verified by the skeleton CMC model. Then, we will prove Theorem 2.5.

THEOREM 4.1. *Suppose that for some $\Delta > 0$, the skeleton CMC verifies the following properties:*

(i) *For every i and j in S, the function*

$$f \mapsto p^f(i, \Delta, j)$$

is continuous on $\mathbb{F}$.

(ii) *For every i in S, the function*

$$f \mapsto \sum_{j \in S} p^f(i, \Delta, j) w(j)$$

is continuous on $\mathbb{F}$.

(iii) *There exists a constant $\gamma < 1$ such that*

$$\sum_{j=1}^{\infty} p^f(i, \Delta, j) w(j) \le \gamma w(i)$$

for every $f \in \mathbb{F}$ and $i \in S$.

Then, there exist constants $H > 0$ and $\beta < 1$ such that

$$\sum_{j \in S} \left| p^f(i, k\Delta, j) - \mu_f(j) \right| w(j) \le H\beta^k w(i) \quad \text{and} \quad \sum_{j \in S} p^f(i, \Delta, j) w(j) \le H w(i)$$

for every $i \in S$, $k \ge 0$ integer, and $f \in \mathbb{F}$. (Recall that μ_f is the invariant probability measure of $\{x^f(t)\}$.)

Proof. For every $f \in \mathbb{F}$, the Markov chain $\{x^f(k\Delta)\}_{k\geq 0}$ is irreducible and aperiodic (recall Assumption C(a)) and, therefore, the stated result directly follows from Key Theorem II in [5, p. 544]. □

Next, we are going to prove that condition (i) in Theorem 4.1 is verified for every $\Delta > 0$. First, we will show it for a model with *bounded* transition rates, and then this will be extended to a model with unbounded transition rates.

We give some previous definitions. Define the 1-norm of the real-valued vector $z := \{z(k)\}_{k\in S}$ as the supremum norm, that is, $||z||_1 := \sup_{k\in S} |z(k)|$. Let $C(1) := \{z \in \mathbb{R}^S : ||z||_1 \leq 1\} \subseteq B_w(S)$. For a given $f \in \mathbb{F}$, recall that the operator Q^f was defined in (3.10) as

$$(Q^f z)(i) := \sum_{k\in S} q_{ik}(f)z(k) \quad \text{for } i \in S,$$

where $z = \{z(k)\}_{k\in S}$.

LEMMA 4.2. *Suppose* $\sup_{i\in S} q(i) =: q < \infty$. *Then, the following results hold.*

(i) *For each* $f \in \mathbb{F}$, Q^f *is a linear operator on* $B_w(S)$ *that satisfies*

$$||Q^f z||_w \leq (q + b)||z||_w, \tag{4.1}$$

where b *is as in Assumption A.*

(ii) *If the sequence* $\{f_n\}_{n\in\mathbb{N}} \subseteq \mathbb{F}$ *converges to* $f \in \mathbb{F}$, *then*

$$\lim_{n\to\infty} \sup_{z\in C(1)} \{||Q^{f_n} z - Q^f z||_w\} = 0.$$

(iii) *If the sequence* $\{z_n\}_{n\in\mathbb{N}} \subseteq C(1)$ *converges pointwise to* $z \in C(1)$, *i.e.,* $z_n(k) \to z(k)$ *for every* $k \in S$, *then*

$$\lim_{n\to\infty} ||Q^f z_n - Q^f z||_w = 0 \quad \textit{for every } f \in \mathbb{F}.$$

Proof. The proof of this lemma is straightforward and we omit it. However, it is worth noting that the following is needed in the proof:

$$\lim_{k\to\infty} \sup_{a\in A(i)} \left\{ \sum_{j=k}^{\infty} q_{ij}(a) \right\} = 0 \quad \text{for every } i \in S,$$

and this result is derived from Dini's theorem. □

Suppose that $f \in \mathbb{F}$ and $j \in S$ are given. Define $z^f(t) \in C(1)$ (we do not make j explicit in the notation) as

$$(z^f(t))(i) := p^f(i, t, j) \quad \text{for } i \in S \text{ and } t \geq 0.$$

LEMMA 4.3. *Suppose that* $\sup_{i\in S} q(i) =: q < \infty$. *Then,*

(i) *For every* $f \in \mathbb{F}$ *and* $j \in S$, *the function* $t \mapsto z^f(t)$ *is* $C^1[0, \infty)$ *and it is a solution of the differential equation*

$$\frac{d}{dt} z^f(t) = Q^f z^f(t) \quad \textit{for } t \geq 0, \tag{4.2}$$

where continuity and differentiability are understood to be in the w*-norm* (*and* not *in a componentwise sense*).

(ii) *For every* $t \geq 0$ *and* $i, j \in S$, $f \mapsto p^f(i, t, j)$ *is continuous on* $\mathbb{F}$.

Proof of (i). Observe that the equality (4.2) corresponds to Kolmogorov's backward equations, which are known to hold in a *componentwise* sense. The fact that $t \mapsto z^f(t)$ is $C^1[0,\infty)$ in the w-norm and that (4.2) holds in the w-norm is deduced from Lemma 4.2(iii), and we skip the details. □

Proof of (ii). It is well known that

$$||y(t)|| \le ||y(0)|| + \int_0^t ||y'(s)||ds \quad \text{for } t \ge 0, \tag{4.3}$$

where y is a $C^1[0,\infty)$ function on a Banach space.

Fix an arbitrary $f \in \mathbb{F}$ and a sequence $\{f_n\}_{n\in\mathbb{N}} \subseteq \mathbb{F}$ that converges to f. By part (i) of this lemma and (4.3),

$$||z^{f_n}(t) - z^f(t)||_w \le \int_0^t ||Q^{f_n} z^{f_n}(s) - Q^f z^f(s)||_w ds \quad \text{for } t \ge 0. \tag{4.4}$$

On the other hand, $||Q^{f_n} z^{f_n}(s) - Q^f z^f(s)||_w$ is bounded by

$$||Q^{f_n} z^{f_n}(s) - Q^{f_n} z^f(s)||_w + ||Q^{f_n} z^f(s) - Q^f z^f(s)||_w,$$

which, by (4.1), is less than or equal to

$$(q+b)||z^{f_n}(s) - z^f(s)||_w + ||Q^{f_n} z^f(s) - Q^f z^f(s)||_w.$$

Now, by Lemma 4.2(ii), given $\varepsilon > 0$ there exists N (not depending on $s \ge 0$) such that

$$||Q^{f_n} z^f(s) - Q^f z^f(s)||_w \le \varepsilon \quad \text{for } n \ge N \text{ and every } s \ge 0.$$

Combining these bounds with (4.4), we deduce that

$$||z^{f_n}(t) - z^f(t)||_w \le (q+b) \int_0^t ||z^{f_n}(s) - z^f(s)||_w ds + \varepsilon t \quad \text{for } n \ge N \text{ and } t \ge 0.$$

Hence, applying Gronwall's inequality, we have that

$$||z^{f_n}(t) - z^f(t)||_w \le \varepsilon T e^{(q+b)T} \quad \text{for } n \ge N \text{ and } 0 \le t \le T,$$

where $T > 0$ is arbitrary. In particular,

$$|p^{f_n}(i,t,j) - p^f(i,t,j)| \le w(i)\varepsilon T e^{(q+b)T} \quad \text{for } n \ge N,$$

thus showing that $f \mapsto p^f(i,t,j)$ is continuous on $\mathbb{F}$, as we wanted to prove. □

Now we drop the boundedness hypothesis on the transition rates made in Lemmas 4.2 and 4.3. Given a state $m \in S$ and $f \in \mathbb{F}$, let

$$T_m := \min_{s\ge 0}\{x^f(s) \ge m\}.$$

Define then the Markov process $\{x_m^f(t)\}$ as

$$x_m^f(t) := x^f(t \wedge T_m)$$

(where "$\wedge$" stands for *minimum*). The transition probabilities of $\{x_m^f\}$ are denoted $p_m^f(i,t,j)$ for $t \ge 0$ and $i, j \in S$.

LEMMA 4.4. *For arbitrary $i, j \in S$ and $t \geq 0$,*

$$\lim_{m\to\infty} \sup_{f\in\mathbb{F}} \{|p^f(i,t,j) - p^f_m(i,t,j)|\} = 0.$$

(*Notice that we do not assume* $\sup_{i\in S} q(i) < \infty$.)

Proof. We have that

$$\begin{aligned} p^f(i,t,j) - p^f_m(i,t,j) &= E^f_i \left[I\{x^f(t) = j\} - I\{x^f_m(t) = j\} \right] \\ &= E^f_i \left[I\{t \geq T_m\} (I\{x^f(t) = j\} - I\{x^f_m(t) = j\}) \right], \end{aligned}$$

where $I\{\cdot\}$ denotes the indicator function. Hence,

$$|p^f(i,t,j) - p^f_m(i,t,j)| \leq 2P^f_{0,i}\{t \geq T_m\}. \tag{4.5}$$

(Recall the notation introduced in section 2.) We also have that

$$\begin{aligned} \{t \geq T_m\} &= \left\{ \sup_{0\leq s\leq t} \{x^f(s)\} \geq m \right\} \\ &\subseteq \left\{ \sup_{0\leq s\leq t} \{w(x^f(s))e^{-bs}\} \geq w(m)e^{-bt} \right\}, \end{aligned}$$

where the constant $b \geq 0$ is taken from Assumption A. Thus, from the nonexplosion condition CD0 in [21, p. 524] and Theorem 2.1 in [21], we deduce that

$$P^f_{0,i}\{t \geq T_m\} \leq w(i)e^{bt}/w(m)$$

(it is important to note that this bound is uniform in $f \in \mathbb{F}$ because the drift constants in [21, condition CD0] do not depend on f), which, together with (4.5), yields that

$$\sup_{f\in\mathbb{F}}\{|p^f(i,t,j) - p^f_m(i,t,j)|\} \leq 2w(i)e^{bt}/w(m).$$

Letting $m \to \infty$ in this expression proves the lemma. □

Now we can show that condition (i) in Theorem 4.1 holds.

PROPOSITION 4.5. *For every $t \geq 0$, and $i, j \in S$ fixed, the function*

$$f \mapsto p^f(i,t,j)$$

is continuous on $\mathbb{F}$.

Proof. Observe that the process $\{x^f_m(t)\}$ verifies the hypotheses of Lemma 4.3 because its transition rates are bounded by $\sup_{0\leq i<m} q(i)$, and then, for fixed $i, j, m \in S$ and $t \geq 0$, $f \mapsto p^f_m(i,t,j)$ is continuous on $\mathbb{F}$. The continuity of $f \mapsto p^f(i,t,j)$ is now a straightforward consequence of Lemma 4.4. □

Condition (ii) in Theorem 4.1, stated in Proposition 4.6, is now easily derived from condition (2.11), and the proof is omitted.

PROPOSITION 4.6. *If condition* (2.11) *is satisfied, then for every $t \geq 0$ and $i \in S$ fixed, the function*

$$f \mapsto \sum_{j\in S} p^f(i,t,j)w(j)$$

is continuous on $\mathbb{F}$.

Remark 4.7. Observe that, taking into account the result of Proposition 4.5, it follows that the uniform integrability condition (2.11) and the continuity of $f \mapsto \sum_{j\in S} p^f(i,t,j)w(j)$ are in fact equivalent. Indeed, the continuous functions

$$f \mapsto \sum_{j=0}^{k} p^f(i,t,j)w(j)$$

form an increasing sequence that converges to the continuous function

$$f \mapsto \sum_{j\in S} p^f(i,t,j)w(j).$$

Hence, by Dini's theorem, convergence is uniform on $\mathbb{F}$, and then (2.11) follows.

Remark 4.8. It could seem that

$$f \mapsto \sum_{j\in S} q_{ij}(f)w(j) \text{ is continuous} \tag{4.6}$$

is the continuous-time version of the discrete-time condition in Theorem 4.1(ii)

$$f \mapsto \sum_{j\in S} p^f(i,t,j)w(j) \text{ is continuous.} \tag{4.7}$$

However, this is not true because (4.6) takes into account only the first transition when starting from state $i \in S$, while (4.7) includes all the transitions that occured in the time interval $[0,t]$.

The following result establishes condition (iii) in Theorem 4.1.

PROPOSITION 4.9. *There exists* $\Delta > 0$ *such that*

$$\sum_{j=1}^{\infty} p^f(i,t,j)w(j) \le e^{-ct/2}w(i)$$

for every $i \in S$, $0 \le t \le \Delta$, *and* $f \in \mathbb{F}$, *where* c *is as in Assumption A.*

Proof. From the drift condition in Assumption A

$$\sum_{j=0}^{\infty} q_{ij}(a)w(j) \le -cw(i) + b\delta_{i0} \quad \text{for every } i \in S \text{ and } a \in A(i),$$

and using Dynkin's formula, we obtain (see, e.g., [6, equation 31])

$$\sum_{j=0}^{\infty} p^f(i,t,j)w(j) \le e^{-ct}w(i) + be^{-ct}\int_0^t p^f(i,s,0)e^{cs}ds, \tag{4.8}$$

which holds for every $f \in \mathbb{F}$, $i \in S$, and $t \ge 0$. The inequality (4.8) may be rewritten as

$$\sum_{j=1}^{\infty} p^f(i,t,j)w(j) \le e^{-ct/2}w(i)R(t,i,f),$$

where

$$R(t,i,f) := e^{-ct/2}\left(1 + \frac{b}{w(i)}\int_0^t p^f(i,s,0)e^{cs}ds - \frac{w(0)}{w(i)}p^f(i,t,0)e^{ct}\right).$$

Hence, to prove the proposition, it suffices to show that there exists $\Delta > 0$ such that $R(t,i,f) \leq 1$ or, equivalently,

$$\frac{b}{w(i)} \int_0^t p^f(i,s,0) e^{cs} ds - \frac{w(0)}{w(i)} p^f(i,t,0) e^{ct} \leq e^{ct/2} - 1 \tag{4.9}$$

for every $0 \leq t \leq \Delta$, $i \in S$, and $f \in \mathbb{F}$. We will proceed in two steps, depending on whether $i = 0$ or $i > 0$.

Case $i = 0$. Proving (4.9) for $i = 0$ is equivalent to show that

$$\frac{b}{w(0)} \int_0^t p^f(0,s,0) e^{cs} ds - p^f(0,t,0) e^{ct} \leq e^{ct/2} - 1,$$

and it suffices that

$$p^f(0,t,0) e^{ct} \geq \frac{b}{w(0)} \int_0^t p^f(0,s,0) e^{cs} ds.$$

Observe now that

$$p^f(0,s,0) e^{cs} \leq e^{ct} \quad \text{and} \quad p^f(0,t,0) \geq e^{-q(0)t}$$

(see (2.3) for the definition of $q(0)$) for every $0 \leq s \leq t$ and $f \in \mathbb{F}$. Hence, choosing $\Delta_1 > 0$ such that

$$e^{-q(0)t} \geq \frac{b}{w(0)} t \quad \text{for } 0 \leq t \leq \Delta_1$$

yields

$$R(t,0,f) \leq 1 \quad \text{for } 0 \leq t \leq \Delta_1 \text{ and } f \in \mathbb{F}.$$

Case $i > 0$. We will show that for $i \geq 1$,

$$\frac{b}{w(i)} \int_0^t p^f(i,s,0) e^{cs} ds \leq e^{ct/2} - 1. \tag{4.10}$$

Note that $e^{ct/2} - 1 \geq ct/2$ and that $\int_0^t p^f(i,s,0) e^{cs} ds \leq t e^{ct}$. As a consequence, there exists some $I \geq 1$ (not depending on f) such that

$$\frac{b}{w(i)} e^{ct} \leq \frac{c}{2} \quad \text{for } 0 \leq t \leq \Delta_1 \text{ and } i > I.$$

Therefore, $R(t,i,f) \leq 1$ for every $0 \leq t \leq \Delta_1$, $i > I$, and $f \in \mathbb{F}$.
In case $1 \leq i \leq I$, observe that

$$p^f(i,s,0) \leq 1 - e^{q_{ii}(f)s} \leq -q_{ii}(f)s \leq q(i)s.$$

Thus, in order to establish (4.10), it is enough to prove that

$$\frac{b e^{ct} q(i) t^2}{2 w(i)} \leq ct/2,$$

which is verified for every $1 \leq i \leq I$, $f \in \mathbb{F}$, and $0 \leq t \leq \Delta_2$, provided that $\Delta_2 > 0$ is small enough.

To conclude, we have proved that $R(t,i,f) \le 1$ for every $0 \le t \le \min\{\Delta_1, \Delta_2\}$, $i \in S$ and $f \in \mathbb{F}$. □

Remark 4.10. The term $e^{-ct/2}$ in the statement of Proposition 4.9 can be replaced with $e^{-\alpha t}$ for any $0 < \alpha < c$, and the proof would be the same.

Finally, we can prove our main result.

Proof of Theorem 2.5. Choosing $\Delta > 0$ as in Proposition 4.9, it follows from Propositions 4.5, 4.6, and 4.9 that conditions (i), (ii), and (iii) in Theorem 4.1 are verified by the skeleton CMC.

Hence, given arbitrary $t \ge 0$, let $0 \le s < \Delta$ and $k \ge 0$ integer satisfy $t = k\Delta + s$. If $i \in S$, $u \in B_w(S)$, and $f \in \mathbb{F}$,

$$\begin{aligned} |E_i^f u(x(t)) - \mu_f(u)| &= \left| \sum_{j \in S} (p^f(i,t,j) - \mu^f(j)) u(j) \right| \\ &\le ||u||_w \sum_{j \in S} w(j) |p^f(i,t,j) - \mu^f(j)| \\ &= ||u||_w \sum_{j \in S} w(j) \left| \sum_{i' \in S} p^f(i,s,i') (p^f(i',k\Delta,j) - \mu^f(j)) \right|, \end{aligned}$$

where we use the Chapman–Kolmogorov equation. Hence,

$$\begin{aligned} |E_i^f u(x(t)) - \mu_f(u)| &\le ||u||_w \sum_{i' \in S} p^f(i,s,i') \sum_{j \in S} w(j) |p^f(i',k\Delta,j) - \mu^f(j)| \\ &\le H\beta^k ||u||_w \sum_{i' \in S} p^f(i,s,i') w(i') \\ &\le H\beta^{-1} (\beta^{1/\Delta})^t ||u||_w \sum_{i' \in S} p^f(i,s,i') w(i') \end{aligned}$$

for some constants H and $\beta < 1$ (see Theorem 4.1). Finally, [10, Lemma 3.2] yields that

$$|E_i^f u(x(t)) - \mu_f(u)| \le H\beta^{-1} (\beta^{1/\Delta})^t ||u||_w (1 + b/c) w(i),$$

where the constants b and c come from Assumption A, thus showing that the uniform w-ergodic property holds, letting

$$R := H\beta^{-1}(1 + b/c) \quad \text{and} \quad \delta := -(\log \beta)/\Delta$$

in Assumption C(b). □

5. Proof of theorems in section 3.

Proof of Theorem 3.3. By (3.4), we only need to show that $\mathbb{F}_{ao} \subseteq \mathbb{F}_{ca}$. Let $f \in \mathbb{F}$ be average optimal, so that (by (2.9) and (3.3)) $g(f) = g^*$. Recalling the Poisson equation (3.7), and if (g^*, h^*) is a solution of the AROE (3.1), then for all $i \in S$,

$$\begin{aligned} g^* &= r(i,f) + \sum_{j \in S} q_{ij}(f) h_f(j) \\ &= \max_{a \in A(i)} \left[r(i,a) + \sum_{j \in S} q_{ij}(a) h^*(j) \right] \\ &\ge r(i,f) + \sum_{j \in S} q_{ij}(f) h^*(j), \end{aligned}$$

and thus, with the notation introduced in (3.13),

$$(Q^f(h_f - h^*))(i) \geq 0 \quad \text{for } i \in S.$$

Using Dynkin's formula, the above yields $E_i^f[h_f(x(t)) - h^*(x(t))] \geq h_f(i) - h^*(i)$, and letting $t \to \infty$ we obtain

$$\mu_f(h_f - h^*) \geq h_f - h^*.$$

Since $\mu_f(h_f - h^*)$ is finite, it follows that there exists a constant $z \in \mathbb{R}$ such that

$$h_f(i) = h^*(i) + z \tag{5.1}$$

for all i in a set of μ_f-probability one, and thus (5.1) holds for every $i \in S$ because the Markov process $\{x^f(t)\}$ is irreducible (recall Assumption C(a)). Finally, as a consequence of Remark 3.2, the pair (g^*, h_f) is a solution of the AROE (3.1), and then

$$\begin{aligned} g^* &= \max_{a \in A(i)} \left[r(i,a) + \sum_{j \in S} q_{ij}(a) h_f(j) \right] \\ &\geq r(i,f) + \sum_{j \in S} q_{ij}(f) h_f(j) = g^* \quad \text{(by (3.7))}. \end{aligned}$$

Hence f is canonical. □

Proof of Theorem 3.6. Let (g^*, h^*, f^*) be a canonical triplet. We will first show that the expressions on the right-hand sides of (i) and (ii) coincide, i.e.,

$$J_T(i, s, f^*, h^*) = (T - s)g^* + h^*(i). \tag{5.2}$$

Indeed, in (3.11)–(3.12) replace $[0, t]$ and (g_k, h_k, f) with $[s, T]$ and (g^*, h^*, f^*), respectively, to obtain (via (3.2))

$$E_{s,i}^{f^*} h^*(x(T)) = h^*(i) + (T - s) \cdot g^* - J_T(i, s, f^*, 0).$$

Thus, rearranging terms we get (5.2).

Now let $v(i, s) := (T - s)g^* + h^*(i)$ for $i \in S$ and $0 \leq s \leq T$, and note that $v(\cdot, s)$ is in $B_w(S)$. Given $f \in \mathbb{F}$, let L^f be the *extended generator* of the state (Markov) process $\{x^f(t), t \geq 0\}$ (see, e.g., [13, p. 2]). Then we obviously have

$$(L^f v)(i, s) = -g^* + \sum_{j \in S} q_{ij}(f) h^*(j)$$

and, therefore,

$$r(i, f) + (L^f v)(i, s) \leq 0 \quad \text{for all } i \in S \text{ and } 0 \leq s \leq T \tag{5.3}$$

with equality if $f = f^*$. Hence, since we also have $v(i, T) = h^*(i)$, we conclude from (5.3) and (3.1)–(3.2) that $v(i, s)$ satisfies the dynamic programming equation for the control problem (3.16) with terminal payoff h^*, namely,

$$\max_{f \in \mathbb{F}} [r(i, f) + (L^f v)(i, s)] = 0 \quad \text{for all } i \in S \text{ and } 0 \leq s \leq T,$$

with boundary condition $v(i,T) = h^*(i)$ for all $i \in S$. Therefore, by a standard dynamic programming "verification theorem" (see [13, Theorem 5.1], for instance), $v(i,s)$ is the control problem's value function (3.16) with $h = h^*$, i.e.,

$$v(i,s) = J_T^*(i,s,h^*).$$

Hence from the definition of $v(i,s)$ and (5.2) we obtain (i), (ii).

Conversely, suppose that (i) and (ii) hold. Given $f \in \mathbb{F}$ we have

$$J_T(i,f) \le Tg^* + h^*(i) - E_i^f h^*(x(T)) \quad \text{for each } i \in S$$

with equality if $f = f^*$. Multiplying by $1/T$ and letting $T \to \infty$ in the above inequality, and also recalling (3.9), we derive that $g(f) \le g^*$ for all $f \in F$ and $g(f^*) = g^*$. Therefore, g^* is the optimal gain and the stationary policy f^* is average optimal. Hence there exists a function $\tilde{h} \in B_w(S)$ such that $(g^*, \tilde{h})$ is a solution of the AROE (3.1), i.e.,

$$g^* = \max_{a \in A(i)} \left[r(i,a) + \sum_{j \in S} q_{ij}(a) \tilde{h}(j) \right] \quad \text{for } i \in S, \tag{5.4}$$

and, by Theorem 3.3, f^* attains the maximum in (5.4) so that, from the proof of Proposition 3.4, h_{f^*} and $\tilde{h}$ differ by a constant. Also, as mentioned above,

$$J_T(i,f^*) = Tg^* + h^*(i) - E_i^{f^*} h^*(x(T)) \quad \text{for } i \in S$$

and, again from the proof of Proposition 3.4, we know that

$$J_T(i,f^*) = Tg^* + h_{f^*}(i) - E_i^{f^*} h_{f^*}(x(T)) \quad \text{for } i \in S.$$

Substracting these equations and letting $T \to \infty$ we derive that h^* and h_{f^*} differ by a constant. This fact, together with (2.2), allows us to replace $\tilde{h}$ with h^* in (5.4):

$$g^* = \max_{a \in A(i)} \left[r(i,a) + \sum_{j \in S} q_{ij}(a) h^*(j) \right] \quad \text{for } i \in S,$$

and we also have

$$g^* = r(i,f^*) + \sum_{j \in S} q_{ij}(f^*) h^*(j) \quad \text{for } i \in S.$$

Hence, (g^*, h^*, f^*) is a canonical triplet. □

Proof of Theorem 3.8. (a) Using (3.15) we may rewrite (3.17) as

$$\hat{h}(i) = h^*(i) + \sup_{f \in \mathbb{F}_{ao}} \left[\sum_{j \in S} -h^*(j) \right] \mu_f(j). \tag{5.5}$$

Therefore, finding bias optimal policies reduces to an average reward problem with reward rate

$$r'(i,a) := -h^*(i) \tag{5.6}$$

and for which, by (5.5), the maximization is over $\mathbb{F}_{ao} = \mathbb{F}_{ca}$. Observe that if $A^*(i)$ is as in (3.21), then a policy f is in $\mathbb{F}_{ca}$ if and only if $f(i) \in A^*(i)$ for all $i \in S$; in other words, f satisfies (3.2). Hence, instead of the control model (2.1) we now consider

$$\mathcal{M} := \{S, (A^*(i), i \in S), q_{ij}(a), r'(i,a)\} \tag{5.7}$$

with r' as in (5.6). Now, by Assumption B, $A^*(i)$ is a (nonempty) compact subset of $A(i)$ and, on the other hand, r' satisfies Assumption A(d) because h^* is in $B_w(S)$. Thus $\mathcal{M}$ satisfies Assumptions A, B, and C, and so the set of corresponding stationary average optimal policies, which is the same as $\mathbb{F}_{bias}$, is nonempty.

(b) To prove the equivalence of (b_1) and (b_2), suppose first that $\hat{f}$ is bias optimal. Then, by definition, $\hat{f}$ is canonical and

$$h_{\hat{f}}(\cdot) = \hat{h}(\cdot).$$

Thus, integration with respect to $\mu_{\hat{f}}$ gives $\mu_{\hat{f}}(\hat{h}) = 0$; that is, (b_1) implies (b_2). Conversely, suppose that (i) $\hat{f}$ is canonical, and (ii) $\mu_{\hat{f}}(\hat{h}) = 0$. By (i), $\hat{f}$ satisfies (3.2), i.e.,

$$g^* = r(i, \hat{f}) + \sum_{j \in S} q_{ij}(\hat{f}) h^*(j),$$

which can be written as

$$g(\hat{f}) = r(i, \hat{f}) + \sum_{j \in S} q_{ij}(\hat{f}) \hat{h}(j),$$

because $g^* = g(\hat{f})$ and the functions h^* and $\hat{h}$ differ by only a constant—see (5.5). Thus, (ii) and Proposition 3.4 imply that $\hat{h} = h_{\hat{f}}$, and (b_1) follows.

$(b_1) \Leftrightarrow (b_3)$. From Remark 3.10,

$$J_T(i,f) = T \cdot g(f) + h_f(i) - E_i^f h_f(x(T))$$

for all $f \in \mathbb{F}$, $i \in S$, and $T > 0$ with $E_i^f h_f(x(T)) \to 0$ as $T \to \infty$. Therefore, if $\hat{f} \in \mathbb{F}$ is any other stationary policy we obtain

$$\begin{aligned} J_T(i,\hat{f}) - J_T(i,f) &= T \cdot [g(\hat{f}) - g(f)] \\ &\quad + h_{\hat{f}}(i) - h_f(i) - (E_i^{\hat{f}} h_{\hat{f}}(x(T)) - E_i^f h_f(x(T))). \end{aligned} \tag{5.8}$$

To prove that (b_1) implies (b_3) suppose that $\hat{f}$ is bias optimal. Then for any $f \in \mathbb{F}$, either $g(\hat{f}) = g^* > g(f)$ or $g(\hat{f}) = g(f)$ and $h_{\hat{f}}(\cdot) = \hat{h}(\cdot) \geq h_f(\cdot)$, and so letting $T \to \infty$ in (5.8) we get (3.18).

To prove the converse, suppose that $\hat{f}$ satisfies (3.18) and that $\hat{f}$ is not bias optimal. If $f \in \mathbb{F}$ is a bias optimal policy, then $g(f) > g(\hat{f})$ or $g(f) = g(\hat{f})$ (that is, $\hat{f}$ is average optimal) and $h_f(i) > h_{\hat{f}}(i)$ for all $i \in S$, as a consequence of Proposition 3.5. In both cases

$$\lim_{T \to \infty} [J_T(i,f) - J_T(i,\hat{f})] > 0,$$

contradicting (3.18). □

Proof of Theorem 3.9. First, let us rewrite (5.5) as

$$\hat{h}(\cdot) = h^*(\cdot) + \hat{g}, \quad \text{with } \hat{g} := \sup_{f \in \mathbb{F}_{ao}} \mu_f(-h^*), \tag{5.9}$$

and let $A^*(i)$ be as in (3.21). It then follows that the AROE (3.1) can be expressed as

$$g^* = \max_{a \in A(i)} \left[r(i,a) + \sum_{j \in S} q_{ij}(a) \hat{h}(j) \right] \quad \text{for } i \in S; \tag{5.10}$$

that is, $(g^*, \hat{h})$ satisfies the *first* bias optimality equation (3.20). Now let $v \in B_w(S)$ be such that $(\hat{g}, v)$ is a *solution of the AROE for the average control model* $\mathcal{M}$ in (5.7), so that [as in (3.1)]

$$\begin{aligned} \hat{g} &= \max_{a \in A^*(i)} \left[-h^*(i) + \sum_{j \in S} q_{ij}(a) v(j) \right] \\ &= -h^*(i) + \max_{a \in A^*(i)} \sum_{j \in S} q_{ij}(a) v(j). \end{aligned}$$

Equivalently, by (5.9),

$$\hat{h}(i) = \max_{a \in A^*(i)} \sum_{j \in S} q_{ij}(a) v(j) \quad \text{for } i \in S, \tag{5.11}$$

which means that $(g^*, \hat{h})$ also satisfies the *second* bias optimality equation (3.21). Moreover (recall Remark 3.2), from (5.9)–(5.11) it follows that g^* and $\hat{h}$ form the unique pair that satisfies the bias optimality equations.

Finally, observe that a stationary policy $f \in \mathbb{F}$ attains the maximum in both (5.10) and (5.11) if and only if f is canonical (by (5.10)) and $\mu_f(\hat{h}) = 0$ (by (5.11)). Therefore, the last statement of Theorem 3.9 follows from the equivalence of (b_1) and (b_2) in Theorem 3.8. □

6. Concluding remarks. In the previous sections we presented an essentially self-contained analysis of the bias optimality criterion for a class of continuous-time CMCs with possibly unbounded reward and transition rates. Our hypotheses allow us to obtain not only the existence of bias optimal policies but also some interesting characterizations. For instance, (3.18)–(3.19) show that among the class of average optimal policies, those with the optimal bias are basically those for which the expected total reward grows faster. In contrast, the characterization in Theorem 3.9 gives that a bias optimal policy is one that lexicographically maximizes the vector criterion $(g(f), h_f)$.

These characterizations together with the relations obtained in [23] between average reward optimality and some "discount-sensitive" criteria give a quite complete picture of what bias optimality is about.

However, there remain many open problems. For instance, the proofs of our results are greatly simplified by the fact that the state space of the CMC is a *countable* set. The obvious question then is, can our results be extended to other classes of continuous-time controlled Markov processes (CMPs), for example, diffusions, general jump processes, and so on?

Finally, it would be interesting to consider some of the problems above for continuous-time *Markov games* rather than CMPs. For instance, there are some connections between average and discounted criteria [22] for discrete-time games that could be naturally extended to the continuous-time case. One should be careful, however, because some results for CMPs cannot be extrapolated to stochastic games. For example, the equivalence between bias optimality and weak overtaking optimality that has been proved in Theorem 3.8 for control problems does not hold for stochastic games [24], thus exhibiting a (perhaps unexpected) discrepancy between control and games models.

REFERENCES

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

[2] G. Carrasco and O. Hernández-Lerma, *Blackwell Optimality for Markov Control Processes in Borel Spaces*, Departamento de Matemáticas, CINVESTAV-IPN, Mexico, 2003, preprint.

[3] M. F. Chen, *Equivalence of exponencial ergodicity and L^2 exponential convergence for Markov chains*, Stoch. Proc. Appl., 87 (2000), pp. 281–297.

[4] P. Coolen-Schrijner and E. A. van Doorn, *The deviation matrix of a continuous-time Markov chain*, Prob. Engrg. Inform. Sci., 16 (2002), pp. 351–366.

[5] R. Dekker, A. Hordijk, and F. M. Spieksma, *On the relation between recurrence and ergodicity properties in denumerable Markov decision chains*, Math. Oper. Res., 19 (1994), pp. 539–559.

[6] D. Down, S. P. Meyn, and R. L. Tweedie, *Exponential and uniform ergodicity of Markov processes*, Ann. Probab., 23 (1995), pp. 1671–1691.

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

[8] D. Gale, *On optimal development in a multi-sector economy*, Rev. Econom. Stud., 34 (1967), pp. 1–19.

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

[10] X. P. Guo and O. Hernández-Lerma, *Drift and monotonicity conditions for continuous-time controlled Markov chains*, IEEE Trans. Automat. Control, 48 (2003), pp. 236–244.

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

[12] M. Haviv and M. L. Puterman, *Bias optimality in controlled queueing systems*, J. Appl. Probab., 35 (1998), pp. 136–150.

[13] O. Hernández-Lerma, *Lectures on Continuous-Time Markov Control Processes*, Sociedad Matemática Mexicana, Mexico City, 1994.

[14] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York, 1996.

[15] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer, New York, 1999.

[16] N. Hilgert and O. Hernández-Lerma, *Bias optimality versus strong* 0*-discount optimality in Markov control processes with unbounded costs*, Acta Appl. Math., 76 (2003), pp. 215–235.

[17] A. Hordijk and A. A. Yushkevich, *Blackwell optimality*, in Handbook of Markov Decision Processes, E. A. Feinberg and A. Shwartz, eds., Internat. Ser. Oper. Res. Management Sci. 40, Kluwer, Boston, 2002, pp. 231–267.

[18] M. E. Lewis and M. L. Puterman, *A probabilistic analysis of bias optimality in unichain Markov decision processes*, IEEE Trans. Automat. Control, 46 (2001), pp. 96–100.

[19] M. E. Lewis and M. L. Puterman, *Bias optimality*, in Handbook of Markov Decision Processes, E. A. Feinberg and A. Shwartz, eds., Internat. Ser. Oper. Res. Management Sci. 40, Kluwer, Boston, 2002, pp. 89–111.

[20] R. B. Lund, S. P. Meyn, and R. L. Tweedie, *Computable exponential convergence rates for stochastically ordered Markov processes*, Ann. Appl. Prob., 6 (1996), pp. 218–237.

[21] S. P. Meyn and R. L. Tweedie, *Stability of Markovian processes* III: *Foster-Lyapunov criteria for continuous-time processes*, Adv. Appl. Prob., 25 (1993), pp. 518–548.

[22] M. A. Muruaga and R. Vélez, *Asymptotic behavior of continuous stochastic games*, Top, 4 (1996), pp. 187–214.

[23] T. Prieto-Rumeau and O. Hernández-Lerma, *The Laurent series, sensitive discount and Blackwell optimality for continuous-time controlled Markov chains*, Math. Methods Oper. Res., 61 (2005), pp. 123–145.
[24] T. Prieto-Rumeau and O. Hernández-Lerma, *Bias and overtaking equilibria for zero-sum continuous-time Markov games*, Math. Methods Oper. Res., 61 (2005), pp. 437–454.
[25] M. L. Puterman, *Sensitive discount optimality in controlled one-dimensional diffusions*, Ann. Probab., 2 (1974), pp. 408–419.
[26] M. L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.
[27] F. P. Ramsey, *A mathematical theory of savings*, Econom. J., 38 (1928), pp. 543–559.
[28] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York, 1999.
[29] R. L. Tweedie, *Criteria for ergodicity, exponential ergodicity and strong ergodicity of Markov processes*, J. Appl. Probab., 18 (1981), pp. 122–130.
[30] A. F. Veinott, Jr., *On finding optimal policies in discrete dynamic programming with no discounting*, Ann. Math. Statist., 37 (1966), pp. 1284–1294.
[31] C. C. von Weizsäcker, *Existence of optimal programs of accumulation for an infinite horizon*, Rev. Econom. Stud., 32 (1965), pp. 85–104.
[32] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications*, Springer, New York, 1998.
[33] A. A. Yushkevich, *On a class of strategies in general Markov decision models*, Theory Probab. Appl., 18 (1973), pp. 777–779.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 74–106

UNIQUENESS RESULTS FOR SECOND-ORDER BELLMAN–ISAACS EQUATIONS UNDER QUADRATIC GROWTH ASSUMPTIONS AND APPLICATIONS*

FRANCESCA DA LIO† AND OLIVIER LEY‡

Abstract. In this paper, we prove a comparison result between semicontinuous viscosity sub- and supersolutions growing at most quadratically of second-order degenerate parabolic Hamilton–Jacobi–Bellman and Isaacs equations. As an application, we characterize the value function of a finite horizon stochastic control problem with unbounded controls as the unique viscosity solution of the corresponding dynamic programming equation.

Key words. degenerate parabolic equations, nonlinear Hamilton–Jacobi equations, nonlinear Isaacs equations, viscosity solutions, unbounded solutions, maximum principle, linear quadratic problems.

AMS subject classifications. 35K65, 49L99, 49L25, 35B50, 49N10

DOI. 10.1137/S0363012904440897

1. Introduction.

In this paper, we are interested in the second-order equation

$$\begin{cases} \dfrac{\partial w}{\partial t} + H(x,t,Dw,D^2w) + G(x,t,Dw,D^2w) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ w(x,0) = \psi(x) & \text{in } \mathbb{R}^N, \end{cases} \tag{1.1}$$

where $N \geq 1$, $T > 0$, the unknown w is a real-valued function defined in $\mathbb{R}^N \times [0,T]$, Dw and D^2w denote, respectively, its gradient and Hessian matrix, and ψ is a given initial condition. The Hamiltonians $H, G : \mathbb{R}^N \times [0,T] \times \mathbb{R}^N \times \mathcal{S}_N(\mathbb{R}) \to \mathbb{R}$ are continuous in all their variables and have the form

$$H(x,t,p,X) = \inf_{\alpha \in A} \left\{ \langle b(x,t,\alpha), p \rangle + \ell(x,t,\alpha) - \operatorname{Tr}\left[\sigma(x,t,\alpha)\sigma^T(x,t,\alpha)X\right] \right\} \tag{1.2}$$

and

$$G(x,t,p,X) = \sup_{\beta \in B} \left\{ -\langle g(x,t,\beta), p \rangle - f(x,t,\beta) - \operatorname{Tr}\left[c(x,t,\beta)c^T(x,t,\beta)X\right] \right\}. \tag{1.3}$$

This kind of equation is of particular interest for applications since it relies on differential game theory (Isaacs equations) or on deterministic and stochastic control problems when either $H \equiv 0$ or $G \equiv 0$ (Hamilton–Jacobi–Bellman equations).

Notations and precise assumptions are given in section 2 but we point out that we allow one of the control set A or B to be unbounded and the solutions to (1.1) may have quadratic growth. Our model case is the well-known stochastic linear quadratic

*Received by the editors February 10, 2004; accepted for publication (in revised form) September 18, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sicon/45-1/44089.html

†Department of Mathematics, University of Padova, via Belzoni, 7 35131 Padova, Italy (dalio@math.unipd.it). This author was partially supported by MIUR project "Viscosity, Metric, and Control Theoretic Methods for Nonlinear Partial Differential Equations."

‡Laboratoire de Mathématiques et Physique Théorique, UMR CNRS 6083, Université de Tours, Faculté des Sciences et Techniques, Parc de Grandmont, 37200 Tours, France (ley@lmpt.univ-tours.fr). This author was partially supported by ACI project "Mouvement d'interfaces avec termes non locaux."

problem. We refer to Bensoussan [10], Fleming and Rishel [18], Fleming and Soner [19], Øksendal [38], Yong and Zhou [41], and the references therein for an overview and to Examples 2.2 and 3.1 below. This problem can be described as follows. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P)$ be a filtered probability space and let $(W_t)_t$ be an $\mathcal{F}_t$-adapted standard M-Brownian motion. The control set is $A = \mathbb{R}^k$ for some $k > 0$ and we consider the linear stochastic differential equation

$$\begin{cases} dX_s = [A(s)X_s + B(s)\alpha_s]ds + [C(s)X_s + D(s)]dW_s \quad \text{for } t \leq s \leq T, \\ X_t = x, \end{cases}$$

where $\alpha_s \in \mathcal{A}_t$, the set of A-valued $\mathcal{F}_t$-progressively measurable controls and the adapted process X_s is the solution. The linear quadratic problem consists in minimizing the quadratic cost

$$V(x,t) = \inf_{\alpha_s \in \mathcal{A}_t} \mathrm{E}\left\{ \int_t^T [\langle X_s, Q(s)X_s\rangle + R|\alpha_s|^2]\, ds + \langle X_T, SX_T\rangle \right\}, \tag{1.4}$$

where $A(\cdot), B(\cdot), C(\cdot), D(\cdot), Q(\cdot)$, and S are deterministic matrix-valued functions of suitable size and $R > 0$ to simplify the presentation. The Hamilton–Jacobi equation associated to this problem is

$$\begin{cases} -\dfrac{\partial w}{\partial t} - \langle A(t)x, Dw\rangle - \langle x, Q(t)x\rangle + \dfrac{1}{4R}|B(t)^T Dw|^2 - \mathrm{Tr}\left[a(x,t)D^2 w\right] = 0, \\ w(x,T) = \langle x, Sx\rangle, \end{cases} \tag{1.5}$$

where $a(x,t) = (C(t)x + D(t))(C(t)x + D(t))^T/2$. Note that this equation is of type (1.1) (with $G \equiv 0$) since

$$\frac{1}{4R}|B(t)^T Dw|^2 = \sup_{\alpha \in \mathbb{R}^k} \Big\{ -\langle B(t)\alpha, Dw\rangle - R|\alpha|^2 \Big\}. \tag{1.6}$$

In this paper, we are concerned with two issues about this problem.

The first question relies on the partial differential equation (1.5). We note that the quadratic cost with unbounded controls leads to a quadratic term with respect to the gradient variable. From the terminal condition, we expect the solutions have quadratic growth. Moreover, the diffusion matrix may be degenerate. Therefore, we cannot hope to obtain smooth solutions in general. We need to consider weak solutions, namely, viscosity solutions. (We refer the reader who is not familiar with this notion of solutions to Crandall, Ishii and Lions [15], Fleming and Soner [19], Bardi and Capuzzo Dolcetta [3], and Barles [6], and all the references therein). We obtain the existence of a unique continuous viscosity solution for (1.5) and for a large class of equations of type (1.1) (see Theorem 2.1 and Corollary 2.1).

We point out that the results obtained in this paper are beyond the classical comparison results for viscosity solutions (see, e.g., [15]) because of the growth of both the solutions and the Hamiltonians. In fact, most of the comparison results in the literature require that either the solutions are uniformly continuous or the Hamiltonian is uniformly continuous with respect to the gradient uniformly in the x variable. (In our case this amounts to assuming that both controls sets are compact.) Let us mention that uniqueness and existence problems for a class of first-order Hamiltonians corresponding to unbounded control sets and under assumptions including deterministic

linear quadratic problems have been addressed by several authors; see, e.g., the book by Bensoussan [10], the papers of Alvarez [2], Bardi and Da Lio [4], Cannarsa and Da Prato [12], and Rampazzo and Sartori [40] in the case of convex operators, and the papers of Da Lio and McEneaney [16] and Ishii [22] for more general operators. As for second-order Hamiltonians under quadratic growth assumptions, Ito [23] obtained the existence of locally Lipschitz solutions to particular equations of the form (1.1) under more regularity conditions on the data, by establishing a priori estimates on the solutions, whereas Crandall and Lions in [14] proved a uniqueness result for very particular operators depending only on the Hessian matrix of the solution. Kobylanski [26] studied equations with the same kind of quadratic nonlinearity in the gradient as ours, but her existence and uniqueness results hold in the class of bounded viscosity solutions. Finally, one can find existence and uniqueness results for viscosity solutions which may have a quadratic growth in [7] for quasilinear degenerate parabolic equations.

The second question we deal with in this paper concerns the link between the control problem (1.4) and equation (1.5). The rigorous connection between the Hamilton–Jacobi–Bellman and optimal control is usually performed by means of a principle of optimality. For deterministic control problems which lead to a first-order Hamilton–Jacobi equation, see Bardi and Capuzzo Dolcetta [3] and Barles [6]; for the connections between stochastic control problems and second-order Hamilton–Jacobi–Bellmann equations see Fleming and Rishel [18], Krylov [27], Lions [31, 32, 33], Fleming and Soner [19], Yong and Zhou [41, Theorem 3.3], and the references therein.

However, for stochastic differential equations with unbounded controls as in stochastic linear quadratic problems, additional difficulties arise. Some results in this direction were obtained for infinite horizon problems, in the deterministic case by Barles [5] and in the stochastic case by Alvarez [1]. In this paper, we characterize the value function (1.4) as the unique solution of (1.5). Actually, our results apply for a larger class of unbounded stochastic control problems

$$V(x,t) = \inf_{\alpha_s \in \mathcal{A}_t} \mathrm{E}\left\{ \int_t^T \ell(X_s, s, \alpha_s)\, ds + \psi(X_T) \right\}, \tag{1.7}$$

where the process X_s is governed by

$$\begin{cases} dX_s = b(X_s, s, \alpha_s)ds + \sigma(X_s, s, \alpha_s)dW_s, \\ X_t = x, \end{cases} \tag{1.8}$$

where A is a possibly unbounded subset of a normed linear space and all the data are continuous with the following restricted growths: b grows at most linearly with respect to both the control and the state, σ grows at most linearly with respect to the state and is bounded in the control variable, ψ can have a quadratic growth, and ℓ grows at most quadratically with respect to both the control and the state with a coercitivity assumption

$$\ell(x, \alpha, t) \geq \frac{\nu}{2}|\alpha|^2 - C(1 + |x|^2). \tag{1.9}$$

In this case, the Hamilton–Jacobi equation looks like (1.1) with $G \equiv 0$. (See section 3 for details.) Because of the unbounded framework, the use of an optimality principle to establish the connection between the control problem and the equation is more delicate than usual. Thus we follow another strategy which consists in comparing

directly the value function with the unique solution of the Hamilton–Jacobi equation as long as this latter exists (see Theorem 3.1).

It is worth noticing that, surprisingly, for the general stochastic linear problem, it is not even clear how to give sense to the partial differential equation associated to the stochastic control problem! For instance, consider again the above linear quadratic problem where $\sigma(x,t,\alpha) = C(t)x + D(t)$ is replaced by $C(t)x + D(t)\alpha$ which depends now both on the state and the control. Taking $A, C, Q \equiv 0$ and $B, R, D \equiv 1$ to simplify, the Hamiltonian in (1.5) becomes

$$\sup_{\alpha\in\mathbb{R}^k}\left\{-\langle\alpha, Dw\rangle - |\alpha|^2 - \frac{|\alpha|^2}{2}\Delta w\right\}, \tag{1.10}$$

which is $+\infty$ as soon as $\Delta w \le -2$. The connection between the control problem and the equation in this case was already investigated (see Yong and Zhou [41] and the references therein). The results need a priori knowledges about the value function. (The value function and its derivatives are supposed to remain in the domain of the Hamiltonian.) We do not consider the case when σ is unbounded in the control variable in this paper; this is the aim of a future work. When finishing this paper, we learned that Krylov [28] succeeded in treating the general stochastic linear regulator. But his assumptions are designed to solve this latter problem (the data are supposed to be polynomials of degree 1 or degree 2 in (x,α)) and the proofs rely heavily on the particular form of the data.

Another important example of equations of type (1.1) where concave and convex Hamiltonians appear is the first-order equation

$$\frac{\partial w}{\partial t} + \min_{\alpha\in\mathbb{R}^k}\left\{\frac{\gamma^2}{2}|\alpha|^2 - \langle\sigma(x)\alpha, Dw\rangle\right\} + \max_{\beta\in B}\{-\langle g(x,\beta), Dw\rangle - f(x,\beta)\} = 0 \tag{1.11}$$

in $\mathbb{R}^N \times (0,T)$. This kind of equation is related to the so-called H_∞-Robust control problem. This problem can be seen as a deterministic differential game. See Example 3.3 and McEneaney [34, 35, 36], Nagai [37], and the references therein for details.

Finally, we point out that one of the main fields of application of these types of equations and problems is mathematical finance; see, e.g., Lamberton and Lapeyre [29], Fleming and Soner [19], Øksendal [38], and the references therein for an introduction. For recent papers which deal with equations we are interested in, see Pham [39] and Benth and Karlsen [11] (see Example 3.2).

Let us now describe how the paper is organized.

Section 2 is devoted to the study of (1.1). More precisely, we prove a uniqueness result for (1.1) in the set of continuous functions growing at most quadratically in the state variable under the assumption that either A or B is an unbounded control set, the functions b, g and ℓ, f grow, respectively, at most linearly and quadratically with respect to both the control and the state. Instead the functions σ, c are assumed to grow at most linearly with respect to the state and bounded in the control variable.

One of the main tools within the theory of viscosity solutions to obtain a uniqueness result is to show a comparison result between viscosity upper semicontinuous subsolutions and lower semicontinuous supersolutions to (1.1); see Theorem 2.1. Indeed, the existence and the uniqueness (Corollary 2.1) follow as a by-product of the comparison result and Perron's method of Ishii [21]. However, under our general assumptions one cannot expect the existence of a solution for all times, as Example 2.2 shows.

The method we use in proving the comparison Theorem 2.1 is similar in spirit to the one applied by Ishii in [22] in the case of first-order Hamilton–Jacobi equations and it is based on a kind of linearization procedure of the equation. Roughly speaking, it consists of three main steps: (1) one computes the equations satisfied by $\Psi_\mu = U - \mu V$ (U, V being, respectively, the sub- and supersolution of the original pde and $0 < \mu < 1$ a parameter); (2) for all $R > 0$ one constructs a strict supersolution χ_R^μ of the linearized equation such that $\chi_R^\mu(x,t) \to 0$ as $R \to \infty$; (3) one shows that $\Psi_\mu(x,t) \le \chi_R^\mu(x,t)$ and then one concludes by letting first $R \to \infty$ and then $\mu \to 1$.

In section 3, we give applications to finite horizon stochastic control problems previously mentioned and we provide some examples.

In section 4, we deal with particular cases where both controls are unbounded but H (or G) is "predominant" in $H + G$ (see Remark 2.2 and Theorem 4.1). For instance, we are able to deal with equations of the form

$$\frac{\partial w}{\partial t} - \frac{|\Sigma_1(x)Dw|^2}{2} + \frac{|\Sigma_2(x)Dw|^2}{2} = 0 \quad \text{in } \mathbb{R}^N \times (0,T),$$

where Σ_1, Σ_2 are $N \times k$ matrices, which corresponds to the case $\alpha, \beta \in \mathbb{R}^k$, $\sigma \equiv c \equiv 0$, $b(x,t,\alpha) = \Sigma_1(x)\alpha$, $g(x,t,\beta) = \Sigma_2(x)\beta$, and $\ell(x,t,\alpha) = |\alpha|^2/2$, $f(x,t,\beta) = |\beta|^2/2$ in (1.2) and (1.3). The comparison result applies if either $(\Sigma_1\Sigma_1^T)(x) > (\Sigma_2\Sigma_2^T)(x)$ or $(\Sigma_1\Sigma_1^T)(x) < (\Sigma_2\Sigma_2^T)(x)$.

When neither H nor G is predominant, the problem seems to be very difficult and our only result takes place in dimension $N = 1$: we have comparison for

$$\frac{\partial w}{\partial t} + h(x,t)|Dw|^2 = 0 \quad \text{in } \mathbb{R}^N \times (0,T), \tag{1.12}$$

where the function h may change sign (see (A5) for details). Finally, we point out that assumptions and proofs in section 4 essentially differ from those of Theorem 2.1 (see Remark 4.1).

2. Comparison result for the Hamilton–Jacobi equation (1.1). In order to give precise assumptions on (1.1) and (1.2), (1.3), we need to introduce some notation. For all integers $N, M \ge 1$ we denote by $\mathcal{M}_{N,M}(\mathbb{R})$ (respectively, $\mathcal{S}_N(\mathbb{R})$, $\mathcal{S}_N^+(\mathbb{R})$) the set of real $N \times M$ matrices (respectively, real symmetric matrices, real symmetric nonnegative $N \times N$ matrices). All the norms which appear in the sequel are denoted by $|\cdot|$. The standard Euclidean inner product in $\mathbb{R}^N$ is written $\langle \cdot, \cdot \rangle$. We recall that a modulus of continuity $m : \mathbb{R} \to \mathbb{R}^+$ is a nondecreasing continuous function such that $m(0) = 0$. Finally, $B(x,r) = \{y \in \mathbb{R}^N : |x - y| < r\}$ is the open ball of center x and radius $r > 0$.

We list the basic assumptions on H, G and ψ. We assume that there exist positive constants $\bar{C}$ and ν such that

(A1) (assumptions on H given by (1.2)):

(i) A is a subset of a separable complete normed space. The main point here is the possible unboundedness of A. Therefore, to emphasize this property in what follows, we take $A = \mathbb{R}^k$ for some $k \ge 1$ (see Remark 2.1 above);

(ii) $b \in C(\mathbb{R}^N \times [0,T] \times \mathbb{R}^k; \mathbb{R}^N)$ satisfying for $x, y \in \mathbb{R}^N, t \in [0,T], \alpha \in \mathbb{R}^k$,

$$|b(x,t,\alpha) - b(y,t,\alpha)| \le \bar{C}(1 + |\alpha|)|x - y|,$$
$$|b(x,t,\alpha)| \le \bar{C}(1 + |x| + |\alpha|);$$

(iii) $\ell \in C(\mathbb{R}^N \times [0,T] \times \mathbb{R}^k; \mathbb{R})$ satisfying for $x \in \mathbb{R}^N, t \in [0,T], \alpha \in \mathbb{R}^k$,

$$\bar{C}(1+|x|^2+|\alpha|^2) \geq \ell(x,t,\alpha) \geq \frac{\nu}{2}|\alpha|^2 + \ell_0(x,t,\alpha) \text{ with } \ell_0(x,t,\alpha) \geq -\bar{C}(1+|x|^2),$$

and for every $R > 0$, there exists a modulus of continuity m_R such that for all $x, y \in B(0,R), t \in [0,T], \alpha \in \mathbb{R}^k$,

$$|\ell(x,t,\alpha) - \ell(y,t,\alpha)| \leq (1+|\alpha|^2)\, m_R(|x-y|); \tag{2.1}$$

(iv) $\sigma \in C(\mathbb{R}^N \times [0,T] \times \mathbb{R}^k; \mathcal{M}_{N,M}(\mathbb{R}))$ is locally Lipschitz continuous with respect to x uniformly for $(t,\alpha) \in [0,T] \times \mathbb{R}^k$ and satisfies for every $x \in \mathbb{R}^N,\ t \in [0,T],\ \alpha \in \mathbb{R}^k$,

$$|\sigma(x,t,\alpha)| \leq \bar{C}(1+|x|).$$

(A2) (assumptions on G given by (1.3)):

(i) B is a bounded subset of a normed space;

(ii) $g \in C(\mathbb{R}^N \times [0,T] \times B; \mathbb{R}^N)$ is locally Lipschitz continuous with respect to x uniformly for $(t,\beta) \in [0,T] \times B$ and satisfies for every $x \in \mathbb{R}^N,\ t \in [0,T],\ \beta \in B$,

$$|g(x,t,\beta)| \leq \bar{C}(1+|x|);$$

(iii) $f \in C(\mathbb{R}^N \times [0,T] \times B; \mathbb{R})$ is locally uniformly continuous with respect to x uniformly in $(t,\beta) \in [0,T] \times B$ and satisfies for every $x \in \mathbb{R}^N,\ t \in [0,T],\ \beta \in B$,

$$|f(x,t,\beta)| \leq \bar{C}(1+|x|^2);$$

(iv) $c \in C(\mathbb{R}^N \times [0,T] \times B; \mathcal{M}_{N,M}(\mathbb{R}))$ is locally Lipschitz continuous with respect to x uniformly for $(t,\beta) \in [0,T] \times B$ and satisfies for every $x \in \mathbb{R}^N,\ t \in [0,T],\ \beta \in B$,

$$|c(x,t,\beta)| \leq \bar{C}(1+|x|).$$

(A3) (assumptions on the initial condition ψ):
$\psi \in C(\mathbb{R}^N; \mathbb{R})$ and

$$|\psi(x)| \leq \bar{C}(1+|x|^2) \quad \text{for every } x \in \mathbb{R}^N.$$

Remark 2.1. (i) Concerning (A1)(i), we choose to take $A = \mathbb{R}^k$ in this section to emphasize the possible unboundedness of A in the notation. Indeed, the calculations when A is any subset of a complete separable normed space are the same and are based on the following inequality: for every $\rho > 0, \gamma \in \mathbb{R}$,

$$\inf_{\alpha \in A} \left\{\rho|\alpha|^2 + \gamma|\alpha|\right\} = \inf_{\alpha \in A} \left\{ \left(\sqrt{\rho}|\alpha|^2 + \frac{\gamma}{2\sqrt{\rho}}\right)^2 - \frac{\gamma^2}{4\rho} \right\} \geq -\frac{\gamma^2}{4\rho}. \tag{2.2}$$

(ii) Note that with respect to the gradient variable, H is a concave function and G is a convex function. Under Assumptions (A1) and (A2), classical computations show that H and G are continuous in all their variables.

Example 2.1. The typical case we have in mind is when H is quadratic in the gradient variable, for instance, $A = \mathbb{R}^k$, $\ell(x,t,\alpha) = |\alpha|^2/2$, $\sigma \equiv 0$, and $b(x,t,\alpha) =$

$a(x)\alpha$, where $a \in C(\mathbb{R}^N; \mathcal{M}_{N,k}(\mathbb{R}))$ is locally Lipschitz continuous and bounded for all $x \in \mathbb{R}^N$. It leads to

$$H(x,p) = \inf_{\alpha\in\mathbb{R}^k} \left\{ \langle a(x)\alpha, p\rangle + \frac{|\alpha|^2}{2} \right\} = -\frac{|a(x)^T p|^2}{2}. \tag{2.3}$$

This particular example is treated both in Ishii [22] and in Da Lio and McEneaney [16] in the case of first-order Hamilton–Jacobi–Bellman equations under more restrictive assumptions than ours. In particular, a has to be a nonsingular matrix in [22]. See also section 4 for some further comments.

For any $O \subseteq \mathbb{R}^K$, we denote by $USC(O)$ the set of upper semicontinuous functions in O and by $LSC(O)$ the set of lower semicontinuous functions in O.

The main result of this section is the following theorem.

THEOREM 2.1. *Assume* (A1)–(A3). *Let* $U \in USC(\mathbb{R}^N \times [0,T])$ *be a viscosity subsolution of* (1.1) *and* $V \in LSC(\mathbb{R}^N \times [0,T])$ *be a viscosity supersolution of* (1.1). *Suppose that* U *and* V *have quadratic growth, i.e., there exists* $\hat{C} > 0$ *such that for all* $x \in \mathbb{R}^N$, $t \in [0,T]$,

$$|U(x,t)|,\ |V(x,t)| \le \hat{C}(1+|x|^2). \tag{2.4}$$

Then $U \le V$ *in* $\mathbb{R}^N \times [0,T]$.

The question of the existence of a continuous solution to (1.1) is not completely obvious. In the framework of viscosity solutions, existence is usually obtained as a consequence of the comparison principle by means of Perron's method as soon as we can build a sub- and a supersolution to the problem. Here, the comparison principle is proved in the class of functions satisfying the quadratic growth condition (2.4). Therefore, to perform the above program of existence, we need to be able to build quadratic sub- and supersolutions to (1.1). In general one can expect to build such sub- and supersolutions only for a short time. (See the following lemma and Corollary 2.1.) In Example 2.2, we see that solutions may not exist for all time.

LEMMA 2.1. *Assume* (A1)–(A3). *If* $K \ge \bar{C} + 1$ *and* ρ *are large enough, then* $\underline{u}(x,t) = -K\mathrm{e}^{\rho t}(1+|x|^2)$ *is a viscosity subsolution of* (1.1) *in* $\mathbb{R}^N \times [0,T]$ *and there exists* $0 < \tau \le T$ *such that* $\overline{u}(x,t) = K\mathrm{e}^{\rho t}(1+|x|^2)$ *is a viscosity supersolution of* (1.1) *in* $\mathbb{R}^N \times [0,\tau]$.

Proof of Lemma 2.1. We only verify that $\overline{u}$ is a supersolution (the proof that $\underline{u}$ is a subsolution being similar and simpler). Since $K \ge \bar{C} + 1$, we have $\overline{u}(x,0) = K(1+|x|^2) \ge \psi(x)$. Moreover, since $\overline{u}$ is smooth and using (A1), (A2), and (2.2), we have

$$\begin{aligned}
&\frac{\partial \overline{u}}{\partial t} + H(x,t,D\overline{u},D^2\overline{u}) + G(x,t,D\overline{u},D^2\overline{u}) \\
&= K\rho\mathrm{e}^{\rho t}(1+|x|^2) + H(x,t,2K\mathrm{e}^{\rho t}x, 2K\mathrm{e}^{\rho t}Id) + G(x,t,2K\mathrm{e}^{\rho t}x, 2K\mathrm{e}^{\rho t}Id) \\
&\ge K\rho\mathrm{e}^{\rho t}(1+|x|^2) + \inf_{\alpha\in\mathbb{R}^k} \Big\{ -\bar{C}(1+|x|+|\alpha|)\}2K\mathrm{e}^{\rho t}|x| + \frac{\nu}{2}|\alpha|^2 \\
&\qquad\qquad - \bar{C}(1+|x|^2) - \bar{C}^2(1+|x|)^2 2K\mathrm{e}^{\rho t} \Big\} \\
&\qquad + \sup_{\beta\in B} \{ -\bar{C}(1+|x|)2K\mathrm{e}^{\rho t}|x| - \bar{C}(1+|x|^2) - \bar{C}^2(1+|x|)^2 2K\mathrm{e}^{\rho t} \} \\
&\ge K\rho\mathrm{e}^{\rho t}(1+|x|^2) - K(10\bar{C} + 12\bar{C}^2)\mathrm{e}^{\rho t}(1+|x|^2) + \inf_{\alpha\in\mathbb{R}^k} \Big\{ -2K\bar{C}\mathrm{e}^{\rho t}|x||\alpha| + \frac{\nu}{2}|\alpha|^2 \Big\} \\
&\ge \left[\rho - 10\bar{C} - 12\bar{C}^2 - \frac{2\bar{C}^2 K\mathrm{e}^{\rho t}}{\nu} \right] K\mathrm{e}^{\rho t}(1+|x|^2).
\end{aligned}$$

We notice that if $t\rho \leq 1$ and $\rho > 0$ is large enough, then the quantity between the brackets is nonnegative. Hence the result follows with $0 < \tau = 1/\rho$. □

As explained above, Theorem 2.1 together with Perron's method implies the following result. We omit its proof since it is standard.

COROLLARY 2.1. *Assume* (A1)–(A3). *Then there is* $\tau > 0$ *such that there exists a unique continuous viscosity solution of* (1.1) *in* $\mathbb{R}^N \times [0,\tau]$ *satisfying the growth condition* (2.4).

Remark 2.2. (i) For global existence results under further regularity assumptions on the data see [23]. For a case where blowup in finite time occurs, see Example 2.2.

(ii) Theorem 2.1 and Corollary 2.1 hold when replacing "inf" by "sup" and "ℓ" by "$-\ell$" in H or/and "sup" by "inf" and "f" by "$-f$" in G. To adapt the proofs, one can use the change of function $w' := -w$ or to consider $\mu U - V$ instead of $U - \mu V$ in the proof of Theorem 2.1. Therefore it is possible to deal either with unbounded controls in the sup in order to have a convex quadratic Hamiltonian or to deal with unbounded control in the inf in order to have a concave quadratic Hamiltonian.

(iii) Up to replace t by $T - t$, all our results hold for

$$\text{(2.5)} \qquad \begin{cases} -\dfrac{\partial w}{\partial t} + H(x,t,Dw,D^2w) + G(x,t,Dw,D^2w) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ w(x,T) = \psi(x) & \text{in } \mathbb{R}^N. \end{cases}$$

This latter equation with terminal condition is the one which arises usually in control theory; see Example 2.2 and section 3.

(iv) In this section, we are not able to consider the case when both α in H and β in G are unbounded controls. Roughly speaking, a reason is that unbounded controls lead to quadratic Hamiltonians. When both controls are unbounded, we then obtain two quadratic-type Hamiltonians, a concave and a convex one. Let us explain the difficulty on a model case where

$$H(x,p) = \inf_{\alpha\in\mathbb{R}^k} \left\{ \langle a_1(x)\alpha, p\rangle + \frac{|\alpha|^2}{2} \right\} = -\frac{|a_1(x)^T p|^2}{2},$$
$$G(x,p) = \sup_{\beta\in\mathbb{R}^k} \left\{ \langle a_2(x)\beta, p\rangle - \frac{|\beta|^2}{2} \right\} = \frac{|a_2(x)^T p|^2}{2},$$

where $a_1, a_2 \in C(\mathbb{R}^N; \mathcal{M}_{N,k}(\mathbb{R}))$ are locally Lipschitz continuous and bounded. The difficulty to treat such a case is related to our strategy of proof which relies on a kind of linearization procedure (see Lemma 2.2 and its proof). In this simple case, this linearization uses in a crucial way the convex inequality

$$\frac{|p|^2}{\mu} - |q|^2 \geq -\frac{|p-q|^2}{1-\mu} \quad \text{for all } p,q \in \mathbb{R}^N \text{ and } 0 < \mu < 1,$$

which does not work at the same time for a concave and a convex Hamiltonian. Of course, in this simple case, there are alternative ways to solve the problem: we have

$$\text{(2.6)} \qquad H(x,p) + G(x,p) = \frac{1}{2}\langle (a_1+a_2)(a_2-a_1)^T p, p\rangle,$$

then we can apply Theorem 2.1 to H given by 2.6 to add some assumptions on a_1 or a_2 (for instance, $(a_1+a_2)(a_2-a_1)^T$ is a nonnegative symmetric matrix with a locally Lipschitz squareroot). In section 4, we provide another approach to solving such equations (see Theorem 4.1 and Remark 4.1).

Example 2.2 (a deterministic linear quadratic control problem). Linear quadratic control problems (see also section 3 for stochastic linear quadratic control problems) are the typical examples we have in mind since they lead to Hamilton–Jacobi equations with quadratic terms. On the other hand, the value function can blow up in finite time. Consider the control problem (in dimension 1 for sake of simplicity)

$$\begin{cases} dX_s = \alpha_s\, ds, \quad s \in [t,T],\ 0 \le t \le T, \\ X_t = x \in \mathbb{R}, \end{cases}$$

where the control $\alpha \in \mathcal{A}_t := L^2([t,T];\mathbb{R})$ and the value function is given by

$$V(x,t) = \inf_{\alpha \in \mathcal{A}_t} \left\{ \rho \int_t^T (|\alpha_s|^2 + |X_s|^2)\, ds - |X_T|^2 \right\} \quad \text{for some } \rho > 0.$$

Then the value function, when it is finite, is the unique viscosity solution of the Hamilton–Jacobi equation of type (1.1) which reads

$$\begin{cases} -w_t + \frac{1}{4\rho}|w_x|^2 = \rho x^2 & \text{in } \mathbb{R} \times (0,T), \\ w(x,T) = -x^2. \end{cases} \tag{2.7}$$

(See Theorem 3.1 for a proof of this result.) Looking for a solution w under the form $w(x,t) = \varphi(t)x^2$, we obtain that φ is a solution of the differential equation

$$-\varphi' + \frac{\varphi^2}{\rho} = \rho \ \text{ in } (0,T), \qquad \varphi(T) = -1. \tag{2.8}$$

We distinguish two cases.

Case 1. If $\rho \ge 1$, then the solution of (2.8) is defined in the whole interval $[0,T]$ for all $T > 0$ and is given by

$$\varphi(t) = \rho\, \frac{(\rho-1)\mathrm{e}^{2(T-t)} - (\rho+1)}{(\rho-1)\mathrm{e}^{2(T-t)} + \rho + 1}, \tag{2.9}$$

which is a function decreasing from $\varphi(0)$ to -1. Therefore (2.7) admits a unique viscosity solution in $\mathbb{R} \times [0,T]$ which is the value function of the control problem

$$V(x,t) = \varphi(t)x^2.$$

Note that if $\rho > 1$ and $T > \ln((\rho+1)/(\rho-1))/2$, then $\varphi(0) > 0$. It follows that the value function satisfies (2.4) but is neither bounded from above nor bounded from below.

Case 2. If $0 < \rho < 1$ and $T > \ln((1+\rho)/(1-\rho))/2$, then the solution of (2.8) is given by (2.9) in $(\bar\tau, T]$, where

$$\bar\tau := T - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right),$$

and blows up at $t = \bar\tau$. Therefore we have existence for (2.7) only in $\mathbb{R} \times (\bar\tau, T]$.

Proof of Theorem 2.1. We divide the proof of the theorem into two steps.

Step 1. We first assume that $\ell_0 \ge 0$, $f \le 0$, and $\psi \le 0$ in (A1), (A2), and (A3).

The proof is based on the two following lemmas, whose proofs are postponed.

LEMMA 2.2 (linearization of (1.1)). *Let $0 < \mu < 1$ and set $\Psi = U - \mu V$. Suppose $\ell_0 \geq 0$, $f \leq 0$ and $\psi \leq 0$. Then Ψ is a USC viscosity subsolution of*

$$\mathcal{L}[w] := \frac{\partial w}{\partial t} - \frac{\bar{C}^2}{2\nu(1-\mu)}|Dw|^2 - 2\bar{C}(1+|x|)|Dw|$$

$$- \sup_{\alpha\in\mathbb{R}^k} \mathrm{Tr}\left[\sigma(x,t,\alpha)\sigma^T(x,t,\alpha)D^2w\right] - \sup_{\beta\in B} \mathrm{Tr}\left[c(x,t,\beta)c^T(x,t,\beta)D^2w\right] = 0 \tag{2.10}$$

in $\mathbb{R}^N \times (0,T)$, with the initial condition

$$w(\cdot,0) \leq (1-\mu)\psi \leq 0. \tag{2.11}$$

LEMMA 2.3. *Consider the parabolic problem*

$$\begin{cases} \varphi_t - r^2\varphi_{rr} - r\varphi_r = 0 & \text{in } [0,+\infty)\times(0,T], \\ \varphi(r,0) = \varphi_R(r) & \text{in } [0,+\infty), \end{cases} \tag{2.12}$$

where $\varphi_R(r) = \max\{0, r-R\}$ for some $R > 0$. Then (2.12) has a unique solution $\varphi \in C([0,+\infty)\times[0,T]) \cap C^\infty([0,+\infty)\times(0,T])$ such that for all $t \in (0,T]$, $\varphi(\cdot,t)$ is positive, nondecreasing, and convex in $[0,+\infty)$. Moreover, for every $(r,t) \in [0,+\infty)\times(0,T]$,

$$\varphi(r,t) \geq \varphi_R(r), \quad 0 \leq \varphi_r(r,t) \leq \mathrm{e}^T \quad \text{and} \quad \varphi(r,t) \underset{R\to+\infty}{\longrightarrow} 0. \tag{2.13}$$

Let $\Phi(x,t) = \varphi(C(1+|x|^2)\mathrm{e}^{Lt}, Mt) + \eta t$, where φ is given by Lemma 2.3, L, M, η are positive constants to be determined, and $C > \max\{\bar{C}, \hat{C}\}$, where $\bar{C}$ and $\hat{C}$ are the constants appearing in the assumptions.

Claim. We can choose the constants L and M such that Φ is a strict supersolution of (2.10) at least in $\mathbb{R}^N \times (0,\tau]$ for small τ.

To prove the claim, we have to show that $\mathcal{L}[\Phi] > 0$ in $\mathbb{R}^N \times (0,\tau]$ for some $\tau > 0$. The function $\Phi \in C(\mathbb{R}^N \times [0,T]) \cap C^\infty(\mathbb{R}^N \times (0,T])$ and we have, for $t > 0$,

$$\Phi_t = \eta + LC(1+|x|^2)\mathrm{e}^{Lt}\varphi_r + M\varphi_t, \quad D\Phi = 2Cx\mathrm{e}^{Lt}\varphi_r,$$
$$\text{and } D^2\Phi = 2C\,Id\,\mathrm{e}^{Lt}\varphi_r + 4C^2\mathrm{e}^{2Lt}\varphi_{rr}\, x\otimes x.$$

Using (A1) and (A2), for all $(x,t) \in \mathbb{R}^N \times (0,T]$, we get

$$\mathcal{L}[\Phi] \geq \eta + LC(1+|x|^2)\mathrm{e}^{Lt}\varphi_r + M\varphi_t - \frac{\bar{C}^2}{2\nu(1-\mu)}|2Cx\mathrm{e}^{Lt}\varphi_r|^2$$
$$- 2\bar{C}(1+|x|)|2Cx\mathrm{e}^{Lt}\varphi_r| - 2\bar{C}^2(1+|x|)^2|2C\,Id\,\mathrm{e}^{Lt}\varphi_r + 4C^2\mathrm{e}^{2Lt}x\otimes x\,\varphi_{rr}|.$$

Setting $r = C(1+|x|^2)\mathrm{e}^{Lt}$ and since $C > \bar{C}$, we obtain

$$\mathcal{L}[\Phi] \geq \eta + M\varphi_t - 16C^2r^2\varphi_{rr} - 8C(C+1)r\varphi_r + \left(L - \frac{2C^3\mathrm{e}^{Lt}\varphi_r}{\nu(1-\mu)}\right) r\varphi_r.$$

Our aim is to fix the parameters M and L in order to make $\mathcal{L}[\Phi]$ positive.

We first choose $M > 16C^2 + 8C$. Since φ is a solution of (2.12) (Lemma 2.3), we obtain

$$\mathcal{L}[\Phi] > \eta + \left(L - \frac{2C^3\mathrm{e}^{Lt}\varphi_r}{\nu(1-\mu)}\right) r\varphi_r. \tag{2.14}$$

Then, taking $L > \frac{2C^3 \mathrm{e}^{T+1}}{\nu(1-\mu)}$, we get $\mathcal{L}[\Phi] > \eta > 0$ for all $x \in \mathbb{R}^N$ and $t \in (0, \tau]$, where $\tau = 1/L$. This proves the claim.

We continue by considering

$$\max_{\mathbb{R}^N \times [0,\tau]} \{\Psi - \Phi\}, \tag{2.15}$$

where Ψ is the function defined in Lemma 2.2 which is a viscosity subsolution of (2.10) and Φ is the strict supersolution of (2.10) in $\mathbb{R}^N \times (0, \tau]$ we built above.

From (2.13), we have $\Phi(x,t) \geq C(1+|x|^2) > \bar{C}(1+|x|^2) \geq \Psi(x,t)$ for $|x| \geq R$. It follows that the maximum (2.15) is achieved at a point $(\bar{x}, \bar{t}) \in \mathbb{R}^N \times [0, \tau]$. We claim that $\bar{t} = 0$. Indeed, suppose by contradiction that $\bar{t} > 0$. Then since Ψ is a viscosity subsolution of (2.10), by taking Φ as a test function, we would have $\mathcal{L}[\Phi](\bar{x}, \bar{t}) \leq 0$, which contradicts the fact that Φ is a strict supersolution.

Thus, for all $(x,t) \in \mathbb{R}^N \times [0, \tau]$,

$$\Psi(x,t) - \Phi(x,t) \leq \Psi(\bar{x}, 0) - \Phi(\bar{x}, 0) \leq (1-\mu)\psi(\bar{x}),$$

where the last inequality follows from (2.11) and the fact that $\Phi \geq 0$. Since we assumed that ψ is nonpositive, for every $(x,t) \in \mathbb{R}^N \times [0, \tau]$, we have $\Psi(x,t) \leq \Phi(x,t)$. Letting η go to 0 and R to $+\infty$, we get by Lemma 2.3, $\Psi \leq 0$ in $\mathbb{R}^N \times [0, \tau]$.

By a step-by-step argument, we prove that $\Psi \leq 0$ in $\mathbb{R}^N \times [0, T]$. Therefore $\Psi = U - \mu V \leq 0$ in $\mathbb{R}^N \times [0, T]$. Letting μ go to 1, we obtain $U \leq V$ as well, which concludes Step 1.

Step 2. *The general case.* The idea is to reduce to the first case by a suitable change of function (see Ishii [22]). Suppose that w is a solution of (1.1). Then, a straightforward computation shows that $\bar{w}(x,t) = w(x,t) - C(1+|x|^2)\mathrm{e}^{\rho t}$ for $C > \bar{C}, \hat{C}$ and $\rho > 0$ is a solution of

$$\begin{cases} \bar{w}_t + \inf\limits_{\alpha \in \mathbb{R}^k} \left\{ \langle b(x,t,\alpha), D\bar{w} \rangle + \bar{\ell}(x,t,\alpha) - \operatorname{Tr}\left[\sigma(x,t,\alpha)\sigma^T(x,t,\alpha) D^2 \bar{w}\right] \right\} \\ \qquad + \sup\limits_{\beta \in B} \left\{ -\langle g(x,t,\beta), D\bar{w} \rangle - \bar{f}(x,t,\beta) - \operatorname{Tr}\left[c(x,t,\beta)c^T(x,t,\beta) D^2 \bar{w}\right] \right\} = 0, \\ \bar{w}(x,0) = \psi(x) - C(1+|x|^2), \end{cases} \tag{2.16}$$

where

$$\begin{aligned} \bar{\ell}(x,t,\alpha) &= \ell(x,t,\alpha) + 2C\mathrm{e}^{\rho t}\langle b(x,t,\alpha), x \rangle - 2C\mathrm{e}^{\rho t} \operatorname{Tr}\left[\sigma(x,t,\alpha)\sigma^T(x,t,\alpha)\right] \\ &\qquad + \frac{1}{2} C\rho \mathrm{e}^{\rho t}(1+|x|^2), \\ \bar{f}(x,t,\beta) &= f(x,t,\beta) + 2C\mathrm{e}^{\rho t}\langle g(x,t,\beta), x \rangle + 2C\mathrm{e}^{\rho t} \operatorname{Tr}\left[c(x,t,\beta)c^T(x,t,\beta)\right] \\ &\qquad - \frac{1}{2} C\rho \mathrm{e}^{\rho t}(1+|x|^2). \end{aligned}$$

We observe that $\bar{\ell}$ and $\bar{f}$ still satisfy, respectively, assumptions (A1)(iii) and (A2)(iii). Moreover from (A3), we can choose $C > \bar{C}$ in order that $\bar{\psi} \leq 0$.

Next we show that if $\rho > 0$ is chosen in a suitable way, then $\bar{\ell}(x,t,\alpha) \geq \bar{\nu}|\alpha|^2/2$ for all $(x,t,\alpha) \in \mathbb{R}^N \times [0, 1/\rho] \times \mathbb{R}^k$ and $f(x,t,\beta) \leq 0$ for all $(x,t,\beta) \in \mathbb{R}^N \times [0, 1/\rho] \times B$.

Indeed, for all $(x,t,\alpha) \in \mathbb{R}^N \times [0,1/\rho] \times \mathbb{R}^k$ by (A1) we have

$$\begin{aligned}\bar{\ell}(x,t,\alpha) &\geq \frac{\nu}{2}|\alpha|^2 - \bar{C}(1+|x|^2) - 2C\bar{C}\mathrm{e}^{\rho t}(1+|x|+|\alpha|)|x| - 2C\bar{C}^2\mathrm{e}^{\rho t}(1+|x|)^2 \\ &\quad + \frac{1}{2}C\rho\mathrm{e}^{\rho t}(1+|x|^2) \\ &\geq \frac{\nu}{2}|\alpha|^2 - (\bar{C}+4C\bar{C}+4C\bar{C}^2)\mathrm{e}^{\rho t}(1+|x|^2) - 2C\bar{C}\mathrm{e}^{\rho t}|\alpha||x| \\ &\quad + \frac{1}{2}C\rho\mathrm{e}^{\rho t}(1+|x|^2).\end{aligned}$$

But

$$2C\bar{C}\mathrm{e}^{\rho t}|\alpha||x| \leq \frac{\nu}{4}|\alpha|^2 + \frac{16C^2\bar{C}^2\mathrm{e}^{2\rho t}}{\nu}|x|^2\,.$$

Therefore, by choosing

$$\rho > 2\frac{\bar{C}}{C} + 8\bar{C} + 8\bar{C}^2 + \frac{32C\bar{C}^2\mathrm{e}}{\nu}, \tag{2.17}$$

we have

$$\bar{\ell}(x,t,\alpha) \geq \frac{\nu}{4}|\alpha|^2 \quad \text{for all } (x,t,\alpha) \in \mathbb{R}^N \times [0,1/\rho] \times \mathbb{R}^k,$$

which is the desired estimate with $\bar{\nu} = \nu/2 > 0$ and $\ell_0 \equiv 0$.

The next step consists in choosing ρ such that $\bar{f} \leq 0$. Using (A2), the same kind of calculation as above shows that taking

$$\rho > 8(\bar{C}+\bar{C}^2) + 2 \tag{2.18}$$

ensures $\bar{f} \leq 0$.

Finally, if we choose $C > \bar{C}$ and ρ as the maximum of the two quantities appearing in (2.17) and (2.18), we are in the framework of Step 1 in $\mathbb{R}^N \times [0,\rho]$. Setting $\bar{U} = U - C(1+|x|^2)\mathrm{e}^{\rho t}$ and $\bar{V} = V - C(1+|x|^2)\mathrm{e}^{\rho t}$, from Step 1 we get $\bar{U} \leq \bar{V}$ in $\mathbb{R}^N \times [0,1/\rho]$; thus $U \leq V$ in $\mathbb{R}^N \times [0,1/\rho]$. Then by a step-by-step argument we obtain the comparison in $\mathbb{R}^N \times [0,T]$. □

Remark 2.3. A key fact in the proof to build a strict supersolution of (2.10) is to use a function which is the solution of the auxiliary—and simpler—pde (2.12). This idea comes from mathematical finance to deal with equations related to the Blake and Scholes formula. See, for instance, Lamberton and Lapeyre [29] and Barles et al. [8].

We turn to the proof of Lemmas 2.2 and 2.3.

Proof of Lemma 2.2. For $0 < \mu < 1$, let $\tilde{V} = \mu V$ and $\Psi = U - \tilde{V}$. We divide the proof into steps.

Step 1. *A new equation for* $\tilde{V}$. It is not difficult to see that if V is a supersolution of (1.1), then $\tilde{V}$ is a supersolution of

$$\begin{aligned}&\tilde{V}_t + \mu\, H\left(x,t,\frac{D\tilde{V}}{\mu},\frac{D^2\tilde{V}}{\mu}\right) + \mu\, G\left(x,t,\frac{D\tilde{V}}{\mu},\frac{D^2\tilde{V}}{\mu}\right) \geq 0 \quad \text{in } \mathbb{R}^N \times (0,T), \\ &\tilde{V}(x,0) \geq \mu\,\psi(x) \qquad\qquad \text{in } \mathbb{R}^N.\end{aligned} \tag{2.19}$$

Step 2. *Viscosity inequalities for* U *and* $\tilde{V}$. This step is classical in viscosity theory. Let $\varphi \in C^2(\mathbb{R}^N \times (0,T])$ and $(\bar{x},\bar{t}) \in \mathbb{R}^N \times (0,T]$ be a local maximum of $\Psi - \varphi$. We

can assume that this maximum is strict in the same ball $\overline{B}(\bar{x},r)\times[\bar{t}-r,\bar{t}+r]$ (see [6] or [3]). Let

$$\Theta(x,y,t)=\varphi(x,t)+\frac{|x-y|^2}{\varepsilon^2}$$

and consider

$$M_\varepsilon:=\max_{x,y\in\overline{B}(\bar{x},r),\,t\in[\bar{t}-r,\bar{t}+r]}\{U(x,t)-\tilde{V}(y,t)-\Theta(x,y,t)\}.$$

This maximum is achieved at a point $(x_\varepsilon,y_\varepsilon,t_\varepsilon)$ and, since the maximum is strict, we know [6], [3] that

$$\frac{|x_\varepsilon-y_\varepsilon|^2}{\varepsilon^2}\to 0\quad\text{as }\varepsilon\to 0$$

and

$$M_\varepsilon=U(x_\varepsilon,t_\varepsilon)-\tilde{V}(y_\varepsilon,t_\varepsilon)-\Theta(x_\varepsilon,y_\varepsilon,t_\varepsilon)\longrightarrow U(\bar{x},\bar{t})-\tilde{V}(\bar{x},\bar{t})-\varphi(\bar{x},\bar{t})=\Psi(\bar{x},\bar{t})-\varphi(\bar{x},\bar{t}).$$

This means that at the limit $\varepsilon\to 0$, we obtain some information on $\Psi-\varphi$ at $(\bar{x},\bar{t})$ which will provide the new equation for Ψ. Before that, we can take Θ as a test function to use the fact that U is a subsolution and $\tilde{V}$ a supersolution. Indeed, $(x,t)\in\overline{B}(\bar{x},r)\times[\bar{t}-r,\bar{t}+r]\mapsto U(x,t)-\tilde{V}(y_\varepsilon,t)-\Theta(x,y_\varepsilon,t)$ achieves its maximum at $(x_\varepsilon,t_\varepsilon)$ and $(y,t)\in\overline{B}(\bar{x},r)\times[\bar{t}-r,\bar{t}+r]\mapsto -U(x_\varepsilon,t)+\tilde{V}(y,t)+\Theta(x_\varepsilon,y,t)$ achieves its minimum at $(y_\varepsilon,t_\varepsilon)$. Thus, by Theorem 8.3 in the user's guide [15], for every $\rho>0$, there exist $a_1,a_2\in\mathbb{R}$ and $X,Y\in\mathcal{S}_N$ such that

$$(a_1,D_x\Theta(x_\varepsilon,y_\varepsilon,t_\varepsilon),X)\in\bar{\mathcal{P}}^{2,+}(U)(x_\varepsilon,t_\varepsilon),(a_2,-D_y\Theta(x_\varepsilon,y_\varepsilon,t_\varepsilon),Y)\in\bar{\mathcal{P}}^{2,-}(\tilde{V})(y_\varepsilon,t_\varepsilon),$$

$a_1-a_2=\Theta_t(x_\varepsilon,y_\varepsilon,t_\varepsilon)=\varphi_t(x_\varepsilon,t_\varepsilon)$ and

$$-\left(\frac{1}{\rho}+|M|\right)I\le\begin{pmatrix}X&0\\0&-Y\end{pmatrix}\le M+\rho M^2,\quad\text{where } M=D^2\Theta(x_\varepsilon,y_\varepsilon,t_\varepsilon).\tag{2.20}$$

Setting $p_\varepsilon=2\dfrac{x_\varepsilon-y_\varepsilon}{\varepsilon^2}$, we have

$$D_x\Theta(x_\varepsilon,y_\varepsilon,t_\varepsilon)=p_\varepsilon+D\varphi(x_\varepsilon,t_\varepsilon)\quad\text{and}\quad D_y\Theta(x_\varepsilon,y_\varepsilon,t_\varepsilon)=-p_\varepsilon,$$

and

$$M=\begin{pmatrix}D^2\varphi(x_\varepsilon,t_\varepsilon)+2I/\varepsilon^2&-2I/\varepsilon^2\\-2I/\varepsilon^2&2I/\varepsilon^2\end{pmatrix}.$$

Thus, from (2.20), it follows

$$\langle Xp,p\rangle-\langle Yq,q\rangle\le\langle D^2\varphi(x_\varepsilon,t_\varepsilon)p,p\rangle+\frac{2}{\varepsilon^2}|p-q|^2+m\left(\frac{\rho}{\varepsilon^4}\right),\tag{2.21}$$

where m is a modulus of continuity which is independent of ρ and ε. In what follows, m will always denote a generic modulus of continuity independent of ρ and ε.

Writing the subsolution viscosity inequality for U and the supersolution inequality for $\tilde{V}$ by means of the semijets and subtracting the inequalities, we obtain

$$\begin{aligned}\varphi_t(x_\varepsilon,t_\varepsilon)+H\big(x_\varepsilon,t_\varepsilon,D\varphi(x_\varepsilon,t_\varepsilon)+p_\varepsilon,X\big)-\mu\,H\left(y_\varepsilon,t_\varepsilon,\frac{p_\varepsilon}{\mu},\frac{Y}{\mu}\right)\\ +G\big(x_\varepsilon,t_\varepsilon,D\varphi(x_\varepsilon,t_\varepsilon)+p_\varepsilon,X\big)-\mu\,G\left(y_\varepsilon,t_\varepsilon,\frac{p_\varepsilon}{\mu},\frac{Y}{\mu}\right)\le 0.\end{aligned}\tag{2.22}$$

Step 3. *Estimate of* $\mathcal{G}:=G\big(x_\varepsilon,t_\varepsilon,D\varphi(x_\varepsilon,t_\varepsilon)+p_\varepsilon,X\big)-\mu\,G\big(y_\varepsilon,t_\varepsilon,\frac{p_\varepsilon}{\mu},\frac{Y}{\mu}\big)$. For simplicity, we set

$$c(x_\varepsilon,t_\varepsilon,\beta)=c_x\quad\text{and}\quad c(y_\varepsilon,t_\varepsilon,\beta)=c_y.$$

We have

$$\begin{aligned}\mathcal{G}&=\sup_{\beta\in B}\left\{-\langle g(x_\varepsilon,t_\varepsilon,\beta),D\varphi(x_\varepsilon,t_\varepsilon)+p_\varepsilon\rangle-f(x_\varepsilon,t_\varepsilon,\beta)-\operatorname{Tr}\left[c_xc_x^TX\right]\right\}\\ &\quad-\sup_{\beta\in B}\left\{-\langle g(y_\varepsilon,t_\varepsilon,\beta),p_\varepsilon\rangle-\mu f(y_\varepsilon,t_\varepsilon,\beta)-\operatorname{Tr}\left[c_yc_y^TY\right]\right\}\\ &\ge\inf_{\beta\in B}\left\{\langle g(y_\varepsilon,t_\varepsilon,\beta)-g(x_\varepsilon,t_\varepsilon,\beta),p_\varepsilon\rangle-\langle g(x_\varepsilon,t_\varepsilon,\beta),D\varphi(x_\varepsilon,t_\varepsilon)\rangle\right.\\ &\qquad\left.-(1-\mu)f(y_\varepsilon,t_\varepsilon,\beta)+f(y_\varepsilon,t_\varepsilon,\beta)-f(x_\varepsilon,t_\varepsilon,\beta)-\operatorname{Tr}\left[c_xc_x^TX-c_yc_y^TY\right]\right\}.\end{aligned}$$

From (A2), if $L_{g,r}$ is the Lipschitz constant of g in $\overline{B}(\bar{x},r)\times[\bar{t}-r,\bar{t}+r]$, then we have

$$\langle g(y_\varepsilon,t_\varepsilon,\beta)-g(x_\varepsilon,t_\varepsilon,\beta),p_\varepsilon\rangle\le L_{g,r}|y_\varepsilon-x_\varepsilon||p_\varepsilon|\le 2L_{g,r}\frac{|y_\varepsilon-x_\varepsilon|^2}{\varepsilon^2}=m(\varepsilon)$$

and

$$-\langle g(x_\varepsilon,t_\varepsilon,\beta),D\varphi(x_\varepsilon,t_\varepsilon)\rangle\ge-\bar{C}(1+|x_\varepsilon|)|D\varphi(x_\varepsilon,t_\varepsilon)|.$$

By assumption, $f\le 0$ thus $-(1-\mu)f(y_\varepsilon,t_\varepsilon,\beta)\ge 0$. Again from (A2) it follows that

$$f(y_\varepsilon,t_\varepsilon,\beta)-f(x_\varepsilon,t_\varepsilon,\beta)\ge-m(|y_\varepsilon-x_\varepsilon|).$$

Let us denote by $(e_i)_{1\le i\le N}$ the canonical basis of $\mathbb{R}^N$. By using (2.21), we obtain

$$\begin{aligned}\operatorname{Tr}\left[c_xc_x^TX-c_yc_y^TY\right]&=\sum_{i=1}^N\langle Xc_xe_i,c_xe_i\rangle-\langle Yc_ye_i,c_ye_i\rangle\\ &\le\operatorname{Tr}\left[c_xc_x^TD^2\varphi(x_\varepsilon,t_\varepsilon)\right]+\frac{2}{\varepsilon^2}|c_x-c_y|^2+m\left(\frac{\rho}{\varepsilon^4}\right)\\ &\le\operatorname{Tr}\left[c_xc_x^TD^2\varphi(x_\varepsilon,t_\varepsilon)\right]+2L_{c,r}^2\frac{|x_\varepsilon-y_\varepsilon|^2}{\varepsilon^2}+m\left(\frac{\rho}{\varepsilon^4}\right)\\ &\le\operatorname{Tr}\left[c_xc_x^TD^2\varphi(x_\varepsilon,t_\varepsilon)\right]+m(\varepsilon)+m\left(\frac{\rho}{\varepsilon^4}\right),\end{aligned}$$

where $L_{c,r}$ is a Lipschitz constant for c in $\bar{B}(x,r)$. Hence, since all the moduli are independent of ε, ρ and the control, we have

$$\begin{aligned}\mathcal{G}\ge-\bar{C}(1+|x_\varepsilon|)|D\varphi(x_\varepsilon,t_\varepsilon)|+\inf_{\beta\in B}\left\{-\operatorname{Tr}\left[c(x_\varepsilon,t_\varepsilon,\beta)c(x_\varepsilon,t_\varepsilon,\beta)^TD^2\varphi(x_\varepsilon,t_\varepsilon)\right]\right\}\\ +m(\varepsilon)+m\left(\frac{\rho}{\varepsilon^4}\right).\end{aligned}$$

Step 4. *Estimate of* $\mathcal{H} := H\big(x_\varepsilon, t_\varepsilon, D\varphi(x_\varepsilon,t_\varepsilon)+p_\varepsilon, X\big) - \mu\, H\big(y_\varepsilon, t_\varepsilon, \frac{p_\varepsilon}{\mu}, \frac{Y}{\mu}\big)$. With the same notation as in Step 3, we have

$$\begin{aligned}\mathcal{H} \geq \inf_{\alpha\in\mathbb{R}^k}\{&\langle b(x_\varepsilon,t_\varepsilon,\alpha)-b(y_\varepsilon,t_\varepsilon,\alpha),p_\varepsilon\rangle + \langle b(x_\varepsilon,t_\varepsilon,\alpha), D\varphi(x_\varepsilon,t_\varepsilon)\rangle\\ &+(1-\mu)\ell(y_\varepsilon,t_\varepsilon,\alpha)+\ell(x_\varepsilon,t_\varepsilon,\alpha)-\ell(y_\varepsilon,t_\varepsilon,\alpha)-\operatorname{Tr}\left[\sigma_x\sigma_x^T X-\sigma_y\sigma_y^T Y\right]\}.\end{aligned}$$

From (A1) these estimates follow:

$$\langle b(x_\varepsilon,t_\varepsilon,\alpha)-b(y_\varepsilon,t_\varepsilon,\alpha),p_\varepsilon\rangle \geq -\bar{C}(1+|\alpha|)|x_\varepsilon-y_\varepsilon||p_\varepsilon| \geq -\bar{C}|\alpha|m(\varepsilon)+m(\varepsilon),$$

$$\langle b(x_\varepsilon,t_\varepsilon,\alpha), D\varphi(x_\varepsilon,t_\varepsilon)\rangle \geq -\bar{C}(1+|x_\varepsilon|)|D\varphi(x_\varepsilon,t_\varepsilon)| - \bar{C}|\alpha||D\varphi(x_\varepsilon,t_\varepsilon)|,$$

$$\ell(x_\varepsilon,t_\varepsilon,\alpha)-\ell(y_\varepsilon,t_\varepsilon,\alpha) \geq -(1+|\alpha|^2)m_r(|x_\varepsilon-y_\varepsilon|) \geq -|\alpha|^2 m(\varepsilon)+m(\varepsilon),$$

$$(1-\mu)\ell(y_\varepsilon,t_\varepsilon,\alpha) \geq (1-\mu)\left(\frac{\nu}{2}|\alpha|^2+\ell_0(x,t,\alpha)\right) \geq \frac{\nu(1-\mu)}{2}|\alpha|^2,$$

where the last inequality follows from the fact that by assumption, $\ell_0(x,t,\alpha)\geq 0$.

By proceeding exactly as in Step 3 one can show that

$$\operatorname{Tr}\left[\sigma_x\sigma_x^T X-\sigma_y\sigma_y^T Y\right] \leq \operatorname{Tr}\left[\sigma_x\sigma_x^T D^2\varphi(x_\varepsilon,t_\varepsilon)\right]+m(\varepsilon)+m\left(\frac{\rho}{\varepsilon^4}\right), \tag{2.23}$$

where m is independent of α. Thus, using (2.2), we have

$$\begin{aligned}\mathcal{H} \geq &\inf_{\alpha\in\mathbb{R}^k}\left\{\left(\frac{\nu(1-\mu)}{2}+m(\varepsilon)\right)|\alpha|^2-\bar{C}(|D\varphi(x_\varepsilon,t_\varepsilon)|+m(\varepsilon))|\alpha|\right\}\\ &+\inf_{\alpha\in\mathbb{R}^k}\left\{-\operatorname{Tr}\left[\sigma(x_\varepsilon,t_\varepsilon,\alpha)\sigma(x_\varepsilon,t_\varepsilon,\alpha)^T D^2\varphi(x_\varepsilon,t_\varepsilon)\right]\right\}\\ &-\bar{C}(1+|x_\varepsilon|)|D\varphi(x_\varepsilon,t_\varepsilon)|+m(\varepsilon)+m\left(\frac{\rho}{\varepsilon^4}\right)\\ \geq &-\frac{(\bar{C}|D\varphi(x_\varepsilon,t_\varepsilon)|+m(\varepsilon))^2}{2\nu(1-\mu)+m(\varepsilon)}-\bar{C}(1+|x_\varepsilon|)|D\varphi(x_\varepsilon,t_\varepsilon)|\\ &+\inf_{\alpha\in\mathbb{R}^k}\left\{-\operatorname{Tr}\left[\sigma(x_\varepsilon,t_\varepsilon,\alpha)\sigma(x_\varepsilon,t_\varepsilon,\alpha)^T D^2\varphi(x_\varepsilon,t_\varepsilon)\right]\right\}+m(\varepsilon)+m\left(\frac{\rho}{\varepsilon^4}\right).\end{aligned} \tag{2.24}$$

Step 5. Finally, from (2.22), (2.23), and (2.24), letting first ρ go to 0 and then sending ε to 0, we obtain

$$\begin{aligned}\mathcal{L}[\varphi](\bar{x},\bar{t}) = \varphi_t(\bar{x},\bar{t}) &- \frac{\bar{C}^2}{2\nu(1-\mu)}|D\varphi(\bar{x},\bar{t})|^2-2\bar{C}(1+|\bar{x}|)|D\varphi(\bar{x},\bar{t})|\\ &+\inf_{\alpha\in\mathbb{R}^k}\left\{-\operatorname{Tr}\left[\sigma(\bar{x},\bar{t},\alpha)\sigma(\bar{x},\bar{t},\alpha)^T D^2\varphi(\bar{x},\bar{t})\right]\right\}\\ &+\inf_{\beta\in B}\left\{-\operatorname{Tr}\left[c(\bar{x},\bar{t},\beta)c(\bar{x},\bar{t},\beta)^T D^2\varphi(\bar{x},\bar{t})\right]\right\}\leq 0,\end{aligned}$$

which means exactly that Ψ is a subsolution of (2.10). □

Proof of Lemma 2.3. Set $\chi(s,t)=\varphi(\mathrm{e}^s,t)$ for $(s,t)\in\mathbb{R}\times[0,+\infty)$. A straightforward calculation shows that φ satisfies (2.12) if and only if χ is a solution of the heat equation

$$\begin{cases}\chi_t-\chi_{ss}=0 & \text{in } \mathbb{R}\times(0,T),\\ \chi(s,0)=\varphi_R(\mathrm{e}^s) & \text{in } \mathbb{R}.\end{cases} \tag{2.25}$$

Since the initial data satisfy the growth estimate $|\chi(s,0)| < \mathrm{e}^{s^2}$, by classical results on the heat equation (John [24], Evans [17]), we know there exists a unique classical solution $\chi \in C(\mathbb{R}\times[0,T]) \times C^\infty(\mathbb{R}\times(0,T])$ of (2.25). It is given by the representation formula: for every $(s,t) \in \mathbb{R}\times[0,T]$,

$$\chi(s,t) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} \mathrm{e}^{-\frac{(s-y)^2}{4t}} \varphi_R(\mathrm{e}^y)\, dy = \frac{1}{\sqrt{4\pi t}} \int_{\log R}^{+\infty} \mathrm{e}^{-\frac{(s-y)^2}{4t}} (\mathrm{e}^y - R)\, dy. \tag{2.26}$$

From the above formula, it follows $\chi(s,t) > 0$ for all $(s,t) \in \mathbb{R}\times(0,T]$. Let $h > 0$. We have $\varphi_R(\mathrm{e}^s) \le \varphi_R(\mathrm{e}^{s+h})$ for all $s \in \mathbb{R}$. Since $\chi(\cdot + h, \cdot)$ is a solution of (2.25) with initial data $\varphi_R(\mathrm{e}^{s+h})$, by the maximum principle, we obtain $\chi(s,t) \le \chi(s+h,t)$. This proves that χ is nondecreasing with respect to s. It follows that $\varphi(r,t) = \chi(\log r, t)$ is the unique solution of (2.12) and $\varphi \in C([0,+\infty)\times[0,T]) \cap C^\infty((0,+\infty)\times(0,T])$ is positive and nondecreasing. Moreover, if the initial data are convex, we know that the solution of a quasi-linear equation like (2.12) is convex in the space variable for every time (see, e.g., Giga et al. [20]). Thus $\varphi(\cdot,t)$ is convex in $[0,+\infty)$ for all $t \in [0,T]$.

It remains to prove the estimates (2.13). Noticing that $\varphi_R(\mathrm{e}^s) \le \mathrm{e}^s$ and that $(s,t) \mapsto \varphi_R(\mathrm{e}^s)$ and $(s,t) \mapsto \mathrm{e}^{s+t}$ are respectively sub- and supersolution (in the viscosity sense, for example) of (2.25), by the maximum principle, we obtain $\varphi_R(\mathrm{e}^s) \le \chi(s,t) \le \mathrm{e}^{s+t} \le \mathrm{e}^{s+T}$ for $(s,t) \in \mathbb{R}\times[0,T]$. It follows

$$\varphi_R(r) \le \varphi(r,t) \le \mathrm{e}^T r \quad \text{for } (r,t) \in [0,+\infty]\times[0,T]. \tag{2.27}$$

This gives the first estimate. To prove the second estimate, we note that $\varphi(\cdot,t)$ is a convex nondecreasing function satisfying (2.27). It follows $0 \le \varphi_r(r,t) \le \mathrm{e}^T$ for $(r,t) \in [0,+\infty]\times(0,T]$. The last assertion is obvious, using the dominated convergence theorem in (2.26). It completes the proof of the lemma. □

3. Applications. This section is divided into two parts. In the first part we consider a finite horizon unbounded stochastic control problem and we characterize the value function as the unique viscosity solution of the corresponding dynamic programming equation, which is a particular case of (1.1). In the second part we list some concrete examples of model cases to which the results of section 1 can be applied.

3.1. Unbounded stochastic control problems. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge0}, P)$ be a filtered probability space, W_t be an $\mathcal{F}_t$-adapted standard M-Brownian motion such that $W_0 = 0$ a.s., and let A be a subset of a separable normed space (possibly unbounded). We consider a *finite horizon unbounded stochastic control problem* for controlled diffusion processes $X_s^{t,x}$ whose dynamic is governed by a stochastic differential equation of the form

$$\begin{cases} dX_s^{t,x} = b(X_s^{t,x}, s, \alpha_s)ds + \sigma(X_s^{t,x}, s, \alpha_s)dW_s, & s \in (t,T),\ 0 \le t \le T, \\ X_t^{t,x} = x \in \mathbb{R}^N, \end{cases} \tag{3.1}$$

where the control $\alpha_s \in A$, $b\colon \mathbb{R}^N \times \mathbb{R} \times A \to \mathbb{R}^N$ is a continuous vector field and σ is a continuous real $N \times M$ matrix. The payoff to be minimized is

$$J(t,x,\alpha) = \mathrm{E}_{tx}\left\{ \int_t^T \ell(X_s^{t,x}, s, \alpha_s)\, ds + \psi(X_T^{t,x}) \right\},$$

where E_{tx} is the expectation with respect to the event $X_t^{t,x} = x$, the functions $\ell : \mathbb{R}^N \times [0,T] \times A \to \mathbb{R}$ and $\psi : \mathbb{R}^N \to \mathbb{R}$ are continuous, $\alpha_s \in \mathcal{A}_t$, the set of A-valued

$\mathcal{F}_t$-progressively measurable controls such that

$$\mathrm{E}_{tx}\left(\int_t^T |\alpha_s|^2\, ds\right) < +\infty, \tag{3.2}$$

and $X_s^{t,x}$ is the solution of (3.1). The value function is defined by

$$V(x,t) = \inf_{\alpha_s \in \mathcal{A}_t} J(t,x,\alpha_s). \tag{3.3}$$

At least formally, the dynamic programming equation associated to this control problem is

$$-\frac{\partial w}{\partial t} + \sup_{\alpha\in A}\left\{-\langle b(x,t,\alpha), Dw\rangle - \ell(x,t,\alpha) - \frac{1}{2}\mathrm{Tr}\left(\sigma(x,t,\alpha)\sigma(x,t,\alpha)^T D^2 w\right)\right\} = 0 \tag{3.4}$$

in $\mathbb{R}^N \times (0,T)$, with the terminal value condition $w(x,T) = \psi(x)$.

Our main goal is to characterize the value function V as the unique continuous viscosity solution of (3.4) with the terminal value condition $V(x,T) = \psi(x)$. We recall that the fact that the value function is a viscosity solution of (3.4) is in general obtained by a direct use of the dynamic programming principle. Since we are in an unbounded control framework, the proof of the dynamic programming principle is rather delicate, and thus we follow another strategy which consists in comparing directly V with the unique viscosity solution U of (3.4) obtained by Corollary 2.1 when this latter exists.

We make the following assumptions on the data.

(S0) A is a subset (possibly unbounded) of a separable complete normed space.

(S1) $b \in C(\mathbb{R}^N \times [0,T] \times A; \mathbb{R}^N)$ and there exists $\bar{C} > 0$ such that for all $x, y \in \mathbb{R}^N$, $t \in [0,T]$, and $\alpha \in A$ we have

$$|b(x,t,\alpha) - b(y,t,\alpha)| \le \bar{C}|x-y|,$$
$$|b(x,t,\alpha)| \le \bar{C}(1+|x|+|\alpha|);$$

(S2) $\sigma \in C(\mathbb{R}^N \times [0,T] \times A; \mathcal{M}_{N,M})$ and there exists $\bar{C} > 0$ such that for all $x, y \in \mathbb{R}^N$, $t \in [0,T]$ and $\alpha \in A$ we have

$$|\sigma(x,t,\alpha) - \sigma(y,t,\alpha)| \le \bar{C}|x-y|,$$
$$|\sigma(x,t,\alpha)| \le \bar{C}(1+|x|).$$

Moreover we assume that ℓ and ψ satisfy, respectively, (A1)(iii) and (A3).

We first observe that under the current assumptions (S1) and (S2) on b and σ, for any control $\alpha \in \mathcal{A}_t$ satifying (3.2) and any random variable Z such that $\mathrm{E}[Z] < \infty$, there exists a unique strong solution of the stochastic differential equation (3.1) which satisfies

$$\mathrm{E}\left\{\sup_{t\le s\le T} |X_s^{t,Z}|^2\right\} < \infty$$

(see, e.g., Appendix D in [19]). Moreover, we have better estimates on the trajectories of (3.1).

LEMMA 3.1. *Assume* (S0), (S1), *and* (S2). *For every* $(x,t) \in \mathbb{R}^N \times [0,T]$ *and every* $\alpha_s \in \mathcal{A}_t$, *the solution* $X_s^{t,x}$ *of* (3.1) *corresponding to* α_s *satisfies the following properties:*

(i) *there exists a constant $C > 0$ such that*

$$\mathrm{E}_{tx}\left\{\sup_{t\le s\le T}|X_s^{t,x}|^2\right\} \le \left(|x|^2 + C(T-t) + C\mathrm{E}_{tx}\int_t^T |\alpha_s|^2\,ds\right)\mathrm{e}^{C(T-t)}\ ; \tag{3.5}$$

(ii) *there exists $C_{x,\alpha} > 0$ which depends on x and on the control α_s such that for all $s, s' \in [t,T]$,*

$$\mathrm{E}_{tx}\{|X_s^{t,x} - X_{s'}^{t,x}|\} \le C_{x,\alpha}|s-s'|^{1/2}. \tag{3.6}$$

In particular for all $\tau \in [t,T]$ we have

$$\mathrm{E}_{tx}\left\{\sup_{t\le s\le \tau}|X_s^{t,x} - x|\right\} \le C_{x,\alpha}|\tau|^{1/2}. \tag{3.7}$$

Proof of Lemma 3.1. We start by proving (i). Let us take an increasing sequence of C^2 functions $\varphi_R : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $R > 0$, $\varphi_R'(r) = 0$ if $r > 2R$, $\varphi_R''(r) \le 0$ and $\varphi_R(r) \uparrow r$, $\varphi_R'(r) \uparrow 1$, as $R \to +\infty$. By applying Ito's formula to the process $\varphi_R(|X_s^{t,x}|^2)$, for a.e. $t \le \tau \le T$, we have (dropping the argument of φ_R and its derivatives)

$$\begin{aligned}\varphi_R(|X_\tau^{t,x}|^2) &= \varphi_R(|x|^2) + \int_t^\tau 2\varphi_R'\langle X_s^{t,x}, b(X_s^{t,x},s,\alpha_s)\rangle\,ds\\ &\quad + \int_t^\tau \left(\varphi_R'\mathrm{Tr}[\sigma\sigma^T(X_s^{t,x},s,\alpha_s)] + 2\varphi_R''|\sigma^T(X_s^{t,x},s,\alpha_s)X_s^{t,x}|^2\right)ds\\ &\quad + \int_t^\tau 2\varphi_R'\langle X_s^{t,x}, \sigma(X_s^{t,x},s,\alpha_s)dW_s\rangle.\end{aligned} \tag{3.8}$$

By using the current assumptions on b and σ the following estimate holds:

$$\begin{aligned}&\int_t^\tau \left(2\varphi_R'\langle X_s^{t,x}, b(X_s^{t,x},s,\alpha_s)\rangle + \varphi_R'\mathrm{Tr}[\sigma\sigma^T(X_s^{t,x},s,\alpha_s)]\right.\\ &\qquad \left. + 2\,\varphi_R''|\sigma^T(X_s^{t,x},s,\alpha_s)X_s^{t,x}|^2\right)ds\\ &\le 2C\int_t^\tau \varphi_R'|X_s^{t,x}|(1+|X_s^{t,x}|+|\alpha_s|)\,ds + C\int_t^\tau \varphi_R'(1+|X_s^{t,x}|^2)\,ds \quad \text{a.s.},\end{aligned}$$

where the constant C depends on neither the control α_s nor on R. Moreover, we observe that since $\varphi_R' = 0$ for $t > 2R$ we have

$$\mathrm{E}_{tx}\left\{\int_t^\tau |\varphi_R'\langle X_s^{t,x}, \sigma(X_s,s,\alpha_s)\rangle|^2\,ds\right\} < +\infty,$$

hence the expectation of the stochastic integral is zero. By taking the expectation in (3.9) and applying Fubini's theorem we obtain

$$\mathrm{E}_{tx}\{\varphi_R(|X_\tau^{t,x}|^2)\} \le \varphi_R(|x|^2) + C\int_t^\tau \mathrm{E}_{tx}\{\varphi_R'[2+5|X_s^{t,x}|^2 + |\alpha_s|^2]\}\,ds.$$

Since φ_R, φ_R' are increasing sequences, we can apply Levi's theorem. Therefore by letting $R \to \infty$ we obtain, for every $t \le \tau \le T$,

$$\begin{aligned}\mathrm{E}_{tx}\{|X_\tau^{t,x}|^2\} &\le |x|^2 + C\int_t^\tau \mathrm{E}_{tx}\{2+5|X_s^{t,x}|^2 + |\alpha_s|^2\}\,ds\\ &\le |x|^2 + 5C\int_t^\tau \mathrm{E}_{tx}|X_s^{t,x}|^2\,ds + C\mathrm{E}_{tx}\left\{\int_t^\tau |\alpha_s|^2\,ds\right\} + 2C(T-t).\end{aligned}$$

Applying Gronwall's inequality we obtain

$$\mathrm{E}_{tx}\{|X_\tau^{t,x}|^2\} \le \left(|x|^2 + 2C(T-t) + C\,\mathrm{E}_{tx}\left\{\int_t^T |\alpha_s|^2\,ds\right\}\right) \mathrm{e}^{5C(T-t)}.$$

We conclude by Doob's maximal inequality (see, e.g., [25]).

The proof of (ii) is an extension of the one in the Appendix D in [19] and we leave it to the reader. □

PROPOSITION 3.1. *Assume* (S0), (S1), (S2), (A1)(iii) *for* ℓ *and* (A3) *for* ψ. *Then there exists* $0 \le \tau < T$ *such that the value function* V *is finite and satisfies the quadratic growth condition* (2.4) *in* $\mathbb{R}^N \times [\tau, T]$.

Proof of Proposition 3.1. We aim to show that if the constants $\rho, C > 0$ are large enough, then there exists $\tau > 0$ depending on ρ such that $|V(x,t)| \le C(1+|x|^2)\mathrm{e}^{\rho(T-t)}$ for all $(x,t) \in \mathbb{R}^N \times [\tau, T]$. The upper estimate is obtained by majorizing directly the value function with the cost functional corresponding to a constant control and using the estimates of the trajectories in Lemma 3.1. The most difficult is to prove the estimate from below since V is defined by an infimum. To this purpose we take any control $\alpha_s \in \mathcal{A}_t$. By applying Ito's formula to the process $(1+|X_s^{t,x}|^2)\mathrm{e}^{\rho(t-s)}$, $X_s^{t,x}$ being the trajectory corresponding to α_s, we have the following estimate:

$$\begin{aligned} d[(1+|X_s^{t,x}|^2)\mathrm{e}^{\rho(T-s)}] &= -\rho\mathrm{e}^{\rho(T-s)}(1+|X_s^{t,x}|^2)ds + \mathrm{e}^{\rho(T-s)}\mathrm{Tr}\,(\sigma\sigma^T(X_s^{t,x},s,\alpha_s))ds \\ &\quad + 2\mathrm{e}^{\rho(T-s)}\langle X_s^{t,x}, b(X_s^{t,x},s,\alpha_s)ds + \sigma(X_s^{t,x},s,\alpha_s)dW_s\rangle. \end{aligned} \tag{3.9}$$

Integrating both sides of (3.9) from t to T and taking the expectation we get

$$\begin{aligned} \mathrm{E}_{tx}\{1+|X_T^{t,x}|^2\} - (1+|x|^2)\mathrm{e}^{\rho(T-t)} = \mathrm{E}_{tx}\Bigg\{\int_t^T \Big(&-\rho(1+|X_s^{t,x}|^2) \\ &+ 2\langle X_s^{t,x}, b(X_s^{t,x},s,\alpha_s)\rangle \\ &+ \mathrm{Tr}(\sigma\sigma^T(X_s,s,\alpha_s)\Big)\mathrm{e}^{\rho(T-s)}\,ds\Bigg\}. \end{aligned} \tag{3.10}$$

We notice that in the above estimate we supposed that the expectation of the stochastic integral is zero. This is false in general but we can overcome such a difficulty by an approximation argument which is similar to the one used in the proof of Lemma 3.1. Now for any ε-optimal control α_s for $V(x,t)$, by using (3.10), we get

$$\begin{aligned} &V(x,t) + C(1+|x|^2)\mathrm{e}^{\rho(T-t)} + \varepsilon \\ &\quad \ge \mathrm{E}_{tx}\left\{\int_t^T \Big(\ell(X_s^{t,x},s,\alpha_s) - 2C\mathrm{e}^{\rho(T-s)}\langle X_s^{t,x}, b(X_s^{t,x},s,\alpha_s)\rangle \right. \\ &\qquad - C\mathrm{e}^{\rho(T-s)}\mathrm{Tr}(\sigma\sigma^T(X_s^{t,x},s,\alpha_s)) + \rho\mathrm{e}^{\rho(T-s)}C(1+|X_s|^2)\Big)\,ds \\ &\qquad \left. + \psi(X_T^{t,x}) + C(1+|X_T^{t,x}|^2)\right\} \\ &\quad = \mathrm{E}_{tx}\left\{\int_t^T \bar{\ell}(X_s^{t,x},s,\alpha_s)\,ds + \bar{\psi}(X_T^{t,x})\right\}, \end{aligned}$$

where $\bar\ell(x,t,\alpha) := \ell(x,t,\alpha) - 2C\mathrm{e}^{\rho(T-t)}\langle b(x,t,\alpha),x\rangle - 2C\mathrm{e}^{\rho(T-t)}\mathrm{Tr}\,a(x,t,\alpha) + C\rho\mathrm{e}^{\rho(T-t)}(1+|x|^2)$ and $\bar\psi(x) := \psi(x) + C(1+|x|^2)$. By analogous arguments as those used in section 2 one can see that for $\rho, C > 0$ large enough there is $\tau > 0$ such that $\bar\ell$ and $\bar\psi$ are nonnegative in $\mathbb{R}^N \times [\tau, T]$. Thus we can conclude since ε is arbitrary. □

Remark 3.1. If ℓ, ψ are bounded from below, namely, they satisfy, for some $C > 0$, the two conditions $\ell(x,t,\alpha) \ge \nu|\alpha|^2 - C$ and $\psi(x) \ge -C$, then V is finite and satisfies the growth condition (2.4) in $\mathbb{R}^N \times [0,T]$ (i.e., for all time).

Next we prove that V is the unique viscosity solution of (3.4). We start with the following proposition.

PROPOSITION 3.2. *Under the assumptions of Proposition* 3.1 *we have*

(i) (*superoptimality principle*) *for all* $t \in (\tau, T]$ *and* $0 < h \le T - t$ *and for all stopping time* $t \le \theta \le T$, *we have*

(3.11)
$$V(x,t) \ge \inf_{\alpha_s \in \mathcal{A}_t} \mathrm{E}_{tx}\left\{\int_t^{(t+h)\wedge\theta} \ell(X_s^{t,x}, s, \alpha_s)\,ds + V_*(X^{t,x}_{(t+h)\wedge\theta}, (t+h)\wedge\theta)\right\};$$

(ii) *the function* V *is a supersolution of* (3.4) *in* $\mathbb{R}^N \times [\tau, T]$.

Proof of Proposition 3.2. The proof of (i) is a standard routine and we refer the reader, for instance, to [41]. The opposite inequality is more delicate; see Krylov [27, 28].

We turn to the proof of (ii), showing that the superoptimality principle implies that V_* is a viscosity supersolution of (3.4). Let $\phi \in C^2(\mathbb{R}^N \times [0,T])$ and $(\bar x, \bar t) \in \mathbb{R}^N \times (\tau, T)$ be a local minimum of $V_* - \phi$. We can assume that $V_*(\bar x, \bar t) = \phi(\bar x, \bar t)$ and that the maximum is strict, i.e., $V_*(x,t) > \phi(x,t)$ for all $(x,t) \in \bar B(\bar x, \varepsilon) \times [\bar t - \varepsilon, \bar t + \varepsilon]$ with $(x,t) \ne (\bar x, \bar t)$ (see [3] or [6]). We assume by contradiction that there exists $\delta_\varepsilon > 0$ such that for all $(x,t) \in \bar B(\bar x, \varepsilon) \times [\bar t - \varepsilon, \bar t + \varepsilon]$, we have

(3.12)
$$-\phi_t(x,t) + \sup_{\alpha\in A}\left\{-\langle b(x,t,\alpha), D\phi(x,t)\rangle - \ell(x,t,\alpha) - \mathrm{Tr}\left[\frac{1}{2}\sigma\sigma^T(x,t,\alpha)D^2\phi(x,t)\right]\right\} \le -\delta_\varepsilon.$$

Since $(\bar x, \bar t)$ is a strict minimum of $V_* - \phi$, it follows that there exists η_ε such that

$$V_*(x,t) \ge \phi(x,t) + \eta_\varepsilon \quad \text{for all } (x,t) \in \partial B(\bar x, \varepsilon) \times [\bar t - \varepsilon, \bar t + \varepsilon]. \tag{3.13}$$

From now on, we fix $0 < h < \varepsilon/2$ such that $h\delta_\varepsilon < \eta_\varepsilon$. Let us denote by $\tau_{t,x}$ the exit time of the trajectory $X_s^{t,x}$ from the ball $B(\bar x, \varepsilon)$. We first observe that by the continuity of the trajectory (see Lemma 3.1), we have $\tau_{t,x} > t$ for all $(x,t) \in B(\bar x, \varepsilon) \times [0,T)$. For every $(x,t) \in B(\bar x, \varepsilon) \times (\bar t - \varepsilon/2, \bar t + \varepsilon/2)$, there exists a control $\alpha_s \in \mathcal{A}_t$ such that

$$V(x,t) + \frac{\delta_\varepsilon h}{2} \ge \mathrm{E}_{tx}\left\{\int_t^{(t+h)\wedge\tau_{t,x}} \ell(X_s^{t,x}, s, \alpha_s)\,ds + V_*(X^{t,x}_{(t+h)\wedge\tau_{t,x}}, (t+h)\wedge\tau_{t,x})\right\}.$$

Since $V_* \ge \phi$ in $\bar B(\bar x, \varepsilon) \times [\bar t - \varepsilon, \bar t + \varepsilon]$, if $\tau_{t,x} < t + h$, then, from (3.13), we have

$$V_*(X^{t,x}_{(t+h)\wedge\tau_{t,x}}, (t+h)\wedge\tau_{t,x}) \ge \phi(X^{t,x}_{\tau_{t,x}}, \tau_{t,x}) + \eta_\varepsilon.$$

Therefore the following estimate holds:

$$
\begin{aligned}
V(x,t)+\frac{\delta_\varepsilon h}{2} &\geq \mathrm{E}_{tx}\left\{\left[\int_t^{\tau_{t,x}} \ell(X_s^{t,x},s,\alpha_s)\,ds+\phi(X_{\tau_{t,x}}^{t,x},\tau_{t,x})+\eta_\varepsilon\right]1_{\{\tau_{t,x}<t+h\}}\right\}\\
&\quad+\mathrm{E}_{tx}\left\{\left[\int_t^{t+h} \ell(X_s^{t,x},s,\alpha_s)\,ds+\phi(X_{t+h}^{t,x},t+h)\right]1_{\{\tau_{t,x}\geq t+h\}}\right\}\\
&\geq \mathrm{E}_{tx}\Big\{[I(\tau_{t,x})+\eta_\varepsilon]1_{\{\tau_{t,x}<t+h\}}\Big\}+\mathrm{E}_{tx}\Big\{I(t+h)1_{\{\tau_{t,x}\geq t+h\}}\Big\}, \qquad (3.14)
\end{aligned}
$$

where for all $\tau'>0$

$$
I(\tau')=\int_t^{\tau'} \ell(X_s^{t,x},s,\alpha_s)\,ds+\phi(X_{\tau'}^{t,x},\tau').
$$

Applying Ito's formula to the process $\phi(X_{\tau'}^{t,x},\tau')$, we obtain

$$
\begin{aligned}
I(\tau')=\int_t^{\tau'}&\Big(\ell(X_s^{t,x},s,\alpha_s)+\phi_t(X_s^{t,x},s)+\langle D\phi(X_s^{t,x},s),b(X_s^{t,x},s,\alpha_s)\rangle\\
&+\frac{1}{2}\mathrm{Tr}\left[\sigma\sigma^T(X_s^{t,x},s,\alpha_s)D^2\phi\right]\Big)\,ds\\
&+\phi(x,t)+\int_t^{\tau'}\langle D\phi(X_s^{t,x},s),\sigma(X_s^{t,x},s,\alpha_s)dW_s\rangle \quad \text{a.s.}
\end{aligned}
$$

Note that the expectation of the above stochastic integral is zero for $\tau'\in[t,t+\varepsilon]$.

Now we can estimate the two last terms in (3.14). For the first term, we have

$$
\begin{aligned}
&\mathrm{E}_{tx}\Big\{[I(\tau_{t,x})+\eta_\varepsilon]1_{\{\tau_{t,x}<t+h\}}\Big\}\\
&\geq -\mathrm{E}_{tx}\left\{\left[\int_t^{\tau_{t,x}}\Big(-\phi_t(X_s^{t,x},s)+\sup_{\alpha\in A}\{-\ell(X_s^{t,x},s,\alpha)-\langle D\phi(X_s^{t,x},s),b(X_s^{t,x},s,\alpha)\rangle\right.\right.\\
&\qquad\left.\left.-\frac{1}{2}\mathrm{Tr}\left[\sigma\sigma^T(X_s^{t,x},s,\alpha)D^2\phi\right]\}\Big)\,ds-\phi(x,t)-\eta_\varepsilon\right]1_{\{\tau_{t,x}<t+h\}}\right\}.
\end{aligned}
$$

Since $X_s^{t,x}\in B(\bar{x},\varepsilon)$ when $s\leq\tau_{t,x}$ and since $t+h<\bar{t}+\varepsilon$, from (3.12), we get

$$
\begin{aligned}
\mathrm{E}_{tx}\Big\{[I(\tau_{t,x})+\eta_\varepsilon]1_{\{\tau_{t,x}<t+h\}}\Big\} &\geq -\mathrm{E}_{tx}\left\{\left[\int_t^{\tau_{t,x}}(-\delta_\varepsilon)\,ds-\phi(x,t)-\eta_\varepsilon\right]1_{\{\tau_{t,x}<t+h\}}\right\}\\
&\geq \delta_\varepsilon\mathrm{E}_{tx}\left[(\tau_{t,x}-t)1_{\{\tau_{t,x}<t+h\}}\right]\\
&\quad+(\eta_\varepsilon+\phi(x,t))P(\{\tau_{t,x}<t+h\})\\
&\geq (\eta_\varepsilon+\phi(x,t))P(\{\tau_{t,x}<t+h\}). \qquad (3.15)
\end{aligned}
$$

For the second term, we proceed in the same way, noting that if $\tau_{t,x}\geq t+h$, then for all $t\leq s\leq t+h$, $X_s^{t,x}\in B(\bar{x},\varepsilon)$, and it allows us to apply (3.12). More precisely we

have

(3.16)

$$
\begin{aligned}
&\mathrm{E}_{tx}\Big\{I(t+h)1_{\{\tau_{t,x}\geq t+h\}}\Big\}\\
&\quad\geq -\mathrm{E}_{tx}\left\{\left[\int_t^{t+h}\Big(-\phi_t(X_s^{t,x},s)+\sup_{\alpha\in A}\{-\ell(X_s^{t,x},s,\alpha)-\langle D\phi(X_s^{t,x},s),b(X_s^{t,x},s,\alpha)\rangle\right.\right.\\
&\qquad\left.\left.-\frac{1}{2}\mathrm{Tr}\,[\sigma\sigma^T(X_s^{t,x},s,\alpha)D^2\phi]\}\Big)\,ds-\phi(x,t)\right]1_{\{\tau_{t,x}\geq t+h\}}\right\}\\
&\quad\geq(\delta_\varepsilon h+\phi(x,t))P(\{\tau_{t,x}\geq t+h\}).
\end{aligned}
$$

Combining (3.14), (3.15), and (3.16), we get

$$
\begin{aligned}
V(x,t)+\frac{\delta_\varepsilon h}{2}&\geq\eta_\varepsilon P(\{\tau_{t,x}<t+h\})+\delta_\varepsilon h\,P(\{\tau_{t,x}\geq t+h\})\\
&\quad+\phi(x,t)\left[P(\{\tau_{t,x}<t+h\})+P(\{\tau_{t,x}\geq t+h\})\right].
\end{aligned}
$$

Since $\eta_\varepsilon>\delta_\varepsilon h$ and $P(\{\tau_{t,x}<t+h\})+P(\{\tau_{t,x}\geq t+h\})=1$, we get

$$
V(x,t)\geq\phi(x,t)+\frac{\delta_\varepsilon h}{2}.
$$

The above inequality is valid for all $(x,t)\in B(\bar{x},\varepsilon)\times(\bar{t}-\varepsilon/2,\bar{t}+\varepsilon/2)$, and thus we have

$$
\liminf_{(x,t)\to(\bar{x},\bar{t})}V(x,t)=V_*(\bar{x},\bar{t})\geq\phi(\bar{x},\bar{t})+\frac{\delta_\varepsilon h}{2},
$$

which is a contradiction with the choice of ϕ. □

THEOREM 3.1. *Under the assumptions of Proposition* 3.1*, the function V is the unique continuous viscosity solution of* (3.4) *in* $\mathbb{R}^N\times[\tau,T]$.

Proof of Theorem 3.1. Let U be the unique solution of (3.4) in $\mathbb{R}^N\times[\tau,T]$ such that $U(x,T)=\psi(x)$ given by Theorem 2.1. Our goal is to prove that $V\equiv U$ in $\mathbb{R}^N\times[\tau,T]$. The inequality $U\leq V_*$ follows by combining Proposition 3.2 and Theorem 2.1. To show that $V^*\leq U$ in $\mathbb{R}^N\times[\tau,T]$, we proceed as follows.

Step 1. We consider the functions $\widetilde{V}(x,t):=V(x,t)-C(1+|x|^2)\mathrm{e}^{\rho(T-t)}$ and $\widetilde{U}(x,t):=U(x,t)-C(1+|x|^2)\mathrm{e}^{\rho(T-t)}$. As it is proved in Step 2 of the proof of Theorem 2.1, $\widetilde{U}$ is the unique solution of

(3.17)

$$
\begin{cases}
-w_t+\sup_{\alpha\in A}\left\{-\frac{1}{2}\mathrm{Tr}(\sigma\sigma^T(x,t,\alpha)D^2w)-\langle b(x,t,\alpha),Dw\rangle-\bar{\ell}(x,t,\alpha)\right\}=0 \text{ in } \mathbb{R}^N\times(\tau,T),\\
w(x,T)=\bar{\psi}(x),
\end{cases}
$$

where $\bar{\ell}(x,t,\alpha):=\ell(x,t,\alpha)+2C\mathrm{e}^{\rho(T-t)}\langle b(x,t,\alpha),x\rangle+C\mathrm{e}^{\rho(T-t)}\mathrm{Tr}\,\sigma\sigma^T(x,t,\alpha)$ $-C\rho\mathrm{e}^{\rho(T-t)}(1+|x|^2)$ and $\bar{\psi}(x):=\psi(x)-C(1+|x|^2)$.

Step 2. Claim: for all $(x,t)\in\mathbb{R}^N\times[\tau,T]$, $\widetilde{V}$ satisfies

$$
\widetilde{V}(x,t)\leq\inf_{\alpha_s\in\mathcal{A}_t}\mathrm{E}_{tx}\left\{\int_t^T\bar{\ell}(X_s,s,\alpha_s)\,ds+\bar{\psi}(X_t)\right\}. \tag{3.18}
$$

To prove the claim let us take any $\alpha_s \in \mathcal{A}_t$. Arguing exactly as in the proof of Proposition 3.1, from (3.10), we have

$$\begin{aligned}\widetilde{V}(x,t) \le \mathrm{E}_{tx}\Bigg\{\int_t^T \Big(&\ell(X_s^{x,t},s,\alpha_s) + 2\mathrm{e}^{\rho(T-s)}\langle X_s^{x,t}, b(X_s^{x,t},t,\alpha_s)\rangle \\ &+ \mathrm{e}^{\rho(T-s)}\,\mathrm{Tr}(\sigma\sigma^T(X_s^{x,t},s,\alpha_s)) - \rho\mathrm{e}^{\rho(T-s)}(1+|X_s^{x,t}|^2)\Big)\,ds + \psi(X_T^{x,t}) \\ &- C(1+|X_T^{x,t}|^2)\Bigg\} \\ = \mathrm{E}_{tx}\Bigg\{\int_t^T &\bar{\ell}(X_s^{x,t},s,\alpha_s)\,ds + \bar{\psi}(X_T^{x,t})\Bigg\}.\end{aligned}$$

Since α_s is arbitrary we get (3.18) and prove the claim.

Step 3. Choose $C, \rho > 0$ large enough so that for all $(x,t,\alpha) \in \mathbb{R}^N \times [\tau,T] \times A$, we have

$$\begin{aligned}-\widetilde{C}\mathrm{e}^{2\rho(T-t)}(1+|x|^2) + \frac{\nu}{4}|\alpha|^2 \le \bar{\ell}(x,t,\alpha) \le -C\mathrm{e}^{\rho(T-t)}(1+|x|^2) + C\mathrm{e}^{\rho(T-t)}(1+|\alpha|^2), \\ -2C(1+|x|^2) \le \bar{\psi}(x) \le 0,\end{aligned}$$

where $\widetilde{C}$ depends only on C and ν. For all real $R > 0$ and all integer $n > 0$, we set $A_n := \{\alpha \in A : |\alpha| \le n\}$, $\bar{\ell}_R(x,t,\alpha) := \max\{\bar{\ell}(x,t,\alpha), -R\}$ and $\bar{\psi}_R(x,t) := \max\{\bar{\psi}(x,t), -R\}$. We observe that $\bar{\ell}_R \colon \mathbb{R}^N \times [\tau,T] \times A_n \to \mathbb{R}$ is bounded and uniformly continuous in $x \in \mathbb{R}^N$ uniformly with respect to $(t,\alpha) \in [\tau,T] \times A_n$ and $\psi_R \colon \mathbb{R}^N \to \mathbb{R}$ is bounded and uniformly continuous in $\mathbb{R}^N$. Set

$$\begin{aligned}\bar{H}(x,t,p,X) &:= \sup_{\alpha\in A}\left\{-\frac{1}{2}\mathrm{Tr}(\sigma\sigma^T(x,t,\alpha)X) - \langle b(x,t,\alpha),p\rangle - \bar{\ell}(x,t,\alpha)\right\}, \\ H_n^R(x,t,p,X) &:= \sup_{\alpha\in A_n}\left\{-\frac{1}{2}\mathrm{Tr}(\sigma\sigma^T(x,t,\alpha)X) - \langle b(x,t,\alpha),p\rangle - \bar{\ell}_R(x,t,\alpha)\right\}\end{aligned}$$

and define

$$V_n^R(x,t) = \inf_{\alpha_s\in\mathcal{A}_t^n} \mathrm{E}_{tx}\left\{\int_t^T \bar{\ell}_R(X_s^{x,t},s,\alpha_s)\,ds + \bar{\psi}_R(X_T^{x,t})\right\},$$

where $\mathcal{A}_t^n$ is the set of A_n-valued $\mathcal{F}_t$-progressively measurable controls such that (3.2) holds. The function V_n^R is now the value function of a stochastic control problem with bounded controls and uniformly continuous datas b, σ, $\bar{\ell}_R$, and $\bar{\psi}_R$. These assumptions enter the framework of Yong and Zhou [41]. We deduce that V_n^R is the unique continuous viscosity solution of

$$-\frac{\partial V_n^R}{\partial t} + H_n^R(x,t,DV_n^R,D^2V_n^R) = 0 \ \text{ in } \mathbb{R}^N \times (\tau,T), \tag{3.19}$$

with terminal condition $V_n^R(x,T) = \bar{\psi}_R(x)$ in $\mathbb{R}^N$. Moreover, for all compact subsets $K \subset \mathbb{R}^N$ there exists $M > 0$ independent on R and n such that $||V_n^R||_\infty \le M$ in $K \times [\tau,T]$. Indeed, take any constant control $\alpha_s = \bar{\alpha} \in A_1$, by definition of V_n^R for all

$R, n > 0$ and for every $(x,t) \in \mathbb{R}^N \times [\tau, T]$ we have

$$V_n^R(x,t) \le \mathrm{E}_{tx}\left\{\int_t^T \bar{\ell}_R(X_s^{x,t}, s, \bar{\alpha})\, ds + \bar{\psi}_R(X_T^{x,t})\right\}$$
$$\le \int_t^T C\mathrm{e}^{\rho(T-s)}(1+\bar{\alpha}^2)\, ds \le \frac{C}{\rho}\mathrm{e}^{\rho(T-t)}(1+\bar{\alpha}^2),$$

and on the other hand we have $V_n^R(x,t) \ge \widetilde{V}(x,t) \ge -\widetilde{C}\mathrm{e}^{2\rho(T-t)}(1+|x|^2)$.

Finally one can readily see that H_n^R converges locally uniformly to $\bar{H}$ as $n, R \to \infty$. Thus by applying the half-relaxed limits method (see Barles and Perthame [9]), the functions

$$\overline{V}(x,t) = \limsup{}^* V_n^R(x,t) = \limsup_{\substack{(y,s)\to(x,t)\\ n,R\to+\infty}} V_n^R(y,s)$$
$$\text{and } \ \underline{V}(x,t) = \liminf{}^* V_n^R(x,t) = \liminf_{\substack{(y,s)\to(x,t)\\ n,R\to+\infty}} V_n^R(y,s)$$

are, respectively, viscosity sub- and supersolution of (3.17). Theorem 2.1 yields $\overline{V}(x,t) \le \widetilde{U}(x,t) \le \underline{V}(x,t)$. On the other hand by construction we have also $\widetilde{V}^*(x,t) \le \limsup^* V_n^R(x,t)$. It follows $\widetilde{V}^*(x,t) \le \widetilde{U}(x,t)$ and we can conclude. □

3.2. Some examples.

Example 3.1. A model case we have in mind is the so-called stochastic linear regulator problem which is a stochastic perturbation of the deterministic linear quadratic problem. In this case, the stochastic differential (3.1) is linear and reads

$$dX_s^{t,x} = [B(s)X_s^{t,x} + C(s)\alpha_s]ds + \sum_j [C_j(s)X_s^{t,x} + D_j(s)]dW_s^j$$

and the expected total cost to minimized is

$$J(x,t,\alpha_s) = \mathrm{E}_{tx}\left\{\int_t^T [\langle X_s^{t,x}, Q(s)X_s^{t,x}\rangle + \langle \alpha_s, R(s)\alpha_s\rangle]\, ds + \langle X_T^{t,x}, GX_T^{t,x}\rangle\right\}.$$

The previous results apply if the functions $B(\cdot), C(\cdot), C_j(\cdot), D_j(\cdot), Q(\cdot), R(\cdot)$, and G are deterministic continuous matrix-valued functions of suitable size and if $R(s)$ is a positive definite symmetric matrix. Deterministic and stochastic linear quadratic problems were extensively studied. For a survey we refer for instance to the books of Bensoussan [10], Fleming and Rishel [18], Fleming and Soner [19], Øksendal [38], and Yong and Zhou [41] and references therein.

Example 3.2. Equations of the type (3.4) are largely considered in mathematical finance. See the introductory books quoted in the introduction or Pham [39]. In particular, recently Benth and Karlsen [11] studied the following semilinear elliptic partial equation

$$-w_t - \frac{1}{2}\beta^2 w_{xx} + F(x, w_x) = 0 \qquad \text{in } \mathbb{R}\times(0,T) \tag{3.20}$$

with the final condition $w(x,T) = 0$. The nonlinear function F is given by

$$F(x,p) = \frac{1}{2}\delta^2 p^2 - \left\{\alpha(x) - \frac{\mu(x)\beta\rho}{\sigma(x)}\right\} p - \frac{\mu^2(x)}{\sigma^2(x)},$$

where δ, β, ρ are real constant and $\alpha(x)$ and $\dfrac{\mu}{\sigma}(x)$ are C^1 functions satisfying

$$|\alpha(x)|, \left|\frac{\mu}{\sigma}(x)\right| \leq C|x| \quad \text{for all } x \in \mathbb{R}.$$

Their main motivation is to determine via the solution of (3.20) the minimal entropy martingale measure in stochastic market. The conditions they assume on data fall within the assumptions of section 2.

Example 3.3. Another application of the results obtained in section 2 is given by the finite time-horizon risk-sensitive limit problem for nonlinear systems. In the stochastic risk-sensitive problem, (3.1) reads

$$dX_s^{t,x,\varepsilon} = g(X_s^{t,x,\varepsilon}, \beta_s)\, dt + \sqrt{\frac{\varepsilon}{\gamma^2}} c(X_s^{t,x,\varepsilon})\, dW_s, \tag{3.21}$$

where $X_s^{t,x,\varepsilon} \in \mathbb{R}^N$ depends on the parameter $\varepsilon > 0$, g represents the nominal dynamics with control $\beta_s \in B$, a compact normed space, and c is an $N \times k$-valued diffusion coefficient, ε is a measure of the risk-sensitivity and γ is the disturbance attenuation level. The cost criterion is of the form

$$J^\varepsilon(x,t,\beta_s) = \mathrm{E} \exp\left\{\frac{1}{\varepsilon}\left[\int_t^T f(X_s^{t,x,\varepsilon}, \beta_s)\, dt + \psi(X_T^{t,x,\varepsilon})\right]\right\} \tag{3.22}$$

and the value function is

$$V^\varepsilon(x,t) \quad = \inf_{\beta_s \in \mathcal{B}_t} \varepsilon \log J^\varepsilon(x,t,\beta_s) = \varepsilon \log \inf_{\beta_s \in \mathcal{B}_t} J^\varepsilon(x,t,\beta_s), \tag{3.23}$$

where $\mathcal{B}_t$ is the set of B-valued, $\mathcal{F}_t$-progressively measurable controls such that there exists a strong solution to (3.21). The dynamic programming equation associated to this problem is

(3.24)
$$\begin{cases} -\dfrac{\partial w}{\partial t} - \dfrac{1}{2\gamma^2}\langle Dw, a(x) Dw\rangle + \tilde{G}(x, Dw) - \dfrac{\varepsilon}{2\gamma^2}\mathrm{Tr}(a(x) D^2 w) = 0 \quad \text{in } \mathbb{R}^N \times [\tau, T], \\ w(x,T) = \psi(x), \end{cases}$$

where $a(x) = c(x)c(x)^T$ and

$$\tilde{G}(x,p) = \max_{\beta \in B}\{\langle -g(x,\beta), p\rangle - f(x,\beta)\}.$$

We note that

$$-\frac{1}{2\gamma^2}\langle p, a(x) p\rangle = \min_{\alpha \in \mathbb{R}^k}\left\{\frac{\gamma^2}{2}|\alpha|^2 - \langle c(x)\alpha, p\rangle\right\}. \tag{3.25}$$

In [16] it is shown that as $\varepsilon \downarrow 0$ (i.e., as the problem becomes infinitely risk averse), the value function of the risk-sensitive problem converges to that of an H_∞ robust control problem. This problem can be considered as a differential game with the cost functional

$$J(x,t,\alpha,\beta) = \int_t^T \Big(f(y_x(s), \beta_s) - \frac{\gamma^2}{2}|\alpha_s|^2\Big) ds + \psi(y_x(T)), \tag{3.26}$$

where $\alpha_s \in \mathcal{A}_t := L^2([t,T],\mathbb{R}^k)$ is the control of the maximizing player, $\beta_s \in \mathcal{B}_t := \{\text{measurable functions } [t,T] \to B\}$ is the control of the minimizing player, and $y_x(\cdot)$ is the unique solution of the following dynamical system

$$\begin{cases} y'(s) = g(y(s),\beta_s) + c(y(s))\alpha_s, \\ y(t) = x. \end{cases}$$

Note that we switch notation from X_t to $y(t)$ to emphasize that the paths are now (deterministic) solutions of ordinary differential equations rather than (stochastic) solutions of stochastic differential equations. The dynamic programing equation associated to the robust control problem is a first-order equation given by

$$-\frac{\partial w}{\partial t} + \min_{\alpha\in\mathbb{R}^k}\left\{\frac{\gamma^2}{2}|\alpha|^2 - \langle c(x)\alpha, Dw\rangle\right\} + \tilde{G}(x,Dw) = 0 \ \text{ in } \mathbb{R}^N \times (\tau,T) \tag{3.27}$$

with the terminal condition $w(T,x) = \psi(x)$. One of the key tools to get this convergence result is the uniqueness property for (3.27). In Da Lio and McEneaney [16], the authors characterized the value function of the H_∞ control as the unique solution to (3.27) in the set of locally Lipschitz continuous functions growing at most quadratically with respect to the state variable. Moreover, the uniqueness result in [16] is obtained by using representation formulas of locally Lipschitz solutions of (3.27). We remark that the comparison theorem, Theorem 2.1, not only improves the uniqueness result for (3.27) obtained in [16] (in the sense it holds in a larger class of functions) but it also should allow us to prove the convergence result in [16] under weaker assumptions, only the equiboundedness estimates of the solutions of (3.24) being enough by means of the half-relaxed limit method.

4. Study of related equations. In this section we focus our attention on Hamilton–Jacobi equations of the form

$$\begin{cases} \dfrac{\partial w}{\partial t} + \langle \Sigma(x,t)Dw, Dw\rangle + G(x,t,Dw,D^2w) = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ w(x,0) = \psi(x) & \text{in } \mathbb{R}^N, \end{cases} \tag{4.1}$$

where $\Sigma(x,t) \in \mathcal{M}_N(\mathbb{R})$ and G is given by (1.3) and

$$\begin{cases} \dfrac{\partial w}{\partial t} + h(x)|Dw|^2 = 0 & \text{in } \mathbb{R}^N \times (0,T), \\ w(x,0) = \psi(x) & \text{in } \mathbb{R}^N, \end{cases} \tag{4.2}$$

where $h : \mathbb{R}^N \to \mathbb{R}$. Our aim is to investigate comparison (and existence) results for (4.1) and (4.2) under assumptions which include Hamiltonians like (2.6) in Remark 2.2(iii) (see Remark 4.1 for further comments). More precisely, we introduce two new assumptions:

(A4) $\Sigma \in C(\mathbb{R}^N \times [0,T]; \mathcal{S}_N^+(\mathbb{R}))$ and for all $x \in \mathbb{R}^N, t \in [0,T]$,

$$0 < \Sigma(x,t) \quad \text{and} \quad |\Sigma(x,t)| \le \bar{C}.$$

(A5) $h \in C(\mathbb{R}^N;\mathbb{R})$, $h \in W^{2,\infty}(\tilde{\Gamma})$, where $\tilde{\Gamma}$ is an open neighborhood of $\Gamma := \{x \in \mathbb{R}^N : h(x) = 0\}$ and for all $x \in \Gamma$,

$$Dh(x) = 0.$$

THEOREM 4.1. *Assume* (A2), (A3) *and* (A4) (*respectively,* (A3) *and* (A5)). *Let* $U \in USC(\mathbb{R}^N \times [0,T])$ *be a viscosity subsolution of* (4.1) (*respectively,* (4.2)) *and* $V \in LSC(\mathbb{R}^N \times [0,T])$ *be a viscosity supersolution of* (4.1) (*respectively,* (4.2)) *satisfying the quadratic growth condition* (2.4). *Then* $U \leq V$ *in* $\mathbb{R}^N \times [0,T]$.

The question of existence faces the same problems as in section 2. We have existence and uniqueness of a continuous viscosity solution for (4.1) and (4.2) in the class of functions with quadratic growth at least for short time (as in Corollary 2.1). But solutions can blow up in finite time.

Before giving the proof of the theorem, we give some comments on the equations and the assumptions.

Remark 4.1. (i) In the same way as in Remark 2.2, the above results hold replacing Σ by $-\Sigma$ in (4.1) and h by $-h$ in (4.2) and when dealing with terminal data in both equations.

(ii) Coming back to Remark 2.2(iii), we note $(a_1+a_2)(a_2-a_1)^T$ is not necessarily symmetric in (2.6), whereas we assume Σ to be symmetric in (A4) (and h is real-valued and therefore symmetric in (A5)). This is not a restriction of generality since comparison for (4.1) with Σ symmetric implies obviously a comparison for (4.1) for any matrix Σ, using that, for any $\Sigma \in \mathcal{M}_N(\mathbb{R})$, $(\Sigma + \Sigma^T)/2 \in \mathcal{S}_N(\mathbb{R})$.

(iii) Coming back to Remark 2.2(iv) again, we see that (A4) corresponds to the case where the convex Hamiltonian is predominant with respect to the concave one in (1.1) with (2.6). Assumption (A5) corresponds to (2.6) when a_1 and a_2 are real valued but h is allowed to change its sign. For example, we have comparison for

$$w_t + \phi(x)^3 |Dw|^2 = 0,$$

where $\phi \in C^2(\mathbb{R}^N;\mathbb{R})$ is any bounded function. Let us compare (1.12) with first-order Hamilton–Jacobi equations whose Hamilonians are Lipschitz continuous both in the state and gradient variables, or, to make it simple, with the Eikonal equation

$$\frac{\partial w}{\partial t} + a(x)|Dw| = 0 \quad \text{in } \mathbb{R}^N \times (0,T),$$

where a is Lipschitz continuous function. We know [13], [30] we have existence and uniqueness of a continuous viscosity solution for any continuous initial data without any restriction on the growth. In this case, the sign of a does not play any role, whereas it seems to be the case for the sign of h in (1.12). We would like to know if comparison for (1.12) is true under weaker assumptions than (A5) and in dimension $N > 1$ (i.e., when Σ is neither positive nor negative definite in (4.1)).

(iv) There are some links between Theorem 2.1 and Theorem 4.1. Nevertheless we point out that if (1.1) under assumptions of section 2 is naturally associated with a control problem, this is not necessarily the case under the assumptions of the current section. To be more precise, let us compare (1.1) with H given by (2.3) and (4.1) under (A4). The matrix a can be singular in (2.3), whereas we impose the nondegeneracy condition $\Sigma > 0$ in (4.1). The counterpart is that the regularity assumption with respect to x on Σ is weaker (Σ is supposed to be merely continuous) than the locally Lipschitz regularity we assume for a. Therefore (4.1) does not enter in the framework of section 2 in general. A natural consequence is that proofs differ: in both proofs of Theorem 2.1 and 4.1, the main argument is a kind of linearization procedure, but while in the proof of Theorem 2.1 we use essentially the convexity (or concavity) of the operator corresponding to the unbounded control set, here we use the locally Lipschitz continuity of the Hamiltonian with respect to the gradient uniformly in the state variable.

(v) Note that we are able to deal with a second-order term in (4.1) but not in (4.2).

Proof of Theorem 4.1. We divide the proof into two parts, corresponding, respectively, to the case of (4.1) and (4.2). The main argument in both proofs is a kind of linearization procedure as the one of Lemma 2.2. The rest of the proof is close to the one of Theorem 2.1.

Part 1. We assume (A2), (A3), and (A4).

The estimates of G are exactly the same than in Lemma 2.2 so, for sake of simplicity, we choose to take $G \equiv 0$. The linearization procedure is the subject of the following lemma, whose proof is postponed to the end of the section.

LEMMA 4.1. *Let* $0 < \lambda < 1$ *and set* $\Psi = \lambda U - V$. *Then* Ψ *is a USC viscosity subsolution of*

$$\begin{cases} \dfrac{\partial w}{\partial t} - \dfrac{2\bar{C}}{1-\lambda}|Dw|^2 \le 0 & \text{in } \mathbb{R}^N \times (0,T), \\ w(x,0) \le (\lambda - 1)\psi(x) & \text{in } \mathbb{R}^N. \end{cases} \tag{4.3}$$

If $\psi \ge 0$, then it follows $\Psi(x,0) \le 0$ in $\mathbb{R}^N$ and using the above lemma with the same supersolution as in the proof of Theorem 2.1, we obtain $\Psi \le 0$ in $\mathbb{R}^N \times [0,T]$. Note that the boundedness of Σ (see (A4)) is crucial to build the supersolution. In the particular case when $G = 0$, we can also choose a simpler supersolution such as, for instance,

$$(x,t) \mapsto K\frac{[(|x| - R)^+]^2}{1 - Lt} + \eta t, \quad K, L, R, \eta > 0. \tag{4.4}$$

Letting λ go to 1, we conclude that $U \le V$ in $\mathbb{R}^N \times [0,T]$.

It remains to prove that we can assume $\psi \ge 0$ without loss of generality. To this end we use an argument similar to the one of Step 2 in the proof of Theorem 2.1: let $\bar{U} = U + M(1+|x|^2)\mathrm{e}^{\rho t}$, $\bar{V} = V + M(1+|x|^2)\mathrm{e}^{\rho t}$, for some positive constants M and ρ and set $\bar{\Psi} = \lambda\bar{U} - \bar{V}$. We can easily prove that for $\rho > 16M^2\bar{C}\mathrm{e}$, $\bar{\Psi}$ is a subsolution of (4.3) (with a larger constant $4\bar{C}$ instead of $2\bar{C}$) in $\mathbb{R}^N \times (0,\tau]$ for $\tau = 1/\rho$. Moreover, $\bar{\Psi}(x,0) \le (\lambda - 1)(\psi(x) + M(1+|x|^2)\mathrm{e}^{\rho t}) \le 0$ for M sufficiently large (since ψ has quadratic growth). Letting λ go to 1, it follows that $U \le V$ in $\mathbb{R}^N \times [0,\tau]$ and we conclude by a step-by-step argument.

Part 2. We assume (A3) and (A5).

Set $\Omega^+ = \{x \in \mathbb{R}^N : h(x) > 0\}$ and $\Omega^- = \{x \in \mathbb{R}^N : h(x) < 0\}$ which are open subsets of $\mathbb{R}^N$. Define $\bar{\Psi} = \lambda\bar{U} - \bar{V}$ for $0 < \lambda < 1$, $\bar{U} = U + M(1+|x|^2)\mathrm{e}^{\rho t}$ and $\bar{V} = V + M(1+|x|^2)\mathrm{e}^{\rho t}$ with M larger than the constants which appear in (A3) and in the growth condition (2.4). Arguing as in Lemma 4.1 and noticing that all arguments are local, we obtain that $\bar{\Psi}$ is a subsolution in Ω^+ of the same kind of equation as (4.3), namely,

$$\begin{cases} \dfrac{\partial w}{\partial t} - \dfrac{C}{1-\lambda}|Dw|^2 \le 0 & \text{in } \Omega^+ \times (0,\tau), \\ w(x,0) \le 0 & \text{in } \Omega^+, \end{cases} \tag{4.5}$$

for some constant $C = 16\,||h||_\infty$, $\rho > 4MC\mathrm{e}$, and $\tau = 1/\rho > 0$ which are independent of λ. Note that $\bar{\Psi}(\cdot,0) \le 0$ in $\mathbb{R}^N$ because of the choice of M.

The sign of $\bar{\Psi}$ on Γ is the subject of the following lemma the proof of which uses (A5) and is postponed.

LEMMA 4.2. *For all* $x \in \Gamma,\ t \in [0,T]$, *we have* $U(x,t) \le \psi(x) \le V(x,t)$.

Noticing that $\{\Gamma, \Omega^+, \Omega^-\}$ is a partition of $\mathbb{R}^N$, we set

$$\hat{\Psi} := \begin{cases} \sup\{\bar{\Psi}, 0\} & \text{in } (\Gamma \cup \Omega^+) \times [0,T], \\ 0 & \text{in } \Omega^- \times [0,T]. \end{cases}$$

From the lemma, $\bar{\Psi} \le 0$ in $\Gamma \times [0,T]$. Therefore the function $\hat{\Psi}$ is continuous in $\mathbb{R}^N \times (0,T)$. Moreover, we claim that $\hat{\Psi}$ is a subsolution of (4.5) in $\mathbb{R}^N \times (0,\tau)$. This claim is clearly true in $\Omega^- \times (0,T)$ (since 0 is clearly a subsolution) and in $\Omega^+ \times (0,\tau)$ (since $\hat{\Psi}$ is the supremum of two subsolutions). It remains to prove the result on Γ. Let $\varphi \in C^1(\mathbb{R}^N \times (0,\tau))$ such that $\hat{\Psi} - \varphi$ achieves a local maximum at a point $(\bar{x},\bar{t}) \in \Gamma \times (0,\tau)$. Let (x,t) be in a neighborhood of $(\bar{x},\bar{t})$. If $x \in \Gamma \cup \Omega^+$, then $0 \le \hat{\Psi}(x,t)$, and if $x \in \Omega^-$, then $0 = \hat{\Psi}(x,t)$. In any case, $(0-\varphi)(x,t) \le (\hat{\Psi}-\varphi)(x,t) \le (\hat{\Psi}-\varphi)(\bar{x},\bar{t})$. But $\hat{\Psi}(\bar{x},\bar{t}) \le 0$ since $(\bar{x},\bar{t}) \in \Gamma \times (0,\tau)$. Therefore $(\bar{x},\bar{t})$ is a local maximum of $0-\varphi$ which ends the proof of the claim. The initial condition $\hat{\Psi} \le 0$ in $\mathbb{R}^N \times \{0\}$ is trivially satisfied. From the comparison principle proved in Part 1 (for example, using (4.4) as a supersolution), we obtain $\hat{\Psi} \le 0$ in $\mathbb{R}^N \times [0,\tau]$. Since τ does not depend on λ, we can send λ to 1. We obtain $U \le V$ in $(\Gamma \cup \Omega^+) \times [0,\tau]$. We conclude in $(\Gamma \cup \Omega^+) \times [0,T]$ by a step-by-step procedure.

Repeating the same kind of arguments replacing Ω^+ by Ω^- and $\bar{\Psi}$ by $U - M(1+|x|^2)\mathrm{e}^{\rho t} - \mu(V - M(1+|x|^2)\mathrm{e}^{\rho t})$, $0 < \mu < 1$, we obtain that $U - V \le 0$ in $(\Gamma \cup \Omega^-) \times [0,T]$, which ends the proof. □

Proof of Lemma 4.1. We proceed as in the proof of Lemma 2.2. We can assume that $G \equiv 0$ since the computations with G are exactly the same than in Lemma 2.2.

Note first that λU is an USC subsolution of

$$\text{(4.6)} \qquad \begin{cases} \dfrac{\partial w}{\partial t} + \dfrac{1}{\lambda}\langle \Sigma(x,t) Dw, Dw\rangle \le 0 & \text{in } \mathbb{R}^N \times (0,T), \\ w(x,0) \le \lambda\psi(x) & \text{in } \mathbb{R}^N. \end{cases}$$

Let $\varphi \in C(\mathbb{R}^N \times [0,T])$ and suppose that $\Psi - \varphi$ reaches a strict local maximum at $(\bar{x},\bar{t}) \in \mathbb{R}^N \times (0,T]$ in some compact subset $K \subset \mathbb{R}^N \times (0,T]$. We have

$$\max_{(x,t)(y,t)\in K} \left\{ \lambda U(x,t) - V(y,t) - \varphi\left(\frac{x+y}{2}, t\right) - \frac{|x-y|^2}{\varepsilon^2} \right\} \underset{\varepsilon \downarrow 0}{\longrightarrow} \Psi(\bar{x},\bar{t}) - \varphi(\bar{x},\bar{t}).$$

The above maximum is achieved at some point $(x_\varepsilon, y_\varepsilon, t_\varepsilon)$. Writing the viscosity inequalities and subtracting them, we obtain

$$\begin{aligned} &\varphi_t\left(\frac{x_\varepsilon + y_\varepsilon}{2}, t_\varepsilon\right) + \frac{1}{\lambda}\left\langle \Sigma(x_\varepsilon, t_\varepsilon)\left(\frac{1}{2} D\varphi\left(\frac{x_\varepsilon + y_\varepsilon}{2}, t_\varepsilon\right) + 2\frac{x_\varepsilon - y_\varepsilon}{\varepsilon^2}\right),\right. \\ &\left. \frac{1}{2} D\varphi\left(\frac{x_\varepsilon + y_\varepsilon}{2}, t_\varepsilon\right) + 2\frac{x_\varepsilon - y_\varepsilon}{\varepsilon^2}\right\rangle \\ &- \left\langle \Sigma(y_\varepsilon, t_\varepsilon)\left(-\frac{1}{2} D\varphi\left(\frac{x_\varepsilon + y_\varepsilon}{2}, t_\varepsilon\right) + 2\frac{x_\varepsilon - y_\varepsilon}{\varepsilon^2}\right),\right. \\ &\left. -\frac{1}{2} D\varphi\left(\frac{x_\varepsilon + y_\varepsilon}{2}, t_\varepsilon\right) + 2\frac{x_\varepsilon - y_\varepsilon}{\varepsilon^2}\right\rangle \le 0. \end{aligned}$$

In what follows, we omit writing the dependence in t_ε and the point $((x_\varepsilon + y_\varepsilon)/2, t_\varepsilon)$ in the derivatives of φ. Set $p_\varepsilon = 2(x_\varepsilon - y_\varepsilon)/\varepsilon^2$, $p_x = D\varphi/2 + p_\varepsilon$, and $p_y = -D\varphi/2 + p_\varepsilon$.

We have

$$\begin{aligned}0 &\geq \varphi_t + \frac{1}{\lambda}\langle\Sigma(x_\varepsilon)p_x, p_x\rangle - \langle\Sigma(y_\varepsilon)p_y, p_y\rangle\\ &\geq \varphi_t + \left(\frac{1}{\lambda} - 1\right)\langle\Sigma(x_\varepsilon)p_x, p_x\rangle + \langle(\Sigma(x_\varepsilon) - \Sigma(y_\varepsilon))p_x, p_x\rangle + \langle\Sigma(y_\varepsilon)p_x, p_x\rangle\\ &\quad -\langle\Sigma(y_\varepsilon)p_y, p_y\rangle\\ &\geq \varphi_t + \left(\frac{1}{\lambda} - 1\right)\langle\Sigma(x_\varepsilon)p_x, p_x\rangle - m_K(|x_\varepsilon - y_\varepsilon|)|p_x|^2 + \langle\Sigma(y_\varepsilon)p_x, p_x\rangle - \langle\Sigma(y_\varepsilon)p_y, p_y\rangle,\end{aligned}$$

where m_K is a modulus of continuity for Σ in the compact subset K. Since $\Sigma(y_\varepsilon)$ is a symmetric matrix, we have

$$\begin{aligned}\langle\Sigma(y_\varepsilon)p_x, p_x\rangle - \langle\Sigma(y_\varepsilon)p_y, p_y\rangle &= \langle\Sigma(y_\varepsilon)(p_x + p_y), p_x - p_y\rangle = 2\,\langle\Sigma(y_\varepsilon)p_\varepsilon, D\varphi\rangle\\ &= -\langle\Sigma(y_\varepsilon)D\varphi, D\varphi\rangle + 2\,\langle\Sigma(y_\varepsilon)p_x, D\varphi\rangle.\end{aligned}$$

For any $\alpha > 0$, denoting by $\sqrt{\Sigma(y_\varepsilon)}$ the positive symmetric squareroot of the positive symmetric matrix $\Sigma(y_\varepsilon)$, we get

$$\begin{aligned}2\langle\Sigma(y_\varepsilon)p_x, D\varphi\rangle = 2\langle\sqrt{\Sigma(y_\varepsilon)}p_x, \sqrt{\Sigma(y_\varepsilon)}D\varphi\rangle &\leq \alpha\langle\Sigma(y_\varepsilon)D\varphi, D\varphi\rangle + \frac{1}{\alpha}\langle\Sigma(y_\varepsilon)p_x, p_x\rangle\\ &\leq \alpha\langle\Sigma(y_\varepsilon)D\varphi, D\varphi\rangle + \frac{1}{\alpha}\langle\Sigma(x_\varepsilon)p_x, p_x\rangle + \frac{1}{\alpha}\langle m_K(|x_\varepsilon - y_\varepsilon|)p_x, p_x\rangle.\end{aligned}$$

It follows

$$\begin{aligned}0 \geq \varphi_t + \left\langle\left[\left(\frac{1}{\lambda} - 1 - \frac{1}{\alpha}\right)\Sigma(x_\varepsilon) - \left(1 + \frac{1}{\alpha}\right)m_K(|x_\varepsilon - y_\varepsilon|)Id\right]p_x, p_x\right\rangle\\ - (1+\alpha)\langle\Sigma(y_\varepsilon)D\varphi, D\varphi\rangle.\end{aligned}$$

Take $\alpha = \frac{2\lambda}{(1-\lambda)} > 0$ in order to have $\frac{1}{\lambda} - 1 - \frac{1}{\alpha} = \frac{1}{2}(\frac{1}{\lambda} - 1) > 0$. We recall that $\Sigma > 0$ by (A4); hence for ε small enough the above scalar product is nonnegative. Thus

$$0 \geq \varphi_t - (1+\alpha)\langle\Sigma(y_\varepsilon, t_\varepsilon)D\varphi, D\varphi\rangle = \varphi_t - \frac{1+\lambda}{1-\lambda}\langle\Sigma(y_\varepsilon, t_\varepsilon)D\varphi, D\varphi\rangle.$$

Letting ε go to 0, we get

$$0 \geq \varphi_t(\bar{x}, \bar{t}) - \frac{2}{1-\lambda}\langle\Sigma(\bar{x}, \bar{t})D\varphi(\bar{x}, \bar{t}), D\varphi(\bar{x}, \bar{t})\rangle,$$

which proves the result. □

Proof of Lemma 4.2. We make the proof for U, the second inequality being similar. Let $x_0 \in \Gamma$ and consider, for $\eta > 0$,

$$\sup_{t\in[0,T]}\{U(x_0, t) - \eta t\}.$$

This supremum is achieved at a point $t_0 \in [0, T]$ and we can assume, up to subtract $|t - t_0|^2$, that it is a strict local maximum. Consider, for $\varepsilon > 0$,

$$\sup_{\bar{B}(x_0,1)\times[0,T]}\left\{U(x, t) - \frac{|x - x_0|^2}{\varepsilon^2} - \eta t\right\}.$$

This supremum is achieved at a point $(x_\varepsilon, t_\varepsilon)$ and it is easy to see that $(x_\varepsilon, t_\varepsilon) \to (x_0, t_0)$ when $\varepsilon \to 0$.

Suppose that $t_\varepsilon > 0$. It allows us to write the viscosity inequality for the subsolution U at $(x_\varepsilon, t_\varepsilon)$ and we get

$$\eta + h(x_\varepsilon)\left|2\frac{x_\varepsilon - x_0}{\varepsilon^2}\right|^2 \leq 0. \tag{4.7}$$

Since $h \in W^{2,\infty}$ in a neighborhood of Γ, we can write a Taylor expansion of h at x_0 for ε small enough

$$h(x_\varepsilon) = h(x_0) + \langle Dh(x_0), x_\varepsilon - x_0\rangle + \frac{1}{2}\langle D^2h(x_0)(x_\varepsilon - x_0), x_\varepsilon - x_0\rangle + o(|x_\varepsilon - x_0|^2).$$

From (4.7) and (A5), it follows

$$\eta - (2C + m(|x_\varepsilon - x_0|^2))\left(\frac{|x_\varepsilon - x_0|^2}{\varepsilon^2}\right)^2 \leq 0,$$

where m is a modulus of continuity. Since $|x_\varepsilon - x_0|^2/\varepsilon^2 \to 0$ as $\varepsilon \to 0$, we obtain a contradiction for small ε.

Therefore $t_\varepsilon = 0$ for ε small enough. It follows that for all $(x,t) \in \bar{B}(x_0,1)\times[0,T]$, we have

$$U(x,t) - \frac{|x - x_0|^2}{\varepsilon^2} - \eta t \leq U(x_\varepsilon, 0) - \frac{|x_\varepsilon - x_0|^2}{\varepsilon^2} \leq \psi(x_\varepsilon).$$

Setting $x = x_0$ and sending ε and then η to 0, we obtain the conclusion. □

Acknowledgment. We are grateful to Martino Bardi and Guy Barles for valuable suggestions during the preparation of this work.

REFERENCES

[1] O. Alvarez, *A quasilinear elliptic equation in* $\mathbf{R}^N$, Proc. Roy. Soc. Edinburgh Sect. A, 126 (1996), pp. 911–921.

[2] O. Alvarez, *Bounded-from-below viscosity solutions of Hamilton-Jacobi equations*, Differential Integral Equations, 10 (1997), pp. 419–436.

[3] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[4] M. Bardi and F. Da Lio, *On the Bellman equation for some unbounded control problems*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 491–510.

[5] G. Barles, *An approach of deterministic control problems with unbounded data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 235–258.

[6] G. Barles, *Solutions de Viscosité des Équations de Hamilton-Jacobi*, Springer-Verlag, Paris, 1994.

[7] G. Barles, S. Biton, M. Bourgoing, and O. Ley, *Uniqueness results for quasilinear parabolic equations through viscosity solutions' methods*, Calc. Var. Partial Differential Equations, 18 (2003), pp. 159–179.

[8] G. Barles, J. Burdeau, M. Romano, and N. Samsoen, *Critical stock price near expiration*, Math. Finance, 5 (1995), pp. 77–95.

[9] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[10] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, Stud. Math. Appl. 11, North–Holland, Amsterdam, 1982.

[11] F. E. Benth and K. H. Karlsen, *A PDE representation of the density of the minimal entropy martingale measure in stochastic volatility markets*, Stoch. Internat. J. Prob. Stoch. Proc., 77 (2005), pp. 109–137.

[12] P. Cannarsa and G. Da Prato, *Nonlinear optimal control with infinite horizon for distributed parameter systems and stationary Hamilton-Jacobi equations*, SIAM J. Control Optim., 27 (1989), pp. 861–875.
[13] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.
[14] M. G. Crandall and P.-L. Lions, *Quadratic growth of solutions of fully nonlinear second order equations in* $\mathbb{R}^n$, Differential Integral Equations, 3 (1990), pp. 601–616.
[15] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[16] F. Da Lio and W. M. McEneaney, *Finite time-horizon risk-sensitive control and the robust limit under a quadratic growth assumption*, SIAM J. Control Optim., 40 (2002), pp. 1628–1661 (electronic).
[17] L. C. Evans, *Partial Differential Equations*, vol. 19, AMS, Providence, RI, 1998.
[18] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Appl. Math. 1, Springer-Verlag, Berlin, 1975.
[19] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[20] Y. Giga, S. Goto, H. Ishii, and M.-H. Sato, *Comparison principle and convexity preserving properties for singular degenerate parabolic equations on unbounded domains*, Indiana Univ. Math. J., 40 (1991), pp. 443–470.
[21] H. Ishii, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.
[22] H. Ishii, *Comparison results for Hamilton-Jacobi equations without growth condition on solutions from above*, Appl. Anal., 67 (1997), pp. 357–372.
[23] K. Ito, *Existence of solutions to Hamilton-Jacobi-Bellman equation under quadratic growth conditions*, J. Differential Equations, 176 (2001), pp. 1–28.
[24] F. John, *Partial Differential Equations*, 4th ed., Springer-Verlag, New York, 1991.
[25] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, Graduate Texts in Math. 113, Springer-Verlag, New York, 1991.
[26] M. Kobylanski, *Backward stochastic differential equations and partial differential equations with quadratic growth*, Ann. Probab., 28 (2000), pp. 558–602.
[27] N. V. Krylov, *Controlled Diffusion Processes*, Appl. Math. 14, Springer-Verlag, New York, 1980.
[28] N. V. Krylov, *Stochastic linear controlled systems with quadratic cost revisited*, in *Stochastics in finite and infinite dimensions*, Trends Math., pp. 207–232. Birkhäuser, Boston, 2001.
[29] D. Lamberton and B. Lapeyre, *Introduction au Calcul Stochastique Appliqué à la Finance*, 2nd ed., Ellipses Édition Marketing, Paris, 1997.
[30] O. Ley, *Lower-bound gradient estimates for first-order Hamilton-Jacobi equations and applications to the regularity of propagating fronts*, Adv. Differential Equations, 6 (2001), pp. 547–576.
[31] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations.* I. *The dynamic programming principle and applications*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174.
[32] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations.* II. *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.
[33] P.-L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations.* III. *Regularity of the optimal cost function*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France seminar, Vol. V (Paris, 1981/1982), Res. Notes Math. 93, Pitman, Boston, 1983, pp. 95–205.
[34] W. M. McEneaney, *Robust control and differential games on a finite time horizon*, Math. Control Signals Systems, 8 (1995), pp. 138–166.
[35] W. M. McEneaney, *Uniqueness for viscosity solutions of nonstationary Hamilton-Jacobi-Bellman equations under some a priori conditions (with applications)*, SIAM J. Control Optim., 33 (1995), pp. 1560–1576.
[36] W. M. McEneaney, *A uniqueness result for the Isaacs equation corresponding to nonlinear* H_∞ *control*, Math. Control Signals Systems, 11 (1998), pp. 303–334.
[37] H. Nagai, *Bellman equations of risk-sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.
[38] B. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer-Verlag, Berlin, 1998.
[39] H. Pham, *Smooth solutions to optimal investment models with stochastic volatilities and portfolio constraints*, Appl. Math. Optim., 46 (2002), pp. 55–78.

[40] F. Rampazzo and C. Sartori, *Hamilton-Jacobi-Bellman equations with fast gradient-dependence*, Indiana Univ. Math. J., 49 (2000), pp. 1043–1077.
[41] J. Yong and X. Y. Zhou, *Stochastic Controls*, in Hamiltonian Systems and HJB Equations, Appl. Math. 43, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 107–122

MINIMUM DISTANCE TO THE RANGE OF A BANDED LOWER TRIANGULAR TOEPLITZ OPERATOR IN ℓ^1 AND APPLICATION IN ℓ^1-OPTIMAL CONTROL*

ZDENĚK HURÁK†, ALBRECHT BÖTTCHER‡, AND MICHAEL ŠEBEK§

Abstract. The subject of the paper is the best approximation in ℓ^1 of a given infinite sequence by sequences in the range of a given banded lower-triangular Toeplitz operator. Necessary and sufficient conditions for the existence of a minimizing solution are established and a numerical algorithm for finding such a solution is designed and theoretically founded. It is also shown that an optimal error sequence is only finitely nonzero. The relevancy of the problem in systems theory is outlined and numerical examples are presented.

Key words. Toeplitz operator, banded, lower-triangular, range, minimum distance problem, ℓ^1-optimal control

AMS subject classifications. 47B35, 90C46, 90C90

DOI. 10.1137/S0363012903437940

1. Introduction. This paper concerns the problem of finding the distance between a given sequence in ℓ^1 and the range of a given infinite lower-triangular Toeplitz band matrix on ℓ^1. Our research was motivated by recent attempts [7], [8], [9], [15], [17], [18] to solve the standard ℓ^1-optimal control problem [24] without using interpolation. The approach of the present paper follows the line of reasoning pursued by Dahleh and Pearson in their seminal paper [10], which launched a hunt for reliable procedures for ℓ^1-optimal control design, but the problem is posed in a different setting, which leads to a different numerical algorithm. Using new theoretical tools, we establish (the already known) necessary and sufficient conditions for the solvability of the problem and design a new numerical algorithm for finding a solution. The convergence of this algorithm is rigorously proved. In essence, our approach shares the spirit of polynomial methods with Casavola's approach [7], [8], [9] in that we formulate the problem using equations with polynomials and power series. In contrast to previous work, which used interpolation techniques (see [11] for a detailed exposition) relying on the solution of numerically ill-conditioned Vandermonde systems, our algorithm is based on finding optimal solutions of overdetermined but numerically better behaved Toeplitz systems. Compared to Casavola's polynomial approach, ours does not introduce any new *structural* parameter (a term coined by Casavola) for the optimization and the Diophantine equation $AX + Y = B$ in the ring of *stable* functions is attacked directly using powerful results from Toeplitz operator theory. The rigorousness of this paper might perhaps seem unnecessary for control of single-input-single-output (SISO) plants, because it is well known that the design of an ℓ^1-optimal controller

*Received by the editors November 21, 2003; accepted for publication (in revised form) September 27, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sicon/45-1/43794.html

†Center for Applied Cybernetics, Czech Technical University, Karlovo náměstí 13/G, 12135 Prague, Czech Republic (z.hurak@c-a-k.cz). The work of this author was supported by the Ministry of Education of the Czech Republic under project 1M6840770004.

‡Fakultät für Mathematik, Technische Universität Chemnitz, 09107 Chemnitz, Germany (aboettch@mathematik.tu-chemnitz.de).

§Department of Control Engineering, Czech Technical University, Karlovo náměstí 13/E, 12135 Prague, Czech Republic (m.sebek@control.felk.cvut.cz).

simply consists in designing a dead-beat controller minimizing the sum of the absolute values of the closed-loop impulse response and that hence not much space is left for innovations. But the theoretical rigorousness pays back in that the extension to the multiple-input-multiple-output (MIMO) case is straightforward; preliminary achievements have recently been presented by the authors in [16].

2. Statement of the problem. We start with some standard definitions [4], [5], [6]. The Wiener algebra W is the Banach algebra of all complex-valued functions on the complex unit circle $\mathbf{T}$ with absolutely convergent Fourier series. Thus, a function $a : \mathbf{T} \to \mathbf{C}$ belongs to W if and only if

$$a(t) = \sum_{j=-\infty}^{\infty} a_j t^j \quad (t = e^{i\theta} \in \mathbf{T}), \quad \|a\|_W := \sum_{j=-\infty}^{\infty} |a_j| < \infty.$$

Wiener's theorem states that if $a \in W$ has no zeros on $\mathbf{T}$, then $1/a$ is also in W. For $a \in W$, the infinite Toeplitz matrix $T(a)$ and the finite Toeplitz matrices $T_k(a)$ are defined by $T(a) = (a_{i-j})_{i,j=1}^{\infty}$ and $T_k(a) = (a_{i-j})_{i,j=1}^{k}$, respectively.

We denote by c_0 the set of all real-valued sequences $x = \{x_j\}_{j=1}^{\infty}$ with $|x_j| \to 0$ as $j \to \infty$ and by ℓ^1 the set of all real-valued sequences $x = \{x_j\}_{j=1}^{\infty}$ satisfying $\sum_{j=1}^{\infty} |x_j| < \infty$. The sets c_0 and ℓ^1 are real Banach spaces under the norms $\|x\|_\infty = \sup_{j\geq 1} |x_j|$ and $\|x\|_1 = \sum_{j=1}^{\infty} |x_j|$, respectively. Moreover, ℓ^1 is the dual space of c_0, $\ell^1 = c_0^*$, under the pairing $\langle z, b\rangle = \sum_{j=1}^{\infty} z_j b_j$, $\{z_j\} \in c_0$, $\{b_j\} \in \ell^1$.

Given a Banach space X, we let $\mathcal{B}(X)$ stand for the Banach algebra of all bounded linear operators on X. For $A \in \mathcal{B}(X)$, the norm $\|A\|_{\mathcal{B}(X)}$ is $\sup \|Ax\|$, the supremum over all $x \in X$ with $\|x\| \leq 1$, and the range and null space of A are defined by $\mathcal{R}(A) = A(X)$ and $N(A) = \{x \in X : Ax = 0\}$.

Let $a \in W$ and suppose the Fourier coefficients of a are all real. Then the infinite Toeplitz matrix $T(a)$ induces a bounded linear operator on c_0 and ℓ^1 via

$$(T(a)x)_i = \sum_{j=1}^{\infty} a_{i-j} x_j \quad (j \geq 1).$$

For $j \in \mathbf{Z}$, define the function χ_j by $\chi_j(t) = t^j$ $(t \in \mathbf{T})$. The operators $T(\chi_1)$ and $T(\chi_{-1})$ are the forward and backward shifts acting by the rules

$$T(\chi_1) : \{x_1, x_2, x_3, \ldots\} \mapsto \{0, x_1, x_2, \ldots\},$$
$$T(\chi_{-1}) : \{x_1, x_2, x_3, \ldots\} \mapsto \{x_2, x_3, x_4, \ldots\}.$$

Clearly, $T(a) = \sum_{j=-\infty}^{\infty} a_j T(\chi_j)$. This implies that $\|T(a)\|_{\mathcal{B}(c_0)} = \|T(a)\|_{\mathcal{B}(\ell^1)} = \|a\|_W$. The adjoint operator of $T(a) : c_0 \to c_0$ is the operator $T(\overline{a}) : \ell^1 \to \ell^1$, where $\overline{a}(t) = \sum_{j=-\infty}^{\infty} a_j t^{-j}$ $(t = e^{i\theta} \in \mathbf{T})$.

Throughout this paper we suppose that $a_+(t) = a_0 + a_1 t + \cdots + a_n t^n$ with real numbers $a_0, a_1, \ldots, a_n$ and with $a_n \neq 0$. Clearly, $T(a_+)$ is a banded lower-triangular Toeplitz matrix, while $T(\overline{a}_+)$ is a banded upper-triangular Toeplitz matrix. We always think of $T(a_+)$ as acting on ℓ^1 and always consider $T(\overline{a}_+)$ as an operator on c_0. Thus, $T(a_+)$ is the adjoint of $T(\overline{a}_+)$.

This paper concerns the following problem of finding the distance between some given real sequence with finite sum of absolute values and the range of a banded lower triangular Toeplitz operator in ℓ^1 space. Given $b \in \ell^1$, determine the distance

$$d := \operatorname{dist}_{\ell^1}(b, \mathcal{R}(T(a_+))) := \inf_{m \in \mathcal{R}(T(a_+))} \|b - m\|_1,$$

find out whether there is an $m_0 \in \mathcal{R}(T(a_+))$ with $\|b - m_0\|_1 = d$, and if yes, compute such an m_0. Note that once m_0 is available, we can easily solve the lower-triangular system $T(a_+)x_0 = m_0$ to get x_0.

3. Two results from functional analysis. We will employ the following two theorems (whose proofs can be found in [21, pp. 121, 156]). Recall that the annihilator $M^\perp$ of a set $M \subset X$ is defined as $M^\perp = \{b \in X^* : \langle z, b\rangle = 0 \text{ for all } z \text{ in } M\}$. Furthermore, two elements $z \in X$ and $b \in X^*$ are said to be aligned if the equality $\|z\|\,\|b\| = \langle z, b\rangle$ holds.

THEOREM 3.1. *Let M be a linear subset of a real normed space X and let $b \in X^*$. Then*

$$\inf_{m \in M^\perp} \|b - m\| = \sup_{z \in M, \|z\| \le 1} \langle z, b\rangle. \tag{3.1}$$

The infimum in (3.1) *is always attained at some $m_0 \in M^\perp$. If the supremum in* (3.1) *is achieved for some $z_0 \in M$ with $\|z_0\| \le 1$, then z_0 and $b - m_0$ are aligned.*

THEOREM 3.2. *Let X be a Banach space and $A \in \mathcal{B}(X)$. Then $\mathcal{R}(A)$ is closed if and only if $\mathcal{R}(A^*)$ is closed, in which case $\mathcal{R}(A^*) = [N(A)]^\perp$.*

4. Toeplitz operators. The product of two Toeplitz operators is in general not a Toeplitz operator. However, this happens in certain special cases. Let W_+ and W_- denote the functions in W whose Fourier coefficients with negative and positive indices vanish, respectively. Thus, if $c_\pm \in W_\pm$, then $T(c_-)$ and $T(c_+)$ are upper and lower triangular, respectively. It is easily seen by direct inspection that if $c_- \in W_-$, $f \in W$, $c_+ \in W_+$, then

$$T(c_-)T(f)T(c_+) = T(c_- f c_+). \tag{4.1}$$

The following results are known to specialists (see, e.g., [4] and [12]). We include the proofs for the reader's convenience.

PROPOSITION 4.1. *The range $\mathcal{R}(T(a_+))$ is a closed subset of ℓ^1 if and only if a_+ has no zeros on $\mathbf{T}$.*

Proof. If a_+ has no zeros on $\mathbf{T}$, then a_+^{-1} belongs to W and has real Fourier coefficients. From (4.1) we obtain that $T(a_+^{-1})T(a_+) = I$. Thus, $T(a_+)$ has a bounded left inverse, which implies that the range of $T(a_+)$ is closed (see, e.g., [12, section I.1.2]).

Now suppose $a_+(\tau) = 0$ for some $\tau \in \mathbf{T}$. Contrary to what we want, we assume that $\mathcal{R}(T(a_+))$ is a closed subset of ℓ^1. We denote by $\ell^1(\mathbf{C})$ the complex Banach space of all complex-valued sequences $x = \{x_j\}_{j=1}^\infty$ for which $\|x\|_1 = \sum_{j=1}^\infty |x_j| < \infty$. The range of $T(a_+)$ on $\ell^1(\mathbf{C})$ is $\mathcal{R}(T(a_+)) + i\mathcal{R}(T(a_+))$, which is closed whenever $\mathcal{R}(T(a_+))$ is closed. From Theorem 3.2 we now infer that $T(\overline{a}_+) : c_0(\mathbf{C}) \to c_0(\mathbf{C})$ has closed range, where $c_0(\mathbf{C})$ is defined in analogy to $\ell^1(\mathbf{C})$. The operator $T(\overline{a}_+)$ is upper-triangular, and it is easily seen that the range of every nonzero upper-triangular Toeplitz operator contains all finitely supported sequences. Consequently, $T(\overline{a}_+)$ must be surjective. We may write

$$\begin{aligned}\overline{a}_+(t) = a(1/t) &= a_0 + a_1\frac{1}{t} + \cdots + a_n\frac{1}{t^n} \\ &= a_n\left(\frac{1}{t} - \tau\right)\left(\frac{1}{t} - z_1\right)\cdots\left(\frac{1}{t} - z_{n-1}\right) = a_n(\chi_{-1}(t) - \tau)d(t).\end{aligned}$$

Since $T(\overline{a}_+) = T(\chi_{-1}-\tau)T(d)$ by (4.1), the operator $T(\chi_{-1}-\tau)$ is surjective together with $T(\overline{a}_+)$. The equation $T(\chi_{-1}-\tau)z = 0$ is satisfied if and only if $z_j = \tau^{j-1}z_1$ ($j \geq 1$), and this is a sequence in $c_0(\mathbf{C})$ only for $z_1 = 0$. Thus, $T(\overline{a}_+)$ is injective on $c_0(\mathbf{C})$. In summary, $T(\chi_{-1}-\tau)$ is invertible on $c_0(\mathbf{C})$. It follows that $T(\chi_1 - 1/\tau)$ is invertible on $\ell^1(\mathbf{C})$. But the solution of $T(\chi_1 - 1/\tau)x = \{1, 0, 0, \ldots\}$ is $x_j = -\tau^j$ ($j \geq 1$), which is not in $\ell^1(\mathbf{C})$. This contradiction proves that $\mathcal{R}(T(a_+))$ cannot be closed. □

The function $a_+(z) = a_0 + a_1 z + \cdots + a_n z^n$ is defined for all $z \in \mathbf{C}$.

PROPOSITION 4.2. *If a_+ has no zeros on $\mathbf{T}$, then the dimension of $N(T(\overline{a}_+))$ in c_0 is equal to the number of zeros of a_+ in the open unit disk $\mathbf{D} := \{z \in \mathbf{C} : |z| < 1\}$.*

Proof. Let a_+ have $\varkappa$ zeros $\delta_1, \ldots, \delta_\varkappa$ in $\mathbf{D}$ and $n - \varkappa$ zeros $\mu_1, \ldots, \mu_{n-\varkappa}$ in $\mathbf{C} \setminus (\mathbf{D} \cup \mathbf{T})$. We then have

$$\begin{aligned}\overline{a}_+(t) = a_+(1/t) &= a_n \prod_{k=1}^{n-\varkappa} \left(\frac{1}{t} - \mu_k\right) \prod_{j=1}^{\varkappa} \left(\frac{1}{t} - \delta_j\right) \\ &= \gamma t^{-\varkappa} \prod_{k=1}^{n-\varkappa} \left(1 - \frac{1}{\mu_k t}\right) \prod_{j=1}^{\varkappa} (1 - \delta_j t)\end{aligned}$$

with $\gamma = a_n(-\mu_1)\ldots(-\mu_{n-\varkappa})$. We consider $T(\overline{a}_+)$ on $c_0(\mathbf{C})$. Let N be the null space of $T(\overline{a}_+)$ on c_0. Then $N + iN$ is the null space of $T(\overline{a}_+)$ on $c_0(\mathbf{C})$. From (4.1) we obtain that

$$T(\overline{a}_+) = \gamma T(\chi_{-\varkappa}) \prod_{k=1}^{n-\varkappa} \left(I - \frac{1}{\mu_k} T(\chi_{-1})\right) \prod_{j=1}^{\varkappa} (I - \delta_j T(\chi_1)).$$

Since $\|(1/\mu_k)T(\chi_{-1})\| = 1/|\mu_k| < 1$ and $\|\delta_j T(\chi_1)\| = |\delta_j| < 1$, we conclude that the operators $I - (1/\mu_k)T(\chi_{-1})$ and $I - \delta_j T(\chi_1)$ are all invertible. Consequently, the dimension of $N + iN$ is the dimension of the null space of $T(\chi_{-\varkappa})$ on $c_0(\mathbf{C})$. It follows that the dimension of $N + iN$ over $\mathbf{C}$ is $\varkappa$, which implies that the dimension of N over $\mathbf{R}$ is also $\varkappa$. □

5. Existence of the solution. Here is our result on the solvability of the problem posed in section 2. This result states that for some sequence b in ℓ^1 and a polynomial a_+, there is a real sequence in the range of the Toeplitz operator generated by a_+ that minimizes distance to b if and only if the polynomial has no zeros on the unit circle. The error sequence then has a finite number of nonzero terms only.

THEOREM 5.1. *The problem*

$$\|b - m\|_1 = \operatorname{dist}_{\ell^1}(b, \mathcal{R}(T(a_+))) =: d \tag{5.1}$$

has a solution $m_0 \in \mathcal{R}(T(a_+))$ for every $b \in \ell^1$ if and only if a_+ has no zeros on $\mathbf{T}$. If $a_+(t) \neq 0$ for $t \in \mathbf{T}$, then for every $b \in \ell^1$ there exists a $z_0 \in N(T(\overline{a}_+))$ such that

$$\|z_0\|_\infty \leq 1 \quad \text{and} \quad d = \langle z_0, b\rangle = \sup_{z \in N(T(\overline{a}_+)), \|z\|_\infty \leq 1} \langle z, b\rangle, \tag{5.2}$$

and if $m_0 \in \mathcal{R}(T(a_+))$ is any sequence satisfying (5.1)*, then the sequence $b - m_0$ has only finitely many nonzero terms.*

Proof. If a_+ has a zero on $\mathbf{T}$, then $\mathcal{R}(T(a_+))$ is not closed due to Proposition 4.1 and hence (5.1) has no solution $m_0 \in \mathcal{R}(T(a_+))$ if b is in the closure of $\mathcal{R}(T(a_+))$ but not in $\mathcal{R}(T(a_+))$.

Now suppose that a_+ has no zeros on $\mathbf{T}$. Then $\mathcal{R}(T(a_+))$ is closed by Proposition 4.1. From Theorem 3.2 we deduce that $\mathcal{R}(T(a_+)) = [N(T(\overline{a}_+))]^\perp$. The existence of an $m_0 \in \mathcal{R}(T(a_+))$ satisfying (5.1) then follows from Theorem 3.1. This theorem also yields the equality

$$d = \sup_{z \in N(T(\overline{a}_+)), \|z\|_\infty \le 1} \langle z, b \rangle,$$

and since $\{z \in N(T(\overline{a}_+)) : \|z\|_\infty \le 1\}$ is compact by virtue of Proposition 4.2 and the map $z \mapsto \langle z, b\rangle$ is continuous, we conclude that the supremum is attained at some $z_0 \in N(T(\overline{a}_+))$ with $\|z_0\|_\infty \le 1$.

The last assertion of the theorem is trivial for $d = 0$. So let $d > 0$, which implies that $\|z_0\|_\infty > 0$. The sequences $b - m_0$ and z_0 are aligned by Theorem 3.1. Consequently, with $b - m_0 = \{e_j\}_{j=1}^\infty$ and $z_0 = \{z_j\}_{j=1}^\infty$,

$$\sum_{j=1}^\infty z_j e_j = \|z_0\|_\infty \sum_{j=1}^\infty |e_j|. \tag{5.3}$$

As $\{z_j\} \in c_0$, there is a j_0 such that $|z_j| < \|z_0\|_\infty$ for all $j \ge j_0$. From (5.3) we infer that $e_j = 0$ for $j \ge j_0$. □

6. Finite sections of Toeplitz operators. In this section, we quote two known theorems that will be needed when proving the convergence of and giving an error estimate for our numerical algorithm. For $k \ge 1$, we denote by P_k the projection on ℓ^1 and c_0 that acts by the rule

$$P_k : \{x_1, x_2, x_3, \ldots\} \mapsto \{x_1, \ldots, x_k, 0, 0, \ldots\}.$$

We identify $\mathcal{R}(P_k)$ and $\mathbf{R}^k$, and hence we may think of vectors in $\mathbf{R}^k$ as elements of ℓ^1 or c_0. The following theorem was established by Reich [23] and Baxter [1]. Full proofs are also in [4, section 3.3] and [12, section II.2].

THEOREM 6.1. *If $f \in W$ and $T(f)$ is invertible on ℓ^1, then the matrices $T_k(f)$ are invertible for all sufficiently large k and $T_k^{-1}(f)P_k y$ converges in ℓ^1 to $T^{-1}(f)y$ for every $y \in \ell^1$.*

The next theorem can be proved using the asymptotic inverses presented in [4, section 3.5] or [6, section 2.3].

THEOREM 6.2. *Let f be a Laurent polynomial, that is, suppose f has only finitely many nonzero Fourier coefficients, and let $T(f)$ be invertible on ℓ^1. Fix a natural number $\varkappa$. Then there exist a natural number k_0 and constants $\alpha > 0$ and $C < \infty$ such that*

$$\|P_\varkappa T_k^{-1}(f) - P_\varkappa T^{-1}(f)\|_{\mathcal{B}(\ell^1)} = \|T_k^{-1}(\overline{f})P_\varkappa - T^{-1}(\overline{f})P_\varkappa\|_{\mathcal{B}(c_0)} \le C\, e^{-\alpha k}$$

for all $k \ge k_0$.

7. Numerical algorithm. Fix $b \in \ell^1$ and a_+ as above. Suppose a_+ has exactly $\varkappa$ zeros in $\mathbf{D}$ and no zeros on $\mathbf{T}$. If $k \geq \varkappa + 1$, the operator $P_k T(a_+) P_{k-\varkappa}$ may be identified with a $k \times (k-\varkappa)$ matrix. The system $P_k T(a_+) P_{k-\varkappa} x^{(k)} = P_k b$ is overdetermined for $\varkappa \geq 1$. However, we can find an $x_0^{(k)} \in \mathbf{R}^{k-\varkappa}$ such that the residue

$$\|P_k T(a_+) P_{k-\varkappa} x^{(k)} - P_k b\|_1 \tag{7.1}$$

assumes its minimum at $x^{(k)} = x_0^{(k)}$. Let $d = \operatorname{dist}_{\ell^1}(b, \mathcal{R}(T(a_+)))$ and let d_k be the minimal value of (7.1). The following theorem reveals that d_k converges to d exponentially fast.

THEOREM 7.1. *There are constants $E < \infty$ and $\beta > 0$ such that $|d_k - d| \leq E\, e^{-\beta k}$ for all $k \geq 1$.*

Proof. Put $f(t) = t^{-\varkappa} a_+(t) = t^{-\varkappa}(a_0 + a_1 t + \cdots + a_n t^n)$. We claim that $T(f)$ is invertible on ℓ^1. Indeed, the proof of Proposition 4.2 shows that

$$T(\overline{f}) = \gamma \prod_{k=1}^{n-\varkappa} \left(I - \frac{1}{\mu_k} T(\chi_{-1}) \right) \prod_{j=1}^{\varkappa} (I - \delta_j T(\chi_1))$$

with all operators on the right being invertible on $c_0(\mathbf{C})$. It follows that $T(\overline{f})$ is invertible on c_0 and hence that $T(f)$ is invertible on ℓ^1.

From (4.1) we deduce that $T(a_+) = T(f)T(\chi_\varkappa)$. Let $x^{(k)} = \{x_1^{(k)}, \ldots, x_{k-\varkappa}^{(k)}, 0, \ldots\}$ and define $w^{(k)} \in \mathcal{R}(P_k)$ by

$$w^{(k)} = \{\underbrace{0, \ldots, 0}_{\varkappa}, x_1^{(k)}, \ldots, x_{k-\varkappa}^{(k)}, 0, \ldots\}.$$

Let Q_k be given on ℓ^1 and c_0 by $Q_k = I - P_k$, that is,

$$Q_k : \{x_1, x_2, x_3, \ldots\} \mapsto \{0, \ldots, 0, x_{k+1}, x_{k+2}, \ldots\}.$$

We have $P_{k-\varkappa} x^{(k)} = T(\chi_{-\varkappa}) P_k w^{(k)}$, and since $T(\chi_{-\varkappa}) T(\chi_\varkappa) = Q_k$ and $Q_\varkappa P_k = P_k Q_\varkappa$, we get

$$\begin{aligned} &\|P_k b - P_k T(a_+) P_{k-\varkappa} x^{(k)}\|_1 = \|P_k b - P_k T(f) T(\chi_\varkappa) T(\chi_{-\varkappa}) P_k w^{(k)}\|_1 \\ &= \|P_k b - P_k T(f) P_k Q_\varkappa w^{(k)}\|_1 = \|P_k b - T_k(f) Q_\varkappa w^{(k)}\|_1. \end{aligned} \tag{7.2}$$

The minimum of (7.2) as $w^{(k)}$ ranges over $\mathbf{R}^k$ is d_k, and the minimum is attained at the $w_0^{(k)}$ corresponding to any $x_0^{(k)}$ that minimizes (7.1). Hence, by Theorems 3.1 and 3.2,

$$d_k = \sup_{z \in N(Q_\varkappa T_k(\overline{f})), \|z\|_\infty \leq 1} \langle z, P_k b \rangle. \tag{7.3}$$

Theorem 6.1 implies that there is a k_0 such that the matrices $T_k(f)$ are invertible for all $k \geq k_0$. Let $k \geq k_0$. We have $Q_\varkappa T_k(\overline{f}) z = 0$ if and only if there is a $y \in \mathbf{R}^\varkappa$ such that $T_k(\overline{f}) z = P_\varkappa y$ or, equivalently, $z = T_k^{-1}(\overline{f}) P_\varkappa y$. (Note that $T_k(\overline{f})$ is simply the transpose of $T_k(f)$). From (7.3) we therefore obtain

$$\begin{aligned} d_k &= \sup_{z = T_k^{-1}(\overline{f}) P_\varkappa y, \|z\|_\infty \leq 1} \langle z, P_k b \rangle \\ &= \sup_{\|T_k^{-1}(\overline{f}) P_\varkappa y\|_\infty \leq 1} \langle T_k^{-1}(\overline{f}) P_\varkappa y, P_k b \rangle \\ &= \sup_{\|T_k^{-1}(\overline{f}) P_\varkappa y\|_\infty \leq 1} \langle P_\varkappa y, P_\varkappa T_k^{-1}(f) P_k b \rangle. \end{aligned}$$

Put $\mathcal{M}_k = \{y \in \mathbf{R}^\varkappa : \|T_k^{-1}(\overline{f})P_\varkappa y\|_\infty \le 1\}$ and define $\varphi_k : \mathcal{M}_k \to \mathbf{R}$ by $\varphi_k(y) = \langle y, P_\varkappa T_k^{-1}(f)P_k b\rangle$. Then

$$d_k = \sup_{y\in\mathcal{M}_k} \varphi_k(y).$$

From Theorem 5.1 we know that

$$d = \sup_{z\in N(T(\overline{a}_+)),\|z\|_\infty\le 1} \langle z, b\rangle.$$

As $T(\overline{a}_+) = T(\chi_{-\varkappa})T(\overline{f})$, the equation $T(\overline{a}_+)z = 0$ is equivalent to the equation $T(\chi_{-\varkappa})T(\overline{f})f = 0$, that is, to the existence of a $y \in \mathbf{R}^\varkappa$ such that $z = T^{-1}(\overline{f})P_\varkappa y$. It follows that

$$\begin{aligned} d &= \sup_{z=T^{-1}(\overline{f})P_\varkappa y,\|z\|_\infty\le 1} \langle z, b\rangle \\ &= \sup_{\|T^{-1}(\overline{f})P_\varkappa y\|_\infty\le 1} \langle T^{-1}(\overline{f})P_\varkappa y, b\rangle \\ &= \sup_{\|T^{-1}(\overline{f})P_\varkappa y\|_\infty\le 1} \langle P_\varkappa y, P_\varkappa T^{-1}(f)b\rangle = \sup_{y\in\mathcal{M}} \varphi(y), \end{aligned}$$

where $\mathcal{M} = \{y \in \mathbf{R}^\varkappa : \|T^{-1}(\overline{f})P_\varkappa y\|_\infty \le 1\}$ and $\varphi : \mathcal{M} \to \mathbf{R}$ is given by $\varphi(y) = \langle y, P_\varkappa T^{-1}(f)b\rangle$. By Theorem 6.2,

$$\varphi_k(y) = \sum_{j=1}^{\varkappa} \gamma_j(k)y_j, \quad \varphi(y) = \sum_{j=1}^{\varkappa} \gamma_j y_j,$$

where $\gamma_j(k)$ converges to γ_j exponentially fast as $k \to \infty$. We remark that if $y \in \mathcal{M}_k$, then

$$\begin{aligned} \|P_\varkappa y\|_\infty &\le \|P_k T(\overline{f})P_k\|_{\mathcal{B}(c_0)}\|T_k^{-1}(\overline{f})P_\varkappa y\|_\infty \\ &\le \|f\|_W\|T_k^{-1}(\overline{f})P_\varkappa y\|_\infty \le \|f\|_W. \end{aligned}$$

Analogously, $\|P_\varkappa y\|_\infty \le \|f\|_W$ for $y \in \mathcal{M}$.

Now take $y_0 = (y_1^{(0)}, \dots, y_\varkappa^{(0)}) \in \mathcal{M}$ so that $\varphi(y_0) = d$. Theorem 6.2 yields

$$\begin{aligned} \|T_k^{-1}(\overline{f})P_\varkappa y_0\|_\infty &\le \|T^{-1}(\overline{f})P_\varkappa y_0\|_\infty + \|T_k^{-1}(\overline{f})P_\varkappa y_0 - T^{-1}(\overline{f})P_\varkappa y_0\|_\infty \\ &\le 1 + C\,e^{-\alpha k}\|P_\varkappa y_0\|_\infty \le 1 + C\,e^{-\alpha k}\|f\|_W =: 1 + \sigma_k. \end{aligned}$$

Thus, $(1+\sigma_k)^{-1}y_0 \in \mathcal{M}_k$. This implies that

$$d_k \ge \varphi_k[(1+\sigma_k)^{-1}y_0] = (1+\sigma_k)^{-1}\sum_{j=1}^{\varkappa} \gamma_j(k)y_j^{(0)}.$$

Since $\{\gamma_j(k) - \gamma_j\}_{k=1}^\infty$ is exponentially decaying for each j, we have

$$\sum_{j=1}^{\varkappa} \gamma_j(k)y_j^{(0)} \ge \sum_{j=1}^{\varkappa} \gamma_j y_j^{(0)} - \tau_k = d - \tau_k$$

with some exponentially decaying sequence $\{\tau_k\}$. In summary, we have shown that $(1+\sigma_k)d_k \geq d - \tau_k$, which gives

$$d - d_k \leq \sigma_k d_k + \tau_k \leq \sigma_k \|b\|_1 + \tau_k. \tag{7.4}$$

Again taking into account that $\{\gamma_j(k) - \gamma_j\}_{k=1}^{\infty}$ is exponentially decaying for each j and using that $\|P_\varkappa y\|_\infty \leq \|f\|_W$ for all $y \in \mathcal{M}_k$, we obtain

$$d_k = \sup_{y\in\mathcal{M}_k} \sum_{j=1}^{\varkappa} \gamma_j(k) y_j \leq \sup_{y\in\mathcal{M}_k} \sum_{j=1}^{\varkappa} \gamma_j y_j + \varrho_k$$

with an exponentially decaying sequence $\{\varrho_k\}$. For $y \in \mathcal{M}_k$, Theorem 6.2 gives

$$\begin{aligned}\|T^{-1}(\overline{f})P_\varkappa y\|_\infty &\leq \|T_k^{-1}(\overline{f})P_\varkappa y\|_\infty + \|T^{-1}(\overline{f})P_\varkappa y - T_k^{-1}(\overline{f})P_\varkappa y\|_\infty \\ &\leq 1 + C\,e^{-\alpha k}\|P_\varkappa y\|_\infty \leq 1 + C\,e^{-\alpha k}\|f\|_W =: 1 + \sigma_k,\end{aligned}$$

and therefore $(1+\sigma_k)^{-1}y \in \mathcal{M}$. It follows that

$$\begin{aligned}d_k &\leq \sup_{(1+\sigma_k)^{-1}y\in\mathcal{M}} \sum_{j=1}^{\varkappa} \gamma_j y_j + \varrho_k = \sup_{v\in\mathcal{M}} \sum_{j=1}^{\varkappa} \gamma_j \cdot (1+\sigma_k) v_j + \varrho_k \\ &= (1+\sigma_k) \sup_{v\in\mathcal{M}} \varphi(v) + \varrho_k \leq (1+\sigma_k)d + \varrho_k,\end{aligned}$$

whence

$$d_k - d \leq \sigma_k d + \varrho_k. \tag{7.5}$$

Combining (7.4) and (7.5) we arrive at the assertion. □

COROLLARY 7.2. *For each $k \geq 1$, let $x_0^{(k)} \in \mathcal{R}(P_{k-\varkappa})$ be an element at which (7.1) attains its minimum d_k. If $k_i \to \infty$ and $\{x_0^{(k_i)}\}_{i=1}^{\infty}$ is any sequence that converges in ℓ^1 to some $x_0 \in \ell^1$, then $\|b - T(a_+)x_0\|_1 = d$.*

Proof. If $\|P_{k_i}b - P_{k_i}T(a_+)P_{k_i-\varkappa}x_0^{(k_i)}\|_1 = d_{k_i}$ and $x_0^{(k_i)} \to x$ as $i \to \infty$, then $\|b - T(a_+)x_0\|_1 = d$ because $d_{k_i} \to d$ by Theorem 7.1. □

8. Error estimate. In practice, we have an $x_0^{(k)}$ with

$$\|P_k b - P_k T(a_+) P_{k-\varkappa} x_0^{(k)}\|_1 = d_k,$$

and $\widetilde{m}_0 = T(a_+)P_{k-\varkappa}x_0^{(k)}$ is taken as an approximate solution. The question is, how far is $\|b - \widetilde{m}_0\|_1$ away from the optimal value d? We have

$$\begin{aligned}\|b - \widetilde{m}_0\|_1 \leq{}& \|b - P_k b\|_1 + \|P_k b - P_k T(a_+) P_{k-\varkappa} x_0^{(k)}\|_1 \\ &+ \|P_k T(a_+) P_{k-\varkappa} x_0^{(k)} - T(a_+) P_{k-\varkappa} x_0^{(k)}\|_1.\end{aligned} \tag{8.1}$$

Clearly, $\|b - P_k b\|_1 = \|Q_k b\|_1 = o(1)$ is given a priori. The second term on the right of (8.1) is just d_k. Let $x_i^{(k)}$ $(i = 1, \ldots, k-\varkappa)$ denote the components of $x_0^{(k)}$. The

vector in the third term on the right of (8.1) is

$$-Q_k T(a_+) P_{k-\varkappa} x_0^{(k)} = - \begin{pmatrix} a_k & \dots & a_{\varkappa+1} \\ a_{k+1} & \dots & a_{\varkappa+2} \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ \vdots \\ x_{k-\varkappa}^{(k)} \end{pmatrix}$$

$$= - \begin{pmatrix} 0 & \dots & 0 & a_n & a_{n-1} & \dots & a_{\varkappa+1} \\ 0 & \dots & 0 & 0 & a_n & \dots & a_{\varkappa+2} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & a_n \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ \vdots \\ x_{k-\varkappa}^{(k)} \end{pmatrix},$$

which implies that

$$\|Q_k T(a_+) P_{k-\varkappa} x_0^{(k)}\|_1 \le (|a_{\varkappa+1}| + \dots + |a_n|) \left(|x_{k-n+1}^{(k)}| + \dots + |x_{k-\varkappa}^{(k)}| \right)$$
$$\le \|a_+\|_W \, \|Q_{k-n} x_0^{(k)}\|_1.$$

The vector $x_0^{(k)}$ is available and $\|Q_{k-n} x_0^{(k)}\|_1$ is the ℓ^1 norm of the last $n-\varkappa$ components of $x_0^{(k)}$. If k is large, then the last $n-\varkappa$ components of $x_0^{(k)}$ are expected to be small. In summary, (8.1) and Theorem 7.1 yield

$$\|b - \widetilde{m}_0\|_1 \le \|Q_k b\|_1 + \|a_+\|_W \, \|Q_{k-n} x_0^{(k)}\|_1 + d + \text{ exponentially small term.}$$

If $\varkappa = n$, then $-Q_k T(a_+) P_{k-\varkappa} x_0^{(k)} = 0$, and hence we even have

$$\|b - \widetilde{m}_0\|_1 \le \|Q_k b\|_1 + d + \text{ exponentially small term.}$$

Finally, if $\varkappa = n$ and b is finitely supported, then

$$\|b - \widetilde{m}_0\|_1 \le d + \text{ exponentially small term.}$$

9. Numerical example. The following example illustrates the algorithm described above. We consider the banded lower triangular Toeplitz matrix $T(a_+)$ with the symbol

$$a_+(t) = -0.1224 - 0.2906t + 0.7122t^2 + 2.7983t^3 + 2.9168t^4 + t^5$$

and we are looking for a sequence $x \in \ell^1$ minimizing $\|b - T(a_+)x\|_1$ for the right-hand side

$$b = \{1.8645, -0.3398, -1.1398, -0.2111, 1.1902, -1.1162, 0, 0, \dots\}.$$

The five zeros of the polynomial $a_+(t)$ are all inside the open unit disk. Thus, we can proceed as in section 7 with $\varkappa = 5$. (Notice that the algorithm of section 7 would be applicable to $\varkappa < 5$ as well.) Accordingly, we approximate $T(a_+)$ by the finite matrices $A_k = P_{5+k} T(a_+) P_k$ $(k = 1, 2, \dots)$. For example,

$$A_3 = \begin{pmatrix} -0.1224 & 0 & 0 \\ -0.2906 & -0.1224 & 0 \\ 0.7122 & -0.2906 & -0.1224 \\ 2.7983 & 0.7122 & -0.2906 \\ 2.9168 & 2.7983 & 0.7122 \\ 1 & 2.9168 & 2.7983 \\ 0 & 1 & 2.9168 \\ 0 & 0 & 1 \end{pmatrix}.$$

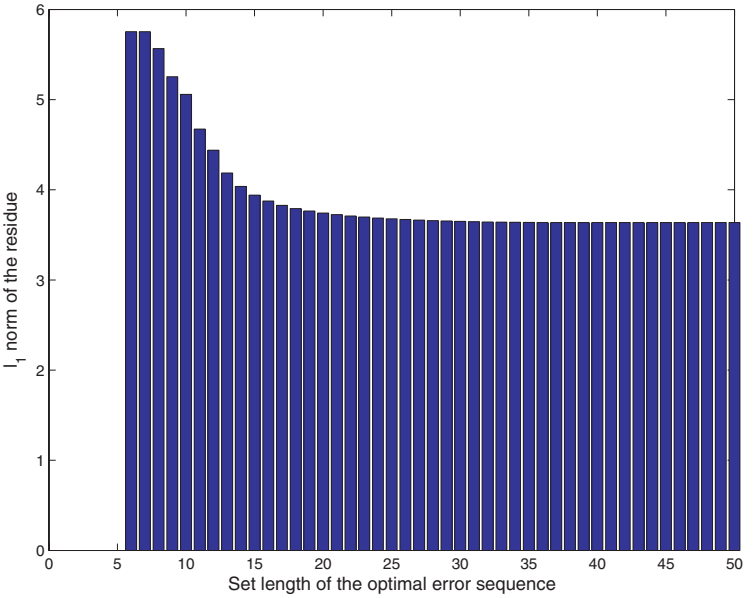


FIG. 9.1. *Evolvement of the ℓ^1-norm of the residue with increasing set length of the optimal error sequence.*

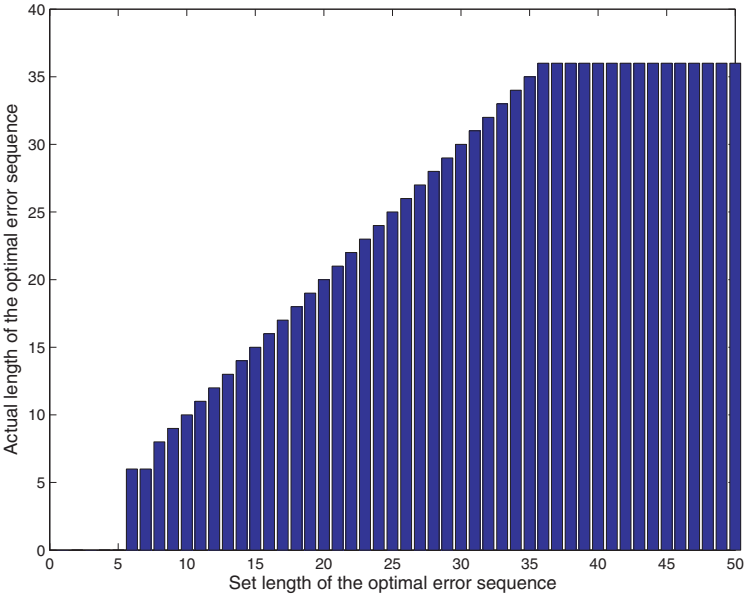


FIG. 9.2. *Evolvement of the actual length of the optimal error sequence with increasing set length of the optimal error sequence.*

Solving the corresponding overdetermined linear system for a solution minimizing the ℓ_1-norm of the residue (using a general LP solver) and repeating this for increasing index k, we obtain Figures 9.1 and 9.2. In Figure 9.1 we nicely see the exponentially fast stabilization of the objective function (that is, the number d_k or, equivalently, the ℓ^1-norm of the residuum) predicted by Theorem 7.1. Figure 9.2 reveals that for

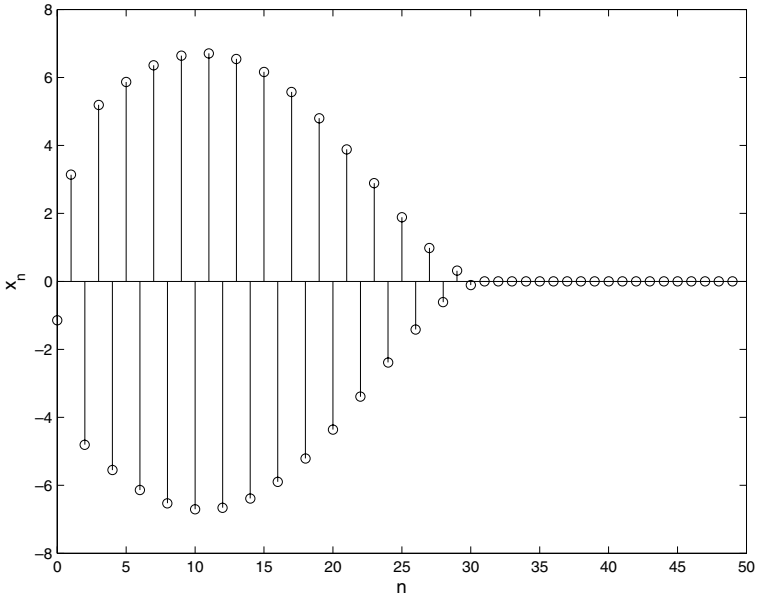


FIG. 9.3. *A sequence $x \in \ell_1$ minimizing $\|b - T(a_+)x\|_1$.*

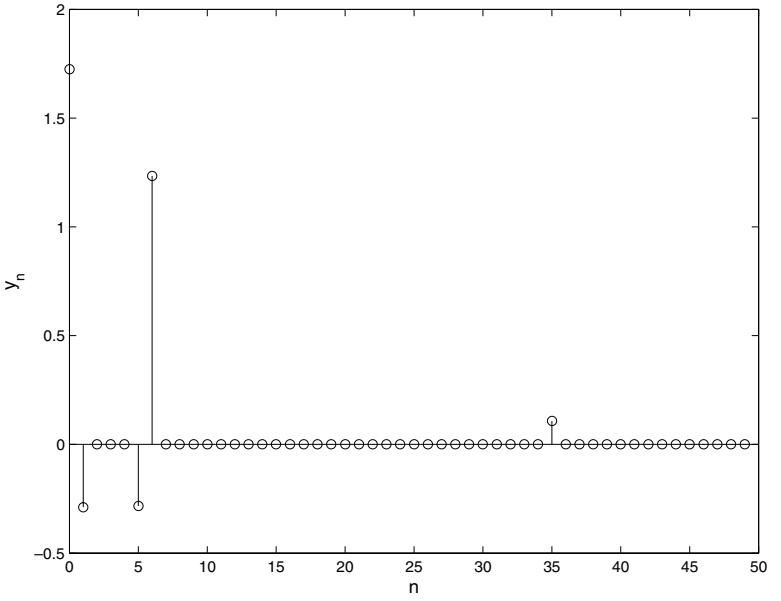


FIG. 9.4. *The optimal error sequence (residue) $y = b - T(a_+)x$.*

$k - \varkappa \geq 36$ the number of nonzero terms in $P_{k-\varkappa}x_0^{(k)}$ is no longer increasing. Thus, although we offer more and more space to the approximate optimal solution sequence $P_{k-\varkappa}x_0^{(k)}$, its actual length settles at 36.

Figures 9.3 and 9.4 show an optimal solution x and the residue sequence $y = b - T(a_+)x$. The very small number of nonzero terms in Figure 9.4 is a mystery we cannot yet explain.

10. Improvement of conditioning. Both our Toeplitz approach and the Vandermonde interpolation approach of [10] lead to linear systems of equations. These can be tackled by invoking an LP solver. Numerical experiments show that the condition numbers of the matrices emerging in our algorithm are much smaller than those of the matrices that result from interpolation. To give a concrete example, consider the task of finding the distance between a given $b \in \ell^1$ and the range of the Toeplitz operator $T(a_+)$ with $a_+(t)$ having its 10 roots equally distributed in the interval $[0.5, 0.9]$, that is, $a_+(t) = 0.0238 - 0.3520t + 2.3334t^2 - 9.1302t^3 + 23.3525t^4 - 40.7975t^5 + 49.3052t^6 - 40.7037t^7 + 21.9685t^8 - 7t^9 + t^{10}$. Let us set the length of the approximate optimal error sequence to 13. The 2-norm condition number, that is, the ratio of the largest and the smallest singular value, of the 10×13 Vandermonde matrix V_{13} built from the roots of $a_+(t)$ equals $\kappa(V_{13}) = 9.5458 \cdot 10^9$. In contrast to this, the 2-norm condition number of the matrix $A_{13} = P_{13}T(a_+)P_3$ is $\kappa(A_{13}) = 14.948$. To make a general conclusion, Vandermonde systems are known to be extremely ill-conditioned unless the roots are distributed along the unit circle. Analytical expressions for the conditioning of some common distributions are in [13, p. 418]. Similar analytical results for Toeplitz systems are not known to the authors.

11. Application to ℓ^1-optimal control. In compliance with control engineering, we now use the variable λ instead of t. (The variables λ, d, and z^{-1} are more common in control theory literature than t. The variable t is dominating in Toeplitz operators theory but it could be confused with time.) It is well known [19] that the achievable internally stable closed-loop maps $y(\lambda)$ of a standard feedback connection are parametrized by $y(\lambda) = b(\lambda) - a(\lambda)x(\lambda)$, where $a(\lambda)$, $b(\lambda)$, $x(\lambda)$, $y(\lambda)$ are power series with coefficient sequences in ℓ^1, $a(\lambda)$ and $b(\lambda)$ are given, and $x(\lambda)$, which is also called the Youla–Kučera parameter, is unknown. For finite-dimensional systems, $a(\lambda)$ is actually a polynomial. A principal task of ℓ^1-control is to design a stabilizing controller that minimizes the ℓ^1-norm of the coefficient sequence of $y(\lambda)$. This controller will guarantee optimal attenuation of peaks in the error signal $y(\lambda)$ because the Wiener norm of $y(\lambda)$ ($= \ell^1$-norm of its coefficient sequence) is equal to the ℓ^∞-induced operator norm of the closed-loop system [24].

Any stabilizing controller is determined by its Youla–Kučera parameter $x(\lambda)$. As the coefficient sequence of $a(\lambda)x(\lambda)$ results from that of $x(\lambda)$ by the action of an infinite lower-triangular Toeplitz band matrix, the problem of designing an optimal controller leads to the minimum distance problem between a given sequence $b \in \ell^1$ and the range $\mathcal{R}(T(a))$ of the infinite lower-triangular Toeplitz matrix $T(a)$ in ℓ^1. A concrete design procedure will be exemplified in the next section.

A noteworthy feature of the approach proposed here is that the Youla–Kučera parameter is explicitly computed and that, consequently, there is no need to extract a controller from the optimal closed-loop transfer function.

12. A concrete ℓ^1-optimal control design. We consider the standard feedback configuration of Figure 12.1 with a discrete-time plant $G(\lambda)$. Our aim is to construct a stabilizing discrete-time controller $C(\lambda)$ that minimizes the Wiener norm of the sensitivity function of the closed-loop system, that is, of the transfer function $1/(1 + C(\lambda)G(\lambda))$ between the disturbance and the error or, equivalently, the ℓ^1 norm of the impulse response.

We suppose that the plant is given as the quotient of two polynomials $p(\lambda)$ and $q(\lambda)$ without common zeros and with no zeros on the unit circle, $G(\lambda) = p(\lambda)/q(\lambda)$.

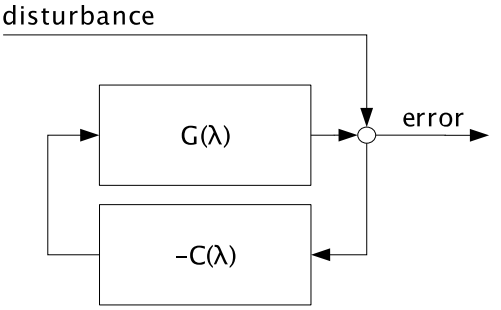


FIG. 12.1. *Standard feedback control configuration.*

The Youla–Kučera parametrization of all stabilizing controllers is

$$C(\lambda) = \frac{v(\lambda) + q(\lambda)x(\lambda)}{w(\lambda) - p(\lambda)x(\lambda)}, \tag{12.1}$$

where $v(\lambda), w(\lambda)$ are polynomials determined by $G(\lambda)$ and $x(\lambda)$ is a function we can freely choose in the Wiener algebra. The entire procedure can be done in four steps.

Step 1. Find stable-unstable factorizations $p(\lambda) = p_s(\lambda)p_u(\lambda)$ and $q(\lambda) = q_s(\lambda)q_u(\lambda)$. Here the indices s and u label polynomials with all zeros inside and outside the unit circle, respectively. Efficient algorithms for stable-unstable factorization are known (see, e.g., [3] and the references cited therein). In particular, reliable FFT-based algorithms are available from [2], [14].

Step 2. Find polynomials $x_0(\lambda)$ and $y_0(\lambda)$ satisfying the Diophantine equation $q(\lambda)x_0(\lambda) + p(\lambda)y_0(\lambda) = 1$. This problem can be conveniently solved using the polynomial toolbox [20].

Step 3. The polynomials $v(\lambda), w(\lambda)$ in (12.1) are

$$v(\lambda) = q_u(\lambda)p_u(\lambda)y_0(\lambda), \quad w(\lambda) = q_u(\lambda)p_u(\lambda)x_0(\lambda).$$

Step 4. Inserting the result of Step 3 in (12.1) we obtain

$$\frac{1}{1+CG} = \frac{1}{1 + \dfrac{q_u p_u y_0 + qx}{q_u p_u y x_0 - px}\,\dfrac{p}{q}} = \frac{qq_u p_u x_0 - qpx}{q_u p_u (qx_0 + py_0)},$$

which equals $qx_0 - q_s p_s x$ by virtue of Step 2. Thus, the final task is to minimize $\|y(\lambda)\|_W = \|q(\lambda)x_0(\lambda) - q_s(\lambda)p_s(\lambda)x(\lambda)\|_W$ or, in terms of the coefficient sequences, to minimize $\|y\|_1 = \|b - T(a_+)x\|_1$, where $b \in \ell^1$ is the coefficient sequence of $q(\lambda)x_0(\lambda)$ and $a_+(\lambda) = q_s(\lambda)p_s(\lambda)$. This problem can be solved using the algorithm of section 7. The desired optimal controller is given by (12.1) with $v(\lambda), w(\lambda)$ from Step 3 and $x(\lambda)$ from Step 4.

To have a numerical example, let

$$G(\lambda) = \frac{p(\lambda)}{q(\lambda)} = \frac{-45\lambda - 132\lambda^2 + 9\lambda^3}{-20 - 48\lambda + 5\lambda^2}.$$

The above procedure yields the Youla–Kučera parameter $x(\lambda) = 0.1321 - 0.0052\lambda$, the sensitivity function $y(\lambda) = 1.0000 - 12.5000\lambda - 37.5000\lambda^2$, and the optimal controller

$$C(\lambda) = \frac{-41.6667 + 4.1667\lambda}{-7.5000 + 113.0000\lambda - 7.5000\lambda^2}.$$

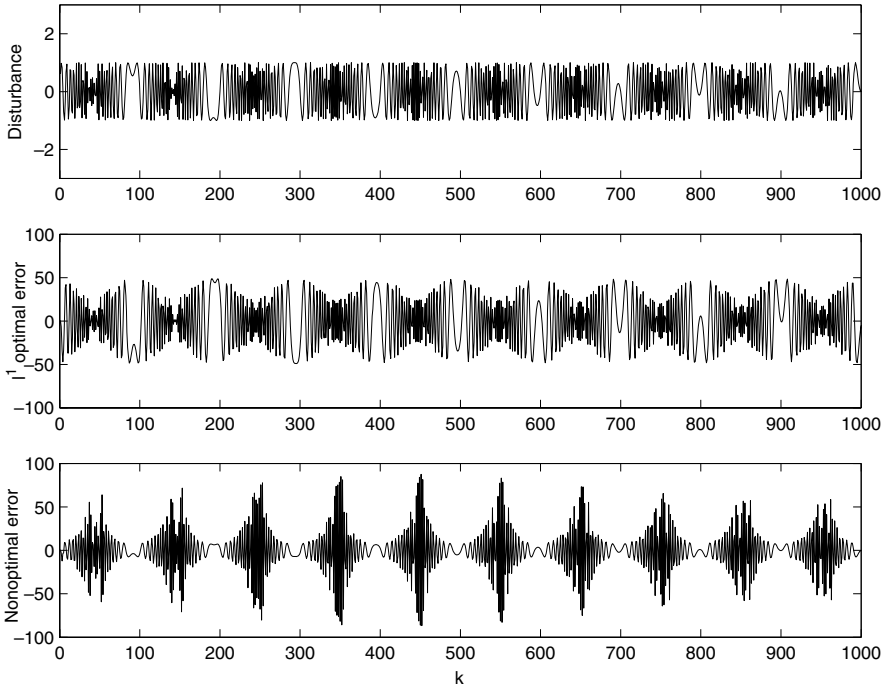


FIG. 12.2. *Simulation of a disturbance rejection with ℓ^1-optimal and nonoptimal controllers.*

A simulation result is shown in Figure 12.2. The horizontal axis represents the discrete time k. The disturbance is only known to be bounded in magnitude. The response of the closed-loop system to a disturbance bounded in magnitude by 1 is compared for the ℓ^1-optimal controller computed above and some random stabilizing controller. Note that with more sophisticated (optimal) controllers like LQG, $\mathcal{H}_2$-, and $\mathcal{H}_\infty$-optimal controllers the difference will not necessarily be this striking, as the system norms in finite-dimensional spaces are equivalent and, loosely speaking, minimizing some norm causes all the other norms to be small as well.

13. Conclusions. This paper contains an analysis of the problem of finding the minimum distance to the range of a banded lower-triangular Toeplitz operator in ℓ^1. Necessary and sufficient conditions for the existence of a minimizing sequence in ℓ^1 are established. It is shown that the optimal error sequences is finitely nonzero. The application of this result in ℓ^1-optimal control is outlined. The paper presents an alternative formulation and a new approach to the well known problem, putting it into operator-theoretic framework and leading to more reliable numerical algorithms.

It is demonstrated by an example that in some situations the interpolation approach may lead to highly ill-conditioned linear systems while the method proposed does not suffer from this unpleasant circumstance. Another striking feature of the present approach is that the (stable) optimal Youla–Kučera parameter is returned as a direct outcome of the linear optimization, so that there is no need for any numerically tricky extraction of an optimal controller from the closed-loop transfer function. Within the interpolation framework, a solution to the truncated problem does not necessarily yield an achievable stable closed-loop transfer function. Hence the controller obtained from it need not stabilize the plant. These complications are more

pronounced in the general MIMO case when only suboptimal solutions can be obtained.

The authors admit that the paper might seem unnecessarily tough to read, but the motivation for this high level of abstraction and rigorousness was to attract researchers from outside the field control theory, especially from operator theory. For instance, at this moment it is not known how to get an estimate of the length of the optimal error sequence (however, it is known that this length is data dependent [22]). Also, it is not known how to build a finite-dimensional approximation to the dual problem such that its optimal solutions are feasible for the original dual problem. This is a vital issue for the development of general primal-dual solvers. The extension of the analysis presented here to operators on vector-valued infinite sequences is very attractive in connection with practical applications, because it would enable control of systems with more than one control/disturbing input and more than one measured and/or regulated input. Some progress recently has been made by the authors, and preliminary results on the straightforward extension to rectangular block Toeplitz operators including analysis of existence, uniqueness, and convergence were presented in [16]. The present paper can be regarded as a building block for this extension.

Acknowledgment. The first author thanks David C. Ullrich for fruitful discussions at the sci.math.research usenet group.

REFERENCES

[1] G. Baxter, *A norm inequality for a finite-section Wiener-Hopf equation,* Illinois J. Math., 7 (1963), pp. 97–103.

[2] D. A. Bini, *Using FFT-based techniques in polynomial and matrix computations: Recent advances and applications,* Numer. Funct. Anal. Optim., 21 (2001), pp. 47–66.

[3] D. A. Bini and A. Böttcher, *Polynomial factorization through Toeplitz matrix computations,* Linear Algebra Appl., 366 (2003), pp. 25–37.

[4] A. Böttcher and S. Grudsky, *Spectral Properties of Banded Toeplitz Matrices,* SIAM, Philadelphia, 2005.

[5] A. Böttcher and B. Silbermann, *Analysis of Toeplitz Operators*, Akademie-Verlag, Berlin, 1989.

[6] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices,* Springer-Verlag, New York, 1999.

[7] A. Casavola, *A polynomial approach to the ℓ_1-mixed sensitivity optimal control problem*, IEEE Trans. Automat. Control, 41 (1996), pp. 751–756.

[8] A. Casavola and D. Famularo, *Q domain sub/super-optimization linear programming methods for MIMO ℓ_1 control problems*, in Proceedings of the 4th IFAC Symposium on Robust Control Design ROCOND'00, Prague, Czech Republic, June 2000.

[9] A. Casavola and D. Famularo, *MIMO ℓ^1 optimal control problems via the polynomial equation approach*, Internat. J. Control, 76 (2003), pp. 823–835.

[10] M. A. Dahleh and J. B. Pearson, Jr., *ℓ^1-optimal feedback controllers for MIMO discrete-time systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 314–322.

[11] M. A.Dahleh and I. J. Diaz-Bobillo, *Control of Uncertain Systems: A Linear Programming Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1995.

[12] I. Gohberg and I. A. Feldman, *Convolution Equations and Projection Methods for Their Solution,* AMS, Providence, R.I., 1974.

[13] N. Higham, *Accuracy and Stability of Numerical Algorithms,* SIAM, Philadelphia, 2002.

[14] M. Hromčík and M. Šebek, *FFT based algorithm for polynomial plus-minus factorization*, in Proceedings of the European Control Conference (ECE'03), Cambridge, UK, Sept. 2003.

[15] Z. Hurák and M. Šebek, *Algebraic approach to the ℓ^1-optimal control*, in Proceedings of the 4th IFAC Symposium on Robust Control Design (ROCOND'03), Milan, Italy, June 2003.

[16] Z. Hurák and A. Böttcher, *MIMO ℓ_1-Optimal Control via Block Toeplitz Operators*, in Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS'04), Katholieke Universiteit Leuven, Belgium, July 2004.

[17] M. Khammash, *A new approach to the solution of the ℓ_1 control problem: The scaled-Q method*, IEEE Trans. Automat. Control, 45 (2000), pp. 180–187.
[18] M. Khammash, M. V. Salapaka, and T. Van Voorhis, *Robust synthesis in ℓ_1: A globally optimal solution*, IEEE Trans. Automat. Control, 46 (2001), pp. 1744–1754.
[19] V. Kučera, *Discrete Linear Control: The Polynomial Approach*, Wiley and Sons, Chichester, UK, 1979.
[20] H. Kwakernaak and M. Šebek, *Polynomial Toolbox for Matlab*, Polyx, Ltd., http://www.polyx.com (1998).
[21] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[22] D. G. Meyer, *Two properties of ℓ_1-optimal controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 876–878.
[23] E. Reich, *On non-Hermitian Toeplitz matrices,* Math. Scand., 10 (1962), pp. 145–152.
[24] M. Vidyasagar, *Optimal rejection of persistent bounded disturbances*, IEEE Trans. Automat. Control, 31 (1986), pp. 527–534.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 123–145

ALGEBRAIC RICCATI EQUATION AND J-SPECTRAL FACTORIZATION FOR $\mathcal{H}_\infty$ SMOOTHING AND DECONVOLUTION*

PATRIZIO COLANERI† AND AUGUSTO FERRANTE‡

Abstract. This paper deals with a general steady-state estimation problem in the $\mathcal{H}_\infty$ setting. The existence of the stabilizing solution of the related algebraic Riccati equation (ARE) and of the solution of the associated J-spectral factorization problem is investigated. The existence of such solutions is well established if the prescribed attenuation level γ is larger than γ_f (the infimum of the values of γ for which a *causal* estimator with attenuation level γ exists). We consider the case when $\gamma \leq \gamma_f$ and show that the stabilizing solution of the ARE still exists (except for a finite number of values of γ) as long as a fixed-lag acausal estimator (smoother) does. The stabilizing solution of the ARE may be employed to derive a state-space realization of a minimum-phase J-spectral factor of the J-spectrum associated with the estimation problem. This J-spectral factor may be used, in turn, to compute the minimum-lag smoothing estimator. Some of the aspects of the J-spectral factorization problem and the properties of its solutions are discussed in correspondence to the (finite number of) values of γ for which the stabilizing solution of the ARE does not exist.

Key words. estimation problems, J-spectral factorization, algebraic Riccati equation, filtering, deconvolution, smoothing

AMS subject classifications. 15A24, 93B36, 93C55, 47A68

DOI. 10.1137/S0363012903434741

1. Introduction and problem statement.

Consider the discrete-time linear system

$$x_{k+1} = Ax_k + Bw_k, \tag{1.1a}$$

$$y_k = Cx_k + Dw_k, \tag{1.1b}$$

$$z_k = Lx_k + Mw_k, \tag{1.1c}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^p$ are the state and the measurement output vector, respectively, and $w_k \in \mathbb{R}^m$ is the vector of inputs and disturbances. The to-be-estimated signal $z_k \in \mathbb{R}^l$ is an unaccessible linear combination of the state and the input. This is very general and includes, as particular cases, the *state-filtering* problem ($L = I$ and $M = 0$) and the *deconvolution* problem ($L = 0$ and $M = I$). See section 1.3 for a brief discussion and references on these problems.

Let $F(z)$ be the transfer function of a causal filter driven by the observations y_k whose output $\hat{z}_k$ is an estimate of z_k, and let $e_k := z_k - \hat{z}_k$. The infinite-horizon $\mathcal{H}_\infty$ *filtering problem* consists in designing an estimator $F(z)$ guaranteeing a prescribed level of attenuation γ between the ℓ_2-norm of w_k and the ℓ_2-norm of the estimation error e_k. Introducing the transfer functions

$$H_1(z) := C(zI - A)^{-1}B + D, \tag{1.2a}$$

*Received by the editors September 15, 2003; accepted for publication (in revised form) October 15, 2005; published electronically February 21, 2006. This work was partially supported by the ministry of higher education of Italy (MIUR), under project *Identification and Control of Industrial Systems*.

http://www.siam.org/journals/sicon/45-1/43474.html

†Dipartimento di Elettronica e Informazione, Politecnico di Milano, P.za L. da Vinci 32, 20133 Milano, Italy (colaneri@elet.polimi.it).

‡Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo 6/B, 35131 Padova, Italy (augusto@dei.unipd.it).

$$H_2(z) := L(zI - A)^{-1}B + M, \tag{1.2b}$$

this problem is equivalent to that of finding a stable causal transfer function $F(z)$ such that

$$F(z)H_1(z) - H_2(z) \in \mathcal{RH}_\infty, \qquad \|F(z)H_1(z) - H_2(z)\|_\infty < \gamma, \tag{1.3}$$

where $\mathcal{RH}_\infty$ denotes the space of real rational matrix functions whose poles lie all in $\{z \in \mathbb{C} : \ |z| < 1\}$. $\mathcal{RH}_\infty$ is endowed with the infinity norm associating with any $G(z) \in \mathcal{RH}_\infty$, $\|G(z)\|_\infty = \sup_{|z|=1} \|G(z)\|$ with $\|G(z)\|$ being the largest singular value of $G(z)$ for each given z.

The $\mathcal{H}_\infty$ *fixed-lag smoothing problem with preview horizon of length* N (or N-*lag smoothing problem*) may be defined as the problem of estimating (in the $\mathcal{H}_\infty$ framework) the signal z_k given the observations up to time $k + N$, $N \geq 0$. In other words, the only difference with the filtering problem is that the transfer function of the estimator is not required to be causal but has to be of the form $E(z) = z^N F(z)$ with $F(z)$ being stable and causal. This may be easily reformulated as the problem of finding a stable causal transfer function $F(z)$ such that

$$F(z)H_1(z) - H_{2,N}(z) \in \mathcal{RH}_\infty, \qquad \|F(z)H_1(z) - H_{2,N}(z)\|_\infty < \gamma \tag{1.4}$$

with

$$H_{2,N}(z) := z^{-N} H_2(z). \tag{1.5}$$

For a general overview on $\mathcal{H}_\infty$ estimation see to [19, 21] and references therein. For the continuous time version of the $\mathcal{H}_\infty$ smoothing problem see [18].

Assumptions. Let[1]

$$B_1 := B(I - D^\top (DD^\top)^{-1} D), \qquad F := A - BD^\top (DD^\top)^{-1} C. \tag{1.6}$$

We make the following standard assumptions:

$$DD^\top > 0, \tag{1.7a}$$

$$(A, C) \text{ is detectable}, \tag{1.7b}$$

$$(F, B_1) \text{ is reachable}, \tag{1.7c}$$

$$F \text{ is invertible}. \tag{1.7d}$$

Moreover, if $MD^\top \neq 0$, we define the auxiliary signal $\xi_k := z_k - MD^\top (DD^\top)^{-1} y_k = \widetilde{L} x_k + \widetilde{M} w_k$ with $\widetilde{L} := L - MD^\top (DD^\top)^{-1} C$ and $\widetilde{M} := M - MD^\top (DD^\top)^{-1} D$ and observe that, since ξ is obtained from z by subtracting a linear combination of the observed signal y, the estimations of z and ξ are equivalent problems (i.e., the estimation error is the same). In addition, a simple computation shows that $\widetilde{M} D^\top = 0$. Hence, without loss of generality, we assume

$$MD^\top = 0 \tag{1.7e}$$

so that the following identity holds:

$$BM^\top = B_1 M^\top. \tag{1.8}$$

Eventually, notice that neither A nor F is assumed to be stable.

[1] We denote by $M^\top$ the transpose of a matrix M.

From the point of view of the motivating $\mathcal{H}_\infty$ estimation problem, we mention that assumptions (1.7a) and (1.7d) do not impair loss of generality. In fact if (1.7a) is not satisfied we may resort to a Silverman transfomation (see [6]) that, employing a spectral interactor matrix, yields an equivalent problem in which $DD^\top > 0$. If (1.7d) is not satisfied, we may perform a preliminary feedback transformation, as described in [5], and obtain an equivalent problem, in which (1.7d) is satisfied.

1.1. *J*-spectral factorization. For a real transfer matrix function $G(z)$, we define $G\tilde{}(z) := G^\top(z^{-1})$. Let

$$H(z) := \begin{bmatrix} H_1(z) & 0 \\ H_2(z) & I_l \end{bmatrix} = \begin{bmatrix} C \\ L \end{bmatrix} (zI - A)^{-1}[B \mid 0\,] + \begin{bmatrix} D & 0 \\ M & I_l \end{bmatrix} \tag{1.9}$$

and $J_{i,j}(\gamma) := \operatorname{diag}\{I_i, -\gamma^2 I_j\}$. (We shall simply write J and I, instead of $J_{i,j}(\gamma)$ and I_i, when there is no risk of confusion.) The transfer function

$$\Psi(z) := H(z)J_{m,l}(\gamma)H\tilde{}(z) = \begin{bmatrix} \Psi_{11}(z) & \Psi_{12}(z) \\ \Psi_{12}\tilde{}(z) & \Psi_{22}(z) \end{bmatrix} \tag{1.10}$$

with

$$\Psi_{11}(z) := H_1(z)H_1\tilde{}(z), \quad \Psi_{12}(z) := H_1(z)H_2\tilde{}(z), \quad \Psi_{22}(z) := H_2(z)H_2\tilde{}(z) - \gamma^2 I \tag{1.11}$$

is called the *J-spectrum* of the system. A *square J-spectral factor* of $\Psi(z)$ is a square matrix function $\Omega(z)$ such that

$$\Omega(z)J_{p,l}(\gamma)\Omega\tilde{}(z) = \Psi(z). \tag{1.12}$$

A square J-spectral factor is said to be *minimum phase* if all its zeros lay in the open unit disk.[2] See [10] and references therein for further details on J-spectral factorization.

The estimation problem is strictly related to the existence of a J-spectral factor. Precisely, the existence of an estimator with prescribed attenuation level γ is equivalent to the existence of a minimum phase J-spectral factor of $\Psi(z)$ having a realization with the same state matrix A and the same output matrix $[C^\top \mid L^\top]^\top$ of the realization (1.9) of $H(z)$; see [3]. Once obtained such a spectral factor, a simple constructive procedure (described in [3]) furnishes a class of estimators $F(z)$ satisfying (1.3).

1.2. Algebraic Riccati equation. Let

$$H_0 := \begin{bmatrix} C \\ L \end{bmatrix}, \quad J_0 := \begin{bmatrix} D \\ M \end{bmatrix}, \quad J_1 := J_0 J_0^\top - \begin{bmatrix} 0 & 0 \\ 0 & \gamma^2 I \end{bmatrix} = \begin{bmatrix} DD^\top & DM^\top \\ MD^\top & MM^\top - \gamma^2 I \end{bmatrix} \tag{1.13}$$

and consider the algebraic Riccati equation (ARE)

$$\Delta = A\Delta A^\top + BB^\top - (A\Delta H_0^\top + BJ_0^\top)(J_1 + H_0\Delta H_0^\top)^{-1}(H_0\Delta A^\top + J_0 B^\top). \tag{1.14}$$

[2] Notice that in some papers and books the definition of minimum phase requires the stability of the spectral factor. Here, since neither A nor F is assumed to be "stable," we are not interested in stable J-spectral factors and we adopt the definition of phase minimality given above.

It is well known (see, e.g., [21]) that the existence of a proper stable filter $F(z)$ satisfying (1.3) is equivalent to the existence of a *symmetric*, *positive definite*, *stabilizing*, *feasible* solution of (1.14), i.e., a solution $\Delta_f = \Delta_f^\top > 0$ such that

$$A_f := A - (A\Delta_f H_0^\top + BJ_0^\top)(J_1 + H_0\Delta_f H_0^\top)^{-1}H_0 \tag{1.15}$$

is stable and $J_1 + H_0\Delta_f H_0^\top$ has the same inertia of $J_{p,l}(\gamma)$. In passing, notice that as long as $(DD^\top + C\Delta_f C^\top) > 0$, the latter condition (feasibility) may be written in the form

$$\gamma^2 I - MM^\top - L\Delta_f L^\top + L\Delta_f C^\top (DD^\top + C\Delta_f C^\top)^{-1} C\Delta_f L^\top > 0. \tag{1.16}$$

If we now define

$$\gamma_f := \inf\{\gamma : \text{ there exists a proper stable filter } F(z) \text{ satisfying (1.3)}\}, \tag{1.17}$$

it is clear that (1.14) admits a feasible symmetric positive semidefinite stabilizing solution if and only if $\gamma > \gamma_f$.

If this is the case, there exists a square J-spectral factor of $\Psi(z)$ having all zeros in the open unit disk and satisfying the stability conditions described in [3, Theorem 3.1], ensuring the existence of a proper stable filter $F(z)$.

On the other hand, it is also well known that there exists a value γ_0 such that for $\gamma < \gamma_0$, the ARE (1.14) does not admit any symmetric solution satisfying the feasibility condition and hence $\Psi(z)$ does not admit any minimal square J-spectral factor, namely, a J-spectral factor having the last possible McMillan degree. The constant γ_0 has the following interpretation from the J-spectral factorization point of view. A necessary condition for the existence of a square spectral factor is that $\Psi(e^{j\omega})$ has p positive eigenvalues and l negative eigenvalues for all $\omega \in (-\pi, \pi]$. Since under the present assumptions $\Psi_{11}(e^{j\omega}) > 0$ for all $\omega \in (-\pi, \pi]$, this condition is equivalent to the negative definiteness of the Schur complement

$$S(z) := \Psi_{22}(z) - \Psi_{12}\tilde{}(z)[\Psi_{11}(z)]^{-1}\Psi_{12}(z) \tag{1.18}$$

on the unit circle, i.e., to

$$S(e^{j\omega}) < 0, \qquad \omega \in (-\pi, \pi]. \tag{1.19}$$

By employing (1.11), we may write $S(z)$ as $S(z) = W(z) - \gamma^2 I$ with

$$\begin{aligned} W(z) &:= H_2(z)H_2\tilde{}(z) - H_2(z)H_1\tilde{}(z)[H_1(z)H_1\tilde{}(z)]^{-1}H_1(z)H_2\tilde{}(z) \\ &= H_2(z)[I - H_1\tilde{}(z)[H_1(z)H_1\tilde{}(z)]^{-1}H_1(z)]H_2\tilde{}(z). \end{aligned} \tag{1.20}$$

It is now easy to see that (1.19) is satified for γ sufficiently large: γ_0 is the infimum of the values of γ for which (1.19) is satisfied, i.e., γ_0^2 is the $\mathcal{L}_\infty$ norm of $W(z)$:

$$\gamma_0^2 = \|W(z)\|_\infty. \tag{1.21}$$

For a general overview on the connections between ARE and spectral and J-spectral factorization see [11, 12, 17, 1, 15] and references therein.

Remark 1.1. From the estimation problem point of view, the constant γ_0 is the infimum of the values of γ for which there exists a fixed-lag smoother (for some preview horizon N) achieving the attenuation level γ. To see this, first recall that under

the detectability assumption of the pair (A, C), without any loss of generality, one can assume from the very beginning the stability of $H_1(z)$ and $H_2(z)$ and transform the given problem in a unilateral model matching problem in a new stable free filter parameter [21]. Now, let $V_1(z)$ be a square spectral factor of $H_1(z)$. A simple computation shows that

$$\|F(z)H_1(z) - H_2(z)\|_\infty^2 = \|F(z)V_1(z) - H_2(z)H_1^\sim(z)V_1^\sim(z)^{-1}\|_\infty^2 + \|W(z)\|_\infty. \tag{1.22}$$

Hence, the attenuation level cannot be lower than γ_0 and this infimum value is achieved by the acausal estimator

$$F_\infty(z) = H_2(z)H_1^\sim(z)\left(H_1(z)H_1^\sim(z)\right)^{-1} = \sum_{i=-\infty}^{\infty} F_i z^i. \tag{1.23}$$

Now the question arises of how to construct a stable fixed-lag smoother when an attenuation level $\gamma > \gamma_0$ is prescribed. This problem can be solved directly from expression (1.22) by taking a finite length expansion of the anticausal part of $F_\infty(z)$ given by (1.23). Indeed, let

$$F_N(z) = F_\infty(z) - \sum_{i=N+1}^{\infty} F_i z^i.$$

If $F_N(z)$ is the (N steps acausal, yet stable) estimator so constructed, then

$$\begin{aligned}\|F_N(z)H_1(z) - H_2(z)\|_\infty^2 &\le \gamma_0^2 + \left\| \sum_{i=N+1}^{\infty} F_i z^i V_1(z) \right\|_\infty^2 \\ &\le \gamma_0^2 + \left(\sum_{i=N+1}^{\infty} \|F_i\| \right)^2 \|H_1(z)\|_\infty^2\end{aligned}$$

can be rendered arbitrarily close to γ_0^2 by selecting a sufficiently large N. The estimator can be realized from the relevant factor following the results in [3] or directly from the solution of a suitable Riccati equation as described in [20].

1.3. Contribution of the paper. In this paper, we investigate the existence of a feasible stabilizing solution of (1.14) and the existence and the construction of a minimum phase square J-spectral factor $\Omega(z)$ for values of γ in the interval $(\gamma_0, \gamma_f]$. Clearly, for such values of γ the filter cannot exist, but there exists a fixed-lag smoother associated with a certain preview horizon length N achieving the desired attenuation level. As shown in [5], the computation of the stabilizing solution Δ of (1.14) and of the corresponding J-spectral factor $\Omega(z)$ are crucial steps to obtain efficiently the minimum-lag smoothing filter achieving the desired attenuation level. Notably, the solution Δ can be directly used to initialize an iterative algorithm to work out a minimum-lag central smoother; see [13, 2].

The contribution of the present paper is to prove that for $\gamma \in (\gamma_0, \gamma_f]$, equation (1.14) still admits a stabilizing feasible solution except for a finite number (at most $2[n + \mathrm{rank}(M)]$) of values of γ. This result generalizes a result in [4] where it was assumed that $DD^\top = I$, $M = 0$, and $DB^\top = 0$. While removing the assumption $DD^\top = I$ impairs only slightly longer formulas, the presence of the matrix M gives rise to a much more difficult problem. Also, the presence of the matrix M implies that

having assumed (1.7e) we cannot assume $DB^\top = 0$ (only one of the two conditions (1.7e) and $DB^\top = 0$ can be assumed without loss of generality) and this adds some other technical difficulties to the problem. On the other hand the estimation problem becomes much more interesting and of practical importance if we consider the presence of the matrix M. Indeed, in this case, not only can we attack deconvolution problems for the importance of which we refer to [9] and references therein, but we can, more generally, address the problem of estimating an arbitrary linear combination of input and state vectors; see the discussion in [7] for the practical importance of the latter problem.

1.4. Paper organization. In section 2 we analyze some auxiliary spectral factorization problems and the corresponding AREs. In section 3 we show that for generic values of γ, the antistabilizing solution of an ARE considered in section 2 is nonsingular. Such preliminary results are employed in section 4, where we prove our main result on the existence of a stabilizing feasible solution of (1.14). In section 5 we explicitly derive a minimum phase J-spectral factor of $\Psi(z)$. In section 6 we analyze some of the issues arising in correspondence of the values of γ for which we cannot guarantee the existence of the stabilizing solution of (1.14) and we discuss some peculiar aspects of the J-spectral factorization.

2. Preliminary spectral factorizations and auxiliary AREs. In this section we compute a state-space realization of a spectral factor $T(z)$ of $W(z)$ defined in (1.20) and a spectral factorization of $-S(z) = \gamma^2 I - W(z)$. These factorizations are related to the stabilizing solutions of a pair of coupled AREs. These solutions will be used to compute the solution (and hence to prove constructively its existence) of a third ARE that is strictly related to (1.14). This procedure will be carried over through various steps.

2.1. Computation of a square spectral factor of $H_1(z)H_1^\sim(z)$. Consider the following standard filtering ARE:

$$P = APA^\top + BB^\top - (APC^\top + BD^\top)(DD^\top + CPC^\top)^{-1}(CPA^\top + DB^\top). \tag{2.1}$$

By assumptions (1.7b) and (1.7c), it admits a positive definite stabilizing solution $P_s = P_s^\top > 0$. Such a solution corresponds to spectral factorization of $H_1(z)H_1^\sim(z)$. In fact, by defining

$$D_1 := (DD^\top + CP_sC^\top)^{1/2}, \qquad G_1 := (AP_sC^\top + BD^\top)D_1^{-1}, \tag{2.2}$$

it is easy to check (see [12]) that

$$H_1(z)H_1^\sim(z) = V_1(z)V_1^\sim(z) \tag{2.3}$$

with

$$V_1(z) := C(zI - A)^{-1}G_1 + D_1. \tag{2.4}$$

2.2. Computation of $V_1(z)^{-1}H_1(z)$. First notice that

$$V_1(z)^{-1} = D_1^{-1} - D_1^{-1}C(zI - \Gamma)^{-1}G_1D_1^{-1} \tag{2.5}$$

with

$$\Gamma := A - G_1D_1^{-1}C \tag{2.6}$$

being the closed-loop matrix corresponding to the stabilizing solution P_s of the ARE (2.1) (so that all the eigenvalues of Γ lay in the open unit disk).

Then we have

$$\begin{aligned} V_2(z) &:= V_1(z)^{-1}H_1(z) \\ &= D_1^{-1}C(zI-A)^{-1}B - D_1^{-1}C(zI-\Gamma)^{-1}G_1D_1^{-1}D + D_1^{-1}D \\ &\quad -D_1^{-1}C(zI-\Gamma)^{-1}\underbrace{G_1D_1^{-1}C}_{(zI-\Gamma)-(zI-A)}(zI-A)^{-1}B \\ &= D_1^{-1}C(zI-\Gamma)^{-1}(B-G_1D_1^{-1}D) + D_1^{-1}D \\ &= C_2(zI-\Gamma)^{-1}G_2 + D_2 \end{aligned} \tag{2.7}$$

with

$$C_2 := D_1^{-1}C, \qquad G_2 := B - G_1D_1^{-1}D, \qquad D_2 := D_1^{-1}D. \tag{2.8}$$

2.3. Computation of $V_2(z)H_2^{\sim}(z)$. We have

$$\begin{aligned} V_2(z)H_2^{\sim}(z) &= [C_2(zI-\Gamma)^{-1}G_2 + D_2][B^\top(z^{-1}I-A^\top)^{-1}L^\top + M^\top] \\ &= C_2(zI-\Gamma)^{-1}G_2M^\top + D_2B^\top(z^{-1}I-A^\top)^{-1}L^\top + D_2M^\top \\ &\quad + C_2(zI-\Gamma)^{-1}G_2B^\top(z^{-1}I-A^\top)^{-1}L^\top. \end{aligned} \tag{2.9}$$

Taking the ARE (2.1) into account, it is easy to check that

$$G_2B^\top = P_s - \Gamma P_sA^\top, \tag{2.10}$$

which may be easily rewritten in the form

$$G_2B^\top = (zI-\Gamma)P_s(z^{-1}I-A^\top) + (zI-\Gamma)P_sA^\top + \Gamma P_s(z^{-1}I-A^\top). \tag{2.11}$$

Plugging this expression in (2.9) we get

$$V_2(z)H_2^{\sim}(z) = C_2(zI-\Gamma)^{-1}G_3 + G_1^\top(z^{-1}I-A^\top)^{-1}L^\top + D_3 \tag{2.12}$$

with $G_3 := \Gamma P_sL^\top + G_2M^\top$ and $D_3 := C_2P_sL^\top + D_2M^\top$ or, taking into account that $DM^\top = 0$,

$$G_3 = \Gamma P_sL^\top + BM^\top, \qquad D_3 = C_2P_sL^\top \tag{2.13}$$

2.4. Computation of $H_2(z)H_1^{\sim}(z)[H_1(z)H_1^{\sim}(z)]^{-1}H_1(z)H_2^{\sim}(z)$. It is clear that $H_2(z)H_1^{\sim}(z)[H_1(z)H_1^{\sim}(z)]^{-1}H_1(z)H_2^{\sim}(z) = H_2(z)V_2^{\sim}(z)V_2(z)H_2^{\sim}(z)$. We have

$$\begin{aligned} H_2(z)V_2^{\sim}(z)V_2(z)H_2^{\sim}(z) &= D_3^\top D_3 + G_3^\top(z^{-1}I-\Gamma^\top)^{-1}C_2^\top C_2(zI-\Gamma)^{-1}G_3 \\ &\quad + D_3^\top C_2(zI-\Gamma)^{-1}G_3 + D_3^\top G_1^\top(z^{-1}I-A^\top)^{-1}L^\top \\ &\quad + [D_3^\top C_2(zI-\Gamma)^{-1}G_3 + D_3^\top G_1^\top(z^{-1}I-A^\top)^{-1}L^\top]^{\sim} \\ &\quad + L(zI-A)^{-1}G_1C_2(zI-\Gamma)^{-1}G_3 \\ &\quad + [L(zI-A)^{-1}G_1C_2(zI-\Gamma)^{-1}G_3]^{\sim} \\ &\quad + L(zI-A)^{-1}G_1G_1^\top(z^{-1}I-A^\top)^{-1}L^\top. \end{aligned} \tag{2.14}$$

The last three terms of such expression may be expanded as follows: first taking into account that

$$G_1C_2 = G_1D_1^{-1}C = A - \Gamma = (zI - \Gamma) - (zI - A) \tag{2.15}$$

we have

$$L(zI - A)^{-1}G_1C_2(zI - \Gamma)^{-1}G_3 = L(zI - A)^{-1}G_3 - L(zI - \Gamma)^{-1}G_3. \tag{2.16}$$

Second, taking into account the ARE (2.1), we have

$$\begin{aligned} G_1G_1^\top = AP_sA^\top - P_s + BB^\top = BB^\top - (zI - A)P_s(z^{-1}I - A^\top) - (zI - A)P_sA^\top \\ -AP_s(z^{-1}I - A^\top) \end{aligned} \tag{2.17}$$

so that

$$\begin{aligned} L(zI-A)^{-1}G_1G_1^\top(z^{-1}I-A^\top)^{-1}L^\top \\ = L(zI - A)^{-1}BB^\top(z^{-1}I - A^\top)^{-1}L^\top - LP_sL^\top \\ - L(zI - A)^{-1}AP_sL^\top - LP_sA^\top(z^{-1}I - A^\top)^{-1}L^\top. \end{aligned} \tag{2.18}$$

2.5. Computation of $W(z)$. From the definition (1.20) of $W(z)$ it follows that

$$W(z) = H_2(z)\tilde{H_2}(z) - H_2(z)\tilde{V_2}(z)V_2(z)\tilde{H_2}(z). \tag{2.19}$$

Moreover, by computing $H_2(z)\tilde{H_2}(z)$, we get the following expression:

$$\begin{aligned} H_2(z)\tilde{H_2}(z) = L(zI - A)^{-1}BB^\top(z^{-1}I - A^\top)^{-1}L^\top + MM^\top \\ + L(zI - A)^{-1}BM^\top + [L(zI - A)^{-1}BM^\top]\tilde{}. \end{aligned} \tag{2.20}$$

By subtracting (2.14) from (2.20), taking (2.16), (2.18), and (2.19) into account, we get

$$\begin{aligned} W(z) = (L - D_3^\top C_2)(zI - \Gamma)^{-1}G_3 + [(L - D_3^\top C_2)(zI - \Gamma)^{-1}G_3]\tilde{} \\ - G_3^\top(z^{-1}I - \Gamma^\top)^{-1}C_2^\top C_2(zI - \Gamma)^{-1}G_3 + LP_sL^\top + MM^\top - D_3^\top D_3 \\ + L(zI - A)^{-1}(BM^\top - G_1D_3 + AP_sL^\top - G_3) \\ + [L(zI - A)^{-1}(BM^\top - G_1D_3 + AP_sL^\top - G_3)]\tilde{}. \end{aligned} \tag{2.21}$$

We now prove that $BM^\top - G_1D_3 + AP_sL^\top - G_3 = 0$, so that the last two terms of the latter expression vanish. We have

$$\begin{aligned} BM^\top - G_1D_3 + AP_sL^\top - G_3 \\ = BM^\top - \underbrace{G_1C_2}_{A-\Gamma}P_sL^\top + AP_sL^\top - \Gamma P_sL^\top - BM^\top = 0. \end{aligned} \tag{2.22}$$

2.6. Spectral factorization of $W(z)$. To factorize $W(z)$ we first need to rewrite the term

$$G_3^\top(z^{-1}I - \Gamma^\top)^{-1}C_2^\top C_2(zI - \Gamma)^{-1}G_3.$$

To this aim we derive a new expression for $C_2^\top C_2$. Standard computations show that the solution P_s of the ARE (2.1) satisfies the following identity [15, p. 271]:

$$P_s = FP_sF^\top - FP_sC^\top(CP_sC^\top + DD^\top)^{-1}CP_sF^\top + B_1B_1^\top, \tag{2.23}$$

where F and B_1 are defined in (1.6). We also have

$$\begin{aligned}\Gamma &= A - G_1 D_1^{-1} C = F + [BD^\top (DD^\top)^{-1} D_1^2 - AP_s C^\top - BD^\top] D_1^{-1} C_2 \\ &= F - FP_s C^\top D_1^{-1} C_2 = F - FP_s C_2^\top C_2 \\ &= F(I - P_s C_2^\top C_2) \end{aligned} \tag{2.24}$$

so that (2.23) may be rewritten in the form

$$P_s - B_1 B_1^\top = \Gamma P_s F^\top = F P_s \Gamma^\top. \tag{2.25}$$

From $DD^\top > 0$ and $P_s > 0$ it easily follows that $(I - P_s C_2^\top C_2) = P_s(P_s^{-1} - C^\top (CP_s C^\top + DD^\top)^{-1} C)$ is nonsingular (in fact, $(P_s^{-1} - C^\top (CP_s C^\top + DD^\top)^{-1} C) > 0$). Thus Γ, and hence $FP_s\Gamma^\top$, are nonsingular so that the left-hand side of (2.25) is nonsingular as well, and we have

$$(P_s - B_1 B_1^\top)^{-1} = \Gamma^{-\top} P_s^{-1} F^{-1}, \tag{2.26}$$

which may be rewritten as

$$P_s^{-1} = \Gamma^\top (P_s - B_1 B_1^\top)^{-1} F, \tag{2.27}$$

and, by multiplying on the right side by $(I - P_s C_2^\top C_2)$,

$$P_s^{-1}(I - P_s C_2^\top C_2) = \Gamma^\top (P_s - B_1 B_1^\top)^{-1} \Gamma. \tag{2.28}$$

Therefore

$$C_2^\top C_2 = P_s^{-1} - \Gamma^\top (P_s - B_1 B_1^\top)^{-1} \Gamma \tag{2.29a}$$

$$= P_s^{-1} - \Gamma^\top P_s^{-1} \Gamma - \Gamma^\top P_s^{-1} B_1 V^{-1} B_1^\top P_s^{-1} \Gamma \tag{2.29b}$$

$$\begin{aligned} &= (z^{-1} I - \Gamma^\top) P_s^{-1} (zI - \Gamma) + \Gamma^\top P_s^{-1} (zI - \Gamma) + (z^{-1} I - \Gamma^\top) P_s^{-1} \Gamma \\ &\quad - \Gamma^\top P_s^{-1} B_1 V^{-1} B_1^\top P_s^{-1} \Gamma, \end{aligned} \tag{2.29c}$$

where

$$V := I - B_1^\top P_s^{-1} B_1. \tag{2.30}$$

Notice that

$$P_s^{-1} - C_2^\top C_2 = P_s^{-1} - C^\top (DD^\top + CP_s C^\top)^{-1} C = P_s^{-1} [P_s^{-1} + C^\top (DD^\top)^{-1} C]^{-1} P_s^{-1} > 0 \tag{2.31}$$

so that from (2.29a) it follows that $P_s - B_1 B_1^\top > 0$, which, together with the positive definiteness of P_s, implies that V is positive definite as well.

Plugging in the expression in the right-hand side of (2.29c) in place of $C_2^\top C_2$ in the third term of (2.21), we get

$$\begin{aligned} W(z) = {} & (L - D_3^\top C_2 - G_3^\top P_s^{-1} \Gamma)(zI - \Gamma)^{-1} G_3 \\ & + [(L - D_3^\top C_2 - G_3^\top P_s^{-1} \Gamma)(zI - \Gamma)^{-1} G_3]^\sim \\ & + G_3^\top (z^{-1} I - \Gamma^\top)^{-1} \Gamma^\top P_s^{-1} B_1 V^{-1} B_1^\top P_s^{-1} \Gamma (zI - \Gamma)^{-1} G_3 \\ & + L P_s L^\top + M M^\top - D_3^\top D_3 - G_3^\top P_s^{-1} G_3. \end{aligned} \tag{2.32}$$

Now we are ready to prove that the following spectral factorization of $W(z)$ holds:

$$W(z) = T(z) T^\sim(z), \tag{2.33}$$

where

$$T(z) := [-G_3^\top (z^{-1}I - \Gamma^\top)^{-1}\Gamma^\top P_s^{-1}B_1 - G_3^\top P_s^{-1}B_1 + M]V^{-1/2}. \tag{2.34}$$

To this aim it is sufficient to show that

$$\begin{aligned}(M - G_3^\top P_s^{-1}B_1)V^{-1}(M - G_3^\top P_s^{-1}B_1)^\top \\ - (LP_sL^\top + MM^\top - D_3^\top D_3 - G_3^\top P_s^{-1}G_3) = 0\end{aligned} \tag{2.35}$$

and

$$\Gamma^\top P_s^{-1}B_1V^{-1}(M - G_3^\top P_s^{-1}B_1)^\top + (L - D_3^\top C_2 - G_3^\top P_s^{-1}\Gamma)^\top = 0. \tag{2.36}$$

As for (2.35) notice that

$$M - G_3^\top P_s^{-1}B_1 = M - LP_s\Gamma^\top P_s^{-1}B_1 - MB_1^\top P_s^{-1}B_1 = MV - LP_s\Gamma^\top P_s^{-1}B_1 \tag{2.37}$$

so that, taking (1.8) into account, we may expand the left-hand side of (2.35) as

$$\begin{aligned}&MVM^\top - MB_1^\top P_s^{-1}\Gamma P_sL^\top - (MB_1^\top P_s^{-1}\Gamma P_sL^\top)^\top \\ &+ LP_s\Gamma^\top P_s^{-1}B_1V^{-1}B_1^\top P_s^{-1}\Gamma P_sL^\top - LP_sL^\top - MM^\top \\ &+ LP_sC_2^\top C_2P_sL^\top + LP_s\Gamma^\top P_s^{-1}\Gamma P_sL^\top + MB_1^\top P_s^{-1}B_1M^\top \\ &+ MB_1^\top P_s^{-1}\Gamma P_sL^\top + (MB_1^\top P_s^{-1}\Gamma P_sL^\top)^\top \\ &= LP_s(\Gamma^\top P_s^{-1}B_1V^{-1}B_1^\top P_s^{-1}\Gamma - P_s^{-1} + C_2^\top C_2 + \Gamma^\top P_s^{-1}\Gamma)P_sL^\top = 0,\end{aligned} \tag{2.38}$$

where the latter equality follows from the expression (2.29b) for $C_2^\top C_2$.

Similarly, we may expand the left-hand side of (2.36) as

$$\begin{aligned}&L^\top - C_2^\top C_2P_sL^\top - \Gamma^\top P_s^{-1}\Gamma P_sL^\top - \Gamma^\top P_s^{-1}B_1V^{-1}B_1^\top P_s^{-1}\Gamma P_sL^\top \\ &= (P_s^{-1} - C_2^\top C_2 - \Gamma^\top P_s^{-1}\Gamma - \Gamma^\top P_s^{-1}B_1V^{-1}B_1^\top P_s^{-1}\Gamma)P_sL^\top = 0.\end{aligned} \tag{2.39}$$

We now rewrite $T(z)$ in a more convenient form. Notice that

$$(z^{-1}I - \Gamma^\top)^{-1} = -\Gamma^{-\top} - \Gamma^{-\top}(zI - \Gamma^{-\top})^{-1}\Gamma^{-\top}$$

so that

$$\begin{aligned}T(z) &= [G_3^\top\Gamma^{-\top}(zI - \Gamma^{-\top})^{-1}P_s^{-1}B_1 + M]V^{-1/2} \\ &= [G_3^\top\Gamma^{-\top}P_s^{-1}(zI - P_s\Gamma^{-\top}P_s^{-1})^{-1}B_1 + M]V^{-1/2} \\ &= L_1(zI - F_a)^{-1}B_1V^{-1/2} + MV^{-1/2}\end{aligned} \tag{2.40}$$

with

$$L_1 := G_3^\top\Gamma^{-\top}P_s^{-1} = L + MB_1^\top\Gamma^{-\top}P_s^{-1}, \qquad F_a := P_s\Gamma^{-\top}P_s^{-1}. \tag{2.41}$$

Notice that since Γ is stable, F_a is antistable. As a direct consequence of (2.26) we have the following relation that will be useful in what follows:

$$F_a = (I - B_1B_1^\top P_s^{-1})^{-1}F. \tag{2.42}$$

2.7. Spectral factorization of $-S(z)$. We have

$$-S(z) = \gamma^2 I - W(z) = \gamma^2 I - T(z)T^{\sim}(z).$$

As we have done for the factorization in section 2.1 consider the following ARE for the spectral factorization of $-S(z)$:

$$X = F_a X F_a^\top - B_1 V^{-1} B_1^\top - (F_a X L_1^\top - B_1 V^{-1} M^\top)(R + L_1 X L_1^\top)^{-1}(L_1 X F_a^\top - M V^{-1} B_1^\top) \tag{2.43}$$

with

$$R := \gamma^2 I - M V^{-1} M^\top. \tag{2.44}$$

In view of (1.21), we have $\|T(z)\|_\infty < \gamma$ for any $\gamma > \gamma_0$ so that as a consequence of a generalized version of the discrete-time bounded real lemma (see [8, Theorem 2.1]), (2.43) admits a stabilizing solution, namely, a solution $X_s = X_s^\top$ such that

$$\Gamma_a := F_a - (F_a X_s L_1^\top - B_1 V^{-1} M^\top)(R + L_1 X_s L_1^\top)^{-1} L_1 \tag{2.45}$$

is a stability matrix; moreover $(R + L_1 X_s L_1^\top) > 0$ and the function

$$T_1(z) := [I + L_1(zI - F_a)^{-1}(F_a X_s L_1^\top - B_1 V^{-1} M^\top)(R + L_1 X_s L_1^\top)^{-1}](R + L_1 X_s L_1^\top)^{1/2} \tag{2.46}$$

is a square spectral factor of $-S(z) = \gamma^2 I_l - W(z)$, namely, it is a square matrix function such that $-S(z) = T_1(z)T_1^{\sim}(z)$. Moreover, the numerator matrix of $T_1(z)$ is given by (2.45) and hence $[T_1(z)]^{-1}$ is stable.

We now show that X_s is positive definite. In fact, since $(R + L_1 X_s L_1^\top) > 0$, we have

$$X_s \le F_a X_s F_a^\top - B_1 V^{-1} B_1^\top. \tag{2.47}$$

But (F, B_1) is, by assumption, reachable, (F_a, B_1) is such (because F_a is obtained from F by state feedback as it is apparent from (2.42)), and then also $(F_a, B_1 V^{-1/2})$ is reachable. Therefore, since F_a is antistable, a standard Lyapunov argument, shows that $X_s > 0$.

2.8. The ARE for $P_s^{-1} - X_s^{-1}$. Define

$$R_1 := I - \frac{M^\top M}{\gamma^2} \tag{2.48}$$

and

$$F_y := A - B J_0' J_1^{-1} H_0, \tag{2.49}$$

where J_0, J_1, H_0 are defined in (1.13). Notice that F_y is well defined if and only if J_1 is nonsingular or, equivalently, if and only if R_1 is nonsingular. Let

$$\mathcal{G}_1 := \{\gamma > \gamma_0 : \text{ at least one of the matrices } R,\ R_1,\ \text{and } F_y,\ \text{is singular}\}. \tag{2.50}$$

Remark 2.1. Notice that since F_y may be written in the form

$$F_y = F - B_1 M^\top (M M^\top - \gamma^2 I)^{-1} L, \tag{2.51}$$

where F is nonsingular, $\mathcal{G}_1$ contains a finite number of points. More precisely, it contains, at most, $n + 2\,\mathrm{rank}(M)$ points.

We consider now the following ARE:

$$\begin{aligned} Y = F^\top Y F + \left(F^\top Y B_1 + \frac{L^\top M}{\gamma^2}\right)(R_1 - B_1^\top Y B_1)^{-1}\left(B_1^\top Y F + \frac{M^\top L}{\gamma^2}\right) \\ - C^\top (DD^\top)^{-1} C + \frac{L^\top L}{\gamma^2} \end{aligned} \tag{2.52}$$

and prove that it admits a solution that can be explicitely computed. To this aim we need a technical lemma.

LEMMA 2.1. *Let $\gamma > \gamma_0$ and $\gamma \notin \mathcal{G}_1$ and define*

$$F_1 := F_a + B_1 V^{-1} M^\top R^{-1} L_1. \tag{2.53}$$

Then the following relation holds:

$$X_s^{-1} = F_1^\top X_s^{-1}(I - B_1 R_1^{-1} B_1^\top (P_s^{-1} - X_s^{-1}))^{-1} F_y - L_1^\top R^{-1} L_1. \tag{2.54}$$

Proof. First we establish the following useful identity:

$$F_y = (I - B_1 R_1^{-1} B_1^\top P_s^{-1}) F_1. \tag{2.55}$$

To show identity (2.55), we write F_1 in the form

$$\begin{aligned} F_1 &= (I + B_1 V^{-1} M^\top R^{-1} M B_1^\top P_s^{-1}) F_a + B_1 V^{-1} M^\top R^{-1} L \\ &= (I + B_1 V^{-1} M^\top R^{-1} M B_1^\top P_s^{-1})(I - B_1 B_1^\top P_s^{-1})^{-1} F + B_1 V^{-1} M^\top R^{-1} L, \end{aligned} \tag{2.56}$$

which gives

$$\begin{aligned} &(I - B_1 R_1^{-1} B_1^\top P_s^{-1}) F_1 \\ &= (I - B_1 R_1^{-1} B_1^\top P_s^{-1})(I + B_1 V^{-1} M^\top R^{-1} M B_1^\top P_s^{-1})(I - B_1 B_1^\top P_s^{-1})^{-1} F \\ &\quad + (I - B_1 R_1^{-1} B_1^\top P_s^{-1}) B_1 V^{-1} M^\top R^{-1} L, \end{aligned} \tag{2.57}$$

so that, taking (2.51) into account, it is sufficient to prove that

$$(I - B_1 R_1^{-1} B_1^\top P_s^{-1})(I + B_1 V^{-1} M^\top R^{-1} M B_1^\top P_s^{-1})(I - B_1 B_1^\top P_s^{-1})^{-1} = I \tag{2.58}$$

and

$$(I - B_1 R_1^{-1} B_1^\top P_s^{-1}) B_1 V^{-1} M^\top R^{-1} = -B_1 M^\top (M M^\top - \gamma^2 I)^{-1}. \tag{2.59}$$

As for (2.58), we have

$$\begin{aligned} &(I - B_1 R_1^{-1} B_1^\top P_s^{-1})(I + B_1 V^{-1} M^\top R^{-1} M B_1^\top P_s^{-1})(I - B_1 B_1^\top P_s^{-1})^{-1} \\ &= (I - B_1 R_1^{-1} B_1^\top P_s^{-1})(I - B_1 B_1^\top P_s^{-1} + B_1 (I + V^{-1} M^\top R^{-1} M) B_1^\top P_s^{-1})(I - B_1 B_1^\top P_s^{-1})^{-1} \\ &= (I - B_1 R_1^{-1} B_1^\top P_s^{-1})\left[I - B_1 B_1^\top P_s^{-1} + B_1 \left(V - \frac{M^\top M}{\gamma^2}\right)^{-1} V B_1^\top P_s^{-1}\right](I - B_1 B_1^\top P_s^{-1})^{-1} \\ &= (I - B_1 R_1^{-1} B_1^\top P_s^{-1})\left[I + B_1 \left(V - \frac{M^\top M}{\gamma^2}\right)^{-1} V B_1^\top P_s^{-1}(I - B_1 B_1^\top P_s^{-1})^{-1}\right] \\ &= (I - B_1 R_1^{-1} B_1^\top P_s^{-1})\left[I + B_1 \left(V - \frac{M^\top M}{\gamma^2}\right)^{-1} B_1^\top P_s^{-1}\right] \\ &= (I - B_1 R_1^{-1} B_1^\top P_s^{-1})(I - B_1 R_1^{-1} B_1^\top P_s^{-1})^{-1} = I. \end{aligned} \tag{2.60}$$

As for (2.59), we have

$$\begin{aligned}(I - B_1R_1^{-1}B_1^\top P_s^{-1})B_1V^{-1}M^\top R^{-1} &= B_1(I - R_1^{-1}B_1^\top P_s^{-1}B_1)V^{-1}M^\top R^{-1}\\ &= B_1R_1^{-1}\underbrace{(R_1 - B_1^\top P_s^{-1}B_1)}_{V-\frac{M^\top M}{\gamma^2}}V^{-1}M^\top R^{-1}\\ &= B_1R_1^{-1}M^\top\left(I - \frac{MV^{-1}M^\top}{\gamma^2}\right)R^{-1}\\ &= -B_1M^\top(MM^\top - \gamma^2 I)^{-1}. \qquad (2.61)\end{aligned}$$

Since X_s is a solution of the ARE (2.43), as a direct consequence of standard equivalence of Riccati equations [15, p. 271], we have the following identity:

$$X_s = F_1X_sF_1^\top - F_1X_sL_1^\top(L_1X_sL_1^\top + R)^{-1}L_1X_sF_1^\top - B_1\left(V - \frac{M^\top M}{\gamma^2}\right)^{-1}B_1^\top \qquad (2.62)$$

with F_1 being defined in (2.53). The closed-loop matrix may thus be written in the form

$$\Gamma_a = F_1 - F_1X_sL_1^\top(L_1X_sL_1^\top + R)^{-1}L_1 = F_1(I + X_sL_1^\top R^{-1}L_1)^{-1}. \qquad (2.63)$$

Notice that $V - \frac{M^\top M}{\gamma^2}$ is nonsingular because V and R are such. In the same way, $X_s^{-1}+L_1^\top R^{-1}L_1$, and hence $(I+X_sL_1^\top R^{-1}L_1)$, are nonsingular because X_s, $L_1X_sL_1^\top + R$, and R are such. From (2.62) we easily get

$$X_s + B_1\left(V - \frac{M^\top M}{\gamma^2}\right)^{-1}B_1^\top = F_1(X_s - X_sL_1^\top(L_1X_sL_1^\top + R)^{-1}L_1X_s)F_1^\top. \qquad (2.64)$$

As an immediate consequence of (2.55) we have that F_1 is nonsingular, so that the right-hand side of (2.64) (which may be written in the form $F_1(I+X_sL_1^\top R^{-1}L_1)^{-1}F_1^\top$) is nonsingular. Hence $X_s + B_1(V - \frac{M^\top M}{\gamma^2})^{-1}B_1^\top$ is nonsingular as well and then $\frac{M^\top M}{\gamma^2} - V - B_1^\top X_s^{-1}B_1 = -\left(R_1 - B_1^\top(P_s^{-1} - X_s^{-1})B_1\right)$ is such. (Actually, a more detailed analysis allows to conclude that $\left(R_1 - B_1^\top(P_s^{-1} - X_s^{-1})B_1\right) > 0$.) Then we have

$$\left[X_s + B_1\left(V - \frac{M^\top M}{\gamma^2}\right)^{-1}B_1^\top\right]^{-1} = [F_1(X_s - X_sL_1^\top(L_1X_sL_1^\top + R)^{-1}L_1X_s)F_1^\top]^{-1}, \qquad (2.65)$$

which yields

$$X_s^{-1} + X_s^{-1}B_1\left(\frac{M^\top M}{\gamma^2} - V - B_1^\top X_s^{-1}B_1\right)^{-1}B_1^\top X_s^{-1} = F_1^{-\top}(X_s^{-1} + L_1^\top R^{-1}L_1)F_1^{-1} \qquad (2.66)$$

and hence

$$X_s^{-1} = F_1^\top X_s^{-1}F_1 - F_1^\top X_s^{-1}B_1\left(R_1 - B_1^\top(P_s^{-1} - X_s^{-1})B_1\right)^{-1}B_1^\top X_s^{-1}F_1 - L_1^\top R^{-1}L_1 \qquad (2.67a)$$

$$= F_1^\top X_s^{-1}(I - B_1R_1^{-1}B_1^\top(P_s^{-1} - X_s^{-1}))^{-1}(I - B_1R_1^{-1}B_1^\top P_s^{-1})F_1 - L_1^\top R^{-1}L_1. \qquad (2.67b)$$

Finally, by plugging (2.55) in (2.67b) we obtain (2.54). □

PROPOSITION 2.1. *Let $\gamma > \gamma_0$ and $\gamma \notin \mathcal{G}_1$. Then the ARE* (2.52) *admits the (necessarily unique) antistabilizing solution. In fact, such solution is given by*[3]

$$Y_a := P_s^{-1} - X_s^{-1}. \tag{2.68}$$

Proof. We have to show that when $Y = Y_a$, (2.52) is an identity and that all the eigenvalues of

$$\Gamma_y := F - B_1(B_1^\top Y_a B_1 - R_1)^{-1}\left(B_1^\top Y_a F + \frac{M^\top L}{\gamma^2}\right) \tag{2.69}$$

lay in $\{z \in \mathbb{C} : \ |z| > 1\}$.

We first prove that Y_a is a solution of the ARE

$$Y = F_y^\top Y F_y + F_y^\top Y B_1(R_1 - B_1^\top Y B_1)^{-1}B_1^\top Y F_y - H_0^\top J_1^{-1} H_0. \tag{2.70}$$

From (2.23) we get

$$(P_s - B_1 B_1^\top)^{-1} = F^{-\top}(P_s - P_s C^\top (C P_s C^\top + DD^\top)^{-1} C P_s)^{-1} F^{-1}, \tag{2.71}$$

which, using the same procedure that lead to (2.67a), yields

$$P_s^{-1} = F^\top P_s^{-1} F - F^\top P_s^{-1} B_1 (B_1^\top P_s^{-1} B_1 - I)^{-1} B_1^\top P_s^{-1} F - C^\top (DD^\top)^{-1} C, \tag{2.72}$$

which, taking (2.42) into account, may be also written in the form

$$P_s^{-1} = F_a^\top P_s^{-1} F - C^\top (DD^\top)^{-1} C \tag{2.73}$$

and, finally, in the form

$$P_s^{-1} = F_1^\top P_s^{-1} F_y + F_a^\top P_s^{-1}(F - F_y) - (F_1 - F_a)^\top P_s^{-1} F_y - C^\top (DD^\top)^{-1} C. \tag{2.74}$$

By subtracting (2.54) from (2.74) we now get

$$Y_a = F_1^\top [P_s^{-1} - X_s^{-1}(I - B_1 R_1^{-1} B_1^\top Y_a)^{-1}] F_y + R_2 \tag{2.75}$$

with

$$R_2 := F_a^\top P_s^{-1}(F - F_y) - (F_1 - F_a)^\top P_s^{-1} F_y - C^\top (DD^\top)^{-1} C + L_1^\top R^{-1} L_1. \tag{2.76}$$

We may rewrite (2.75) as follows:

$$\begin{aligned} Y_a &= F_1^\top [(Y_a - P_s^{-1} B_1 R_1^{-1} B_1^\top Y_a)(I - B_1 R_1^{-1} B_1^\top Y_a)^{-1}] F_y + R_2 \\ &= F_1^\top (I - P_s^{-1} B_1 R_1^{-1} B_1^\top) Y_a (I - B_1 R_1^{-1} B_1^\top Y_a)^{-1} F_y + R_2 \\ &= F_y^\top Y_a (I - B_1 R_1^{-1} B_1^\top Y_a)^{-1} F_y + R_2 \\ &= F_y^\top Y_a F_y + F_y^\top Y_a B_1 (R_1 - B_1^\top Y_a B_1)^{-1} B_1^\top Y_a F_y + R_2, \end{aligned} \tag{2.77}$$

where we have taken (2.55) into account.

[3] We shall denote such a solution by $Y_a(\gamma)$ when we want to stress its dependence upon γ.

To prove that Y_a is a solution of the ARE (2.70), it remains only to show that

$$R_2 = -H_0^\top J_1^{-1} H_0. \tag{2.78}$$

Indeed, taking into account (2.51), (2.53), and the first of (2.41), we have

$$\begin{aligned}
&R_2 + H_0^\top J_1^{-1} H_0 \\
&= F_a^\top P_s^{-1}(F - F_y) - (F_1 - F_a)^\top P_s^{-1} F_y + L_1^\top R^{-1} L_1 + L^\top (MM^\top - \gamma^2 I)^{-1} L \\
&= F_a^\top P_s^{-1} B_1 M^\top (MM^\top - \gamma^2 I)^{-1} L - L_1^\top R^{-1} M V^{-1} B_1^\top P_s^{-1} F \\
&\quad + L_1^\top R^{-1} M \underbrace{V^{-1} B_1^\top P_s^{-1} B_1}_{V^{-1} - I} M^\top (MM^\top - \gamma^2 I)^{-1} L \\
&\quad + L_1^\top R^{-1} L_1 + L^\top (MM^\top - \gamma^2 I)^{-1} L \\
&= F_a^\top P_s^{-1} B_1 M^\top (MM^\top - \gamma^2 I)^{-1} L - L_1^\top R^{-1} M V^{-1} B_1^\top P_s^{-1} F \\
&\quad + L_1^\top R^{-1} \underbrace{M V^{-1} M^\top}_{-R + \gamma^2 I} (MM^\top - \gamma^2 I)^{-1} L \\
&\quad - L_1^\top R^{-1} \underbrace{MM^\top}_{(MM^\top - \gamma^2 I) + \gamma^2 I} (MM^\top - \gamma^2 I)^{-1} L \\
&\quad + L_1^\top R^{-1} L_1 + L^\top (MM^\top - \gamma^2 I)^{-1} L \\
&= F_a^\top P_s^{-1} B_1 M^\top (MM^\top - \gamma^2 I)^{-1} L - L_1^\top R^{-1} M V^{-1} B_1^\top P_s^{-1} F \\
&\quad + \underbrace{(L^\top - L_1^\top)}_{-F_a^\top P_s^{-1} B_1 M^\top} (MM^\top - \gamma^2 I)^{-1} L + L_1^\top R^{-1} \underbrace{(L_1 - L)}_{M B_1^\top P_s^{-1} F_a} \\
&= L_1^\top R^{-1} M (B_1^\top P_s^{-1} F_a - V^{-1} B_1^\top P_s^{-1} F) \\
&= L_1^\top R^{-1} M (B_1^\top P_s^{-1} F_a - B_1^\top P_s^{-1} \underbrace{(I - B_1 B_1^\top P_s^{-1})^{-1} F}_{F_a}) = 0.
\end{aligned} \tag{2.79}$$

We now prove that Y_a is a solution of (2.52) as well. In fact, we have

$$\begin{aligned}
F_y &= F - B_1 M^\top (MM^\top - \gamma^2 I)^{-1} L = F + B_1 M^\top \left(\frac{I}{\gamma^2} + \frac{M R_1^{-1} M^\top}{\gamma^4} \right) L \\
&= F + B_1 R_1^{-1} \frac{M^\top L}{\gamma^2}
\end{aligned} \tag{2.80}$$

and

$$\begin{aligned}
H_0^\top J_1^{-1} H_0 &= C^\top (DD^\top)^{-1} C + L^\top (MM^\top - \gamma^2 I)^{-1} L \\
&= C^\top (DD^\top)^{-1} C - \frac{L^\top L}{\gamma^2} - \frac{L^\top M}{\gamma^2} R_1^{-1} \frac{M^\top L}{\gamma^2}
\end{aligned} \tag{2.81}$$

so that, by using standard manipulations [15, p. 271], (2.70) may be rewritten in the form (2.52).

It remains to prove that the solution Y_a is indeed antistabilizing, i.e., that all the eigenvalues of Γ_y lay in $\{z \in \mathbb{C} : \ |z| > 1\}$. Notice that Γ_y may be written in the form

$$\Gamma_y = F_y + B_1 (R_1 - B_1^\top Y_a B_1)^{-1} B_1^\top Y_a F_y = (I - B_1 R_1^{-1} B_1^\top Y_a)^{-1} F_y. \tag{2.82}$$

Moreover, we rewrite (2.54) in the form

$$X_s^{-1} + L_1^\top R^{-1} L_1 = F_1^\top X_s^{-1} \Gamma_y, \tag{2.83}$$

which, together with (2.63) yields

$$I = X_s(I + L_1^\top R^{-1} L_1 X_s)^{-1} F_1^\top X_s^{-1} \Gamma_y = X_s \Gamma_a^\top X_s^{-1} \Gamma_y, \tag{2.84}$$

so that Γ_y is clearly antistable. □

3. A monotonicity result. In this section we prove that the values of γ for which $Y_a(\gamma)$ is singular are finitely many. This will be usefull in order to prove existence of the stabilizing solution of (1.14).

PROPOSITION 3.1. *Let $Y_a(\gamma)$ be as in* (2.68). *The set*

$$\mathcal{G}_2 := \{\gamma > \gamma_0 : \ Y_a(\gamma) \text{ is singular}\} \tag{3.1}$$

contains, at most, n points.

Proof. The solution X_s of (2.43) is a continuous function of γ and its first derivative with respect to γ^2 exists and is continuous; see, e.g., [15, Theorem 14.2.2]. Since P_s does not depend on γ, the derivative $\frac{dY_a}{d\gamma^2}$ exists and is continuos as well. Consider an open set $\mathcal{I} = (a, b)$ such that $a \geq \gamma_0$, $b > a$ (possibly $b = \infty$) and $\mathcal{I} \cap \mathcal{G}_1 = \emptyset$. (Notice that the set $\{\gamma \in \mathbb{R} : \ \gamma > \gamma_0\}$ may be written as the union of a finite number of sets of this form and of a finite number of isolated points.) For $\gamma \in \mathcal{I}$, we may take derivatives in both sides of (2.52) and, by defining

$$Y_\gamma := \frac{dY_a}{d\gamma^2}, \tag{3.2}$$

we get

$$\begin{aligned}
Y_\gamma = {} & F^\top Y_\gamma F - \left(F^\top Y_\gamma B_1 - \frac{L^\top M}{\gamma^4}\right)(R_1 - B_1^\top Y_a B_1)^{-1}\left(B_1^\top Y_a F + \frac{M^\top L}{\gamma^2}\right) \\
& - \left(F^\top Y_a B_1 + \frac{L^\top M}{\gamma^2}\right)(R_1 - B_1^\top Y_a B_1)^{-1}\left(B_1^\top Y_\gamma F - \frac{M^\top L}{\gamma^4}\right) \\
& - \left(F^\top Y_a B_1 + \frac{L^\top M}{\gamma^2}\right)(R_1 - B_1^\top Y_a B_1)^{-1}\frac{M^\top M}{\gamma^4} \\
& \quad (R_1 - B_1^\top Y_a B_1)^{-1}\left(B_1^\top Y_a F + \frac{M^\top L}{\gamma^2}\right) \\
& + \left(F^\top Y_a B_1 + \frac{L^\top M}{\gamma^2}\right)(R_1 - B_1^\top Y_a B_1)^{-1} B_1^\top Y_\gamma B_1 \\
& \quad (R_1 - B_1^\top Y_a B_1)^{-1}\left(B_1^\top Y_a F + \frac{M^\top L}{\gamma^2}\right) \\
& - \frac{L^\top L}{\gamma^4}
\end{aligned} \tag{3.3}$$

and, taking (2.69) into account,

$$Y_\gamma = \Gamma_y^\top Y_\gamma \Gamma_y - \frac{L_2^\top L_2}{\gamma^4} \tag{3.4}$$

with

$$L_2 := L - M(R_1 - B_1^\top Y_a B_1)^{-1}\left(B_1^\top Y_a F + \frac{M^\top L}{\gamma^2}\right). \tag{3.5}$$

Taking (2.82) into account, we may rewrite (2.70) (with $Y + Y_a$) in the form

$$Y_a = \Gamma_y^\top Y_a \Gamma_y - \Gamma_y^\top Y_a B_1 R_1^{-1} B_1^\top Y_a \Gamma_y - H_0^\top J_1^{-1} H_0. \tag{3.6}$$

Define

$$Z := Y_a + \gamma^2 Y_\gamma. \tag{3.7}$$

By adding (3.4) multiplied by γ^2 to (3.6), we get

$$Z = \Gamma_y^\top Z \Gamma_y - \Gamma_y^\top Y_a B_1 R_1^{-1} B_1^\top Y_a \Gamma_y - H_0^\top J_1^{-1} H_0 - \frac{L_2^\top L_2}{\gamma^2}. \tag{3.8}$$

By taking into account the definitions (2.69) and (3.5) of Γ_y and L_2, respectively, and identity (2.81), it is not difficult to see that

$$-\Gamma_y^\top Y_a B_1 R_1^{-1} B_1^\top Y_a \Gamma_y - H_0^\top J_1^{-1} H_0 - \frac{L_2^\top L_2}{\gamma^2} = -C_1^\top C_1 - C^\top (DD^\top)^{-1} C \tag{3.9}$$

with

$$C_1 := (R_1 - B_1^\top Y_a B_1)^{-1} \left(B_1^\top Y_a F + \frac{M^\top L}{\gamma^2} \right). \tag{3.10}$$

Therefore, Z satisfies

$$Z = \Gamma_y^\top Z \Gamma_y - C_1^\top C_1 - C^\top (DD^\top)^{-1} C. \tag{3.11}$$

Notice that since Γ_y is antistable, (3.4) and (3.11) imply that

$$Y_\gamma \geq 0, \tag{3.12a}$$

$$Z \geq 0. \tag{3.12b}$$

Moreover, by multiplying (3.4) and (3.11) on the left side by $\Gamma_y^{-\top}$ and on the right side by Γ_y^{-1}, it is easy to see that both $\ker Y_\gamma$ and $\ker Z$ are invariant for Γ_y^{-1} and hence for Γ_y. Then

$$\ker Y_\gamma \subseteq \ker L_2 \tag{3.13}$$

and

$$\ker Z \subseteq \ker \begin{bmatrix} C_1 \\ C \end{bmatrix}. \tag{3.14}$$

Therefore,

$$\mathcal{K} := \ker \begin{bmatrix} Y_a \\ Y_\gamma \end{bmatrix} = \ker \begin{bmatrix} Z \\ Y_\gamma \end{bmatrix} \tag{3.15}$$

is Γ_y-invariant and satisfies

$$\mathcal{K} \subseteq \ker \begin{bmatrix} C \\ C_1 \\ L_2 \end{bmatrix}. \tag{3.16}$$

We now prove that $\mathcal{K} = \{0\}$. To this end it is sufficient to show that

$$\ker \begin{bmatrix} \Gamma_y - \lambda I \\ C \\ C_1 \\ L_2 \end{bmatrix} = \{0\} \quad \forall \lambda \in \mathbb{C}. \tag{3.17}$$

Assume by contradiction that there exist $v \neq 0$ and $\lambda \in \mathbb{C}$ such that

$$\begin{bmatrix} \Gamma_y - \lambda I \\ C \\ C_1 \\ L_2 \end{bmatrix} v = 0. \tag{3.18}$$

Then, since Γ_y is antistable, $|\lambda| > 1$. Moreover, the definitions (1.6) of F and (2.69) of Γ_y yield

$$\Gamma_y = A - BD^\top (DD^\top)^{-1} C - B_1 C_1 \tag{3.19}$$

so that $\Gamma_y v = Av$ and we get

$$\begin{bmatrix} A - \lambda I \\ C \end{bmatrix} v = 0, \quad |\lambda| > 1, \tag{3.20}$$

which is in contradiction with detectability of the pair (A, C). Therefore $\mathcal{K} = 0$ and then, when one eigenvalue of Y_a (that is a continuous function of Y_a and hence of γ) is zero, its derivative with respect to γ is positive. Thus, if $Y_a(\bar\gamma)$ is singular and (counting with multiplicity) has, say, k_+ positive, k_0 zero, and k_- negative eigenvalues, then there exists a positive value δ such that $Y_a(\gamma)$ has k_+ positive and $k_0 + k_-$ negative eigenvalues for $\gamma \in (\bar\gamma - \delta, \bar\gamma)$ and $k_+ + k_0$ positive and k_- negative eigenvalues for $\gamma \in (\bar\gamma, \bar\gamma + \delta)$. Clearly, this may happen for, at most, n different values of γ.

So far we have assumed that $\gamma \in \mathcal{I}$. As already observed, the set $\{\gamma \in \mathbb{R} : \ \gamma > \gamma_0\}$ may be written as the union of a finite number of sets of the same type of $\mathcal{I}$ and of a finite number of values of γ. In correspondence of such values, Y_a remains a continuous function of γ so that we can extend the conclusion to the whole set $\{\gamma \in \mathbb{R} : \ \gamma > \gamma_0\}$. □

4. Existence of a stabilizing solution of the ARE (1.14). Next we show that when Y_a is nonsingular, then Y_a^{-1} is the (necessarily unique) stabilizing solution of (1.14), so that, in view of Propositions 2.1 and 3.1, we get that, except for a finite number of values of $\gamma > \gamma_0$, the ARE (1.14) admits the stabilizing solution. We shall prove in the next section that such solution is also feasible.

Consider the set $\mathcal{G}_1 \cup \mathcal{G}_2$ and observe that it contains, at most, $2[n + \operatorname{rank}(M)]$ points. Then, the set of *regular* values of γ defined as

$$\mathcal{G}_r := \{\gamma > \gamma_0 : \ \gamma \notin \mathcal{G}_1 \cup \mathcal{G}_2\} \tag{4.1}$$

is generic in $\{\gamma \in \mathbb{R} : \ \gamma > \gamma_0\}$.

The following theorem is our main result.

THEOREM 4.1. *Let $\gamma \in \mathcal{G}_r$ and let $Y_a(\gamma)$ be the corresponding antistabilizing solution of* (2.52). *Then* (1.14) *admits a unique symmetric stabilizing solution. Such solution is given by*

$$\Delta_s = Y_a^{-1}. \tag{4.2}$$

Proof. As shown in Proposition 2.1, Y_a is a solution of (2.70) so that we have

$$\begin{aligned} Y_a + H_0^\top J_1^{-1} H_0 &= F_y^\top Y_a F_y + F_y^\top Y_a B_1 (R_1 - B_1^\top Y_a B_1)^{-1} B_1^\top Y_a F_y \\ &= F_y^\top (Y_a + Y_a B_1 (R_1 - B_1^\top Y_a B_1)^{-1} B_1^\top Y_a) F_y. \end{aligned} \tag{4.3}$$

Since $\gamma \in \mathcal{G}_r$, Y_a, R_1 and $(R_1 - B_1^\top Y_a B_1)$ are nonsingular so that $Y_a + Y_a B_1 (R_1 - B_1^\top Y_a B_1)^{-1} B_1^\top Y_a$ is such. Moreover, F_y is nonsingular so that the left-hand side of (4.3) is nonsingular and the same procedure that led to (2.67a) gives

$$Y_a^{-1} = F_y Y_a^{-1} F_y^\top - F_y Y_a^{-1} H_0^\top (J_1 + H_0 Y_a^{-1} H_0^\top)^{-1} H_0 Y_a^{-1} F_y^\top + B_1 R_1^{-1} B_1^\top, \tag{4.4}$$

which, taking (4.2) into account, yields, after some standard manipulations [15, p. 271]

$$\Delta_s = A \Delta_s A^\top + B B^\top - (A \Delta_s H_0^\top + B J_0^\top)(J_1 + H_0 \Delta_s H_0^\top)^{-1} (H_0 \Delta_s A^\top + J_0 B^\top). \tag{4.5}$$

It remains to show that

$$\Gamma_s := A - (A \Delta_s H_0^\top + B J_0^\top)(J_1 + H_0 \Delta_s H_0^\top)^{-1} H_0 = F_y - F_y \Delta_s H_0^\top (J_1 + H_0 \Delta_s H_0^\top)^{-1} H_0 \tag{4.6}$$

is stable. To this aim we rewrite (4.4) in the form

$$Y_a^{-1} - B_1 R_1^{-1} B_1^\top = \Gamma_s Y_a^{-1} F_y^\top, \tag{4.7}$$

which, taking (2.82) into account, yields

$$I = \Gamma_s Y_a^{-1} F_y^\top (Y_a^{-1} - B_1 R_1^{-1} B_1^\top)^{-1} = \Gamma_s Y_a^{-1} \Gamma_y^\top Y_a, \tag{4.8}$$

so that, since Γ_y is antistable, Γ_s is stable. □

Remark 4.1. It is worth noticing that the computation of the stabilizing solution of the Riccati equation (1.14) can be numerically obtained by resorting to standard routines available in most control packages without the need of computing the solution the auxiliary Riccati equations (2.1) and (2.43). Indeed, those Riccati equations were only instrumental to the purpose of proving the existence of the stabilizing solution of (1.14).

5. Construction of the minimum phase J-spectral factor and $\mathcal{H}_\infty$ estimator design. The following important theorem gives a constructive procedure to obtain, from the solution Δ_s, a minimum-phase J-spectral factor of $\Psi(z)$ having a realization with the same state matrix A and the same output matrix H_0 of the realization (1.9) of $H(z)$.

THEOREM 5.1. *Let*

$$\gamma \in \mathcal{G}_r \tag{5.1}$$

and let Δ_s be the corresponding stabilizing solution of equation (1.14). *Then,*

1. *The solution Δ_s is feasible, i.e., $J_1 + H_0 \Delta_s H_0^\top$ has the same inertia of $J_{p,l}(\gamma)$ and hence there exists a nonsingular matrix Λ such that*

$$\Lambda J_{p,l}(\gamma) \Lambda^\top = J_1 + H_0 \Delta_s H_0^\top. \tag{5.2}$$

2. *The transfer matrix*

$$\Omega_s(z) := H_0(zI - A)^{-1}(A\Delta_s H_0^\top + BJ_0^\top)(J_1 + H_0\Delta_s H_0^\top)^{-1}\Lambda + \Lambda \tag{5.3}$$

is a square J-spectral factor of the J-spectral density $\Psi(z)$.

3. *All the zeros of $\Omega_s(z)$ lay in the open unit disk, i.e., $\Omega_s(z)^{-1}$ is stable.*

Proof. The proof of this theorem may be obtained following the same lines of Theorem 4.1 in [4]. □

We recall from [3] that there is a simple procedure that, from $\Omega_s(z)$, furnishes a realization of an $\mathcal{H}_\infty$ smoothing filter with attenuation level γ.

6. The critical values of γ: A peculiar feature of the J-spectral factorization. In this section, we analyze the ARE (1.14) and the associated J-spectral factorization problem in the case when $\gamma \in \mathcal{G}_s := \mathcal{G}_1 \cup \mathcal{G}_2$. To this aim we write the set $\mathcal{G}_s$ as the union of disjoint sets $\mathcal{G}_s = \mathcal{G}_1' \cup \mathcal{G}_2$ with $\mathcal{G}_1' := \mathcal{G}_1 \cap \overline{\mathcal{G}_2}$ (usually $\mathcal{G}_1' = \mathcal{G}_1$) and we consider the cases $\gamma \in \mathcal{G}_1'$ and $\gamma \in \mathcal{G}_2$ separately.

For $\gamma \in \mathcal{G}_1'$, $Y_a(\gamma)$ is still nonsingular and we can define $\Delta_s := Y_a(\gamma)^{-1}$. In all the manifold examples that we have worked out, it turns out that, even for $\gamma \in \mathcal{G}_1'$, Δ_s defined in this way continues to be the stabilizing solution of (1.14). This leads us to conjecture that Theorem 4.1 holds (possibly with the Moore–Penrose pseudoinverse in place of the inverse in the ARE (1.14)) for all $\gamma \in \mathcal{G}_r \cup \mathcal{G}_1'$. In view of continutity, to prove this generalization it would be sufficient to show that for $\gamma \in \mathcal{G}_1'$, $(J_1 + H_0\Delta_s H_0^\top)$ is invertible or (using the pseudoinverse in place of the inverse) that $(A\Delta_s H_0^\top + BJ_0^\top)(J_1 + H_0\Delta_s H_0^\top)^\sharp(H_0\Delta_s A^\top + J_0 B^\top)$ is a continuous function of γ for each $\gamma \in \mathcal{G}_1'$.

Much more interesting is the behavior associated with $\gamma \in \mathcal{G}_2$. In this case, as shown in the following example, the ARE does not admit a stabilizing solution nor does the J-spectral density admit a minimum phase J-spectral factor having a realization with state matrix equal to A. This is a peculiar (and in the authors' opinion, rather counterintuitive) feature of the J-spectral factorization. In fact, in the standard (positive) spectral factorization this phenomenon cannot occur.

Let us consider the following very simple example: $A = 2$, $C = L = 1$, $B = [1 \mid 0]$, $D = [0 \mid 1]$, and $M = [0 \mid 0]$. In this case $W(z)$ defined in (1.20) is given by

$$W(z) = \frac{1}{2(3 - z - z^{-1})} \tag{6.1}$$

so that

$$\gamma_0^2 = \sup_{\vartheta} \left| \frac{1}{2(3 - 2\cos(\vartheta))} \right| = \frac{1}{2}. \tag{6.2}$$

Moreover, the stabilizing solution of the ARE (2.1) is easily computed to be $P_s = 2 + \sqrt{5}$ and the ARE (2.43) assumes the form

$$X^2 + \left(\frac{3 + \sqrt{5} - \gamma^2(10 + 6\sqrt{5})}{4} \right) X + \gamma^2 \left(\frac{3 + \sqrt{5}}{4} \right) = 0, \tag{6.3}$$

whose stabilizing solution $X_s(\gamma)$ equals P_s for $\gamma = 1$. Moreover, $X_s(\gamma) > P_s$ for $\gamma > 1$ and $X_s(\gamma) < P_s$ for $\gamma < 1$. Thus, in this case, $\mathcal{G}_2 = \{1\}$. The ARE (1.14) assumes the form

$$3\Delta + 1 - \frac{4\Delta^2}{\Delta(1 - \gamma^2) - \gamma^2}(1 - \gamma^2) = 0. \tag{6.4}$$

For $\gamma < 1$, (6.4) admits a negative stabilizing solution Δ_s that tends to $-\infty$ as $\gamma \to 1_-$. For $\gamma > 1$, (6.4) admits a positive stabilizing solution Δ_s that tends to $+\infty$ as $\gamma \to 1_+$.

For $\gamma = 1$, the order of such equation collapses and the stabilizing solution does not exist any more: the only solution is $\Delta_a = -1/3$ and the corresponding closed-loop matrix is $A - 0 = A = 2$. Notice that $\Delta_a = -1/3$ is feasible. In fact

$$J_1 + H_0 \Delta_a H_0^\top = \frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & -4 \end{bmatrix} \tag{6.5}$$

has a positive and a negative eigenvalue so that there exists a matrix Λ such that[4]

$$J_1 + H_0 \Delta_a H_0^\top = \Lambda J_{1,1}(1) \Lambda^\top. \tag{6.6}$$

The same computation used in the proof of Theorem 5.1 shows that

$$\begin{aligned} \Omega_a(z) &:= H_0(zI - A)^{-1}(A\Delta_a H_0^\top + B J_0^\top)(J_1 + H_0 \Delta_a H_0^\top)^{-1}\Lambda + \Lambda \\ &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} (z-2)^{-1} \left[-\frac{2}{3} \mid \frac{2}{3} \right] \Lambda + \Lambda \end{aligned} \tag{6.7}$$

is a J-spectral factor of the J-spectrum $\Psi(z)$ that, in this case, is given by

$$\Psi(z) = \begin{bmatrix} \frac{1}{(z-2)(z^{-1}-2)} + 1 & \frac{1}{(z-2)(z^{-1}-2)} \\ \frac{1}{(z-2)(z^{-1}-2)} & \frac{1}{(z-2)(z^{-1}-2)} - 1 \end{bmatrix}. \tag{6.8}$$

The numerator matrix of $\Omega_a(z)$ is $2 - \left[-\frac{2}{3} \mid \frac{2}{3} \right] \Lambda\Lambda^{-1} \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right] = 2$: the unique zero of $\Omega_a(z)$ lies outside the closed unit disk. (Such an $\Omega_a(z)$ is said to be a *maximum phase* J-spectral factor.) The set of all J-spectral factors of $\Psi(z)$ having a realization with state matrix $A = 2$ may be obtained as follows. We set

$$\Omega(z) = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} (z-2)^{-1} [g_1 \mid g_2] + \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \tag{6.9}$$

and impose

$$\Omega(z) J_{1,1}(1) \widetilde{\Omega}(z) = \Psi(z). \tag{6.10}$$

The corresponding solutions, up to a noninteresting change of basis in the state space, may be parametrizied as follows:

$$\Omega(z) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} (z-2)^{-1} [g \mid \pm g] + \begin{bmatrix} \frac{-2-3\,g^2}{6\,g} & \pm\frac{2-3\,g^2}{6\,g} \\ \frac{-1+3\,g^2}{3\,g} & \pm\frac{1+3\,g^2}{3\,g} \end{bmatrix}, \qquad g \in \mathbb{R} \setminus \{0\}, \tag{6.11}$$

which is readily seen to be equal to the right-hand side of (6.7) as Λ varies among the solutions of (6.6). In conclusion, we have produced a maximum-phase J-spectral

[4]The set of solutions of (6.6) may be parametrized as

$$\Lambda = \begin{bmatrix} \frac{-2-3\,g^2}{6\,g} & \pm\frac{2-3\,g^2}{6\,g} \\ \frac{-1+3\,g^2}{3\,g} & \pm\frac{1+3\,g^2}{3\,g} \end{bmatrix}, g \in \mathbb{R} \setminus \{0\}$$

(where the $\pm$ signs are either both $+$ or both $-$).

factor and have shown that the corresponding minimum-phase one (i.e., a minimum-phase J-spectral factor having a realization with the same state-space matrix) does not exist. This is a very peculiar behavior that has no counterpart in the regular spectral factorization.

Notice that to obtain (6.11) we have imposed the McMillan degree[5] of $\Omega(z)$ to be equal to 1 or, equivalently, we have restricted our search to *minimal* J-spectral factors, namely, J-spectral factors having the last possible McMillan degree. (Such McMillan degree is clearly equal to one-half of the McMillan degree of the J-spectrum $\Psi(z)$.) See [16] for a discussion on the minimality of spectral factors. It is interesting to observe that extending the search to nonminimal J-spectral factors (i.e., J-spectral factors having larger McMillan degree), a minimum-phase J-spectral factor having a unique pole in $z = 2$ does exist. Indeed, it is not difficult to check that

$$\Omega_m(z) = (zI_2 - 2I_2)^{-1} \begin{bmatrix} \frac{-11}{2\sqrt{2}} & \frac{-3}{2\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{bmatrix} + \begin{bmatrix} \frac{-5\sqrt{2}}{3} & \frac{-\sqrt{2}}{3} \\ \frac{-1}{3\sqrt{2}} & \frac{7}{3\sqrt{2}} \end{bmatrix} \tag{6.12}$$

is a minimal realization of a J-spectral factor of $\Psi(z)$. As $H(z)$, $\Omega_m(z)$ has a unique pole in $z = 2$ and the corresponding numerator matrix

$$2I_2 - \begin{bmatrix} \frac{-11}{2\sqrt{2}} & \frac{-3}{2\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{3}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{-5\sqrt{2}}{3} & \frac{-\sqrt{2}}{3} \\ \frac{-1}{3\sqrt{2}} & \frac{7}{3\sqrt{2}} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} \\ -\frac{1}{6} & \frac{2}{3} \end{bmatrix} \tag{6.13}$$

has a unique eigenvalue (of multiplicity 2) in $z = 1/2$. Again, this fact has no counterpart in classical spectral factorization.

Eventually, notice that these discrepancies between J-spectral factorization and classical spectral factorization are associated to very particular J-spectra. In fact they may occur only for a finite number of values of γ.

6.1. Example with $\gamma \in \mathcal{G}_r$. In the following we consider the previous example in the case when $\gamma_0 < \gamma < \gamma_f$ and $\gamma \in \mathcal{G}_r$. We design a smoothing filter starting from the stabilizing solution Δ_s of (1.14). Let $\gamma^2 = 3/4$. The stabilizing solution Δ_s is then given by $-4 - \sqrt{13}$ and the corresponding J-spectral factor is given by

$$\Omega_s(z) = \left[\begin{array}{c} 1 \\ 1 \end{array}\right] (z-2)^{-1} \left[\begin{array}{c|c} 2.942 & 6.374 \end{array}\right] + \begin{bmatrix} -0.3124 & 2.99 \\ 0.2785 & 3.353 \end{bmatrix}. \tag{6.14}$$

From $\Omega_s(z)$, by following the procedure described in [3], we easily get the following transfer function of a 1-step lag smoother,

$$S_1(z) = z[0.2063(z - 0.02824)^{-1}2.958 + 0.1906], \tag{6.15}$$

and it is easy to check that

$$\|S_1(z)H_1(z) - H_2(z)\| = 0.8383 < \gamma \simeq 0.8660. \tag{6.16}$$

7. Conclusions. In this paper a general J-spectral factorization problem was considered and its relation with the existence of the stabilizing solution of the associated Riccati equation was investigated. The stabilizing solution depends on a positive

[5]We recall that the *McMillan degree* of a rational proper matrix function $P(z)$ is the state-space dimension of a minimal realization of $P(z)$; see [14] for more details.

parameter which represents the prescribed attenuation level for the underlying estimation problem. We have shown that the stabilizing solution of the ARE still exists (except for a finite number of values of γ) as long as a fixed-lag acausal estimator (smoother) does. A few aspects of the J-spectral factorization problem and the properties of its solutions are discussed in correspondence to the (finite number of) values of γ for which the stabilizing solution of the ARE does not exist.

REFERENCES

[1] T. Başar and T. P. Bernhard, *$\mathcal{H}_\infty$ Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser, Boston, 1995.

[2] P. Bolzern, P. Colaneri, and G. De Nicolao, *On discrete-time $\mathcal{H}_\infty$ fixed-lag smoothing*, IEEE Trans. Signal Process., 52 (2004), pp. 132–141.

[3] P. Colaneri and A. Ferrante, *A J-spectral factorization approach for $\mathcal{H}_\infty$ estimation problems in discrete-time*, IEEE Trans. Automat. Control AC-47 (2000), pp. 2108–2113.

[4] P. Colaneri and A. Ferrante, *Algebraic Riccati equation and J-spectral factorization for $\mathcal{H}_\infty$ estimation*, Systems Control Lett., 51 (2004), pp. 383–393.

[5] P. Colaneri, M. Maroni, and U. Shaked, *$\mathcal{H}_\infty$ prediction and smoothing for discrete-time systems: A J-spectral factorization approach*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 2836–2842.

[6] P. Colaneri and M. Maroni, *The J-Spectral Interactor Matrix n the Discrete-Time Singular H_∞ Filtering Problem,* in Proceedings of the 7th Mediterranean Conference on Control and Automation (MED99), Haifa, Israel, 1999, pp. 1311–1317.

[7] F. Cuzzola and A. Ferrante, *Explicit formula for LMI based $\mathcal{H}_2$ filtering and deconvolution*, Automatica, 37 (2001), pp. 1443–1449.

[8] C. de Souza and L. Xie, *On discrete-time bounded real lemma with application in the characterization of static state feedback $\mathcal{H}_\infty$ controllers*, System Control Lett., 18 (1992), pp. 61–71.

[9] F. Fagnani, V. Maksimov, and L. Pandolfi, *A recursive deconvolution approach to disturbance reduction*, IEEE Trans. Automat. Control AC-49 (2002), pp. 907–921.

[10] I. Gohberg, M. A. Kaashoek, and A. C. M. Ran, *Factorizations of and extensions to J-unitary matrix functions on the unit circle*, Integral Equations Operator Theory, 15 (1992), pp. 262–300.

[11] B. Hassibi, A. H. Sayed, and T. Kailath, *Indefinite-Quadratic Estimation and Control*, SIAM, Philadelphia, 1999.

[12] V. Ionescu, C. Oara, and M. Weiss, *Generalized Riccati Theory and Robust Control*, Wiley, New York, 1999.

[13] A. C. Kahane, L. Mirkin, and Z. J. Palmor, *On the discrete-time $\mathcal{H}_\infty$ fixed-lag smoothing*, in Proceedings of the 15th IFAC World Congress, Barcelona, 2002.

[14] T. Kailath, *Linear System Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[15] P. Lancaster and R. Rodman, *Algebraic Riccati Equation*, Oxford Science Publications, London, 1995.

[16] A. Lindquist and G. Picci, *A geometric approach to modelling and estimation of linear stochastic systems*, J. Math. Systems Estimation Control, 1 (1991), pp. 241–333.

[17] A. Lindquist and G. Picci, *Geometric Methods for State Space Identification*, in Identification, Adaptation, Learning. The Science of Learning Models from Data, S. Bittanti and G. Picci, eds., Springer Verlag, New York, 1996, pp. 1–69.

[18] L. Mirkin, *On the $\mathcal{H}_\infty$ fixed-lag smoothing: How to exploit the information preview*, Automatica, 39 (2003), pp. 1495–1504.

[19] U. Shaked and Y. Theodor, *$\mathcal{H}_\infty$-optimal estimation: A tutorial*, in Proceedings of the 31 IEEE Conference on Decision and Control, 1992, pp. 2278–2286.

[20] G. Tadmor and L. Mirkin, *$\mathcal{H}_\infty$ control and fixed-lag smoothing* II: *Fixed size ARE solutions in discrete-time systems*, 42nd IEEE Conference on Decision and Control, 2003, pp. 6521–6526.

[21] K. Takaba and T. Katayama, *Discrete-time $\mathcal{H}_\infty$ Algebraic Riccati Equation and parametrization of all $\mathcal{H}_\infty$ filters*, Internat. J. Control, 64 (1996), pp. 1129–1149.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 146–173

SOME CONTROLLABILITY RESULTS FOR THE N-DIMENSIONAL NAVIER–STOKES AND BOUSSINESQ SYSTEMS WITH $N-1$ SCALAR CONTROLS*

ENRIQUE FERNÁNDEZ-CARA†, SERGIO GUERRERO†, OLEG YU. IMANUVILOV‡, AND JEAN-PIERRE PUEL§

Abstract. In this paper we deal with some controllability problems for systems of the Navier–Stokes and Boussinesq kind with distributed controls supported in small sets. Our main aim is to control N-dimensional systems ($N+1$ scalar unknowns in the case of the Navier–Stokes equations) with $N-1$ scalar control functions. In a first step, we present some global Carleman estimates for suitable adjoint problems of linearized Navier–Stokes and Boussinesq systems. In this way, we obtain null controllability properties for these systems. Then, we deduce results concerning the local exact controllability to the trajectories. We also present (global) null controllability results for some (truncated) approximations of the Navier–Stokes equations.

Key words. Navier–Stokes system, exact controllability, Carleman inequalities

AMS subject classifications. 34B15, 35Q30, 93B05, 93C10

DOI. 10.1137/04061965X

1. Introduction and examples.

Let $\Omega \subset \mathbf{R}^N$ ($N = 2$ or 3) be a bounded connected open set whose boundary $\partial\Omega$ is regular enough (for instance of class C^2). Let $\mathcal{O} \subset \Omega$ be a (small) nonempty open subset and let $T > 0$. We will use the notation $Q = \Omega \times (0,T)$ and $\Sigma = \partial\Omega \times (0,T)$ and we will denote by $n(x)$ the outward unit normal to Ω at the point $x \in \partial\Omega$.

On the other hand, we will denote by C, C_1, C_2, ... various positive constants (usually depending on Ω and $\mathcal{O}$).

We will be concerned with the following controlled Navier–Stokes and Boussinesq systems:

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)y + \nabla p = v1_{\mathcal{O}}, \quad \nabla\cdot y = 0 & \text{in } Q,\\ y = 0 & \text{on } \Sigma,\\ y(0) = y^0 & \text{in } \Omega \end{cases} \tag{1}$$

and

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)y + \nabla p = v1_{\mathcal{O}} + \theta\, e_N, \quad \nabla\cdot y = 0 & \text{in } Q,\\ \theta_t - \Delta\theta + y\cdot\nabla\theta = h1_{\mathcal{O}} & \text{in } Q,\\ y = 0, \quad \theta = 0 & \text{on } \Sigma,\\ y(0) = y^0, \quad \theta(0) = \theta^0 & \text{in } \Omega \end{cases} \tag{2}$$

(in both dimensions $N = 2$ and $N = 3$).

*Received by the editors November 25, 2004; accepted for publication (in revised form) October 20, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sicon/45-1/61965.html

†Dpto. E.D.A.N., Universidad de Sevilla, Aptdo. 1160, 41080 Sevilla, Spain (cara@us.es, sguerrero@us.es). Partially supported by grant BFM2003–06446 of the D.G.E.S. (Spain).

‡Department of Mathematics, Iowa State University, 400 Carver Hall, Ames, IA 50011-2064 (vika@iastate.edu). This work is supported by NSF grant DMS 0205148.

§Laboratoire de Mathématiques Appliquées, Université de Versailles - St. Quentin, 45 Avenue des Etats Unis, 78035 Versailles, France (jppuel@cmapx.polytechnique.fr).

For $N=2$, we will also consider the following approximation of the Navier–Stokes system with boundary conditions of the Navier kind:

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)\mathbf{T}_M(y) + \nabla p = v1_{\mathcal{O}}, \quad \nabla\cdot y = 0 & \text{in } Q,\\ y\cdot n = 0, \quad \nabla\times y = 0 & \text{on } \Sigma,\\ y(0) = y^0 & \text{in } \Omega, \end{cases} \tag{3}$$

where $M>0$, $\mathbf{T}_M(y) = (T_M(y_1), T_M(y_2))$ and T_M is given by

$$T_M(s) = \begin{cases} -M & \text{if } s \le -M,\\ s & \text{if } -M \le s \le M,\\ M & \text{if } s \ge M. \end{cases}$$

In systems (1), (2) and (3), $v = v(x,t)$ and $h = h(x,t)$ stand for the control functions. They act during the whole time interval $(0,T)$ over the set $\mathcal{O}$. The symbol $1_{\mathcal{O}}$ stands for the characteristic function of $\mathcal{O}$ and e_N is the Nth vector of the canonical basis of $\mathbf{R}^N$.

The controllability of Navier–Stokes systems has been the objective of considerable work over the last years. Up to our knowledge, the strongest results have been given in [7], where a strategy based on the methods in [13] and [14] has been followed. Recently, the techniques in [7] have been adapted in [12] to cover Boussinesq systems (see also [3], [4], [8] and [10] for other results).

This paper can be viewed as a continuation of [7]. We will present some new results which show that the N-dimensional systems (1) and (2) can be controlled, at least under some geometrical assumptions, with only $N-1$ scalar controls in $L^2(\mathcal{O}\times(0,T))$. In particular, the Boussinesq system (2) in dimension $N=2$ can be controlled by an action performed only on the temperature equation. We will also prove that the two-dimensional system (3) can be controlled with controls of the form $v1_{\mathcal{O}}$ where v is the curl of a function in $L^2(0,T;H^1(\mathcal{O}))$.

In this paper, we will have to impose some regularity assumptions on the initial data. To this purpose, we introduce the spaces H, E and V, with

$$H = \{w \in L^2(\Omega)^N : \nabla\cdot w = 0 \text{ in } \Omega,\ w\cdot n = 0 \text{ on } \partial\Omega\}, \tag{4}$$

$$E = \begin{cases} H & \text{if } N=2,\\ L^4(\Omega)^3 \cap H & \text{if } N=3 \end{cases}$$

and

$$V = \{w \in H_0^1(\Omega)^N : \nabla\cdot w = 0 \text{ in } \Omega\}.$$

For system (1), we will assume that the control region $\mathcal{O}$ is adjacent to the boundary $\partial\Omega$ (see assumption (11) below) and we will deal with the *local exact controllability to the trajectories*. More precisely, our task will be to prove that, for any bounded and sufficiently regular solution $(\overline{y},\overline{p})$ of the uncontrolled Navier–Stokes equations

$$\begin{cases} \overline{y}_t - \Delta\overline{y} + (\overline{y}\cdot\nabla)\overline{y} + \nabla\overline{p} = 0, \quad \nabla\cdot\overline{y} = 0 & \text{in } Q,\\ \overline{y} = 0 & \text{on } \Sigma,\\ \overline{y}(0) = \overline{y}^0 & \text{in } \Omega, \end{cases} \tag{5}$$

there exists $\delta > 0$ such that, whenever $y^0 \in E$ and

$$\|y^0 - \overline{y}^0\|_E \leq \delta,$$

we can find L^2 controls v with $v_k \equiv 0$ for at least one k and associated states (y, p) satisfying

$$y(T) = \overline{y}(T) \text{ in } \Omega. \tag{6}$$

Notice that, under these circumstances, after time $t = T$ we can switch off the control and let the system follow the "ideal" trajectory $(\overline{y}, \overline{p})$.

For the Boussinesq system (2), we will assume that $\mathcal{O}$ is adjacent to $\partial\Omega$ near a point x^0 such that $n_k(x^0) \neq 0$ for some $k < N$. We will also be concerned with the local exact controllability to the trajectories. Now, a trajectory is a bounded and sufficiently regular solution $(\overline{y}, \overline{p}, \overline{\theta})$ of

$$\begin{cases} \overline{y}_t - \Delta\overline{y} + (\overline{y} \cdot \nabla)\overline{y} + \nabla\overline{p} = \overline{\theta}\, e_N, \quad \nabla \cdot \overline{y} = 0 & \text{in } Q, \\ \overline{\theta}_t - \Delta\overline{\theta} + \overline{y} \cdot \nabla\overline{\theta} = 0 & \text{in } Q, \\ \overline{y} = 0, \quad \overline{\theta} = 0 & \text{on } \Sigma, \\ \overline{y}(0) = \overline{y}^0, \quad \overline{\theta}(0) = \overline{\theta}^0 & \text{in } \Omega. \end{cases} \tag{7}$$

The goal will be to prove that there exists $\delta > 0$ such that, whenever $(y^0, \theta^0) \in E \times L^2(\Omega)$ and

$$\|(y^0, \theta^0) - (\overline{y}^0, \overline{\theta}^0)\|_{E \times L^2} \leq \delta,$$

we can find L^2 controls v and h with $v_k \equiv v_N \equiv 0$ and associated states (y, p, θ) satisfying

$$y(T) = \overline{y}(T) \text{ and } \theta(T) = \overline{\theta}(T) \text{ in } \Omega. \tag{8}$$

In this context, the results established in [12] will be fundamental.

Notice that, in particular, when $N = 2$, we try to control the whole system (2) with just one scalar control h.

As far as (3) is concerned, our goal will be to prove the *(global) null controllability*. That is to say, for each $y^0 \in H$, we will try to find controls of the form $v1_{\mathcal{O}}$, where v belongs to the Hilbert space

$$W = \{\nabla \times z = (\partial_2 z, -\partial_1 z) : z \in L^2(0, T; H^1(\mathcal{O}))\}, \tag{9}$$

such that the associated solutions (y, p) satisfy

$$y(T) = 0 \text{ in } \Omega. \tag{10}$$

Approximate controllability results have been established for analogous systems in [4].

Observe that in this system the boundary conditions are of the Navier kind as in [3] (for their physical meaning, see, for instance, [11]). This and the fact that $N = 2$ will be essential in the arguments presented below.

Similarly to the previous situation, an extension by zero of the control after time $t = T$ will keep (y, p) at rest.

As mentioned above, some hypotheses will be imposed on the control domain and the trajectories. More precisely, we will frequently assume that

$$\exists x^0 \in \partial\Omega,\ \exists \varepsilon > 0 \text{ such that } \overline{\mathcal{O}} \cap \partial\Omega \supset B(x^0;\varepsilon) \cap \partial\Omega \tag{11}$$

($B(x^0;\varepsilon)$ is the ball centered at x^0 of radius ε),

$$\overline{y} \in L^\infty(Q)^N, \quad \overline{y}_t \in L^2(0,T;L^\sigma(\Omega)^N) \quad \begin{pmatrix} \sigma > 1 & \text{if } N = 2 \\ \sigma > 6/5 & \text{if } N = 3 \end{pmatrix} \tag{12}$$

and

$$\overline{\theta} \in L^\infty(Q), \quad \overline{\theta}_t \in L^2(0,T;L^\sigma(\Omega)) \quad \begin{pmatrix} \sigma > 1 & \text{if } N = 2 \\ \sigma > 6/5 & \text{if } N = 3 \end{pmatrix}. \tag{13}$$

Let us now present our main results in a precise form. The first one concerns the local exact controllability to the trajectories of system (1).

THEOREM 1. *Assume that $\mathcal{O}$ satisfies* (11). *Then, for any $T > 0$,* (1) *is locally exactly controllable at time T to the trajectories $(\overline{y}, \overline{p})$ satisfying* (12) *with controls $v \in L^2(\mathcal{O} \times (0,T))^N$ having one component identically zero.*

The second main result concerns the controllability of (2).

THEOREM 2. *Assume that $\mathcal{O}$ satisfies* (11) *with $n_k(x^0) \neq 0$ for some $k < N$. Then, for any $T > 0$,* (2) *is locally exactly controllable at time T to the trajectories $(\overline{y}, \overline{p}, \overline{\theta})$ satisfying* (12)–(13) *with L^2 controls v and h such that $v_k \equiv v_N \equiv 0$. In particular, if $N = 2$, we have local exact controllability to the trajectories with controls $v \equiv 0$ and $h \in L^2(\mathcal{O} \times (0,T))$.*

The last main result we present in this paper follows in Theorem 3.

THEOREM 3. *Let $N = 2$. Then, for any $T > 0$ and any $M > 0$,* (3) *is null controllable at time T with controls of the form $v1_{\mathcal{O}}$, where $v \in W$.*

For the proofs of these results, following a standard approach, we will first deduce null controllability results for suitable linearized versions of (1), (2) and (3), namely,

$$\begin{cases} y_t - \Delta y + (\overline{y} \cdot \nabla) y + (y \cdot \nabla)\overline{y} + \nabla p = f + v1_{\mathcal{O}}, \quad \nabla \cdot y = 0 & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y^0 & \text{in } \Omega, \end{cases} \tag{14}$$

$$\begin{cases} y_t - \Delta y + (\overline{y} \cdot \nabla) y + (y \cdot \nabla)\overline{y} + \nabla p = f + v1_{\mathcal{O}} + \theta\, e_N & \text{in } Q, \\ \nabla \cdot y = 0 & \text{in } Q, \\ \theta_t - \Delta\theta + \overline{y} \cdot \nabla\theta + y \cdot \nabla\overline{\theta} = k + h1_{\mathcal{O}} & \text{in } Q, \\ y = 0, \quad \theta = 0 & \text{on } \Sigma, \\ y(0) = y^0, \quad \theta(0) = \theta^0 & \text{in } \Omega \end{cases} \tag{15}$$

and

$$\begin{cases} y_t - \Delta y + (y \cdot \nabla)\overline{y} + \nabla p = v1_{\mathcal{O}}, \quad \nabla \cdot y = 0 & \text{in } Q, \\ y \cdot n = 0, \quad \nabla \times y = 0 & \text{on } \Sigma, \\ y(0) = y^0 & \text{in } \Omega. \end{cases} \tag{16}$$

Then, appropriate arguments will be used to deduce the controllability of the nonlinear systems (1)–(3).

Remark 1. When $N = 3$, it is very natural to ask whether a result similar to Theorem 1 holds with controls having two zero components. In general, the answer is no. In fact, it seems difficult to identify the open sets Ω and $\mathcal{O}$ such that one has null controllability for all $T > 0$ with controls of this kind. This is unknown even for the classical Stokes equations for which, up to now, the only known results concern *approximate controllability*; see [16].

Remark 2. Assume that $N = 2$. The arguments in [7] implicitly show that, under hypotheses (12), we can find controls $v1_{\mathcal{O}}$ with $v \in W$ such that the associated solutions to (1) satisfy $y(T) = \overline{y}(T)$. Observe that the assumption (11) on the control domain is not necessary here.

This paper is organized as follows. We will first establish all the technical results needed in this work in section 2. Section 3 will deal with null controllability results for the linear control systems (14)–(16). Finally, the proofs of Theorems 1, 2 and 3 will be given in section 4.

2. Some previous results. In this section we will establish all the technical results needed in this paper. More precisely, we will present and prove the required *Carleman estimates* for the backward systems (19), (20) and (21), given below.

To do this, let us first introduce some weight functions:

$$\alpha(x,t) = \frac{e^{5/4\lambda m\|\eta^0\|_\infty} - e^{\lambda(m\|\eta^0\|_\infty + \eta^0(x))}}{t^4(T-t)^4},$$

$$\xi(x,t) = \frac{e^{\lambda(m\|\eta^0\|_\infty + \eta^0(x))}}{t^4(T-t)^4},$$

$$\widehat{\alpha}(t) = \min_{x\in\overline{\Omega}} \alpha(x,t) = \frac{e^{5/4\lambda m\|\eta^0\|_\infty} - e^{\lambda(m+1)\,\|\eta^0\|_\infty}}{t^4(T-t)^4}, \tag{17}$$

$$\alpha^*(t) = \max_{x\in\overline{\Omega}} \alpha(x,t) = \frac{e^{5/4\lambda m\|\eta^0\|_\infty} - e^{\lambda m\|\eta^0\|_\infty}}{t^4(T-t)^4},$$

$$\widehat{\xi}(t) = \max_{x\in\overline{\Omega}} \xi(x,t) = \frac{e^{\lambda(m+1)\|\eta^0\|_\infty}}{t^4(T-t)^4}, \quad \xi^*(t) = \min_{x\in\overline{\Omega}} \xi(x,t) = \frac{e^{\lambda m\|\eta^0\|_\infty}}{t^4(T-t)^4},$$

where $m > 4$ is a fixed real number. Here, η^0 is a function verifying

$$\eta^0 \in C^2(\overline{\Omega}), \quad |\nabla\eta^0| > 0 \text{ in } \overline{\Omega\setminus\mathcal{O}_0}, \quad \eta^0 > 0 \text{ in } \Omega \text{ and } \eta^0 \equiv 0 \text{ on } \partial\Omega \tag{18}$$

with $\mathcal{O}_0$ a nonempty open subset of $\mathcal{O}$ that will be determined below. For any $\mathcal{O}_0$, the existence of such a function η^0 is proved in [9]. Note that these weights have already been used in [7] and [12].

We will be dealing in this section with the adjoint systems to (14) and (15), that is to say,

$$\begin{cases} -\varphi_t - \Delta\varphi - (D\varphi)\,\overline{y} + \nabla\pi = g, \quad \nabla\cdot\varphi = 0 & \text{in } Q, \\ \varphi = 0 & \text{on } \Sigma, \\ \varphi(T) = \varphi^0 & \text{in } \Omega \end{cases} \tag{19}$$

and

$$
(20)\qquad \begin{cases} -\varphi_t - \Delta\varphi - (D\varphi)\,\overline{y} + \nabla\pi = g + \overline{\theta}\,\nabla\psi, \quad \nabla\cdot\varphi = 0 & \text{in } Q,\\ -\psi_t - \Delta\psi - \overline{y}\cdot\nabla\psi = q + \varphi_N & \text{in } Q,\\ \varphi = 0, \quad \psi = 0 & \text{on } \Sigma,\\ \varphi(T) = \varphi^0, \quad \psi(T) = \psi^0 & \text{in } \Omega \end{cases}
$$

(where $D\varphi = \nabla\varphi + \nabla\varphi^t$) as well as with the adjoint system of $\omega := \nabla\times y$ (where y is the solution of (16)), which is

$$
(21)\qquad \begin{cases} -\rho_t - \Delta\rho - \nabla\times((\overline{y}\cdot\nabla\times)\nabla\gamma) = 0, \quad \Delta\gamma = \rho & \text{in } Q,\\ \gamma = 0, \quad \rho = 0 & \text{on } \Sigma,\\ \rho(T) = \rho^0 & \text{in } \Omega. \end{cases}
$$

Here, $g \in L^2(Q)^N$, $q \in L^2(Q)$, $\varphi^0 \in H$, $\psi^0 \in L^2(\Omega)$ and $\rho^0 \in H^{-1}(\Omega)$ (of course, φ_N stands for the last component of the vector field φ).

2.1. New Carleman estimates for system (19). We will establish some new Carleman estimates for the solutions of (19). We will assume that $\mathcal{O}$ and $\overline{y}$ satisfy (11)–(12). To fix ideas, we will also assume for the moment that $N = 3$ and $n_1(x^0) \neq 0$ (x^0 appears in assumption (11)).

The desired Carleman inequalities will have the form

$$
I(\varphi) \le C\left(\iint_Q \rho_1^2\,|g|^2\,dx\,dt + \iint_{\mathcal{O}\times(0,T)} \rho_2^2\left(|\varphi_2|^2 + |\varphi_3|^2\right)\,dx\,dt\right),
$$

where $I(\varphi)$ contains global weighted integrals of $|\varphi|^2$, $|\nabla\varphi|^2$, etc. and ρ_1 and ρ_2 are appropriate weights that vanish exponentially as $t \to T$. This will suffice to prove in section 3 the null controllability of (14) with controls $v1_{\mathcal{O}}$ satisfying $v_1 \equiv 0$.

LEMMA 1. *Assume that $N = 3$, $n_1(x^0) \neq 0$ and $\mathcal{O}$ and $\overline{y}$ verify* (11)–(12). *Then there exists a positive constant C such that, for any $g \in L^2(Q)^3$ and any $\varphi^0 \in H$, the associated solution to* (19) *satisfies:*

$$
(22)\qquad \begin{cases} I(\varphi) := \displaystyle\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\,t^{-12}(T-t)^{-12}\,|\varphi|^2\,dx\,dt\\ \qquad + \displaystyle\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\,t^{-4}(T-t)^{-4}\,|\nabla\varphi|^2)\,dx\,dt\\ \qquad + \displaystyle\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\,t^{4}(T-t)^{4}\,\left(|\Delta\varphi|^2 + |\varphi_t|^2\right)\,dx\,dt\\ \qquad \le C\left(\displaystyle\iint_Q e^{\frac{-4\boldsymbol{\alpha}+2\overline{\alpha}}{t^4(T-t)^4}}t^{-30}(T-t)^{-30}\,|g|^2\,dx\,dt\right.\\ \qquad \left. + \displaystyle\iint_{\mathcal{O}\times(0,T)} e^{\frac{-16\boldsymbol{\alpha}+14\overline{\alpha}}{t^4(T-t)^4}}\,t^{-132}(T-t)^{-132}\,(|\varphi_2|^2 + |\varphi_3|^2)\,dx\,dt\right). \end{cases}
$$

Here, $\overline{\alpha}$ and $\widetilde{\alpha}$ are constants only depending on Ω, $\mathcal{O}$, T and $\overline{y}$ satisfying $0 < \widetilde{\alpha} < \overline{\alpha}$ and $8\widetilde{\alpha} - 7\overline{\alpha} > 0$.

Proof. Let us first recall a Carleman inequality for the solutions of (19) which has been proved in [7] whenever (12) is fulfilled:

$$
\begin{aligned}
& s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\varphi|^2\,dx\,dt + s\lambda^2 \iint_Q e^{-2s\alpha}\xi|\nabla\varphi|^2\,dx\,dt \\
& \quad +s^{-1}\iint_Q e^{-2s\alpha}\xi^{-1}\left(|\varphi_t|^2+|\Delta\varphi|^2\right)dx\,dt \\
& \quad \le C_0(1+T^2)\left(s^{15/2}\lambda^{20}\iint_Q e^{-4sb+2s\alpha^*}\widehat{\xi}^{15/2}|g|^2\,dx\,dt\right. \\
& \quad \left. +s^{16}\lambda^{40}\iint_{\mathcal{O}_0\times(0,T)} e^{-8sb+6s\alpha^*}\widehat{\xi}^{16}|\varphi|^2\,dx\,dt\right).
\end{aligned}
\tag{23}
$$

Here, $s \ge s_0$ and $\lambda \ge \lambda_0$ are arbitrarily large and C_0, s_0 and λ_0 are suitable constants depending on Ω, $\mathcal{O}_0$, T and $\overline{y}$; see Theorem 1 in [7].

Recall that an inequality like (23) had already been proved in [13] using stronger properties on $\overline{y}$ than (12).

It is immediate from (23) that, for some C_1, $\overline{\alpha}$ and $\widetilde{\alpha}$ depending on Ω, $\mathcal{O}_0$, T and $\overline{y}$, we have:

$$
\left\{
\begin{array}{l}
\displaystyle\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\left(t^{-12}(T-t)^{-12}\,|\varphi|^2+t^{-4}(T-t)^{-4}\,|\nabla\varphi|^2\right)dx\,dt \\
\displaystyle\quad +\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\,t^4(T-t)^4\,\left(|\Delta\varphi|^2+|\varphi_t|^2\right)dx\,dt \\
\displaystyle\quad \le C_1\left(\iint_Q e^{\frac{-4b+2\overline{\alpha}}{t^4(T-t)^4}}\,t^{-30}(T-t)^{-30}\,|g|^2\,dx\,dt\right. \\
\displaystyle\quad \left. +\iint_{\mathcal{O}_0\times(0,T)} e^{\frac{-8b+6\overline{\alpha}}{t^4(T-t)^4}}\,t^{-64}(T-t)^{-64}\,|\varphi|^2\,dx\,dt\right).
\end{array}
\right.
\tag{24}
$$

Indeed, it suffices to choose

$$
\left\{
\begin{array}{l}
\overline{\alpha} = s_0\left(e^{5/4\lambda_0 m\|\eta^0\|_\infty}-e^{\lambda_0 m\|\eta^0\|_\infty}\right), \\
\widetilde{\alpha} = s_0\left(e^{5/4\lambda_0 m\|\eta^0\|_\infty}-e^{\lambda_0 (m+1)\|\eta^0\|_\infty}\right)
\end{array}
\right.
\tag{25}
$$

and $C_1 = C_0(1+T^2)s_0^{17}\lambda_0^{40}e^{17\lambda_0(m+1)\|\eta^0\|_\infty}$. Notice that $0<\widetilde{\alpha}<\overline{\alpha}$. Moreover, it can be assumed that $8\widetilde{\alpha}-7\overline{\alpha}>0$ (it suffices to notice that λ_0 is large enough in (25)).

We will apply (24) for the open set $\mathcal{O}_0 \subset \mathcal{O}$ defined as follows. We choose $\kappa>0$ such that

$$
n_1(x) \ne 0 \quad \forall x \in B(x^0;\kappa)\cap\partial\mathcal{O}\cap\partial\Omega
$$

and we denote this set by Γ_κ. Then, we define

$$
\mathcal{O}_0 = \{x\in\Omega : x = w+\tau\,e_1,\ w\in\Gamma_\kappa,\ |\tau|<\tau^0\},
\tag{26}
$$

with $\kappa, \tau^0 > 0$ small enough so that we still have

$$
\mathcal{O}_0 \subset \mathcal{O} \quad \text{and} \quad d_0 := \operatorname{dist}(\overline{\mathcal{O}_0}, \partial\mathcal{O}\cap\Omega) > 0.
\tag{27}
$$

Observe that, with this choice, each $P \in \mathcal{O}_0$ verifies that one of the two points where the straight line $\{P + \mathbf{R}\, e_1\}$ intersects $\partial\Omega$ belongs to $\partial\mathcal{O}_0$.

Once $\mathcal{O}_0$ is defined, we apply inequality (24) in this open set and we try to bound the term

$$\iint_{\mathcal{O}_0\times(0,T)} e^{\frac{-8\mathbf{a}+6\overline{\alpha}}{t^4(T-t)^4}}\, t^{-64}(T-t)^{-64}\,|\varphi_1|^2\,dx\,dt$$

in terms of local integrals of φ_2 and φ_3.

To this end, for each $(x,t) \in \mathcal{O}_0 \times (0,T)$ we denote by $l(x,t)$ (resp., $\widetilde{l}(x,t)$) the segment that starts from (x,t) with direction e_1 in the positive (resp. negative) sense and ends at $\partial\mathcal{O}_0$. Then, since φ is divergence-free, it is not difficult to see that

$$\varphi_1(x,t) = \int_{l(x,t)} (\partial_2\varphi_2 + \partial_3\varphi_3)(y_1, x_2, x_3, t)\,dy_1$$

for each $(x,t) \in \mathcal{O}_0 \times (0,T)$. For simplicity, let us introduce the notation

$$\beta(t) = e^{\frac{-8\mathbf{a}+6\overline{\alpha}}{t^4(T-t)^4}}\, t^{-64}(T-t)^{-64} \quad \forall t \in (0,T).$$

Applying at this point Hölder's inequality and Fubini's formula, we obtain

$$\begin{aligned}
&\iint_{\mathcal{O}_0\times(0,T)} \beta(t)\,|\varphi_1|^2\,dx\,dt \\
&\quad\le C_2 \iint_{\mathcal{O}_0\times(0,T)} \beta(t) \left(\int_{l(x,t)} (|\partial_2\varphi_2|^2 + |\partial_3\varphi_3|^2)\,dy_1\right) dx\,dt \\
&\quad= C_2 \iint_{\mathcal{O}_0\times(0,T)} (|\partial_2\varphi_2|^2 + |\partial_3\varphi_3|^2) \left(\int_{\widetilde{l}(y_1)} \beta(t)\,dx_1\right) dy_1\,dx_2\,dx_3\,dt \\
&\quad\le C_3 \iint_{\mathcal{O}_0\times(0,T)} \beta(t)(|\partial_2\varphi_2|^2 + |\partial_3\varphi_3|^2)\,dx\,dt,
\end{aligned} \tag{28}$$

where $\widetilde{l}(y_1)$ stands for the segment $\widetilde{l}(y_1, x_2, x_3, t)$. Then, we introduce a function $\zeta \in C^2(\overline{\mathcal{O}})$ such that

$$\zeta \equiv 1 \text{ in } \mathcal{O}_0, \quad 0 \le \zeta \le 1$$

and $\zeta(x) = 0$ at any point $x \in \mathcal{O}$ satisfying $\operatorname{dist}(x, \partial\mathcal{O} \cap \Omega) \le d_0/2$ (d_0 was defined in (27)). This and the fact that $\varphi|_\Sigma \equiv 0$ imply

$$\begin{aligned}
\iint_{\mathcal{O}_0\times(0,T)} \beta(t)\,|\partial_i\varphi_i|^2\,dx\,dt &\le \iint_{\mathcal{O}\times(0,T)} \zeta\,\beta(t)\,|\partial_i\varphi_i|^2\,dx\,dt \\
&= \frac{1}{2}\iint_{\mathcal{O}\times(0,T)} \partial_{ii}^2\zeta\,\beta(t)\,|\varphi_i|^2\,dx\,dt - \iint_{\mathcal{O}\times(0,T)} \zeta\,\beta(t)\,\partial_{ii}^2\varphi_i\,\varphi_i\,dx\,dt
\end{aligned}$$

for $i = 2, 3$. Finally, in view of Young's inequality and regularity estimates for φ_i in

Ω ($\varphi_i \in H^2(\Omega)$ and $\|\varphi_i\|_{H^2} \leq C\|\Delta\varphi_i\|_{L^2}$), we also have:

$$\begin{aligned}
&\iint_{\mathcal{O}_0\times(0,T)} \beta(t)\,|\partial_i\varphi_i|^2\,dx\,dt \\
&\leq C_4 \iint_{\mathcal{O}\times(0,T)} e^{\frac{-16\mathbf{a}+14\overline{\alpha}}{t^4(T-t)^4}}\, t^{-132}(T-t)^{-132}\,|\varphi_i|^2\,dx\,dt \\
&+\frac{1}{2C_1C_3}\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\, t^4(T-t)^4\,|\Delta\varphi_i|^2\,dx\,dt,
\end{aligned}$$

which, combined with (24) and (28), yields (22). □

Let us now present another Carleman inequality for (19) with weight functions not vanishing at time $t = 0$.

LEMMA 2. *Assume that $N = 3$, $n_1(x^0) \neq 0$ and $\mathcal{O}$ and $\overline{y}$ verify (11)–(12). Then there exist positive constants C, $\overline{\alpha}$ and $\widetilde{\alpha}$ with $0 < \widetilde{\alpha} < \overline{\alpha}$ and $8\widetilde{\alpha} - 7\overline{\alpha} > 0$ depending on Ω, $\mathcal{O}$, T and $\overline{y}$ such that, for any $g \in L^2(Q)^3$ and any $\varphi^0 \in H$, the associated solution to (19) satisfies:*

$$\left\{\begin{array}{l}
\displaystyle\iint_Q e^{\frac{-2\overline{\alpha}}{\ell(t)^4}}\left(\ell(t)^{-12}\,|\varphi|^2 + \ell(t)^{-4}\,|\nabla\varphi|^2\right)\,dx\,dt \\
\displaystyle\quad \leq C\left(\iint_Q e^{\frac{-4\mathbf{a}+2\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-30}\,|g|^2\,dx\,dt\right. \\
\displaystyle\quad \left.+\iint_{\mathcal{O}\times(0,T)} e^{\frac{-16\mathbf{a}+14\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-132}\,(|\varphi_2|^2+|\varphi_3|^2)\,dx\,dt\right),
\end{array}\right. \tag{29}$$

where ℓ is the C^1 function given by

$$\ell(t) = \left\{\begin{array}{ll} \dfrac{T^2}{4} & \text{for } 0 \leq t \leq T/2, \\ t(T-t) & \text{for } T/2 \leq t \leq T. \end{array}\right. \tag{30}$$

To prove (29), it suffices to use (22) and the classical parabolic estimates for the Stokes system satisfied by φ. The argument has already been used in [9], [13] and [7] in several similar situations, so we omit it for simplicity.

For completeness, let us state the similar result that can be established when $N = 2$. Here, we assume again that $n_1(x^0) \neq 0$.

LEMMA 3. *Assume that $N = 2$, $n_1(x^0) \neq 0$ and $\mathcal{O}$ and $\overline{y}$ verify (11)–(12). Then there exist positive constants C, $\overline{\alpha}$ and $\widetilde{\alpha}$ with $0 < \widetilde{\alpha} < \overline{\alpha}$ and $8\widetilde{\alpha} - 7\overline{\alpha} > 0$ depending on Ω, $\mathcal{O}$, T and $\overline{y}$ such that, for any $g \in L^2(Q)^2$ and any $\varphi^0 \in H$, the associated solution to (19) satisfies:*

$$\left\{\begin{array}{l}
\displaystyle\iint_Q e^{\frac{-2\overline{\alpha}}{\ell(t)^4}}\left(\ell(t)^{-12}\,|\varphi|^2 + \ell(t)^{-4}\,|\nabla\varphi|^2\right)\,dx\,dt \\
\displaystyle\quad \leq C\left(\iint_Q e^{\frac{-4\mathbf{a}+2\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-30}\,|g|^2\,dx\,dt\right. \\
\displaystyle\quad \left.+\iint_{\mathcal{O}\times(0,T)} e^{\frac{-16\mathbf{a}+14\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-132}\,|\varphi_2|^2\,dx\,dt\right),
\end{array}\right. \tag{31}$$

where ℓ is the function given by (30).

2.2. New Carleman estimates for system (20). We will establish suitable Carleman inequalities for the solutions of (20). To this end, our approach will be similar to the one in subsection 2.1.

Thus, we will assume again that $N=3$ and $n_1(x^0)\neq 0$ and we will prove an estimate of the form

$$K(\varphi,\psi)\leq C\left(\iint_Q \rho_3^2\left(|g|^2+|q|^2\right)dx\,dt+\iint_{\mathcal{O}\times(0,T)}\rho_4^2\left(|\varphi_2|^2+|\psi|^2\right)\,dx\,dt\right),$$

where $K(\varphi,\psi)=I(\varphi)+I(\psi)$ ($I(\varphi)$ has been given in (22)) and ρ_3 and ρ_4 are appropriate weights. This will be used in section 3 to find controls $v1_{\mathcal{O}}$ and $h1_{\mathcal{O}}$ with $v_1\equiv v_3\equiv 0$ leading to the null controllability of (15).

LEMMA 4. *Assume that $N=3$, $n_1(x^0)\neq 0$ and $\mathcal{O}$ and $(\overline{y},\overline{\theta})$ satisfy* (11)–(13). *Then, there exist positive constants C, $\overline{\alpha}$ and $\widetilde{\alpha}$ depending on Ω, $\mathcal{O}$, T, $\overline{y}$ and $\overline{\theta}$ with $0<\widetilde{\alpha}<\overline{\alpha}$ and $16\widetilde{\alpha}-15\overline{\alpha}>0$ such that, for any $g\in L^2(Q)^3$, $q\in L^2(Q)$, $\varphi^0\in H$ and $\psi^0\in L^2(\Omega)$, the associated solution to* (20) *satisfies:*

$$\left\{\begin{array}{l} I(\varphi)+I(\psi)\leq C\left(\displaystyle\iint_Q e^{\frac{-4\boldsymbol{a}+2\widetilde{\alpha}}{t^4(T-t)^4}}t^{-30}(T-t)^{-30}\,|g|^2\,dx\,dt\right.\\ \qquad+\displaystyle\iint_Q e^{\frac{-32\boldsymbol{a}+30\overline{\alpha}}{t^4(T-t)^4}}\,t^{-252}(T-t)^{-252}\,|q|^2\,dx\,dt\\ \qquad+\displaystyle\iint_{\mathcal{O}\times(0,T)} e^{\frac{-16\boldsymbol{a}+14\overline{\alpha}}{t^4(T-t)^4}}\,t^{-132}(T-t)^{-132}\,|\varphi_2|^2\,dx\,dt\\ \qquad\left.+\displaystyle\iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\boldsymbol{a}+30\overline{\alpha}}{t^4(T-t)^4}}\,t^{-268}(T-t)^{-268}\,|\psi|^2\,dx\,dt\right).\end{array}\right. \tag{32}$$

Proof. Let us first recall a Carleman inequality for the solutions of (20) which has recently been proved in [12] (Proposition 1) whenever (12)–(13) are fulfilled:

$$\begin{array}{l} s^3\lambda^4\displaystyle\iint_Q e^{-2s\alpha}\xi^3(|\varphi|^2+|\psi|^2)\,dx\,dt\\ +s\lambda^2\displaystyle\iint_Q e^{-2s\alpha}\xi(|\nabla\varphi|^2+|\nabla\psi|^2)\,dx\,dt\\ +s^{-1}\displaystyle\iint_Q e^{-2s\alpha}\xi^{-1}\left(|\varphi_t|^2+|\psi_t|^2+|\Delta\varphi|^2+|\Delta\psi|^2\right)dx\,dt\\ \leq C_5(1+T^2)\left(s^{15/2}\lambda^{24}\displaystyle\iint_Q e^{-4s\boldsymbol{b}+2s\alpha^*}\widehat{\xi}^{15/2}(|g|^2+|q|^2)\,dx\,dt\right.\\ \left.+s^{16}\lambda^{48}\displaystyle\iint_{\mathcal{O}_0\times(0,T)} e^{-8s\boldsymbol{b}+6s\alpha^*}\widehat{\xi}^{16}(|\varphi|^2+|\psi|^2)\,dx\,dt\right).\end{array} \tag{33}$$

Here, $s\geq s_1$ and $\lambda\geq\lambda_1$ are arbitrarily large and C_5, s_1 and λ_1 are suitable constants depending on Ω, $\mathcal{O}_0$, T, $\overline{y}$ and $\overline{\theta}$; see Proposition 1 in [12]. The proof of this inequality follows the same arguments employed in [7] to prove (23) and can be achieved without any further regularity on $\overline{y}$ or $\overline{\theta}$.

It is clear from (33) that, for some C_6, $\overline{\alpha}$ and $\widetilde{\alpha}$ depending on Ω, $\mathcal{O}_0$, T, $\overline{y}$ and $\overline{\theta}$, we have:

$$
\begin{aligned}
&\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}} t^{-12}(T-t)^{-12}\left(|\varphi|^2+|\psi|^2\right)dx\,dt\\
&+\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}} t^{-4}(T-t)^{-4}\left(|\nabla\varphi|^2+|\nabla\psi|^2\right)dx\,dt\\
&+\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\, t^4(T-t)^4\,\left(|\Delta\varphi|^2+|\Delta\psi|^2+|\varphi_t|^2+|\psi_t|^2\right)dx\,dt\\
&\le C_6\left(\iint_Q e^{\frac{-4\mathbf{a}+2\overline{\alpha}}{t^4(T-t)^4}} t^{-30}(T-t)^{-30}\left(|g|^2+|q|^2\right)dx\,dt\right.\\
&\left.+\iint_{\mathcal{O}_0\times(0,T)} e^{\frac{-8\mathbf{a}+6\overline{\alpha}}{t^4(T-t)^4}}\, t^{-64}(T-t)^{-64}\left(|\varphi|^2+|\psi|^2\right)dx\,dt\right).
\end{aligned}
\tag{34}
$$

Indeed, it suffices to take $\overline{\alpha}$ and $\widetilde{\alpha}$ as in (25) and

$$C_6 = C_5(1+T^2)s_1^{17}\lambda_1^{48}e^{17\lambda_1(m+1)\|\eta^0\|_\infty}.$$

We thus obtain $0<\widetilde{\alpha}<\overline{\alpha}$ and, noticing that λ_1 is large enough, $16\widetilde{\alpha}-15\overline{\alpha}>0$.

We apply (34) for the open set $\mathcal{O}_0$ defined in (26). Then we can argue as in subsection 2.1 and deduce that

$$
\begin{aligned}
&\iint_{\mathcal{O}_0\times(0,T)} e^{\frac{-8\mathbf{a}+6\overline{\alpha}}{t^4(T-t)^4}}\, t^{-64}(T-t)^{-64}\,|\varphi_1|^2\,dx\,dt\\
&\le C_7\iint_{\mathcal{O}_1\times(0,T)} e^{\frac{-16\mathbf{a}+14\overline{\alpha}}{t^4(T-t)^4}} t^{-132}(T-t)^{-132}\left(|\varphi_2|^2+|\varphi_3|^2\right)dx\,dt\\
&+\varepsilon\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}} t^4(T-t)^4\left(|\Delta\varphi_2|^2+|\Delta\varphi_3|^2\right)dx\,dt,
\end{aligned}
$$

where $\mathcal{O}_1$ is an appropriate nonempty open set verifying

$$\mathcal{O}_0\subset\mathcal{O}_1\subset\mathcal{O},\quad d_1:=\operatorname{dist}(\overline{\mathcal{O}_1},\partial\mathcal{O}\cap\Omega)>0.$$

This inequality combined with (34) yields:

$$
\begin{aligned}
&\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}} t^{-12}(T-t)^{-12}\left(|\varphi|^2+|\psi|^2\right)dx\,dt\\
&+\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}} t^{-4}(T-t)^{-4}\left(|\nabla\varphi|^2+|\nabla\psi|^2\right)dx\,dt\\
&+\iint_Q e^{\frac{-2\overline{\alpha}}{t^4(T-t)^4}}\, t^4(T-t)^4\,\left(|\Delta\varphi|^2+|\Delta\psi|^2+|\varphi_t|^2+|\psi_t|^2\right)dx\,dt\\
&\le C_8\left(\iint_Q e^{\frac{-4\mathbf{a}+2\overline{\alpha}}{t^4(T-t)^4}} t^{-30}(T-t)^{-30}\left(|g|^2+|q|^2\right)dx\,dt\right.\\
&+\iint_{\mathcal{O}_1\times(0,T)} e^{\frac{-16\mathbf{a}+14\overline{\alpha}}{t^4(T-t)^4}}\, t^{-132}(T-t)^{-132}\left(|\varphi_2|^2+|\varphi_3|^2\right)dx\,dt\\
&\left.+\iint_{\mathcal{O}\times(0,T)} e^{\frac{-8\mathbf{a}+6\overline{\alpha}}{t^4(T-t)^4}}\, t^{-64}(T-t)^{-64}\,|\psi|^2\,dx\,dt\right).
\end{aligned}
\tag{35}
$$

Our last task will be to estimate the integral

$$\iint_{\mathcal{O}_1\times(0,T)} e^{\frac{-16\boldsymbol{\alpha}+14\overline{\alpha}}{t^4(T-t)^4}}\, t^{-132}(T-t)^{-132}\,|\varphi_3|^2\,dx\,dt$$

in terms of $\varepsilon I(\varphi_3)$ and local integrals of ψ and q. To do this, we set

$$\beta_1(t) = e^{\frac{-16\boldsymbol{\alpha}+14\overline{\alpha}}{t^4(T-t)^4}}\, t^{-132}(T-t)^{-132}$$

and we introduce a function $\zeta_0 \in C^2(\overline{\mathcal{O}})$ such that

$$\zeta_0 \equiv 1 \text{ in } \mathcal{O}_1, \quad 0 \le \zeta \le 1$$

and $\zeta_0(x) = 0$ at any point $x \in \mathcal{O}$ satisfying $\operatorname{dist}(x, \partial\mathcal{O} \cap \Omega) \le d_1/2$. From the differential equation satisfied by ψ (see (20)), we have

$$\begin{aligned}\iint_{\mathcal{O}_1\times(0,T)} \beta_1(t)\,|\varphi_3|^2\,dx\,dt &\le \iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta_0\,|\varphi_3|^2\,dx\,dt\\ &= \iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta_0\,\varphi_3(-\psi_t - \Delta\psi - \overline{y}\cdot\nabla\psi - q)\,dx\,dt.\end{aligned} \tag{36}$$

To end the proof, we perform integrations by parts in the last integral and pass all the derivatives from ψ to φ_3.

First, we integrate by parts in time taking into account that $\beta_1(0) = \beta_1(T) = 0$:

$$\begin{aligned}&-\iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta_0\,\varphi_3\,\psi_t\,dx\,dt\\ &\quad= \iint_{\mathcal{O}\times(0,T)} \beta_{1,t}(t)\,\zeta_0\,\varphi_3\,\psi\,dx\,dt + \iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta_0\,\varphi_{3,t}\,\psi\,dx\,dt\\ &\quad\le \varepsilon I(\varphi_3) + C_9(\varepsilon)\iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\boldsymbol{\alpha}+30\overline{\alpha}}{t^4(T-t)^4}}\, t^{-268}(T-t)^{-268}\,|\psi|^2\,dx\,dt.\end{aligned} \tag{37}$$

Next, we integrate by parts twice in space. Here, we use the properties of the cut-off function ζ and the Dirichlet boundary conditions for φ_3 and ψ:

$$\begin{aligned}&-\iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta_0\,\varphi_3\,\Delta\psi\,dx\,dt\\ &\quad= \iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,(-\Delta\zeta_0\,\varphi_3 - 2\nabla\zeta_0\cdot\nabla\varphi_3 - \zeta_0\,\Delta\varphi_3)\,\psi\,dx\,dt\\ &\quad\le \varepsilon I(\varphi_3) + C_{10}(\varepsilon)\iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\boldsymbol{\alpha}+30\overline{\alpha}}{t^4(T-t)^4}}\, t^{-268}(T-t)^{-268}\,|\psi|^2\,dx\,dt.\end{aligned} \tag{38}$$

We also integrate by parts in the third term with respect to x and we use the incompressibility condition on $\overline{y}$:

$$\begin{aligned}&-\iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta_0\,\varphi_3\,\overline{y}\cdot\nabla\psi\,dx\,dt\\ &\quad= \iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\overline{y}\cdot(\varphi_3\,\nabla\zeta + \zeta\,\nabla\varphi_3)\psi\,dx\,dt\\ &\quad\le \varepsilon I(\varphi_3) + C_{11}(\varepsilon)\iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\boldsymbol{\alpha}+30\overline{\alpha}}{t^4(T-t)^4}}\, t^{-260}(T-t)^{-260}\,|\psi|^2\,dx\,dt.\end{aligned} \tag{39}$$

We finally apply Young's inequality in the last term and we have:

$$
\begin{aligned}
&-\iint_{\mathcal{O}\times(0,T)} \beta_1(t)\,\zeta\,\varphi_3\,q\,dx\,dt \\
&\qquad \le \varepsilon I(\varphi_3) + C_{12}(\varepsilon)\iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\mathbf{a}+30\overline{\alpha}}{t^4(T-t)^4}}\,t^{-252}(T-t)^{-252}\,|q|^2\,dx\,dt.
\end{aligned}
\tag{40}
$$

From (35), (36) and (37)–(40), it is easy to deduce the desired inequality (32). □

Arguing as in subsection 2.1, that is to say, combining the previous result and the classical energy estimates satisfied by φ and ψ, we can deduce the following Carleman inequality.

LEMMA 5. *Assume that $N = 3$, $n_1(x^0) \neq 0$ and $\mathcal{O}$ and $(\overline{y}, \overline{\theta})$ satisfy* (11)–(13). *Then, there exist positive constants C, $\overline{\alpha}$ and $\widetilde{\alpha}$ depending on Ω, $\mathcal{O}$, T, $\overline{y}$ and $\overline{\theta}$ with $0 < \widetilde{\alpha} < \overline{\alpha}$ and $16\widetilde{\alpha} - 15\overline{\alpha} > 0$ such that, for any $g \in L^2(Q)^3$, $q \in L^2(Q)$, $\varphi^0 \in H$ and $\psi^0 \in L^2(\Omega)$, the associated solution to* (20) *satisfies:*

$$
\left\{
\begin{aligned}
&\iint_Q e^{\frac{-2\overline{\alpha}}{\ell(t)^4}}\left(\ell(t)^{-12}\left(|\varphi|^2 + |\psi|^2\right) + \ell(t)^{-4}\left(|\nabla\varphi|^2 + |\nabla\psi|^2\right)\right)dx\,dt \\
&\qquad \le C\left(\iint_Q e^{\frac{-4\mathbf{a}+2\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-30}\,|g|^2\,dx\,dt\right. \\
&\qquad + \iint_Q e^{\frac{-32\mathbf{a}+30\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-252}\,|q|^2\,dx\,dt \\
&\qquad + \iint_{\mathcal{O}\times(0,T)} e^{\frac{-16\mathbf{a}+14\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-132}\,|\varphi_2|^2\,dx\,dt \\
&\qquad \left. + \iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\mathbf{a}+30\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-268}\,|\psi|^2\,dx\,dt\right),
\end{aligned}
\right.
\tag{41}
$$

where the function ℓ was defined in (30).

The similar result that can be established when $N = 2$ follows.

LEMMA 6. *Assume that $N = 2$, $n_1(x^0) \neq 0$ and $\mathcal{O}$ and $(\overline{y}, \overline{\theta})$ satisfy* (11)–(13). *Then, there exist positive constants C, $\overline{\alpha}$ and $\widetilde{\alpha}$ depending on Ω, $\mathcal{O}$, T, $\overline{y}$ and $\overline{\theta}$ with $0 < \widetilde{\alpha} < \overline{\alpha}$ and $16\widetilde{\alpha} - 15\overline{\alpha} > 0$ such that, for any $g \in L^2(Q)^2$, $q \in L^2(Q)$, $\varphi^0 \in H$ and $\psi^0 \in L^2(\Omega)$, the associated solution to* (20) *satisfies:*

$$
\left\{
\begin{aligned}
&\iint_Q e^{\frac{-2\overline{\alpha}}{\ell(t)^4}}\left(\ell(t)^{-12}\left(|\varphi|^2 + |\psi|^2\right) + \ell(t)^{-4}\left(|\nabla\varphi|^2 + |\nabla\psi|^2\right)\right)dx\,dt \\
&\qquad \le C\left(\iint_Q e^{\frac{-4\mathbf{a}+2\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-30}\,|g|^2\,dx\,dt\right. \\
&\qquad + \iint_Q e^{\frac{-32\mathbf{a}+30\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-252}\,|q|^2\,dx\,dt \\
&\qquad \left. + \iint_{\mathcal{O}\times(0,T)} e^{\frac{-32\mathbf{a}+30\overline{\alpha}}{\ell(t)^4}}\,\ell(t)^{-268}\,|\psi|^2\,dx\,dt\right).
\end{aligned}
\right.
\tag{42}
$$

2.3. An observability estimate for system (21). We will prove an observability estimate for the system

$$\begin{cases} -\rho_t - \Delta\rho - \nabla\times((\overline{y}\cdot\nabla\times)\nabla\gamma) = 0, \quad \Delta\gamma = \rho & \text{in } Q,\\ \gamma = 0, \quad \rho = 0 & \text{on } \Sigma,\\ \rho(T) = \rho^0 & \text{in } \Omega. \end{cases} \tag{43}$$

This estimate will be implied by a Carleman inequality of the form

$$S(\nabla\gamma) \le C \iint_{\mathcal{O}\times(0,T)} |\nabla\gamma|^2\, dx\, dt,$$

where $S(\nabla\gamma)$ contains several global weighted integrals involving $\nabla\gamma$ (see (44)).

LEMMA 7. *Assume that $N = 2$ and $\overline{y} \in L^\infty(Q)^2$. There exist three positive constants C, $\overline{s}$ and $\overline{\lambda}$ depending on Ω, $\mathcal{O}$, T and $\overline{y}$ such that, for any $\rho^0 \in H^{-1}(\Omega)$, the associated solution to* (43) *satisfies:*

$$\begin{cases} S(\nabla\gamma) := s^4\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^4|\nabla\gamma|^2\, dx\, dt \\ \qquad\qquad +s\lambda^2 \iint_Q e^{-2s\alpha}\,\xi|\nabla\rho|^2\, dx\, dt + s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\rho|^2\, dx\, dt \\ \qquad\qquad \le C\, s^5\lambda^6 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^5|\nabla\gamma|^2\, dx\, dt, \end{cases} \tag{44}$$

for any $s \ge \overline{s}$ and any $\lambda \ge \overline{\lambda}$. Recall that α and ξ were defined in (17).

Proof. For the proof, s_j and λ_j ($j \ge 2$) will denote various positive constants that can eventually depend on Ω, $\mathcal{O}$, T and $\overline{y}$.

Let $\mathcal{O}_0$ be a nonempty open set satisfying $\mathcal{O}_0 \subset\subset \mathcal{O}$ and let us apply to ρ a Carleman inequality for parabolic systems with right-hand sides in $L^2(0,T;H^{-1}(\Omega))$, originally proved in [15] (this version can be found in Lemma 2.1 of [6]):

$$\begin{aligned} &s\lambda^2 \iint_Q e^{-2s\alpha}\,\xi|\nabla\rho|^2\, dx\, dt + s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\rho|^2\, dx\, dt \\ &\qquad \le C_{13}\left(s^2\lambda^2\|\overline{y}\|_\infty^2 \iint_Q e^{-2s\alpha}\xi^2|\nabla(\nabla\times\gamma)|^2\, dx\, dt\right. \\ &\qquad\quad \left. +s^3\lambda^4 \iint_{\mathcal{O}_0\times(0,T)} e^{-2s\alpha}\xi^3|\rho|^2\, dx\, dt\right), \end{aligned} \tag{45}$$

for any $s \ge s_2$ and $\lambda \ge \overline{\lambda}_2$.

Observe that, here, the assumption $\rho^0 \in H^{-1}(\Omega)$ may seem too weak to apply this result. Indeed, (45) can be proved as in [15] whenever $\rho \in C^1(\overline{Q})$ and, by a continuity argument, also for the solutions of problem (43) for which the left-hand side of (45) is finite. This is our case, since one can ensure that $\rho \in L^2(Q)$ as soon as $\rho^0 \in H^{-1}(\Omega)$ (for instance, taking into account the definition of ρ as the solution by transposition of (43)).

Once (45) has been justified, let us first estimate the last integral in its right-hand side. Thus, let $\zeta \in C^2(\overline{\mathcal{O}})$ be a cut-off function satisfying

$$\zeta \equiv 1 \text{ in } \mathcal{O}_0, \quad 0 \le \zeta \le 1 \quad \text{and} \quad \zeta = 0 \text{ on } \partial\mathcal{O}.$$

We have:

$$
\begin{aligned}
s^3\lambda^4 \iint_{\mathcal{O}_0\times(0,T)} e^{-2s\alpha}\xi^3|\Delta\gamma|^2\,dx\,dt &\le s^3\lambda^4 \iint_{\mathcal{O}\times(0,T)} \zeta\, e^{-2s\alpha}\xi^3|\Delta\gamma|^2\,dx\,dt \\
&= -s^3\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^3(\nabla\zeta\cdot\nabla\gamma)\Delta\gamma\,dx\,dt \\
&\quad -3s^3\lambda^5 \iint_{\mathcal{O}\times(0,T)} \zeta\, e^{-2s\alpha}\xi^3(\nabla\eta^0\cdot\nabla\gamma)\Delta\gamma\,dx\,dt \\
&\quad +2s^4\lambda^5 \iint_{\mathcal{O}\times(0,T)} \zeta\, e^{-2s\alpha}\xi^4(\nabla\eta^0\cdot\nabla\gamma)\Delta\gamma\,dx\,dt \\
&\quad -s^3\lambda^4 \iint_{\mathcal{O}\times(0,T)} \zeta\, e^{-2s\alpha}\xi^3(\nabla\Delta\gamma\cdot\nabla\gamma)\,dx\,dt.
\end{aligned}
$$

Now, we apply Young's inequality several times and we obtain

$$
\begin{aligned}
&s^3\lambda^4 \iint_{\mathcal{O}_0\times(0,T)} e^{-2s\alpha}\xi^3|\Delta\gamma|^2\,dx\,dt \\
&\le C_{14}(\varepsilon)\, s^5\lambda^6 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^5|\nabla\gamma|^2\,dx\,dt \\
&+\varepsilon\left(s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\Delta\gamma|^2\,dx\,dt + s\lambda^2 \iint_Q e^{-2s\alpha}\xi|\nabla\Delta\gamma|^2\,dx\,dt\right),
\end{aligned}
$$

for $s \ge s_3$ and $\lambda \ge \lambda_3$ and for any small positive constant ε. Combining this, the fact that $\rho = \Delta\gamma$, and (45), we get

$$
\begin{aligned}
&s\lambda^2 \iint_Q e^{-2s\alpha}\,\xi|\nabla\rho|^2\,dx\,dt + s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\rho|^2\,dx\,dt \\
&\qquad \le C_{15}\left(s^2\lambda^2\|\overline{y}\|_\infty^2 \iint_Q e^{-2s\alpha}\xi^2|\nabla(\nabla\times\gamma)|^2\,dx\,dt\right. \\
&\qquad\quad \left. +s^5\lambda^6 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^5|\nabla\gamma|^2\,dx\,dt\right)
\end{aligned} \tag{46}
$$

for any $s \ge s_4$ and $\lambda \ge \lambda_4$.

Finally, we are going to estimate the first integral in the right-hand side of (45). To this end, let us notice that, for $j = 1$ and 2 and almost every $t \in (0,T)$, the function $\partial_j\gamma(t)$ satisfies:

$$
\Delta(\partial_j\gamma)(t) = \partial_j\rho(t) \ \text{ in } \Omega.
$$

Let us apply the main result in [14] to $\partial_j\gamma$. This yields the existence of two numbers $\widetilde{\tau} > 1$ and $\widetilde{\lambda} > 1$ such that

$$
\begin{aligned}
&\tau^4\lambda^4 \int_\Omega e^{2\tau\eta}\eta^4|\partial_j\gamma|^2(t)\,dx + \tau^2\lambda^2 \int_\Omega e^{2\tau\eta}\eta^2|\nabla(\partial_j\gamma)|^2(t)\,dx \\
&\le C_{16}\left(\tau \int_\Omega e^{2\tau\eta}\eta|\partial_j\rho|^2(t)\,dx + \tau^4\lambda^4 \int_{\mathcal{O}} e^{2\tau\eta}\eta^4|\partial_j\gamma|^2(t)\,dx\right. \\
&\quad \left. +\tau^{5/2}\lambda^2\, e^{2\tau}\|\partial_j\gamma(t)\|^2_{H^{1/2}(\partial\Omega)}\right)
\end{aligned} \tag{47}
$$

for $\tau \geq \widetilde{\tau}$ and $\lambda \geq \widetilde{\lambda}$. Here, we have introduced the function η, with

$$\eta(x) = e^{\lambda \eta^0(x)}.$$

In fact, the inequality one can find in [14] contains local integrals of $|\partial_j \gamma|^2$ and $|\nabla(\partial_j \gamma)|^2$ in the right-hand side. But it can be written for a smaller set $\mathcal{O}' \subset\subset \mathcal{O}$. Using localizing arguments together with the fact that we actually have a global weighted integral of $|\Delta(\partial_j \gamma)|^2$ in the left-hand side, (47) is easily found.

Following the same steps of [7], we set

$$\tau = \frac{s}{t^4(T-t)^4}\, e^{\lambda\, m \|\eta^0\|_\infty},$$

we multiply (47) by

$$\exp\left\{-2s\frac{e^{5/4\lambda m\|\eta^0\|_\infty}}{t^4(T-t)^4}\right\}$$

and we integrate in time over $(0,T)$. This gives

$$\begin{aligned}
&s^4\lambda^4 \iint_Q e^{-2s\alpha}\xi^4|\partial_j\gamma|^2\,dx\,dt + s^2\lambda^2 \iint_Q e^{-2s\alpha}\xi^2|\nabla(\partial_j\gamma)|^2\,dx\,dt \\
&\quad \leq C_{17}\left(s\iint_Q e^{-2s\alpha}\xi|\partial_j\rho|^2\,dx\,dt + s^4\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^4|\partial_j\gamma|^2\,dx\,dt\right. \\
&\quad \left. + s^{5/2}\lambda^2 \int_0^T e^{-2s\alpha^*}(\xi^*)^{5/2}\|\partial_j\gamma\|^2_{H^{1/2}(\partial\Omega)}\right)
\end{aligned}$$

for $s \geq s_5$ and $\lambda \geq \widetilde{\lambda}$. Combining this estimate and (46), we have

$$\begin{aligned}
&s^4\lambda^4 \iint_Q e^{-2s\alpha}\xi^4|\partial_j\gamma|^2\,dx\,dt + s\lambda^2 \iint_Q e^{-2s\alpha}\,\xi|\nabla\rho|^2\,dx\,dt \\
&\quad + s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\rho|^2\,dx\,dt \leq C_{18}\left(s^{5/2}\lambda^2\int_0^T e^{-2s\alpha^*}(\xi^*)^{5/2}\|\partial_j\gamma\|^2_{H^{1/2}(\partial\Omega)}\right. \\
&\quad \left. + s^5\lambda^6 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^5|\nabla\gamma|^2\,dx\,dt\right)
\end{aligned}$$

for any $s \geq s_6$ and $\lambda \geq \lambda_5$. On the other hand, the boundary term can readily be bounded using the continuity of the trace operator:

$$\|\partial_j\gamma(t)\|^2_{H^{1/2}(\partial\Omega)} \leq C_{19}(\|\partial_j\gamma(t)\|^2_{L^2} + \|\nabla(\partial_j\gamma)(t)\|^2_{L^2}).$$

Furthermore, since $\gamma|_\Sigma \equiv 0$, we know that there exists a positive constant C_{20} such that

$$\|\nabla(\partial_j\gamma)(t)\|_{L^2} \leq C_{20}\|\Delta\gamma(t)\|_{L^2} \quad \text{a.e. in } (0,T) \text{ for } j = 1,2.$$

Consequently,

$$\begin{aligned}
&s^4\lambda^4 \iint_Q e^{-2s\alpha}\xi^4|\partial_j\gamma|^2\,dx\,dt + s\lambda^2 \iint_Q e^{-2s\alpha}\,\xi|\nabla\rho|^2\,dx\,dt \\
&\quad + s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|\rho|^2\,dx\,dt \leq C_{21}\, s^5\lambda^6 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\xi^5|\nabla\gamma|^2\,dx\,dt
\end{aligned}$$

for $s \geq s_6$.

This implies (44) and ends the proof of Lemma 7. □

Remark 3. An almost immediate consequence of the Carleman estimate (44) is the following observability inequality:

$$\|(\nabla\gamma)(0)\|^2_{L^2} \le C \iint_{\mathcal{O}\times(0,T)} |\nabla\gamma|^2 \, dx\, dt. \tag{48}$$

Proof. All comes to prove a dissipation result for the L^2 norm of $\nabla\gamma$. Indeed, if we can prove that

$$\|\nabla\gamma(t_1)\|^2_{L^2} \le C\|\nabla\gamma(t_2)\|^2_{L^2} \quad \forall 0 \le t_1 < t_2 \le T, \tag{49}$$

then using the properties of the weight function $e^{-2s\alpha}$ and estimate (44), we readily deduce (48).

Thus, we multiply the equation in (43) by $-\gamma$ and we integrate in Ω. Taking into account that γ and ρ vanish on $\partial\Omega$, this yields:

$$-\frac{1}{2}\frac{d}{dt}\int_\Omega |\nabla\gamma|^2\, dx + \int_\Omega |\Delta\gamma|^2\, dx - \int_\Omega ((\overline{y}\cdot\nabla\times)\nabla\gamma)\cdot\nabla\times\gamma\, dx = 0,$$

from which the dissipation estimate (49) follows. □

In fact, this is what will be used in section 3 to prove the null controllability of system (16).

3. Null controllability of the linearized systems (14), (15) and (16).

3.1. Null controllability of (14). We are dealing here with the following system:

$$\begin{cases} y_t - \Delta y + (\overline{y}\cdot\nabla)y + (y\cdot\nabla)\overline{y} + \nabla p = f + v1_{\mathcal{O}}, \quad \nabla\cdot y = 0 & \text{in } Q,\\ y = 0 & \text{on } \Sigma,\\ y(0) = y^0 & \text{in } \Omega, \end{cases} \tag{50}$$

where $\mathcal{O}$ satisfies (11) and $\overline{y}$ satisfies (12). Our goal will be to find a control v such that $y(T) = 0$ in Ω.

Let us introduce some weight functions:

$$\beta_2(t) = \exp\left\{\frac{\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^6, \quad \beta_3(t) = \exp\left\{\frac{2\widetilde{\alpha}-\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^{15}$$

and

$$\beta_4(t) = \exp\left\{\frac{8\widetilde{\alpha}-7\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^{66}$$

(recall that ℓ was defined in (30)), where $\overline{\alpha}$ and $\widetilde{\alpha}$ are the constants provided by Lemma 2 when $N = 3$ and Lemma 3 when $N = 2$. Recall that, in particular, $0 < \widetilde{\alpha} < \overline{\alpha}$ and $8\widetilde{\alpha} - 7\overline{\alpha} > 0$.

Of course, we will need some specific conditions on f and y^0 to get the null controllability of (50). We will use the arguments in [7].

Thus, let us set

$$Ly = y_t - \Delta y + (\overline{y},\nabla)y + (y,\nabla)\overline{y} \tag{51}$$

and let us introduce the spaces

$$E_2 = \{(y,p,v) : (y,v) \in E_0,\ \ell^{-4}\beta_2(Ly + \nabla p - v1_{\mathcal{O}}) \in L^2(0,T;H^{-1}(\Omega)^2)\}$$

when $N = 2$ and

$$E_3 = \{(y,p,v) : (y,v) \in E_0,\ \ell^{-2}\beta_2^{1/2} y \in L^4(0,T;L^{12}(\Omega)^3),\\ \ell^{-4}\beta_2(Ly + \nabla p - v1_{\mathcal{O}}) \in L^2(0,T;W^{-1,6}(\Omega)^3)\}$$

when $N = 3$, where

$$E_0 = \{(y,v) : \beta_3\, y,\ \beta_4\, v1_{\mathcal{O}} \in L^2(Q)^N,\ v_1 \equiv 0,\\ \ell^{-2}\beta_2^{1/2} y \in L^2(0,T;V) \cap L^\infty(0,T;H)\}.$$

It is clear that E_N is a Banach space for the norm $\|\cdot\|_{E_N}$, where

$$\|(y,p,v)\|_{E_2} = \Big(\|\beta_3\, y\|^2_{L^2} + \|\beta_4\, v1_{\mathcal{O}}\|^2_{L^2} + \|\ell^{-2}\beta_2^{1/2} y\|^2_{L^2(0,T;V)}\\ + \|\ell^{-2}\beta_2^{1/2} y\|^2_{L^\infty(0,T;H)} + \|\ell^{-4}\beta_2(Ly + \nabla p - v1_{\mathcal{O}})\|^2_{L^2(0,T;H^{-1})}\Big)^{1/2}$$

and

$$\|(y,p,v)\|_{E_3} = \Big(\|\beta_3\, y\|^2_{L^2} + \|\beta_4\, v1_{\mathcal{O}}\|^2_{L^2} + \|\ell^{-2}\beta_2^{1/2} y\|^2_{L^2(0,T;V)}\\ + \|\ell^{-2}\beta_2^{1/2} y\|^2_{L^\infty(0,T;H)} + \|\ell^{-2}\beta_2^{1/2}\, y\|^2_{L^4(0,T;L^{12})}\\ + \|\ell^{-4}\beta_2(Ly + \nabla p - v1_{\mathcal{O}})\|^2_{L^2(0,T;W^{-1,6})}\Big)^{1/2}.$$

Remark 4. The spaces E_j $(j = 0, 2, 3)$ are natural spaces where solutions of the null controllability of (50) must be found in order to preserve these properties for the nonlinear term $(y \cdot \nabla)y$. More details are provided in subsection 4.1.

PROPOSITION 1. *Assume that $n_1(x^0) \neq 0$ and $\mathcal{O}$ and $\overline{y}$ verify* (11)–(12). *Let $y^0 \in E$ and let us assume that*

$$\ell^{-4}\beta_2 f \in \begin{cases} L^2(0,T;H^{-1}(\Omega)^2) & \text{if } N = 2,\\ L^2(0,T;W^{-1,6}(\Omega)^3) & \text{if } N = 3. \end{cases}$$

Then, we can find a control v such that the associated solution (y,p) to (50) *satisfies $(y,p,v) \in E_N$. In particular, $v_1 \equiv 0$ and $y(T) = 0$.*

Sketch of the proof. The proof of this proposition is very similar to the one of Proposition 2 in [7], so we will just give the main ideas. For simplicity, we will only consider the case $N = 3$. When $N = 2$, the proof is even easier.

Following the arguments in [9] and [13], let us introduce the auxiliary optimal control problem

$$\begin{cases} \inf\ \dfrac{1}{2}\left(\displaystyle\iint_Q |\beta_3\, y|^2 dxdt + \iint_{\mathcal{O}\times(0,T)} |\beta_4\, v|^2 dxdt\right)\\ \text{subject to } v \in L^2(Q)^3,\ \text{supp } v \subset \mathcal{O}\times(0,T),\ v_1 \equiv 0 \text{ and}\\ \qquad \begin{cases} Ly + \nabla p = f + v1_{\mathcal{O}} & \text{in } Q,\\ \nabla \cdot y = 0 & \text{in } Q,\\ y = 0 & \text{on } \Sigma,\\ y(0) = y^0, \quad y(T) = 0 & \text{in } \Omega. \end{cases} \end{cases} \tag{52}$$

Notice that a solution $(\widehat{y},\widehat{p},\widehat{v})$ to (52) is a good candidate to satisfy $(\widehat{y},\widehat{p},\widehat{v}) \in E_3$.

For the moment, let us assume that (52) possesses a solution $(\widehat{y},\widehat{p},\widehat{v})$. Then, by virtue of Lagrange's principle, there must exist dual variables $\widehat{z}$ and $\widehat{q}$ such that

$$\left\{\begin{array}{ll} \widehat{y}=\beta_3^{-2}(L^*\widehat{z}+\nabla\widehat{q}), \quad \nabla\cdot\widehat{z}=0 & \text{in } Q,\\ \widehat{v}_1\equiv 0, \quad \widehat{v}_i=-\beta_4^{-2}\,\widehat{z}_i \quad (i=2,3) & \text{in } \mathcal{O}\times(0,T),\\ \widehat{z}=0 & \text{on } \Sigma, \end{array}\right. \tag{53}$$

where L^* is the adjoint operator of L, i.e.,

$$L^*z=-z_t-\Delta z-(Dz)\,\overline{y}.$$

At least formally, the couple $(\widehat{z},\widehat{q})$ satisfies

$$a((\widehat{z},\widehat{q}),(w,h))=\langle G,(w,h)\rangle \quad \forall (w,h)\in P_0, \tag{54}$$

where P_0 is the space

$$P_0=\{(w,h)\in C^2(\overline{Q})^4 : \nabla\cdot w=0,\ w=0 \text{ on } \Sigma,\ \int_{\mathcal{O}} h(x,t)\,dx=0\}$$

and we have used the notation

$$\begin{aligned} a((\widehat{z},\widehat{q}),(w,h))=&\iint_Q \beta_3^{-2}\,(L^*\widehat{z}+\nabla\widehat{q})\cdot(L^*w+\nabla h)\,dx\,dt\\ &+\iint_{\mathcal{O}\times(0,T)}\beta_4^{-2}\,(\widehat{z}_2\,w_2+\widehat{z}_3\,w_3)\,dx\,dt \end{aligned}$$

and

$$\langle G,(w,h)\rangle=\int_0^T \langle f(t),w(t)\rangle_{H^{-1},H_0^1}\,dt+\int_\Omega y^0\cdot w(0)\,dx.$$

Conversely, if we are able to "solve" (54) and then use (53) to define $(\widehat{y},\widehat{p},\widehat{v})$, we will probably have found a solution to (52).

Thus, let us consider the linear space P_0. It is clear that $a(\cdot\,,\cdot):P_0\times P_0\mapsto\mathbf{R}$ is a symmetric, definite positive bilinear form on P_0. We will denote by P the completion of P_0 for the norm induced by $a(\cdot\,,\cdot)$. Then $a(\cdot\,,\cdot)$ is well-defined, continuous and again definite positive on P. Furthermore, in view of the Carleman estimate (29), the linear form $(w,h)\mapsto\langle G,(w,h)\rangle$ is well-defined and continuous on P. Hence, from Lax-Milgram's lemma, we deduce that the variational problem

$$\left\{\begin{array}{l} a((\widehat{z},\widehat{q}),(w,h))=\langle G,(w,h)\rangle\\ \forall (w,h)\in P, \quad (\widehat{z},\widehat{q})\in P, \end{array}\right. \tag{55}$$

possesses exactly one solution $(\widehat{z},\widehat{q})$.

Let $\widehat{y}$ and $\widehat{v}$ be given by (53). Then, it is readily seen that they verify

$$\iint_Q \beta_3^2|\widehat{y}|^2dx\,dt+\iint_{\mathcal{O}\times(0,T)}\beta_4^2|\widehat{v}|^2dx\,dt<+\infty$$

and, also, that $\widehat{y}$ is, together with some pressure $\widehat{p}$, the weak solution (belonging to $L^2(0,T;V)\cap L^\infty(0,T;H)$) of the Stokes system in (52) for $v=\widehat{v}$.

In order to prove that $(\widehat{y},\widehat{p},\widehat{v})\in E_3$, it only remains to check that $\ell^{-2}\,\beta_2^{1/2}\,\widehat{y}$ is, together with $\ell^{-2}\,\beta_2^{1/2}\,\widehat{p}$, a weak solution of a Stokes problem of the kind (50) with a right-hand side in $L^2(0,T;W^{-1,6}(\Omega)^3)$ that belongs to $L^4(0,T;L^{12}(\Omega)^3)$. To this end, we define the functions $y^*=\ell^{-2}\,\beta_2^{1/2}\,\widehat{y}$, $p^*=\ell^{-2}\,\beta_2^{1/2}\,\widehat{p}$ and $f^*=\ell^{-2}\,\beta_2^{1/2}(f+\widehat{v}1_{\mathcal{O}})$. Then (y^*,p^*) satisfies

$$\left\{\begin{array}{ll} Ly^*+\nabla p^*=f^*+(\ell^{-2}\beta_2^{-1/2})_t\,\widehat{y},\quad \nabla\cdot y^*=0 & \text{in } Q,\\ y^*=0 & \text{on } \Sigma,\\ y^*(0)=\ell^{-2}(0)\beta_2^{1/2}(0)y^0 & \text{in } \Omega. \end{array}\right. \tag{56}$$

From the fact that $f^*\in L^2(0,T;H^{-1}(\Omega)^3)$ and $y^0\in H$, we have indeed

$$y^*\in L^2(0,T;V)\cap L^\infty(0,T;H).$$

Finally, we deduce that $y^*\in L^4(0,T;L^{12}(\Omega)^3)$ from Lemma 2 in [7]. This ends the sketch of the proof of Proposition 1. □

3.2. Null controllability of system (15). We will establish the null controllability of the linear system

$$\left\{\begin{array}{ll} y_t-\Delta y+(\overline{y}\cdot\nabla)y+(y\cdot\nabla)\overline{y}+\nabla p=f+v1_{\mathcal{O}}+\theta\,e_N & \text{in } Q,\\ \nabla\cdot y=0 & \text{in } Q,\\ \theta_t-\Delta\theta+\overline{y}\cdot\nabla\theta+y\cdot\nabla\overline{\theta}=k+h1_{\mathcal{O}} & \text{in } Q,\\ y=0,\quad \theta=0 & \text{on } \Sigma,\\ y(0)=y^0,\quad \theta(0)=\theta^0 & \text{in } \Omega, \end{array}\right. \tag{57}$$

where $\mathcal{O}$ satisfies (11) and $\overline{y}$ and $\overline{\theta}$ satisfy (12) and (13), for suitable right-hand sides f and k.

The arguments we present here are completely analogous to those in [12] and subsection 3.1 of this paper, so that we will only give a sketch. Thus, we restrict ourselves again to the three-dimensional case with $n_1(x^0)\neq 0$.

Let us introduce the weight functions

$$\beta_5(t)=\exp\left\{\frac{\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^6,\quad \beta_6(t)=\exp\left\{\frac{2\widetilde{\alpha}-\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^{15},$$

$$\beta_7(t)=\exp\left\{\frac{16\widetilde{\alpha}-15\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^{126},\quad \beta_8(t)=\exp\left\{\frac{8\widetilde{\alpha}-7\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^{66}$$

and

$$\beta_9(t)=\exp\left\{\frac{16\widetilde{\alpha}-15\overline{\alpha}}{\ell(t)^4}\right\}\ell(t)^{134},$$

where the constants $\overline{\alpha}$ and $\widetilde{\alpha}$ are furnished by Lemma 5 when $N=3$ and Lemma 6 when $N=2$ (and, in particular, $0<\widetilde{\alpha}<\overline{\alpha}$ and $16\widetilde{\alpha}-15\overline{\alpha}>0$).

Let us set

$$P\theta=\theta_t-\Delta\theta+\overline{y}\cdot\nabla\theta \tag{58}$$

and let us introduce the spaces

$$\begin{aligned}\widetilde{E}_2 = \{&(y,p,\theta,v,h) : (y,\theta,v,h) \in \widetilde{E}_0,\\ &\ell^{-4}\beta_5(Ly + \nabla p - v1_{\mathcal{O}}) \in L^2(0,T;H^{-1}(\Omega)^2),\\ &\ell^{-4}\beta_5(P\theta + y\cdot\nabla\overline{\theta} - h1_{\mathcal{O}}) \in L^2(0,T;H^{-1}(\Omega))\}\end{aligned}$$

when $N = 2$ and

$$\begin{aligned}\widetilde{E}_3 = \{&(y,p,\theta,v,h) : (y,\theta,v,h) \in \widetilde{E}_0,\\ &\ell^{-2}\beta_5^{1/2} y \in L^4(0,T;L^{12}(\Omega)^3),\\ &\ell^{-4}\beta_5(Ly + \nabla p - v1_{\mathcal{O}}) \in L^2(0,T;W^{-1,6}(\Omega)^3),\\ &\ell^{-4}\beta_5(P\theta + y\cdot\nabla\overline{\theta} - h1_{\mathcal{O}}) \in L^2(0,T;H^{-1}(\Omega))\}\end{aligned}$$

when $N = 3$, where

$$\begin{aligned}\widetilde{E}_0 = \{&(y,\theta,v,h) : (\beta_6\, y)_i,\ \beta_7\,\theta,\ (\beta_8\, v1_{\mathcal{O}})_i,\ \beta_9\, h1_{\mathcal{O}} \in L^2(Q)\ (1\le i\le N),\\ &v_1 \equiv v_N \equiv 0,\ \ell^{-2}\beta_5^{1/2} y \in L^2(0,T;V)\cap L^\infty(0,T;H),\\ &\ell^{-2}\beta_5^{1/2}\theta \in L^2(0,T;H_0^1(\Omega))\cap L^\infty(0,T;L^2(\Omega))\}.\end{aligned}$$

It can be readily seen now that $\widetilde{E}_0$, $\widetilde{E}_2$ and $\widetilde{E}_3$ are Banach spaces for the norms

$$\begin{aligned}\|(y,\theta,v,h)\|_{\widetilde{E}_0} = \big(&\|\beta_6\, y\|^2_{L^2} + \|\beta_7\,\theta\|^2_{L^2} + \|\beta_8\, v\|^2_{L^2}\\ &+\|\beta_9\, h\|^2_{L^2} + \|\ell^{-2}\beta_5^{1/2} y\|^2_{L^2(0,T;V)} + \|\ell^{-2}\beta_5^{1/2} y\|^2_{L^\infty(0,T;H)}\\ &+\|\ell^{-2}\beta_5^{1/2}\theta\|^2_{L^2(0,T;H_0^1)} + \|\ell^{-2}\beta_5^{1/2}\theta\|^2_{L^\infty(0,T;L^2)}\big)^{1/2},\\ \|(y,p,\theta,v,h)\|_{\widetilde{E}_2} = \Big(&\|(y,\theta,v,h)\|^2_{\widetilde{E}_0}\\ &+\|\ell^{-4}\beta_5(Ly+\nabla p - v1_{\mathcal{O}})\|^2_{L^2(0,T;H^{-1})}\\ &+\|\ell^{-4}\beta_5(P\theta + y\cdot\nabla\overline{\theta} - h1_{\mathcal{O}})\|^2_{L^2(0,T;H^{-1})}\Big)^{1/2}\end{aligned}$$

and

$$\begin{aligned}\|(y,p,\theta,v,h)\|_{\widetilde{E}_3} = \Big(&\|(y,\theta,v,h)\|^2_{\widetilde{E}_0} + \|\ell^{-2}\beta_5^{1/2} y\|^2_{L^4(0,T;L^{12})}\\ &+\|\ell^{-4}\beta_5(Ly+\nabla p - v1_{\mathcal{O}})\|^2_{L^2(0,T;W^{-1,6})}\\ &+\|\ell^{-4}\beta_5(P\theta + y\cdot\nabla\overline{\theta} - h1_{\mathcal{O}})\|^2_{L^2(0,T;H^{-1})}\Big)^{1/2}.\end{aligned}$$

PROPOSITION 2. *Assume that $n_1(x^0)\neq 0$ and $\mathcal{O}$ and $(\overline{y},\overline{\theta})$ satisfy* (11)–(13). *Let $y^0 \in E$, $\theta^0 \in L^2(\Omega)$ and let us assume that*

$$\ell^{-4}\beta_1(f,k) \in \begin{cases} L^2(0,T;H^{-1}(\Omega)^2)\times L^2(0,T;H^{-1}(\Omega)) & \text{if } N=2,\\ L^2(0,T;W^{-1,6}(\Omega)^3)\times L^2(0,T;H^{-1}(\Omega)) & \text{if } N=3.\end{cases}$$

Then, we can find controls v and h such that the associated solution to (57) *satisfies $(y,p,\theta,v,h)\in\widetilde{E}_N$. In particular, $v_1\equiv v_N\equiv 0$ and $y(T)=\theta(T)=0$.*

We omit the proof of this proposition, since it is essentially the same as the one of Proposition 2 in [12] and follows the steps of Proposition 1 above. As we have already indicated, the main ideas come from [13].

3.3. Null controllability of system (16). We will prove the null controllability of the linear system

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)\overline{y} + \nabla p = v1_{\mathcal{O}}, \quad \nabla\cdot y = 0 & \text{in } Q,\\ y\cdot n = 0, \quad \nabla\times y = 0 & \text{on } \Sigma,\\ y(0) = y^0 & \text{in } \Omega, \end{cases} \tag{59}$$

where $N=2$ and $\overline{y}\in L^\infty(Q)^2$.

For this purpose, we first rewrite this system using the streamline-vorticity formulation. Thus, setting $\omega = \nabla\times y$, we have

$$\begin{cases} \omega_t - \Delta\omega + \nabla\times((\nabla\times\psi\cdot\nabla)\overline{y}) = \nabla\times(v1_{\mathcal{O}}), \quad \Delta\psi = \omega & \text{in } Q,\\ \psi = 0, \quad \omega = 0 & \text{on } \Sigma,\\ \omega(0) = \nabla\times y^0 & \text{in } \Omega. \end{cases} \tag{60}$$

PROPOSITION 3. *Assume that $y^0\in H$ and $\overline{y}\in L^\infty(Q)^2$. Then, there exists a constant $C(\Omega,\mathcal{O},T)>0$ and controls $v1_{\mathcal{O}}$ with $v\in W$ (W was defined in* (9)*), such that*

$$\|v\|_{L^2} \le C\|y^0\|_H \tag{61}$$

and the associated solutions of (59) *satisfy*

$$y\in L^2(0,T;H^1(\Omega)^2)\cap C^0([0,T];L^2(\Omega)^2), \quad y_t\in L^2(0,T;H^{-1}(\Omega)^2), \tag{62}$$

and $y(T)=0$, with

$$\|y\|_{L^2(0,T;H^1)} + \|y\|_{C^0([0,T];L^2)} + \|y_t\|_{L^2(0,T;H^{-1})} \le C\|y^0\|_H. \tag{63}$$

Proof. We first establish the null controllability property for y. This can be done in several ways. One of them is the following. We first define for each $\varepsilon>0$ the functional

$$\begin{cases} J_\varepsilon(\gamma^0) = \dfrac{1}{2}\displaystyle\iint_{\mathcal{O}\times(0,T)} |\nabla\times\gamma|^2\,dx\,dt + \varepsilon\|\nabla\gamma^0\|_{L^2} + ((\nabla\times\gamma)(0), y^0)_{L^2}\\ \forall\gamma^0\in H_0^1(\Omega), \end{cases}$$

where γ is given by (43) with $\rho^0 = \Delta\gamma^0\in H^{-1}(\Omega)$.

It is not difficult to see from the observability inequality (48) that this functional possesses a unique minimizer $\gamma_\varepsilon^0\in H_0^1(\Omega)$ (see Proposition 2.1 in [5]). Now, from the necessary conditions for J_ε to reach a minimum, we have

$$\begin{aligned} &\iint_Q ((\nabla\times\gamma_\varepsilon)1_{\mathcal{O}})\cdot(\nabla\times\gamma)\,dx\,dt + \varepsilon\Big(\frac{\nabla\times\gamma_\varepsilon^0}{\|\nabla\times\gamma_\varepsilon^0\|_{L^2}}, \nabla\times\gamma^0\Big)_{L^2}\\ &\quad + ((\nabla\times\gamma)(0), y^0)_{L^2} = 0 \quad \forall\gamma^0\in H_0^1(\Omega). \end{aligned} \tag{64}$$

Thus, setting $v_\varepsilon = (\nabla\times\gamma_\varepsilon)1_{\mathcal{O}}$ and putting $\gamma^0=\gamma_\varepsilon^0$, we find from (48) and (64) that (61) holds for v_ε for some C independent of ε:

$$\|\nabla\times\gamma_\varepsilon\|_{L^2(\mathcal{O}\times(0,T))^2} = \|v_\varepsilon\|_{L^2(\mathcal{O}\times(0,T))^2} \le C. \tag{65}$$

Let us denote by $(\omega_\varepsilon, \psi_\varepsilon)$ the solution to (60) for $v = v_\varepsilon$. Then, taking into account the systems satisfied by (ρ, γ) and $(\omega_\varepsilon, \psi_\varepsilon)$, we deduce that

$$\iint_Q \nabla \times (v_\varepsilon 1_{\mathcal{O}})\, \gamma \, dx\, dt + (\nabla \times \gamma^0, (\nabla \times \psi_\varepsilon)(T))_{L^2} - ((\nabla \times \gamma)(0), y^0)_{L^2} = 0 \quad \forall \gamma^0 \in H_0^1(\Omega).$$

Combining this and (64), we obtain

$$\|(\nabla \times \psi_\varepsilon)(T)\|_{L^2} \leq \varepsilon. \tag{66}$$

From (65) and (66) written for each $\varepsilon > 0$, we deduce that, at least for a subsequence, $v_\varepsilon \to v$ weakly in $L^2(\mathcal{O} \times (0,T))^2$, where the control $v1_{\mathcal{O}}$ is such that the corresponding solution (ω, ψ) to (60) satisfies

$$(\nabla \times \psi)(T) = y(T) = 0 \quad \text{in } \Omega.$$

Since $v \in L^2(\mathcal{O} \times (0,T))^2$ and $\nabla \cdot v = 0$ in $\mathcal{O} \times (0,T)$, we necessarily have $v \in W$ (from De Rham's lemma applied to $(v_2, -v_1)$).

In order to obtain the desired regularity for y, we will consider again the equations satisfied by ψ and ω and we will check that

$$\psi \in L^2(0,T; H^2(\Omega)) \cap C^0([0,T]; H_0^1(\Omega)) \ \text{ and } \ \psi_t \in L^2(Q), \tag{67}$$

with appropriate estimates.

For simplicity, we will only present the estimates. The rigorous argument relies on introducing a standard Galerkin approximation of (60) with a "special" basis of $H_0^1(\Omega)$ (more precisely, the basis formed by the eigenfunctions of the Laplacian-Dirichlet operator in Ω) and deducing for the associated approximate solutions the estimates below.

Thus, let us multiply the first equation in (60) by ψ and let us integrate by parts. We find that

$$\frac{1}{2}\int_\Omega |\nabla\psi(t)|^2\, dx + \int_0^t\int_\Omega |\Delta\psi|^2\, dx\, d\tau = \int_0^t\int_{\mathcal{O}} v \cdot (\nabla \times \psi)\, dx\, d\tau - \int_0^t\int_\Omega (((\nabla \times \psi), \nabla)\, \overline{y}) \cdot (\nabla \times \psi)\, dx\, d\tau + \frac{1}{2}\|(\nabla\psi)(0)\|_{L^2}^2$$

for all $t \in (0,T)$. If we integrate by parts in the last integral, we also have

$$-\int_0^t\int_\Omega (((\nabla \times \psi) \cdot \nabla)\, \overline{y}) \cdot (\nabla \times \psi)\, dx\, d\tau = \int_0^t\int_\Omega ((\nabla \times \psi) \cdot \nabla)(\nabla \times \psi) \cdot \overline{y}\, dx\, d\tau.$$

Since $\psi|_\Sigma \equiv 0$, we deduce that

$$\psi \in L^2(0,T; H^2(\Omega)) \cap L^\infty(0,T; H_0^1(\Omega)) \tag{68}$$

and

$$\|\psi\|_{L^2(0,T;H^2)} + \|\psi\|_{L^\infty(0,T;H_0^1)} \leq C\|y^0\|_{L^2}. \tag{69}$$

Now, let us introduce for each t the function $\psi^*(t) = \Delta^{-1}\psi_t(t)$, i.e., the solution to

$$\begin{cases} -\Delta\psi^*(t) = \psi_t(t) & \text{in } \Omega \\ \psi^*(t) = 0 & \text{on } \partial\Omega. \end{cases}$$

Observe that, whenever $\psi_t(t) \in L^2(\Omega)$, this function satisfies $\psi^*(t) \in H^2(\Omega) \cap H_0^1(\Omega)$ and

$$\|\psi^*(t)\|_{H^2} \leq C\|\psi_t(t)\|_{L^2}. \tag{70}$$

Then, we multiply the first equation of (60) by ψ^* and we integrate by parts. This gives

$$\iint_Q |\psi_t|^2\,dx\,dt = \iint_Q (\Delta\psi)\,\psi_t\,dx\,dt - \iint_Q ((\nabla\times\psi)\cdot\nabla)(\nabla\times\psi^*)\cdot\overline{y}\,dx\,dt + \iint_{\mathcal{O}\times(0,T)} v\cdot(\nabla\times\psi^*)\,dx\,dt.$$

Using that $v \in L^2(Q)^2$ and we already have (69) and (70), we conclude that $\psi_t \in L^2(Q)$ and

$$\|\psi_t\|_{L^2} \leq C\|y^0\|_{L^2}. \tag{71}$$

From (69) and (71), we immediately obtain (67), (62) and (63).

This ends the proof of Proposition 3. □

4. Proofs of the controllability results for the nonlinear systems. In this last section, we will give the proofs of Theorems 1, 2 and 3. For the proofs of Theorems 1 and 2 we employ an inverse mapping theorem, while a fixed point argument is used for Theorem 3.

4.1. Proof of Theorem 1. We also follow here the steps in [7].

Thus, we set $y = \overline{y} + z$ and $p = \overline{p} + \chi$ and we use these identities in (1). Taking into account that $(\overline{y}, \overline{p})$ solves (5), we find:

$$\begin{cases} Lz + (z\cdot\nabla)z + \nabla\chi = v1_{\mathcal{O}}, \quad \nabla\cdot z = 0 & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(0) = y^0 - \overline{y}^0 & \text{in } \Omega \end{cases} \tag{72}$$

(recall that L was defined in (51)).

This way, we have reduced our problem to a local null controllability result for the solution (z, χ) to the *nonlinear* problem (72).

We will use the following inverse mapping theorem (see [1]).

THEOREM 4. *Let B_1 and B_2 be two Banach spaces and let $\mathcal{A} : B_1 \mapsto B_2$ satisfy $\mathcal{A} \in C^1(B_1; B_2)$. Assume that $b_0 \in B_1$, $\mathcal{A}(b_0) = d_0$ and also that $\mathcal{A}'(b_0) : B_1 \mapsto B_2$ is surjective. Then there exists $\delta > 0$ such that, for every $d \in B_2$ satisfying $\|d - d_0\|_{B_2} < \delta$, there exists a solution of the equation*

$$\mathcal{A}(b) = d, \quad b \in B_1.$$

We will apply this result with $B_1 = E_N$,

$$B_2 = \begin{cases} L^2(\ell^{-4}\beta_2; 0, T; H^{-1}(\Omega)^2) \times H & \text{if } N = 2, \\ L^2(\ell^{-4}\beta_2; 0, T; W^{-1,6}(\Omega)^3) \times (H \cap L^4(\Omega)^3) & \text{if } N = 3 \end{cases}$$

and

$$\mathcal{A}(z, \chi, v) = (Lz + (z \cdot \nabla)z + \nabla\chi - v1_{\mathcal{O}}, z(0)) \quad \forall (z, \chi, v) \in E_N.$$

From the facts that $\ell^{-2}\beta_2^{1/2} y \in L^4(0, T; L^{12}(\Omega)^3)$ and $\mathcal{A}$ is bilinear, it is not difficult to check that $\mathcal{A} \in C^1(B_1; B_2)$; more details can be found in [13] or [7].

Let b_0 be the origin in B_1. Notice that $\mathcal{A}'(0,0,0) : B_1 \mapsto B_2$ is given by

$$\mathcal{A}'(0,0,0)(z, \chi, v) = (Lz + \nabla\chi - v1_{\mathcal{O}}, z(0)) \quad \forall (z, \chi, v) \in E_N$$

and is surjective, in view of the null controllability result for (14) given in Proposition 1.

Consequently, we can indeed apply theorem 4 with these data and there exists $\delta > 0$ such that, if $\|z(0)\|_E \leq \delta$, then we find a control v satisfying $v_1 \equiv 0$ such that the associated solution to (72) verifies $z(T) = 0$ in Ω.

This concludes the proof of Theorem 1.

4.2. Proof of Theorem 2. Again, we follow here the ideas of [12].

Therefore, we set $y = \overline{y} + z$, $p = \overline{p} + \chi$ and $\theta = \overline{\theta} + \rho$, so from (2) and (7), we find:

$$(73) \qquad \begin{cases} Lz + (z \cdot \nabla)z + \nabla\chi = v1_{\mathcal{O}} + \rho\, e_N, \quad \nabla \cdot z = 0 & \text{in } Q, \\ P\rho + (z \cdot \nabla)\rho + z \cdot \nabla\overline{\theta} = h1_{\mathcal{O}} & \text{in } Q, \\ z = 0, \quad \rho = 0 & \text{on } \Sigma, \\ z(0) = y^0 - \overline{y}^0, \quad \rho(0) = \theta^0 - \overline{\theta}(0) & \text{in } \Omega \end{cases}$$

(L and P were respectively defined in (51) and (58)).

We are thus led to prove the local null controllability of (73). To this end, we will use again Theorem 4, which was presented in subsection 4.1. Using the same notation as there, we set $B_1 = \widetilde{E}_N$,

$$B_2 = L^2(\ell^{-4}\beta_5; 0, T; H^{-1}(\Omega)^3) \times H \times L^2(\Omega)$$

if $N = 2$ and

$$B_2 = L^2(\ell^{-4}\beta_5; 0, T; W^{-1,6}(\Omega)^3 \times H^{-1}(\Omega)) \times (L^4(\Omega)^3 \cap H) \times L^2(\Omega)$$

if $N = 3$.

Let us introduce $\mathcal{A}$, with

$$\mathcal{A}(z, \chi, \rho, v, h) = (\mathcal{A}_1(z, \chi, \rho, v), \mathcal{A}_2(z, \rho, h), z(0), \rho(0)),$$

$$\mathcal{A}_1(z, \chi, \rho, v) = Lz + (z \cdot \nabla)z + \nabla\chi - v1_{\mathcal{O}} - \rho e_N$$

and

$$\mathcal{A}_2(z, \rho, h) = P\rho + (z \cdot \nabla)\rho + z \cdot \nabla\overline{\theta} - h1_{\mathcal{O}}$$

for every $(z, \chi, \rho, v, h) \in \widetilde{E}_N$.

Using the fact that $\ell^{-2}\beta_5^{1/2} z \in L^4(0,T;L^{12}(\Omega)^3)$, it can be checked that $\mathcal{A}_1$ is C^1. Then, since $\ell^{-2}\beta_5^{1/2}\rho \in L^2(0,T;H^1(\Omega)) \cap L^\infty(0,T;L^2(\Omega))$ and this space is continuously embedded in $L^4(0,T;L^3(\Omega))$, we deduce that

$$\ell^{-4}\beta_5(z,\nabla)\rho = \nabla\cdot(z\,\rho) \in L^2(0,T;W^{-1,12/5}(\Omega)) \subset L^2(0,T;H^{-1}(\Omega))$$

and, consequently, $\mathcal{A}$ is well-defined and satisfies $\mathcal{A} \in C^1(B_1;B_2)$.

The fact that $\mathcal{A}'(0,0,0,0,0) : B_1 \longmapsto B_2$ is surjective is an immediate consequence of the result given in Proposition 2.

As a conclusion, we can apply Theorem 4 and the null controllability for system (73) holds.

4.3. Proof of Theorem 3. Let us recall the nonlinear system we are dealing with:

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)\mathbf{T}_M(y) + \nabla p = v1_{\mathcal{O}} & \text{in } Q,\\ \nabla\cdot y = 0 & \text{in } Q,\\ y\cdot n = 0, \quad \nabla\times y = 0 & \text{on } \Sigma,\\ y(0) = y^0 & \text{in } \Omega. \end{cases}$$

In this case, we are going to apply Kakutani's fixed point theorem (see, for instance, [2]).

THEOREM 5. *Let Z be a Banach space and let $\Lambda : Z \mapsto Z$ be a set-valued mapping satisfying the following assumptions:*

- *$\Lambda(z)$ is a nonempty closed convex set of Z for every $z \in Z$.*
- *There exists a convex compact set $K \subset Z$ such that $\Lambda(K) \subset K$.*
- *Λ is upper-hemicontinuous in Z, i.e., for each $\sigma \in Z'$ the single-valued mapping*

$$z \mapsto \sup_{y\in\Lambda(z)} \langle \sigma, y\rangle_{Z',Z} \tag{74}$$

is upper-semicontinuous.

Then Λ possesses a fixed point in the set K, i.e., there exists $z \in K$ such that $z \in \Lambda(z)$.

In order to apply this result, we set $Z = L^2(Q)^2$ and, for each $z \in Z$, we consider the following system:

$$\begin{cases} y_t - \Delta y + (y\cdot\nabla)\mathbf{T}_M(z) + \nabla p = v1_{\mathcal{O}} & \text{in } Q,\\ \nabla\cdot y = 0 & \text{in } Q,\\ y\cdot n = 0, \quad \nabla\times y = 0 & \text{on } \Sigma,\\ y(0) = y^0 & \text{in } \Omega. \end{cases} \tag{75}$$

Then, for each $z \in Z$, we denote by $A(z)$ the set of controls $v1_{\mathcal{O}}$ with $v \in W$ that drive system (75) to zero and satisfy (61). Finally, our set-valued mapping is given as follows: for each $z \in Z$, $\Lambda(z)$ is the set of functions y that solve, together with some p, the linear system (75) corresponding to a control $v \in A(z)$.

Let us check that the assumptions of Theorem 5 are satisfied in this setting. The first one holds easily, so we omit the proof. Next, the estimates (62) and (63) tell us that the whole space Z is actually mapped into a compact set.

Let us finally see that Λ is upper-hemicontinuous in Z. Assume that $\sigma \in Z'$ and let $\{z_n\}$ be a sequence in Z such that $z_n \to z$ in Z. We have to prove that

$$\limsup_{n\to\infty} \sup_{y\in\Lambda(z_n)} \langle \sigma, y\rangle_{Z',Z} \leq \sup_{y\in\Lambda(z)} \langle \sigma, y\rangle_{Z',Z}. \tag{76}$$

Let us choose a subsequence $\{z_{n'}\}$ such that

$$\limsup_{n\to\infty} \sup_{y\in\Lambda(z_n)} \langle \sigma, y\rangle_{Z',Z} = \lim_{n'\to\infty} \sup_{y\in\Lambda(z_{n'})} \langle \sigma, y\rangle_{Z',Z}. \tag{77}$$

From the fact that $\Lambda(z_{n'})$ is a compact set of Z, for each n' we have

$$\sup_{y\in\Lambda(z_{n'})} \langle \sigma, y\rangle_{Z',Z} = \langle \sigma, y_{n'}\rangle_{Z',Z}$$

for some $y_{n'} \in \Lambda(z_{n'})$. Obviously, it can be assumed that

$$z_{n'}(x,t) \to z(x,t) \text{ a.e. } (x,t) \in Q \tag{78}$$

and

$$v_{n'} \rightharpoonup v \text{ weakly in } L^2(Q)^2 \tag{79}$$

with $v \in A(z)$. Furthermore, since all the $y_{n'}$ belong to a fixed compact set, we can also assume that

$$y_{n'} \to y \text{ in } Z$$

(after extraction of a subsequence). This, together with (77)–(79) implies that $y \in \Lambda(z)$, since we have a Stokes system with a right-hand side weakly converging in L^2 and a coefficient converging almost everywhere. As a conclusion, (76) holds and the proof of Theorem 3 is achieved.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Translated from the Russian by V. M. Volosov, Contemporary Soviet Mathematics, Consultants Bureau, New York, 1987.

[2] J.-P. Aubin, *L'Analyse non Linéaire et ses Motivations Économiques*, Masson, Paris, 1984.

[3] J.-M. Coron, *On the controllability of the 2-D incompressible Navier-Stokes equations with the Navier slip boundary conditions*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 35–75.

[4] C. Fabre, *Uniqueness results for Stokes equations and their consequences in linear and nonlinear control problems*, ESAIM Control Optim. Calc. Var., 1 (1995/96), pp. 267–302.

[5] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[6] E. Fernández-Cara and S. Guerrero, *Global Carleman inequalities for parabolic systems and applications to controllability*, SIAM J. Control Optim., to appear.

[7] E. Fernández-Cara, S. Guerrero, O. Yu. Imanuvilov and J.-P. Puel, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl., 83/12 (2004), pp. 1501–1542.

[8] A. Fursikov and O. Yu. Imanuvilov, *Local exact controllability for 2-D Navier-Stokes equations*, Sb. Math. 187, 9 (1996), pp. 1355–1390.

[9] A. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes 34, Seoul National University, Korea, 1996.

[10] A. Fursikov and O. Yu. Imanuvilov, *Exact controllability of the Navier-Stokes and Boussinesq equations*, Russian Math. Surveys, 54 (1999), pp. 565–618.

[11] G. Geymonat and E. Sánchez-Palencia, *On the vanishing viscosity limit for acoustic phenomena in a bounded region*, Arch. Ration. Mech. Anal., 75 (1981), pp. 257–268.

[12] S. Guerrero, *Local exact controllability to the trajectories of the Boussinesq system*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.

[13] O. Yu. Imanuvilov, *Remarks on exact controllability for the Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 39–72.

[14] O. Yu. Imanuvilov and J.-P. Puel, *Global Carleman estimates for weak elliptic non homogeneous Dirichlet problem,* Int. Math. Res. Not., 16 (2003), pp. 883–913.

[15] O. Yu. Imanuvilov and M. Yamamoto, *Carleman estimate for a parabolic equation in a Sobolev space of negative order and its applications*, in Control of Nonlinear Distributed Parameter Systems, G. Chen, I. Lasiecka, and J. Zhou, eds., Lecture Notes in Pure and Appl. Math. 218, Dekker, New York 2001, pp. 113–137.

[16] J.-L. Lions and E. Zuazua, *A generic uniqueness result for the Stokes system and its control theoretical consequences*, Partial Differential Equations and Appl., 177 (1996), pp. 221–235.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 174–206

STABILITY ANALYSIS OF DETERMINISTIC AND STOCHASTIC SWITCHED SYSTEMS VIA A COMPARISON PRINCIPLE AND MULTIPLE LYAPUNOV FUNCTIONS*

DEBASISH CHATTERJEE† AND DANIEL LIBERZON†

Abstract. This paper presents a general framework for analyzing stability of nonlinear switched systems, by combining the method of multiple Lyapunov functions with a suitably adapted comparison principle in the context of stability in terms of two measures. For deterministic switched systems, this leads to a unification of representative existing results and an improvement upon the current scope of the method of multiple Lyapunov functions. For switched systems perturbed by white noise, we develop new results which may be viewed as natural stochastic counterparts of the deterministic ones. In particular, we study stability of deterministic and stochastic switched systems under average dwell-time switching.

Key words. switched systems, stability analysis, stochastic stability, multiple Lyapunov functions, comparison principle, two measures

AMS subject classifications. 93B12, 93D20, 93E15

DOI. 10.1137/040619429

1. Introduction.

A family of continuous-time systems, together with a *switching signal* that chooses an *active subsystem* from the family at every instant of time, constitute a *switched system* [30]. Compared to *hybrid systems* [44], which currently are the focus of a large and growing interdisciplinary area of research, switched systems enable a more abstract modeling of continuous time systems with isolated switching events, which is suitable from a control-theoretic viewpoint. The abstraction is the result of modeling the switching signal as a purely time-dependent function, regardless of the mechanism of its generation. However, results obtained in this framework are then applicable to more specific hybrid systems; see, e.g., [18, 30] for a discussion. This paper is concerned with stability analysis of switched systems whose continuous dynamics are described by ordinary or stochastic differential equations.

Stability analysis by Lyapunov's direct method, in the simplest case of a single system, involves seeking a positive definite function of the states—called a Lyapunov function—that decreases along solution trajectories; see, e.g., [13, 22] for details. In case of switched systems, there are essentially two approaches to analyzing stability using Lyapunov's direct method; one involves investigating the existence of a *common Lyapunov function*, and the other utilizes *multiple Lyapunov functions*; see, e.g., [30, Chapters 2, 3] for an extensive account. The former approach is usually more challenging, although once a common Lyapunov function is found, the subsequent analysis is simple. The latter approach is usually more amenable to applications; typically, to check stability, one needs to allocate one Lyapunov function to each subsystem, trace through the sequence of values of these functions at switching instants, and verify certain monotonicity requirements on this sequence. We elaborate further on this method below.

*Received by the editors November 22, 2004; accepted for publication (in revised form) August 4, 2005; published electronically March 3, 2006. This work was supported by NSF ECS-0134115 CAR and DARPA/AFOSR MURI F49620-02-1-0325 grants.
http://www.siam.org/journals/sicon/45-1/61942.html
†Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801 (dchatter@control.csl.uiuc.edu, liberzon@control.csl.uiuc.edu).

The *comparison principle* [22, 28] helps in stability analysis by acting upon Lyapunov's direct method. It characterizes the time evolution of a Lyapunov function along system trajectories in terms of the solution of a scalar differential equation—the *comparison system*. The stability characteristics of the original higher dimensional system can then be inferred from those of the comparison system.

Our purpose is to establish a framework for stability analysis of (a) deterministic and (b) stochastic switched systems, by combining the method of multiple Lyapunov functions with the comparison principle. Specific motivations and contributions are elaborated below.

Deterministic switched systems. Stability analysis of deterministic switched systems using multiple Lyapunov functions first appeared in [37] and has evolved over a series of articles, for instance, [6, 19, 38]. The basic idea behind this method is to utilize stability properties of individual subsystems to infer stability properties of a switched system, thereby characterizing switching signals that ensure stability. A typical result, e.g., [30, Theorem 3.1], involves verifying two conditions to check global asymptotic stability of a switched system: first, each Lyapunov function is to monotonically decay when the corresponding subsystem is active, and second, the sequence formed by the values of each Lyapunov function at the instants when the corresponding subsystem becomes active is to be monotonically decreasing. (Henceforth we shall refer to this as the *fixed-index monotonicity condition*.) The verification of the second condition apparently requires quantitative knowledge of system trajectories. However, in situations where the switching is triggered by the state crossing some switching surfaces, on which the values of relevant Lyapunov functions match, the second condition follows if the first holds. Also, slow switching with a suitable *dwell-time* (see, e.g., [30]) allows each Lyapunov function to decay sufficiently before a switching occurs, thereby satisfying the second condition. A generalization of dwell-time switching is provided by the scheme of *average dwell-time switching* [16], which has proved to be a fruitful analysis tool in supervisory control; see [30, Chapter 6] and the references therein. This scheme requires that the number of switches over an arbitrary time interval should increase at most linearly with the length of the interval but places no specific restrictions on monotonicity of Lyapunov function values at switching instants (thereby allowing violation of the second condition above). We know that under suitable hypotheses, average dwell-time switching guarantees global asymptotic stability of a switched system [16], but the original proof of this result (provided in [30, Chapter 3]) does not utilize Lyapunov functions alone. We propose an alternative approach to stability analysis, based on the observation that a time trace of Lyapunov functions corresponding to active subsystems, in a typical trajectory of a switched system, shows impulsive behavior. Trajectories of suitable scalar impulsive differential equations may be used to generate such traces and are thus natural choices for comparison systems (cf. [29, 42], where impulsive comparison systems were utilized in stability analysis of impulsive differential equations). By employing various types of impulsive differential equations as comparison systems, in our results we relax the fixed-index sequence monotonicity condition, allowing oscillations and overshoots in the sequence, and also letting unstable subsystems participate in the dynamics. Additionally, our proofs retain conventional characteristics of Lyapunov's direct method in the context of switched systems, without resorting to independent arguments as in the proof of the average dwell-time result [16]. We propose our results on deterministic switched systems in section 2.

Stochastic switched systems. There is an enormous body of literature considering effects of noise and disturbances in systems, particularly from control and communication viewpoint; see, e.g., [7, 24, 34, 45]. The use of Lyapunov functions in stability analysis of stochastic systems is a classical idea, discussed extensively in [1, 11, 14, 25, 46]. The literature on stability of stochastic *switched* systems—constituted by subsystems perturbed by a standard Wiener process—is much less extensive compared to that on deterministic switched systems. In recent times, modeling and analysis of stochastic hybrid systems have appeared in, e.g., [17, 20]. Ergodic control of switched diffusion processes appears in [12]; stabilization methods that involve arguments similar to multiple Lyapunov functions appear in [2]. Some straightforward results involving common Lyapunov functions for stochastic switched systems may be found in [8]. We propose a framework for stochastic stability analysis by utilizing statistical estimates of Lyapunov functions at switching instants and during active periods of each subsystem. We consider some general stability definitions, for instance, global asymptotic stability in the mean and global asymptotic stability in probability, and employ the method of multiple Lyapunov functions adapted to the stochastic context. Much like the deterministic case, in a typical trajectory of a stochastic switched system, a time trace of expected values of Lyapunov functions corresponding to active subsystems shows impulsive behavior. We utilize impulsive differential equations as comparison systems to build a general framework for stability analysis of stochastic switched systems. This allows for very general behavior of the expected values of Lyapunov functions corresponding to active subsystems between switching intervals. In particular, replacing the fixed-index sequence monotonicity condition in the deterministic case with a stochastic analogue involving statistical estimates of Lyapunov functions at switching instants, we obtain a natural stochastic counterpart of [30, Theorem 3.1]. In addition, our results provide sufficient conditions for global asymptotic stability in the mean under average dwell-time switching and some more specific hypotheses. We propose our results on stochastic switched systems in section 3.

The concept of *stability analysis in terms of two measures* generalizes analysis of the norm of the state vector to analysis of the behavior of more general functions of the states; see, e.g., [29, 35]. We incorporate stability analysis in terms of two measures in the framework that we build for switched systems and gain greater flexibility for our results.

We study representative notions of deterministic and stochastic stability for a reasonably large class of switched systems. Naturally, not every type of stability can be described here. However, the framework of stability analysis we propose is applicable, as it stands, to more general stability notions, as will be indicated subsequently at appropriate places.

Some Notations. For notational convenience and brevity, we adopt the following conventions. For M_1, M_2, M_3 subsets of Euclidean space, let $\mathrm{C}[M_1, M_2]$ denote the set of all continuous functions $f : M_1 \longrightarrow M_2$, and $\mathrm{C}^1[M_1, M_2]$ denote the set of all continuously differentiable functions $f : M_1 \longrightarrow M_2$. We also use $\mathrm{C}^{1,2}[M_1 \times M_2, M_3]$ to denote the set of functions $f : M_1 \times M_2 \longrightarrow M_3$ that are continuously differentiable once and twice in the first and second arguments, respectively. We denote by $|\cdot|$ the standard Euclidean norm and by $\mathbb{R}_{\geqslant 0}$ the interval $[0, \infty[$. As usual, $\circ$ between two functions denotes their composition.

We say that a function $\alpha \in \mathrm{C}[\mathbb{R}_{\geqslant 0}, \mathbb{R}_{\geqslant 0}]$ is of class $\mathcal{K}$ if α is strictly increasing with $\alpha(0) = 0$, is of class $\mathcal{K}_\infty$ if in addition $\alpha(r) \to \infty$ as $r \to \infty$; and we write $\alpha \in \mathcal{K}$

and $\alpha \in \mathcal{K}_\infty$, respectively. A function $\beta \in \mathrm{C}\big[\mathbb{R}^2_{\geqslant 0}, \mathbb{R}_{\geqslant 0}\big]$ is said to be of class $\mathcal{KL}$ if $\beta(\cdot, t)$ is a function of class $\mathcal{K}$ for every fixed t and $\beta(r, t) \to 0$ as $t \to \infty$ for every fixed r; and we write $\beta \in \mathcal{KL}$.

2. Deterministic switched systems. In this section, we study stability of deterministic switched nonlinear nonautonomous systems. In section 2.1, we describe a switched system and the stability notions that we study and define the comparison systems that we employ in our analysis. We propose our comparison theorem for deterministic switched systems in section 2.2 and illustrate it in sections 2.3 and 2.4. We discuss two other stability notions in section 2.5 that are different from Lyapunov stability and demonstrate that their analysis can be similarly carried out in the proposed framework.

2.1. Preliminaries.

System description and stability definitions. We consider a family of nonlinear nonautonomous systems,

$$\dot{x} = f_p(t, x), \qquad p \in \mathcal{P}, \tag{2.1}$$

where $x \in \mathbb{R}^n$, $\mathcal{P}$ is an index set, $f_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}^n]$ is a vector field measurable in the first argument and locally Lipschitz in the second (see, e.g., [41] for further details), $f_p(\cdot, 0) \equiv 0$, for every $p \in \mathcal{P}$. Let there exist a piecewise constant function (continuous from the right by convention) $\sigma : \mathbb{R}_{\geqslant 0} \longrightarrow \mathcal{P}$, which specifies at every time t the index $\sigma(t) = p \in \mathcal{P}$ of the active subsystem. A *switched system* generated by the family (2.1) and such a *switching signal* σ is

$$\dot{x} = f_\sigma(t, x), \qquad x(t_0) = x_0, \quad t \geqslant t_0, \tag{2.2}$$

where $t_0 \in \mathbb{R}_{\geqslant 0}$. It is assumed that there is no jump in the state x at the switching instants and that there is a finite number of switches on every bounded interval of time. We denote the switching instants by τ_i, $i = 1, 2, \ldots$, $\tau_0 := t_0$, and the sequence $\{\tau_i\}_{i \geqslant 0}$ is strictly increasing. The *solution of* (2.2) *as a function of time* t*, interpreted in the sense of Carathéodory, initialized at a given pair* (t_0, x_0)*, and under a given switching signal* σ, is denoted by $x(t)$.

To perform analysis in terms of two measures, we utilize functions belonging to the class defined by

$$\Gamma := \left\{ h \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}] \;\middle|\; \inf_{(t,x)} h(t, x) = 0 \right\}. \tag{2.3}$$

We focus on the following notion of stability in terms of two measures; see, e.g., [29] for further details on other related concepts of stability. This stability definition coincides in spirit with the class $\mathcal{KL}$ stability defined for differential inclusions in the paper [43], where the authors consider autonomous systems on an extended state space.

Definition 2.1. *Let* $h^\circ, h \in \Gamma$. *The switched system* (2.2) *is said to be* (h°, h)-globally uniformly asymptotically stable ((h°, h)-GUAS) *if there exists a class* $\mathcal{KL}$ *function* β *such that for every* $(t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$*, the inequality*

$$h(t, x(t)) \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0 \tag{2.4}$$

holds.

In this paper we consider stability notions that are uniform with respect to the initial time t_0, e.g., global uniform asymptotic stability. The occurrences of "uniform" in the sequel convey this particular sense. In contrast, in much of the existing literature "uniform" global asymptotic stability is used to signify uniformity over a class of switching signals; see, e.g., [30]. In situations where there is uniformity in this sense, it will be explicitly indicated at appropriate places.

Remark 2.2. Under the special case of $h^\circ(t,x) = h(t,x) = |x|$, we recover usual *global uniform asymptotic stability* (GUAS), and for autonomous systems the corresponding specializations lead to *global asymptotic stability* (GAS); see, e.g., [22] for further details on GUAS and GAS. Other examples include stability with respect to arbitrary "tubes" in $\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ with the measures h°, h being the Hausdorff point-to-set distances, stability of prescribed motion $x_{\mathrm{r}}(\cdot)$ measured by $h^\circ(t,x) = h(t,x) := |x(t) - x_{\mathrm{r}}(t)|$, and partial stability. More examples may be found in [29].

Remark 2.3. The (h°, h)-GUAS property can be rephrased in traditional $\varepsilon - \delta$ form as follows. The (h°, h)-GUAS property expressed by (2.4) holds if and only if the following properties hold simultaneously:

(S1) ((h°, h)-uniform Lyapunov stability) there exists a class $\mathcal{K}_\infty$ function δ such that for every $\varepsilon > 0$ and $t_0 \in \mathbb{R}_{\geqslant 0}$, we have $h^\circ(t_0, x_0) < \delta(\varepsilon) \implies h(t, x(t)) < \varepsilon \quad \forall\, t \geqslant t_0$;

(S2) ((h°, h)-uniform global asymptotic convergence) for every $r, \varepsilon > 0$, there exists a number $T(r,\varepsilon) \geqslant 0$ such that for every $t_0 \in \mathbb{R}_{\geqslant 0}$, we have $h^\circ(t_0, x_0) < r \implies h(t, x(t)) < \varepsilon \quad \forall\, t \geqslant t_0 + T(r,\varepsilon)$.

A similar equivalence is established in [43, Proposition 1]; the proof of the above can be easily constructed from the proof of this proposition, and for completeness is provided in section A.

We introduce properties that will later be required from Lyapunov functions, cf. [29]. Let $V \in \mathrm{C}\big[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}\big]$. The function V is said to be *h-positive definite* if for $(t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ there exists a function $\alpha_1 \in \mathcal{K}_\infty$ such that $\alpha_1 \circ h(t,x) \leqslant V(t,x)$, and *$h^\circ$-decrescent* if for $(t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ there exists a function $\alpha_2 \in \mathcal{K}_\infty$ such that $V(t,x) \leqslant \alpha_2 \circ h^\circ(t,x)$. In results that follow, we shall require a family of functions $\{V_p \mid p \in \mathcal{P}\}$ to be *$\mathcal{P}$-uniformly h-positive definite and h° decrescent*; i.e., there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that we have

$$\alpha_1 \circ h(t,x) \leqslant V_p(t,x) \leqslant \alpha_2 \circ h^\circ(t,x) \qquad \forall\, (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \quad \forall\, p \in \mathcal{P}. \tag{2.5}$$

With h° and h specialized to Euclidean norms, usual positive definiteness and decrescence of V_p are recovered.

Remark 2.4. We point out that if $\mathcal{P}$ is finite, or if $\mathcal{P}$ is compact and suitable continuity assumptions hold true, then (2.5) is no loss of generality.

In this section, we present our results in the absence of classical differentiability assumptions, and the directional upper right Dini derivative is utilized. For example, along the vector field of a member with index p of the family of systems (2.1) the derivative of a function $V_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ is defined as

$$\mathrm{D}^+_{f_p} V_p(t,x) := \limsup_{\epsilon \downarrow 0} \frac{1}{\epsilon} \big(V_p(t+\epsilon, x + \epsilon f_p(t,x)) - V_p(t,x) \big);$$

for further details, see, e.g., [10]. We require that the functions V_p, $p \in \mathcal{P}$ are locally Lipschitz in the second argument.

Remark 2.5. For continuously differentiable functions, the upper right Dini derivative derivative reduces to the ordinary derivative. In particular, we note that

for $V_p \in \mathrm{C}^1[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$, the expression simplifies to

$$\mathrm{D}_{f_p} V_p(t,x) = \frac{\mathrm{d}}{\mathrm{d}t} V_p(t,x) = \left(\frac{\partial V_p}{\partial t} + \frac{\partial V_p}{\partial x} f_p \right)(t,x),$$

which is the total derivative of V_p along solutions of the system with index p in the family (2.1).

Comparison systems and deterministic comparison principle. Let the switched system (2.2) be given, and for a given switching signal σ, let $\{\tau_i\}_{i \geqslant 1}$ be the sequence of switching instants. Let $\mathbb{R}^n \times \mathcal{P} \ni (x,p) \longmapsto y(x,p) \in \mathbb{R}_{\geqslant 0}$ be a function continuous in x. We consider scalar nonlinear nonautonomous impulsive differential equations of the type

$$(2.6) \quad \begin{cases} \dot{\xi} = \phi(t,\xi), \quad t \neq \tau_i, \\ \xi(\tau_i) = \psi_i\left(\xi\left(\tau_i^-\right), y\left(x(\tau_i), \sigma(\tau_i)\right)\right), \end{cases} \qquad \xi(t_0) = \xi_0 \geqslant 0, \quad i \geqslant 1, \quad t \geqslant t_0,$$

where $\xi \in \mathbb{R}$, the field $\phi \in \mathrm{C}\big[\mathbb{R}^2_{\geqslant 0}, \mathbb{R}\big]$ and the *reset map* $\psi_i \in \mathrm{C}\big[\mathbb{R}^2_{\geqslant 0}, \mathbb{R}_{\geqslant 0}\big]$ are such that $\phi(\cdot,0) \equiv 0$ and $\psi_i(0,\cdot) \equiv 0$, $t_0 \in \mathbb{R}_{\geqslant 0}$, τ_i, $i = 1,2,\ldots$, are the instants of the impulses, $\tau_0 := t_0$, *the sequence* $\{\tau_i\}_{i \geqslant 1}$ *is identified with the sequence of switching instants generated by* σ. Systems of the type (2.6) are utilized as *comparison systems*, as well as various special cases—with or without impulses, time-variation, and y. In this paper we shall not be utilizing time-varying ϕ; an example of this may be found in [9].

We denote by $\xi(t)$ *the solution* (2.6) *as a function of time, with the sequence* $\{\tau_i\}_{i \geqslant 0}$ *specified, and initialized at* (t_0, ξ_0). In all cases, we shall tacitly assume the existence of a unique solution to differential equations of the type (2.6). If there are multiple solutions, however, the results hold true with the largest solution in place of the (unique) solution of the comparison systems.

The GUAS property of equations of the type (2.6) is defined similarly to Definition 2.1, with h°, h specialized to the scalar norm. Formally, a system of the type (2.6), with a given sequence $\{\tau_i\}_{i \geqslant 0}$, is said to be *globally uniformly asymptotically stable* (GUAS) if there exists a function $\beta_\xi \in \mathcal{KL}$ such that for every $(t_0, \xi_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}_{\geqslant 0}$ the inequality

$$(2.7) \qquad |\xi(t)| \leqslant \beta_\xi(|\xi_0|, t - t_0) \qquad \forall\, t \geqslant t_0$$

holds. The properties required from ϕ and ψ_i in (2.6) ensure that $\xi(\cdot) \geqslant 0$; we shall therefore omit the absolute values on ξ in the sequel.

Remark 2.6. Remark 2.3 still applies if a system of the type (2.6) replaces (2.2), with h°, h specialized to the scalar norm.

The need to compare different Lyapunov functions at switching instants, which is inherent in the multiple Lyapunov functions method, prompts us to have a function y of the system states in the reset equation of a comparison system of the type (2.6). This function will be utilized in the construction of a suitable comparison system that will render [30, Theorem 3.1] a special case of our Theorem 2.8. For yet other applications, y will not be required.

The following well-known comparison lemma for nonswitched deterministic systems is needed for subsequent developments in this section; see, e.g., [28] for a proof.

LEMMA 2.7. *Consider the system with index* p *in the family* (2.1). *Suppose that there exist a function* $V_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ *and a comparison system*

$$\Sigma: \quad \dot{\xi} = \phi(t,\xi), \qquad \xi(t_0) = \xi_0 \geqslant 0, \quad t \geqslant t_0,$$

such that $\forall (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, *the differential inequality*

$$\mathrm{D}^+_{f_p} V_p(t,x) \leqslant \phi(t, V_p(t,x)) \tag{2.8}$$

holds. Then $V_p(\tau, x(\tau)) \leqslant \xi(\tau)$ *implies* $V_p(t,x(t)) \leqslant \xi(t) \quad \forall\, t \geqslant \tau$, *where* $x(t)$ *and* $\xi(t)$ *are the solutions of the system with index* p *in* (2.1) *and* Σ, *respectively.*

2.2. Comparison theorem for deterministic switched systems. The following result establishes a general framework for testing stability of deterministic switched systems using multiple Lyapunov functions and a comparison system.

THEOREM 2.8. *Consider the switched system* (2.2) *with a fixed switching signal* σ *generating a sequence of switching instants* $\{\tau_i\}_{i \geqslant 1}$, *and two functions* $h^\circ, h \in \Gamma$. *Suppose that there exist functions* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$, $p \in \mathcal{P}$, *locally Lipschitz in the second argument, and a system* Σ *of the type* (2.6), *such that*

(i) *the family* $\{V_p \mid p \in \mathcal{P}\}$ *is* $\mathcal{P}$*-uniformly* h*-positive definite and* h°*-decrescent in the sense of* (2.5);

(ii) $\forall\, (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ *and* $\forall\, p \in \mathcal{P}$, *we have* $\mathrm{D}^+_{f_p} V_p(t,x) \leqslant \phi(t, V_p(t,x))$;

(iii) $\forall\, (t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, *there exists* $\xi_0 \in \mathbb{R}_{\geqslant 0}$ *such that* $V_{\sigma(\tau_i)}(\tau_i, x(\tau_i)) \leqslant \xi(\tau_i)$ $\forall\, i \geqslant 0$, *where* $x(t)$ *and* $\xi(t)$ *are the corresponding solutions of* (2.2) *and* Σ, *respectively;*

(iv) Σ *is* GUAS *in the sense of* (2.7).

Then (2.2) *is* (h°, h)-GUAS *in the sense of* Definition 2.1.

Proof. Consider the interval $[\tau_\ell, \tau_{\ell+1}[$, with ℓ an arbitrary nonnegative integer. From hypotheses (iii) and (ii), and Lemma 2.7 applied with $\tau = \tau_\ell$, we have

$$V_{\sigma(\tau_\ell)}(t, x(t)) \leqslant \xi(t) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[.$$

The above estimate in conjunction with hypothesis (iv) leads to

$$V_{\sigma(\tau_\ell)}(t, x(t)) \leqslant \xi(t) \leqslant \beta_\xi(\xi_0, t - t_0) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[.$$

In view of hypothesis (i), we have

$$\alpha_1 \circ h(t, x(t)) \leqslant \beta_\xi(\alpha_2 \circ h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[. \tag{2.9}$$

With $\beta(r,s) := \alpha_1^{-1} \circ \beta_\xi(\alpha_2(r), s)$, the estimate in (2.9) is equivalent to

$$h(t, x(t)) \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[.$$

Clearly $\beta \in \mathcal{KL}$. The arbitrariness of ℓ implies that

$$h(t, x(t)) \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0.$$

The (h°, h)-GUAS property of (2.2) follows. ∎

Theorem 2.8 does not provide a direct method for analyzing stability of a given switched system; we need to look for a suitable comparison system and check its stability properties first. We will now demonstrate how to proceed with such a scheme of analysis and how our framework stands in relation to existing results involving multiple Lyapunov functions. In section 2.3, we generalize [30, Theorem 3.1] in terms of two measures, which involves verification of a fixed-index sequence monotonicity condition to establish the (h°, h)-GUAS property of a switched system. The comparison system utilized in the proof makes use of quantitative information of system trajectories. In section 2.4 we construct comparison systems to rederive the sufficient conditions for GUAS of switched system under average dwell-time switching given in [30, Theorem 3.2] or [16]. Although the fixed-index sequence monotonicity condition is violated, the comparison framework of Theorem 2.8 works; also we do not require explicit information of system trajectories.

2.3. Stability under fixed-index sequence monotonicity condition. For this subsection, we let $\mathcal{P}$ be a finite set of N elements.

COROLLARY 2.9. *Consider the switched system* (2.2) *and* $h^\circ, h \in \Gamma$. *Suppose that there exist functions* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $\rho, U \in \mathcal{K}$, α *positive definite,* $V_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ *for each* $p \in \mathcal{P}$ *locally Lipschitz in the second argument, such that*

(i) *the family* $\{V_p \mid p \in \mathcal{P}\}$ *is* $\mathcal{P}$*-uniformly* h*-positive definite and* h°*-decrescent in the sense of* (2.5)*;*

(ii) $\forall\, (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ *and* $\forall\, p \in \mathcal{P}$, *we have* $\mathrm{D}^+_{f_p} V_p(t,x) \leqslant -\alpha \circ h^\circ(t,x)$*;*

(iii) *for every pair of switching time* (τ_i, τ_j), $i < j$ *such that* $\sigma(\tau_i) = \sigma(\tau_j) = p \in \mathcal{P}$ *and* $\sigma(\tau_k) \neq p$ *for* $\tau_i < \tau_k < \tau_j$, *the inequality*

$$V_p(\tau_j, x(\tau_j)) - V_p(\tau_i, x(\tau_i)) \leqslant -U \circ h^\circ(\tau_i, x(\tau_i)) \tag{2.10}$$

holds, where $x(t)$ *is the solution of* (2.2) *initialized at* (t_0, x_0)*;*

(iv) $\forall\, (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, *we have* $\alpha_2 \circ h^\circ(t,x) \leqslant \rho \circ \alpha_1 \circ h(t,x)$.

Then (2.2) *is* (h°, h)-GUAS.

Proof. We define a candidate impulsive differential comparison system of the type (2.6):

$$\Sigma: \quad \begin{cases} \dot{\xi} = -\alpha \circ \alpha_2^{-1}(\xi), & t \neq \tau_i, \\ \xi(\tau_i) = V_{\sigma(\tau_i)}(\tau_i, x(\tau_i)), & \end{cases} \qquad i \geqslant 0, \quad t \geqslant t_0. \tag{2.11}$$

By its very definition, Σ satisfies hypotheses (ii)–(iii) of Theorem 2.8. Hypothesis (i) of Theorem 2.8 is satisfied by our hypothesis (i). To verify hypothesis (iv) of Theorem 2.8, we shall first prove uniform Lyapunov stability of Σ and then prove its global uniform asymptotic convergence, in view of Remark 2.6.

Consider the interval $[\tau_0, \tau_1[$. From hypothesis (ii) we have

$$V_{\sigma(\tau_0)}(\tau_1, x(\tau_1)) \leqslant V_{\sigma(\tau_0)}(\tau_0, x(\tau_0)).$$

Combining with hypothesis (i), we reach

$$\alpha_1 \circ h(\tau_1, x(\tau_1)) \leqslant \alpha_2 \circ h^\circ(t_0, x_0). \tag{2.12}$$

For every $p \in \mathcal{P}$, we have from hypothesis (i)

$$V_p(\tau_1, x(\tau_1)) \leqslant \alpha_2 \circ h^\circ(\tau_1, x(\tau_1)),$$

and therefore by hypothesis (iv),

$$V_p(\tau_1, x(\tau_1)) \leqslant \rho \circ \alpha_1 \circ h(\tau_1, x(\tau_1)).$$

In view of (2.12), we get

$$V_p(\tau_1, x(\tau_1)) \leqslant \rho \circ \alpha_2 \circ h^\circ(t_0, x_0). \tag{2.13}$$

Consider now the interval $[\tau_1, \tau_2[$. From hypothesis (ii) we have

$$V_{\sigma(\tau_1)}(\tau_2, x(\tau_2)) \leqslant V_{\sigma(\tau_1)}(\tau_1, x(\tau_1)).$$

Combining with hypothesis (i) and applying (2.13) with $p = \sigma(\tau_2)$, we get

$$\alpha_1 \circ h(\tau_2, x(\tau_2)) \leqslant \rho \circ \alpha_2 \circ h^\circ(t_0, x_0). \tag{2.14}$$

Now, for all $p \in \mathcal{P}$, we have

$$V_p(\tau_2, x(\tau_2)) \leqslant \alpha_2 \circ h^\circ(\tau_2, x(\tau_2)),$$

so by hypothesis (iv),

$$V_p(\tau_2, x(\tau_2)) \leqslant \rho \circ \alpha_1 \circ h(\tau_2, x(\tau_2)).$$

Now (2.14) gives

$$V_p(\tau_2, x(\tau_2)) \leqslant \rho \circ \rho \circ \alpha_2 \circ h^\circ(t_0, x_0).$$

It is not difficult to see that the worst-case situation for maximum possible overshoot of the function V_σ occurs when the switching signal σ visits every element of the set $\mathcal{P}$ without repetition until $\mathcal{P}$ is exhausted. Let $\tau_{j^\star}$ be the first switching instant after all the subsystems that participate in the dynamics have become active at least once since initialization at $t = t_0$. Define the function

$$\rho^j := \underbrace{\rho \circ \ldots \circ \rho}_{j \text{ times}}.$$

From the above computations, it is easy to see that

$$\xi(\tau_{j^\star}) = V_{\sigma(\tau_{j^\star})}(\tau_{j^\star}, x(\tau_{j^\star})) \leqslant \rho^{\mathrm{N}-1} \circ \alpha_2 \circ h^\circ(t_0, x_0).$$

Clearly $\rho^{\mathrm{N}-1} \in \mathcal{K}$. Define the function

$$\gamma(\cdot) := \max\left\{\alpha_2(\cdot), \rho \circ \alpha_2(\cdot), \ldots, \rho^{\mathrm{N}-1} \circ \alpha_2(\cdot)\right\}.$$

From (2.10) it follows that $\xi(t) \leqslant \gamma \circ h^\circ(t_0, x_0)\, \forall\, t \geqslant t_0$. Therefore, by hypothesis (i) and the definition of ξ_0 in (2.11),

$$\xi(t) \leqslant \gamma \circ \alpha_1^{-1}(\xi_0) \qquad \forall\, t \geqslant t_0. \tag{2.15}$$

It remains to prove uniform global asymptotic convergence of Σ.

We distinguish two cases.

Case 1. *Switching stops in finite time.* Since σ eventually attains a constant value, say, from the κth switching instant, it follows that there are no impulses after $t = \tau_\kappa$. That is to say, the system (2.11) becomes an autonomous scalar ordinary differential equation after $t = \tau_\kappa$, with negative right-hand side for nonzero $\xi(\tau_\kappa)$. Therefore, $\xi(t)$ monotonically decreases to $0\, \forall\, t \geqslant \tau_\kappa$. In conjunction with (2.15) which shows uniform Lyapunov stability of Σ, we conclude that (2.11) is GUAS. Theorem 2.8 now guarantees that (2.2) is (h°, h)-GUAS.

Case 2. *Switching continues indefinitely.* Consider the restatement of the inequality (2.10) with $\xi(\tau_i)$ as defined in (2.11)

$$\begin{aligned}\xi(\tau_j) - \xi(\tau_i) &\leqslant -U \circ h^\circ(\tau_i, x(\tau_i)) \\ &\leqslant -U \circ \alpha_2^{-1} \circ V_p(\tau_i, x(\tau_i)) \\ &= -U \circ \alpha_2^{-1} \circ \xi(\tau_i).\end{aligned}$$

The pair (τ_i, τ_j) satisfies the condition in hypothesis (iii). Clearly, $\{\xi(\tau_i)\}_{\{i \geqslant 0 | \sigma(\tau_i) = p\}}$ is a positive (for nonzero x_0) monotonically decreasing sequence and must attain a limit, say, $c \geqslant 0$. If $c \neq 0$, then

$$\xi(\tau_j) - \xi(\tau_i) \leqslant -U \circ \alpha_2^{-1}(c)$$

for all (τ_i, τ_j) satisfying hypothesis (iii), which means that for some large enough j, say, j', $\xi(\tau_{j'}) < 0$. In view of (2.11) this means $V_{\sigma(\tau_{j'})}(\tau_{j'}, x(\tau_{j'})) < 0$, contradicting the hypothesis. Therefore, the subsequence $\{\xi(\tau_i)\}_{\{i \geqslant 0 | \sigma(\tau_i) = p\}}$ attains the limit 0 as $i \uparrow \infty$. For all time t between any two switching instants (τ_i, τ_j) satisfying hypothesis (iii), there is a uniform bound on $\xi(t)$ given by (2.15). For each $p \in \mathcal{P}$, the subsequence $\{\xi(\tau_i)\}_{\{i \geqslant 0 | \sigma(\tau_i) = p\}}$ attains the limiting value of 0, implying global asymptotic convergence. Combining with Lyapunov stability proved in (2.15) above, we conclude that Σ is GUAS.

By Theorem 2.8 now we conclude that (2.2) is (h°, h)-GUAS. □

Remark 2.10. We note that in the above proof, Σ makes explicit use of state information of (2.2)—in the notation of (2.6), we use

$$\psi_i\left(\xi\left(\tau_i^-\right), y(x(\tau_i), \sigma(\tau_i))\right) = V_{\sigma(\tau_i)}(\tau_i, x(\tau_i))$$

in the reset equation of Σ. On the other hand, $\xi(\tau_i^-)$ is not utilized.

Remark 2.11. Hypothesis (iv) in Corollary 2.9 essentially is a technical requirement to ensure that the measure h is nontrivial. This guarantees that the different Lyapunov functions at switching instants can be estimated from the initial condition. For $h^\circ(t, x) = h(t, x) = |x|$, this property is automatic (just let $\rho := \alpha_2 \circ \alpha_1^{-1}$), and then Corollary 2.9 becomes identical to [30, Theorem 3.1].

2.4. Stability under average dwell-time switching. In this subsection we rederive an existing result on global asymptotic stability under average dwell-time switching via our Theorem 2.8. We no longer retain the assumption that $\mathcal{P}$ is finite. Although we specialize to Euclidean norms and autonomous switched systems—to be able to use the aforesaid result in situ—the analysis can be readily generalized to two measures and nonautonomous switched systems.

Let us consider the autonomous switched system

$$\dot{x} = f_\sigma(x), \qquad x(t_0) = x_0, \quad t \geqslant t_0, \tag{2.16}$$

where $x \in \mathbb{R}^n$, $f_p \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}^n]$ is locally Lipschitz for every $p \in \mathcal{P}$, $f_p(0) = 0$. The switching signal σ is said to have *average dwell-time* $\tau_a > 0$ [16] if there exists a positive number $N_\circ$ such that the number of switches $N_\sigma(T, t)$ on the interval $[t, T[$ satisfies

$$N_\sigma(T, t) \leqslant N_\circ + \frac{T - t}{\tau_a} \qquad \forall\, T \geqslant t \geqslant t_0. \tag{2.17}$$

We investigate the conditions on the average dwell-time of the switching signal σ such that (2.16) is GAS. The available result is as follows. (For a detailed discussion and proof, see, e.g., [30, Theorem 3.2].)

THEOREM 2.12 (see [16]). *Consider the switched system* (2.16). *Let there exist functions* $V_p \in \mathrm{C}^1[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ *for every* $p \in \mathcal{P}$, $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ *and a positive number* $\lambda_\circ$ *such that*

$$\alpha_1(|x|) \leqslant V_p(x) \leqslant \alpha_2(|x|) \qquad \forall\, p \in \mathcal{P} \tag{2.18}$$

and

$$\frac{\partial V_p}{\partial x}(x) f_p(x) \leqslant -\lambda_\circ V_p(x) \qquad \forall\, x \in \mathbb{R}^n. \tag{2.19}$$

Suppose also that there exists a positive constant μ such that

$$V_p(x) \leqslant \mu V_q(x) \qquad \forall\, x \in \mathbb{R}^n, \quad \forall\, p, q \in \mathcal{P}. \tag{2.20}$$

Then (2.16) *is* GAS *for every switching signal σ with average dwell-time $\tau_a > \frac{\ln \mu}{\lambda_\circ}$.*

Remark 2.13. The condition (2.20) imposes a restriction on permissible Lyapunov functions. Since this is a global result, it does not hold if, for example, some of the Lyapunov functions are quadratic and some others are quartic. Also, in view of interchangeability of p and q in (2.20), it follows that $\mu > 1$, excluding the trivial case of $\mu = 1$, which implies that there is a *common Lyapunov function* for the switched system [30].

For a fixed index $p \in \mathcal{P}$, the values of the Lyapunov function V_p at every switching instant τ_j with $\sigma(\tau_j) = p$ form a sequence $\{V_p(x(\tau_j))\}_{\{j \geqslant 0 | \sigma(\tau_j) = p\}}$. As discussed in section 1, results like [30, Theorem 3.1] provide sufficient conditions for stability of the switched system under the assumption that the sequences $\{V_p(x(\tau_j))\}_{\{j \geqslant 0 | \sigma(\tau_j) = p\}}$ are monotonically decreasing for every $p \in \mathcal{P}$. But σ with an average dwell-time permits overshoots and oscillations in each of these sequences; thus [30, Theorem 3.1] is inapplicable. We mentioned in section 1 that currently the problem is tackled by independent arguments utilizing auxiliary functions, as in [30]; the proof does not utilize the Lyapunov functions alone. In the framework of Theorem 2.8, we can dispense with such auxiliary functions and easily rederive Theorem 2.12, as we now demonstrate.

To the end of this subsection, we assume that the hypotheses of Theorem 2.12 hold.

An impulsive differential equation as a comparison system. Consider an impulsive differential system of the type (2.6) with

$$\psi_i\left(\xi\left(\tau_i^-\right), y(x(\tau_i), \sigma(\tau_i))\right) := \mu \xi(\tau_i^-), \quad i \geqslant 1, \quad \mu > 0,$$

as the reset equation and

$$\phi(t, \xi) := -\lambda_\circ \xi, \quad \lambda_\circ > 0.$$

The complete system stands as

$$\Sigma' : \quad \begin{cases} \dot{\xi} = -\lambda_\circ \xi, & t \neq \tau_i, \\ \xi(\tau_i) = \mu \xi(\tau_i^-), & \mu > 0, \end{cases} \qquad \xi(t_0) = V_{\sigma(t_0)}(x_0), \quad i \geqslant 1, \quad t \geqslant t_0. \tag{2.21}$$

From (2.18) it follows that hypothesis (i) of Theorem 2.8 is satisfied with $h^\circ(t, x) = h(t, x) = |x|$. Further, from (2.19) and (2.20) together with the initial condition in (2.21), it follows that hypotheses (ii) and (iii) of Theorem 2.8 are satisfied, respectively. Now we investigate stability of Σ'.

Let $T > 0$ be arbitrary. Consider the evolution of the system (2.21) from $t = t_0$ through $t = T$. Let there be $N_\sigma(T, t_0)$ switches on this interval, and let $\nu := N_\sigma(T, t)$, where $N_\sigma(T, t)$ is as defined in (2.17)

$$\xi(\tau_{i+1}^-) = \xi(\tau_i) \mathrm{e}^{-\lambda_\circ(\tau_{i+1} - \tau_i)}, \quad 0 \leqslant i \leqslant \nu,$$

and

$$\xi(T) = \xi(\tau_\nu) \mathrm{e}^{-2\lambda_\circ(T - \tau_\nu)}.$$

Combining with the reset equation of Σ' and iterating over i, it follows that

$$\xi(T) = \xi(t_0)\mu^{\nu}\mathrm{e}^{-\lambda_\circ(T-t_0)}. \tag{2.22}$$

Using the definition of ν, (2.22) leads to

$$\xi(T) = \xi(t_0)\mu^{N_\circ}\mathrm{e}^{\lambda_\circ t_0}\mathrm{e}^{-(\lambda_\circ - \ln\mu/\tau_a)T}. \tag{2.23}$$

To ensure $\xi(T) \to 0$ as $T \uparrow \infty$, it is sufficient to have $\tau_a > \frac{\ln\mu}{\lambda_\circ}$. This guarantees the convergence of the impulsive differential system (2.21) to zero as time increases to infinity. Stability of Σ' follows directly from (2.23)—the estimate $\xi(t) \leqslant \xi(t_0)\mu^{N_\circ}\mathrm{e}^{\lambda_\circ t_0}$ holds if $\tau_a > \frac{\ln\mu}{\tau_a}$. We conclude that Σ' is GAS, considering Remark 2.6. Therefore, hypothesis (iv) of Theorem 2.8 is also satisfied.

Intuitively,

- the minimum rate of decay of the Lyapunov function corresponding to each active subsystem is captured by the continuous dynamics of Σ'; and
- the maximum jump in the values of two Lyapunov functions corresponding to two consecutively active subsystems is captured by the reset equation of Σ'.

We conclude that by Theorem 2.8 the switched system (2.16) is GAS for switching signals with $\tau_a > \frac{\ln\mu}{\lambda_\circ}$.

An ordinary differential equation as a comparison system. Consider the following scalar autonomous differential equation, a special case of (2.6), for a candidate comparison system:

$$\Sigma'' : \quad \dot{\xi} = \left(\frac{\ln\mu}{\tau_a} - \lambda_\circ\right)\xi, \qquad \xi(t_0) = \mu^{N_\circ}\mathrm{e}^{\lambda_\circ t_0}V_{\sigma(t_0)}(x_0), \quad t \geqslant t_0. \tag{2.24}$$

The solution of Σ'' is

$$\xi(t) = \mu^{N_\circ}V_{\sigma(t_0)}(x_0)\mathrm{e}^{-(\lambda_\circ - \ln\mu/\tau_a)(t-t_0)} \qquad \forall\, t \geqslant t_0. \tag{2.25}$$

Let average dwell-time of σ be τ_a, and $\nu := N_\sigma(T, t_0)$ be the number of switches on $[t_0, T[$. Considering the least rate of decay for Lyapunov functions corresponding to active subsystems, we have for an arbitrary $T \geqslant 0$,

$$V_{\sigma(\tau_i)}(x(\tau_i^-)) \leqslant V_{\sigma(\tau_i)}(x(\tau_i))\mathrm{e}^{-\lambda_\circ(\tau_{i+1}-\tau_i)}, \quad 0 \leqslant i \leqslant \nu,$$

and

$$V_{\sigma(\tau_\nu)}(x(T)) \leqslant V_{\sigma(\tau_\nu)}(x(T))\mathrm{e}^{-\lambda_\circ(T-\tau_\nu)}.$$

Combining with (2.20) at switching instants and iterating over i, we reach the estimate

$$V_{\sigma(\tau_\nu)}(x(T)) \leqslant \mu^{\nu}V_{\sigma(t_0)}(x_0)\mathrm{e}^{-\lambda_\circ T} = \mu^{N_\circ}\mathrm{e}^{\lambda_\circ t_0}V_{\sigma(t_0)}(x_0)\mathrm{e}^{-(\lambda_\circ - \ln\mu/\tau_a)T}. \tag{2.26}$$

Clearly, $\lambda_\circ > \frac{\ln\mu}{\tau_a}$ ensures global asymptotic stability of Σ''. By Theorem 2.8 with $h^\circ(t,x) = h(t,x) = |x|$, it follows that (2.16) is GAS for switching signals with $\tau_a > \frac{\ln\mu}{\lambda_\circ}$.

From (2.18) it follows that hypothesis (i) of Theorem 2.8 is satisfied with $h^\circ(t,x) = h(t,x) = |x|$. Further, from the discussion above, it follows that hypotheses (ii)–(iv) of Theorem 2.8 are also satisfied

Intuitively,

- the initial condition of Σ'' captures the maximum possible overshoot in V_σ—this corresponds to the situation when all $N_\circ$ switches occur very close to $t = t_0$;
- $\xi(\cdot)$ forms an envelope of the sequence $\left\{V_{\sigma(\tau_i)}(x(\tau_i))\right\}_{i\geqslant 0}$ over the interval $[t_0, T]$.

We conclude that by Theorem 2.8 the switched system (2.16) is GAS for switching signals with $\tau_a > \frac{\ln \mu}{\lambda_\circ}$. This agrees with the assertion of Theorem 2.12.

Remark 2.14. We note that in contrast to (2.11), the comparison systems (2.21) and (2.24) do not utilize state information in the function y directly. However (2.21) utilizes $\xi(\tau_i^-)$ in the reset equation; cf. Remark 2.10.

Remark 2.15. It is easy to see from (2.22) and (2.25) that for switching signals with average dwell-time bounded away from $\frac{\ln \mu}{\lambda_\circ}$, we have GAS of (2.16). For instance, for $\lambda \in \,]0, \lambda_\circ[$ if $\tau_a \geqslant \frac{\ln \mu}{\lambda_\circ - \lambda}$, then the GAS property of (2.16) follows. Under this situation, (2.16) is GAS *uniformly* over all switching signals with $\tau_a \geqslant \frac{\ln \mu}{\lambda_\circ - \lambda}$. In prevailing literature, (2.16) is said to be *globally "uniformly" asymptotically stable* over all such switching signals.

Our comparison-based approach enables us to work with simple scalar differential equations which provide upper bounds of Lyapunov functions, rather than analyze the complicated evolution of the Lyapunov functions themselves. This provides new insights into average dwell-time switching, as is illustrated above, and moreover we can derive new switching rules that extend average dwell-time, as the following two remarks illustrate.

Remark 2.16. Let $N_\sigma(T, t)$ denote the number of switches on the interval $[t, T[$. As another illustration of Theorem 2.8, consider the switched system (2.2), and suppose that there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$, $p \in \mathcal{P}$, and real numbers $m > 1, \lambda_\circ > 0, \mu > 1$, such that

(i) the family $\{V_p \mid p \in \mathcal{P}\}$ is $\mathcal{P}$-uniformly h-positive definite and h°-decrescent in the sense of (2.5);
(ii) $\forall\, (t, x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ and $\forall\, p \in \mathcal{P}$, we have $\mathrm{D}^+_{f_p} V_p(t, x) \leqslant -\lambda_\circ V_p^m(t, x)$;
(iii) $\forall\, (t, x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ and $\forall\, p, q \in \mathcal{P}$, we have $V_p(t, x) \leqslant \mu V_q(t, x)$.

Let us find conditions on $N_\sigma(T, t_0)$ such that (2.2) is (h°, h)-GUAS.

Consider the impulsive differential system of the type (2.6):

$$\overline{\Sigma}: \quad \begin{cases} \dot{\xi} = -\lambda_\circ \xi^m, & t \neq \tau_i, \\ \xi(\tau_i) = \mu \xi(\tau_i^-), & \mu > 0, \end{cases} \qquad \xi(t_0) = V_{\sigma(t_0)}(x(t_0)), \quad i \geqslant 1, \quad t \geqslant t_0.$$

A straightforward analysis leads to

$$\xi(T) = \frac{\mu^{N_\sigma(T,t_0)} \xi(t_0)}{\left(1 + \lambda_\circ (m-1)(\xi(t_0))^{m-1} \left(\sum_{i=1}^{N_\sigma(T,t_0)+1} \mu^{(m-1)(i-1)} (\tau_i - \tau_{i-1})\right)\right)^{1/(m-1)}},$$

where we let $\tau_{N_\sigma(T,t_0)+1} := T$. Since $\mu > 1$, we obtain

$$\xi(T) \leqslant \frac{\mu^{N_\sigma(T,t_0)} \xi(t_0)}{\left(1 + \lambda_\circ (m-1)(\xi(t_0))^{m-1} (T - t_0)\right)^{1/(m-1)}}.$$

A little calculation shows that if there exists $\epsilon > 0$ and $K > 0$ such that

(2.27)
$$N_\sigma(T, t_0) \leqslant \frac{1}{\ln \mu} \ln \left(\frac{K \left(1 + \lambda_\circ (m-1) \left(\alpha_2 \circ h^\circ(t_0, x_0)\right)^{m-1} (T - t_0)\right)^{\left(\frac{1}{m-1} - \epsilon\right)}}{\alpha_2 \circ h^\circ(t_0, x_0)} \right),$$

then $\overline{\Sigma}$ is GUAS. Just as in the case of Σ' and Σ'' above, it is easy to verify that the hypotheses of Theorem 2.8 hold with $\Sigma = \overline{\Sigma}$. It then follows by Theorem 2.8 that (2.2) is (h°, h)-GUAS. Note that unlike the case of average dwell-time, the switching law in (2.27) that guarantees (h°, h)-GUAS of (2.2) depends on the initial conditions (t_0, x_0).

Remark 2.17. In the context of Theorem 2.8, it is not necessary to assume that each subsystem in the family (2.1) is (h°, h)-GUAS for the switched system (2.2) to be (h°, h)-GUAS. This is evident from hypothesis (ii), which does not require ϕ to be negative definite in the second argument. It is possible that none of the individual subsystems is (h°, h)-GUAS but the switched system is, provided each subsystem is active for a small enough time.

For instance, consider the switched system (2.16) and suppose that switching is such that the number of switches $N_\sigma(T, t_0)$ on the interval $[t_0, T[$ is given by

$$N_\sigma(T, t_0) \geqslant \frac{T - t_0}{\delta_a} - N_\circ, \tag{2.28}$$

where $N_\circ, \delta_a > 0$. Such switching signals were called "reverse average dwell-time switching signals" in [15]. Let there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}^1[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$, $p \in \mathcal{P}$, and real numbers $\mu' \in]0, 1[$ and $\lambda_\circ > 0$, such that

(i) the estimate (2.18) holds;
(ii) $\frac{\partial V_p}{\partial x} f_p(x) \leqslant \lambda_\circ V_p(x) \qquad \forall\, x \in \mathbb{R}^n, \quad \forall\, p \in \mathcal{P}$;
(iii) $V_{\sigma(\tau_i)}(x(\tau_i)) \leqslant \mu' V_{\sigma(\tau_i^-)}(x(\tau_i^-)) \quad \forall\, i \geqslant 1$;
(iv) $\delta_a < -\ln \mu' / \lambda_\circ$.

Then (2.16) is GAS. Indeed, consider the comparison system

$$\widetilde{\Sigma}: \quad \begin{cases} \dot{\xi} = \lambda_\circ \xi, & t \neq \tau_i, \\ \xi(\tau_i) = \mu' \xi(\tau_i^-), \end{cases} \qquad \xi(t_0) = V_{\sigma(t_0)}(t_0, x_0), \quad i \geqslant 1, \quad t \geqslant t_0. \tag{2.29}$$

It is easy to see that hypothesis (i) with $h^\circ(t, x) = h(t, x) = |x|$, and hypotheses (ii) and (iii) of Theorem 2.8 are satisfied. Following the constructions for Σ' in (2.21) above, it is not difficult to show that the solution of $\widetilde{\Sigma}$ is given by

$$\xi(T) = (\mu')^{N_\circ} V_{\sigma(t_0)}(x(t_0)) \mathrm{e}^{(\lambda_\circ + \ln \mu'/\delta_a)(T - t_0)}, \tag{2.30}$$

where $T > t_0$. Clearly, from (2.30) it follows that $\widetilde{\Sigma}$ is GAS if $\lambda_\circ + \ln \mu'/\delta_a < 0$. In other words, if the condition (iv) above holds, hypothesis (iv) of Theorem 2.8 is also satisfied; Theorem 2.8 now guarantees that (2.16) is GAS. However, this situation is admittedly restrictive; the condition (iii) above holds only under special situations.

Remark 2.18. Consider a hybrid system described by a partition of its continuous state space into regions via fixed switching surfaces (*guards*) and fixed continuous dynamics in each region [30, 44]. Every trajectory of such a hybrid system corresponding to a fixed initial condition can be realized as a trajectory of the switched system (2.2),

for a suitable time-dependent switching signal σ. For hybrid system trajectories corresponding to different initial conditions, the resulting switching signals are in general different, and stability properties of the corresponding switched system realizations are different. Therefore, we cannot conclude stability of the hybrid system from stability of the associated switched system. However, it is possible to conclude stability of the hybrid system from the switched system provided we have uniform stability with respect to a suitable class of switching signals—namely, the class of switching signals obtained from the hybrid system by varying its initial condition.

We have stated our Theorem 2.8 for a fixed switching signal σ; no uniformity with respect to σ is claimed. However, we have noted in Remark 2.15 that under specific hypotheses, we do have uniform stability over the class of switching signals with sufficiently large average dwell-time. It turns out that in supervisory control algorithms based on state-dependent hysteresis switching (utilizing guards with memory), the switching signal is effectively constrained to precisely such a class, thereby ensuring uniform stability; see, e.g., [30] and the references therein. Identification of other useful classes of switching signals for which we can conclude stability for hybrid systems in this way remains to be studied.

2.5. Remarks on other stability notions. Although not covered by the results presented so far, the comparison framework of Theorem 2.8 is general enough to describe various other stability behavior. The classical comparison principle has been successfully applied to describe *strict stability*, *total stability*, *practical stability*, and *finite time stability*, among others. The idea has also been applied in the context of partial, impulsive, stochastic, and functional differential equations and integral equations; see [29] for further details. In this subsection we study two other notions of stability of switched systems in our framework of Theorem 2.8, namely, uniform practical stability and finite time stability.

Practical stability. We briefly study one representative notion of practical stability, defined below; see, e.g., [27] for other definitions and details.

DEFINITION 2.19. *Let $h^\circ, h \in \Gamma$ and the pair (λ, A), $\lambda \in]0, A[$ be given. The system* (2.2) *is said to be* (h°, h)-uniformly practically stable with respect to (λ, A) *if for every $t_0 \in \mathbb{R}_{\geqslant 0}$, the property*

$$h^\circ(t_0, x_0) < \lambda \implies h(t, x(t)) < A \qquad \forall\, t \geqslant t_0 \tag{2.31}$$

holds for all solutions of (2.2).

With h°, h specialized to Euclidean norms, we recover usual uniform practical stability.

Remark 2.20. Uniform practical stability of impulsive differential equations of the type (2.6), for a given sequence $\{\tau_i\}_{i\geqslant 0}$, is identical to Definition 2.19, with h°, h specialized to absolute values.

The following result, which we state without proof, provides sufficient conditions for (h°, h) uniform practical stability of the switched system (2.2) with respect to a given pair (λ, A); see [9] for a proof. In section 3.5 we state and prove a stochastic version of this result. We define the "open tube" (or cylinder) of radius $r > 0$ in terms of a measure $h \in \Gamma$ to be the set

$$\mathcal{B}(h, r) = \{(t, x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n \mid h(t, x) < r\}.$$

PROPOSITION 2.21. *Consider the switched system* (2.2) *with a given σ, $h^\circ, h \in \Gamma$, and let the pair (λ, A), $\lambda \in]0, A[$ be given. Suppose that there exist functions $\alpha_1, \alpha_2 \in$*

$\mathcal{K}_\infty$, $V_p \in \mathrm{C}\big[\mathcal{B}(h,A) \cap \big((\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n) \backslash \mathcal{B}(h^\circ, \lambda)\big),\ \mathbb{R}_{\geqslant 0}\big]$ *for each* $p \in \mathcal{P}$ *locally Lipschitz in the second argument, and a system* Σ *of the type* (2.6)*, such that*

(i) *the family* $\{V_p \mid p \in \mathcal{P}\}$ *is* $\mathcal{P}$*-uniformly* h*-positive definite and* h°*-decrescent in the sense of* (2.5)*;*

(ii) $\forall\, (t,x) \in \mathcal{B}(h,A) \cap \big((\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n) \backslash \mathcal{B}(h^\circ, \lambda)\big)$ *and* $\forall\, p \in \mathcal{P}$*, we have* $\mathrm{D}^+_{f_p} V_p(t,x) \leqslant \phi(t, V_p(t,x))$*;*

(iii) $V_{\sigma(\tau_i)}(\tau_i, x(\tau_i)) \leqslant \xi(\tau_i) \quad \forall\, i \geqslant 0$*, where* $x(t)$ *and* $\xi(t)$ *are solutions of* (2.2) *and* Σ*, respectively;*

(iv) Σ *is uniformly practically stable with respect to* $(\alpha_2(\lambda), \alpha_1(A))$.

Then (2.2) *is* (h°, h)*-uniformly practically stable with respect to* (λ, A).

Finite time stability. We provide sufficient conditions for *finite time stability* in case of autonomous switched systems only. The definition of finite-time stability in terms of two measures is proposed in Definition 2.22; for further details on finite-time stability with Euclidean norms, see [3].

For autonomous switched systems, we specialize the class of functions Γ in (2.3) to the corresponding autonomous version $\Gamma_a := \{h \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}] \mid \inf_x h(x) = 0\}$. For simplicity, we only consider special cases of h° and h below.

Definition 2.22. *Let* $h^\circ, h \in \Gamma_a$ *such that* $\ker h^\circ = \ker h = \{0\}$ *and* h°, h *are positive definite, radially unbounded. The switched system* (2.16) *is said to be* (h°, h)-finite time stable *if for every* $\varepsilon > 0$ *the following two properties hold simultaneously:*

- *The finite time convergence property holds; i.e., there exists a function* $T : \mathbb{R}^n \backslash \{0\} \longrightarrow \mathbb{R}_{>0}$ *called the settling time function such that for every* $x_0 \in \mathbb{R}^n \backslash \{0\}$*,* $x(t)$ *is defined on the interval* $[0, T(x_0)[$ *with* $x(t) \in \mathbb{R}^n \backslash \{0\}$ *and* $\lim_{t \uparrow T(x_0)} x(t) = 0$*, where* $x(t)$ *is the solution of* (2.16) *with initial condition* x_0.
- *Lyapunov stability holds; i.e., there exists a function* $\delta \in \mathcal{K}_\infty$ *such that* $h^\circ(x_0) < \delta(\varepsilon) \implies h \circ x(t) < \varepsilon \quad \forall\, t \in [0, T(x_0)[$.

With h°, h specialized to Euclidean norms, we recover finite time stability in the sense of [3].

Remark 2.23. Finite time stability of impulsive differential equations of the type (2.6) is identical to Definition 2.22, with h°, h specialized to absolute values.

Proposition 2.24. *Consider the switched system* (2.2) *with a given* σ*, and* $h^\circ, h \in \Gamma_a$*. Suppose that there exist functions* $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$*, locally Lipschitz* $V_p \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ *for each* $p \in \mathcal{P}$ *and a system* Σ *of the type* (2.6)*, such that*

(i) $\forall\, x \in \mathbb{R}^n$ *and* $\forall\, p \in \mathcal{P}$*, we have* $\alpha_1 \circ h(x) \leqslant V_p(x) \leqslant \alpha_2 \circ h^\circ(x)$*;*

(ii) $\forall\, x \in \mathbb{R}^n$ *and* $\forall\, p \in \mathcal{P}$*, we have* $\mathrm{D}^+_{f_p} V_p(x) \leqslant \phi(t, V_p(x))$*;*

(iii) $\forall\, x_0 \in \mathbb{R}^n$*, there exists* $\xi_0 \in \mathbb{R}_{\geqslant 0}$ *such that* $V_{\sigma(\tau_i)}(\tau_i, x(\tau_i)) \leqslant \xi(\tau_i) \quad \forall\, i \geqslant 0$*, where* $x(t)$ *and* $\xi(t)$ *are the corresponding solutions of* (2.16) *and* Σ*, respectively;*

(iv) Σ *is finite time stable.*

Then (2.16) *is finite time stable.*

The proof of this Proposition is not difficult and is omitted.

3. Stochastic switched systems. In this section, we study stability of stochastic switched nonlinear nonautonomous systems of the Itô type. In section 3.1, we describe a stochastic switched system and define the notions of stability that we study. We propose our comparison theorem for stochastic switched systems in section 3.2 and illustrate it in sections 3.3 and 3.4. We discuss two other stability notions in section 3.5 and demonstrate that their analysis can be similarly carried out in the framework of the proposed results.

3.1. Preliminaries.

System description and stability definitions. Let $\mathbf{\Omega} := (\Omega, \mathcal{F}, \mathsf{P})$ be a complete probability space, where Ω is the sample space, $\mathcal{F}$ is the Borel σ-algebra on Ω, and P is a probability measure on the measurable space $(\Omega, \mathcal{F})$.

We consider a family of nonlinear nonautonomous Itô systems,

$$\mathrm{d}x = f_p(t,x)\mathrm{d}t + G_p(t,x)\mathrm{d}w, \qquad p \in \mathcal{P}, \tag{3.1}$$

where $x \in \mathbb{R}^n$, $\mathcal{P}$ is an index set, $f_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}^n]$ is a vector field, $G_p \in \mathrm{C}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}^{n\times m}]$ is a diffusion rate matrix function, $f_p(\cdot, 0) \equiv 0$ and $G_p(\cdot, 0) \equiv 0$ for every $p \in \mathcal{P}$, w is an m-dimensional normalized Wiener process defined on the probability space $\mathbf{\Omega}$, and $\mathrm{d}x$ is a stochastic differential of x. We assume that f_p and G_p are smooth enough to ensure existence and uniqueness of the corresponding solution process; for precise conditions see, e.g., [36]. A *switched system* generated by the family (3.1) and a *switching signal* σ, defined similarly to section 2.1, is

$$\mathrm{d}x = f_\sigma(t,x)\mathrm{d}t + G_\sigma(t,x)\mathrm{d}w, \qquad x(t_0) = x_0, \quad t \geqslant t_0, \tag{3.2}$$

where $t_0 \in \mathbb{R}_{\geqslant 0}$. It is assumed, just as in section 2.1, that there is no jump in the state x at the switching instants, and there is a finite number of switches on every bounded interval of time. We denote the switching instants by τ_i, $i = 1, 2, \ldots$, with $\tau_0 := t_0$, and the sequence $\{\tau_i\}_{i\geqslant 0}$ is strictly increasing. The *solution process of* (3.2) *as a function of time* t*, initialized at a given pair* (t_0, x_0)*, and under a given switching signal* σ, is denoted by $x(t)$. In what follows, expected values at the (deterministic) initial condition are to be identified with their actual values; see also Remark 3.3.

Let $|x|_q := (\mathsf{E}[|x|^q])^{1/q}$ denote the qth mean of a random variable x defined on $\mathbf{\Omega}$. We will have occasion to use *Jensen's inequality*: if $\varphi \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}]$ is concave and x is a random variable on $\mathbf{\Omega}$, then $\mathsf{E}[\varphi(x)] \leqslant \varphi(\mathsf{E}[x])$. Also, we need *Chebyshev's inequality*: for $\varepsilon > 0$, $\psi \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$, x a random variable on $\mathbf{\Omega}$, we have $\mathsf{P}[\psi(x) \geqslant \varepsilon] \leqslant \mathsf{E}[\psi(x)]\,/\varepsilon$.

Consider the system with index p in the family (3.1). Let $V \in \mathrm{C}^{1,2}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$. By Itô's formula we have the stochastic differential of V as

$$\mathrm{d}V(t,x) = \mathscr{L}_p V(t,x)\mathrm{d}t + V_x(t,x)G_p(t,x)\mathrm{d}w(t),$$

where

$$\mathscr{L}_p V(t,x) := V_t(t,x) + V_x(t,x)f_p(t,x) + \frac{1}{2}\operatorname{tr}\left(V_{xx}(t,x)G_p(t,x)G_p^{\mathrm{T}}(t,x)\right) \tag{3.3}$$

is the *infinitesimal generator* for the system with index p in (3.1) acting on the function V, and V_t, V_x, V_{xx} denote the partial differentials of $V(t,x)$ with respect to t, x, and twice with respect to x, respectively, and tr denotes the trace of a matrix; see, e.g., [14].

We focus on the following two general notions of stochastic stability in terms of two measures, which belong to the set Γ in (2.3).

Definition 3.1. *Let* $h^\circ, h \in \Gamma$. *The stochastic switched system* (3.2) *is said to be* (h°, h)-globally uniformly asymptotically stable in the mean ((h°, h)-GUAS-M) *if there exists a function* $\beta \in \mathcal{KL}$ *such that for every* $(t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$*, the inequality*

$$\mathsf{E}[h(t, x(t))] \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0 \tag{3.4}$$

holds.

DEFINITION 3.2. *The stochastic switched system* (3.2) *is said to be* (h°, h)-globally uniformly asymptotically stable in probability ((h°, h)-GUAS-P) *if for every* $\eta \in]0,1[$, *there exists a function* $\beta \in \mathcal{KL}$ *such that for every* $(t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, *the inequality*

$$\mathsf{P}\Big[h(t, x(t)) \geqslant \beta(h^\circ(t_0, x_0), t - t_0)\Big] < \eta \qquad \forall\, t \geqslant t_0 \tag{3.5}$$

holds.

Remark 3.3. We have stated our stability definitions in terms of deterministic initial condition. However, the results of this section hold true for the general case of stochastic initial condition as well. For instance, if we consider replacing $h^\circ(t_0, x_0)$ with $\mathsf{E}[h^\circ(t_0, x_0)]$ or $h^\circ(t_0, \mathsf{E}[x_0])$ in either (3.4) or (3.5), all the results hold true, with minor straightforward modifications in the proofs.

Remark 3.4. The (h°, h)-GUAS-M property is equivalent to the simultaneous verification of the following properties:

(SM1) there exists a class $\mathcal{K}_\infty$ function δ such that for every $\varepsilon > 0$, $t_0 \in \mathbb{R}_{\geqslant 0}$, we have $h^\circ(t_0, x_0) < \delta(\varepsilon) \implies \mathsf{E}[h(t, x(t))] < \varepsilon \quad \forall\, t \geqslant t_0$;

(SM2) for every $r, \varepsilon > 0$, there exists a number $T(r, \varepsilon) \geqslant 0$ such that for every $t_0 \in \mathbb{R}_{\geqslant 0}$, we have $h^\circ(t_0, x_0) < r \implies \mathsf{E}[h(t, x(t))] < \varepsilon \quad \forall\, t \geqslant t_0 + T(r, \varepsilon)$.

The proof follows directly from the proof of the equivalence in Remark 2.3 presented in Appendix A, with $\mathsf{E}[h(t, x(t))]$ replacing $h(t, x(t))$.

The (h°, h)-GUAS-P property is equivalent to the simultaneous verification of the following properties:

(WP1) for every $\eta' \in]0,1[$, there exists a function $\delta \in \mathcal{K}_\infty$ such that for every $\varepsilon > 0$, $t_0 \in \mathbb{R}_{\geqslant 0}$, we have $h^\circ(t_0, x_0) < \delta(\varepsilon) \implies \mathsf{P}[h(t, x(t)) \geqslant \varepsilon] < \eta' \quad \forall\, t \geqslant t_0$; and

(WP2) for every $\eta'' \in]0,1[$, $r, \varepsilon' > 0$, there exists a number $T(r, \varepsilon) \geqslant 0$ such that for every $t_0 \in \mathbb{R}_{\geqslant 0}$, we have $h^\circ(t_0, x_0) < r \implies \mathsf{P}[h(t, x(t)) \geqslant \varepsilon'] < \eta'' \quad \forall\, t \geqslant t_0 + T(r, \varepsilon')$.

Establishing this equivalence takes some more work; an outline may be found at the end of Appendix B.

Remark 3.5. We recover a notion essentially equivalent to *global uniform asymptotic stability in the qth mean* (GUAS-M$_q$) from Definition 3.1 with $h^\circ(t, x) = h(t, x) = |x|^q$, $q \geqslant 1$; see, e.g., [14]. With the same h° and h, Definition 3.2 yields *global uniform asymptotic stability in probability* (GUAS-P) in the sense of [23].

Remark 3.6. The (h°, h)-GUAS-P property of (3.2) follows from its (h°, h)-GUAS-M property. To see this, pick $\eta \in]0,1[$ and let there exist a function $\beta \in \mathcal{KL}$ such that (3.4) is satisfied. Consider a second function $\overline{\beta} \in \mathcal{KL}$ such that $\overline{\beta}(r, s) > \beta(r, s)/\eta$ for all $(r, s) \in \mathbb{R}^2_{\geqslant 0}$. Utilizing Chebyshev's inequality, we now have for every $t \geqslant t_0$

$$\begin{aligned}\mathsf{P}\big[h(t, x(t)) \geqslant \overline{\beta}(h^\circ(t_0, x_0), t - t_0)\big] &\leqslant \frac{\mathsf{E}[h(t, x(t))]}{\overline{\beta}(h^\circ(t_0, x_0), t - t_0)} \\ &\leqslant \frac{\beta(h^\circ(t_0, x_0), t - t_0)}{\overline{\beta}(h^\circ(t_0, x_0), t - t_0)} < \eta,\end{aligned}$$

which is the (h°, h)-GUAS-P property.

Comparison systems and stochastic comparison principle. We utilize the comparison system (2.6); see section 2.1 for the definitions. We need the following stochastic version of the comparison principle; see, e.g., [26] for a proof, where the authors also consider stochastic comparison systems.

LEMMA 3.7. *Consider the stochastic system with index p in the family* (3.1). *Suppose that there exist a function $V \in \mathrm{C}^{1,2}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ and a comparison system*

$$\Sigma : \quad \dot{\xi} = \phi(t, \xi), \qquad \xi(t_0) = \xi_0 \geqslant 0, \quad t \geqslant t_0,$$

where ϕ is concave in the second argument, such that for all $(t, x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, the differential inequality

$$\mathscr{L}_p V(t, x) \leqslant \phi(t, V(t, x))$$

is valid. Then $\mathsf{E}[V(\tau, x(\tau))] \leqslant \xi(\tau)$ implies $\mathsf{E}[V(t, x(t))] \leqslant \xi(t) \quad \forall\, t \geqslant \tau$, where $x(t)$ and $\xi(t)$ are the solution process of the system with index p in (3.1) *and the solution of Σ, respectively.*

3.2. Comparison theorem for stochastic switched systems. The following result establishes a general framework for testing stability of stochastic switched systems using multiple Lyapunov functions and a comparison system.

THEOREM 3.8. *Consider the stochastic switched system* (3.2) *with a fixed switching signal σ generating a sequence of switching instants $\{\tau_i\}_{i \geqslant 1}$ and two functions $h^\circ, h \in \Gamma$. Suppose that there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}^{1,2}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ for each $p \in \mathcal{P}$, and a system Σ of the type* (2.6), *such that*

(i) *α_1 is convex, and ϕ is concave in the second argument;*
(ii) *the family $\{V_p \mid p \in \mathcal{P}\}$ is $\mathcal{P}$-uniformly h-positive definite and h°-decrescent in the sense of* (2.5)*;*
(iii) *$\forall\, (t, x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ and $\forall\, p \in \mathcal{P}$, the estimate $\mathscr{L}_p V_p(t, x) \leqslant \phi(t, V_p(t, x))$ holds;*
(iv) *$\forall\, (t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, there exists $\xi_0 \in \mathbb{R}_{\geqslant 0}$ such that $\mathsf{E}\big[V_{\sigma(\tau_i)}(\tau_i, x(\tau_i))\big] \leqslant \xi(\tau_i)$ for all $i \geqslant 0$, where $x(t)$ and $\xi(t)$ are the solution process of* (3.2) *and the solution of Σ, respectively, for these initial conditions;*
(v) *Σ is* GUAS *in the sense of* (2.7).

Then (3.2) *is (h°, h)-*GUAS-M *in the sense of Definition* 3.1.

Proof. Consider the interval $[\tau_\ell, \tau_{\ell+1}[$, with ℓ an arbitrary nonnegative integer. Combining hypotheses (iv), (iii), and (i), and Lemma 3.7 with $\tau = \tau_\ell$, we have

$$\mathsf{E}\big[V_{\sigma(\tau_\ell)}(t, x(t))\big] \leqslant \xi(t) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[.$$

The above estimate, in conjunction with hypothesis (v) and (2.7), yields

$$\mathsf{E}\big[V_{\sigma(\tau_\ell)}(t, x(t))\big] \leqslant \beta_\xi(\xi_0, t - t_0) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[. \tag{3.6}$$

The function α_1 being convex by hypothesis (i), taking expectations and using Jensen's inequality in (2.5) and considering (3.6), we reach the estimate

$$\alpha_1 \circ \mathsf{E}[h(t, x(t))] \leqslant \beta_\xi(\xi_0, t - t_0) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[. \tag{3.7}$$

The arbitrariness of ℓ in (3.7) leads to

$$\mathsf{E}[h(t, x(t))] \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0,$$

where $\beta(r, s) := \alpha_1^{-1} \circ \beta_\xi(\alpha_2(r), s)$. The function β being of class $\mathcal{KL}$, it follows that (3.2) is (h°, h)-GUAS-M. ∎

The following obvious Corollary, which we merely state, follows almost immediately from Theorem 3.8 in the light of Remark 3.6.

COROLLARY 3.9. *Suppose that the hypotheses of Theorem* 3.8 *hold true. Then* (3.2) *is* (h°, h)-GUAS-P *in the sense of* (3.5).

Remark 3.10. The hypothesis on convexity of α_1 in Corollary 3.9 actually is not necessary for (h°, h)-GUAS-P of (3.2). In Theorem 3.24 we will prove a stronger property without this convexity assumption, which will imply the (h°, h)-GUAS-P property of (3.2). For the moment, however, we shall work with (h°, h)-GUAS-M.

Theorem 3.8, like its deterministic counterpart Theorem 2.8, does not provide a direct method for analyzing stability of a given stochastic switched system. We will now demonstrate how to proceed with such a scheme of analysis and propose a few more specific results. In section 3.3 we provide sufficient conditions for (h°, h)-GUAS-M under a fixed-index sequence monotonicity condition. The comparison system utilized in the proof of this result utilizes quantitative information of system trajectories. In section 3.4 we provide sufficient conditions for GUAS-M_q of a stochastic switched system under average dwell-time switching, where the fixed-index sequence monotonicity assumption imposed in section 3.3 is violated. We construct two comparison systems that do not require quantitative information of system trajectories.

3.3. Stochastic stability under fixed-index sequence monotonicity condition. For this subsection we let $\mathcal{P}$ be a finite set with N elements. The following result provides a stochastic version of Corollary 2.9.

COROLLARY 3.11. *Consider the stochastic switched system* (3.2) *and* $h^\circ, h \in \Gamma$. *Suppose there exist functions* $\alpha, \alpha_1, \alpha_2, \rho, U \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}^{1,2}[\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ *for each* $p \in \mathcal{P}$, *such that*

(i) α_1, $\alpha \circ \alpha_2^{-1}$ *and* $U \circ \alpha_2^{-1}$ *are convex, and* ρ *is concave;*
(ii) *the family* $\{V_p \mid p \in \mathcal{P}\}$ *is* $\mathcal{P}$*-uniformly* h*-positive definite and* h°*-decrescent in the sense of* (2.5);
(iii) $\forall\, (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ *and* $\forall\, p \in \mathcal{P}$, *we have* $\mathscr{L}_p V_p(t,x) \leqslant -\alpha \circ h^\circ(t,x)$;
(iv) *for every pair of switching times* (τ_i, τ_j), $i < j$ *such that* $\sigma(\tau_i) = \sigma(\tau_j) = p$ *and* $\sigma(\tau_k) \neq p$ *for* $\tau_i < \tau_k < \tau_j$, *the inequality*

$$\mathsf{E}[V_p(\tau_j, x(\tau_j))] - \mathsf{E}[V_p(\tau_i, x(\tau_i))] \leqslant -\mathsf{E}[U \circ h^\circ(\tau_i, x(\tau_i))] \tag{3.8}$$

holds, where $x(t)$ *is the solution process of* (3.2) *initialized at* (t_0, x_0);
(v) $\forall\, (t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, *we have* $\alpha_2 \circ h^\circ(t,x) \leqslant \rho \circ \alpha_1 \circ h(t,x)$.

Then (3.2) *is* (h°, h)-GUAS-M.

Proof. We define a candidate impulsive differential comparison system of the type (2.6):

$$\Sigma: \quad \begin{cases} \dot{\xi} = -\alpha \circ \alpha_2^{-1}(\xi), & t \neq \tau_i, \\ \xi(\tau_i) = \mathsf{E}\left[V_{\sigma(\tau_i)}(\tau_i, x(\tau_i))\right], & \end{cases} \qquad i \geqslant 0, \quad t \geqslant t_0. \tag{3.9}$$

By its very definition, Σ satisfies hypotheses (iii)–(iv) of Theorem 3.8. Hypotheses (i)–(ii) of Theorem 3.8 are satisfied by our hypotheses (i)–(ii). To verify hypothesis (v) of Theorem 3.8, we shall first prove Lyapunov stability of Σ and then prove its global asymptotic convergence, in view of Remark 2.6.

Consider the time interval $[\tau_0, \tau_1[$. From hypothesis (iii), we have

$$\mathsf{E}\left[V_{\sigma(\tau_0)}(\tau_1, x(\tau_1))\right] \leqslant \mathsf{E}\left[V_{\sigma(\tau_0)}(\tau_0, x_0)\right].$$

Combining with hypothesis (ii), using Jensen's inequality we reach

$$\mathsf{E}[\alpha_1(h(\tau_1, x(\tau_1)))] \leqslant \alpha_2 \circ h^\circ(t_0, x_0). \tag{3.10}$$

For every $p \in \mathcal{P}$, we have from hypothesis (i)

$$\mathsf{E}[V_p(\tau_1, x(\tau_1))] \leqslant \mathsf{E}[\alpha_2 \circ h^\circ(\tau_1, x(\tau_1))]\,,$$

and therefore by hypothesis (v),

$$\mathsf{E}[V_p(\tau_1, x(\tau_1))] \leqslant \mathsf{E}[\rho \circ \alpha_1 \circ h(\tau_1, x(\tau_1))]\,. \tag{3.11}$$

In view of hypothesis (ii), we apply Jensen's inequality to (3.11) and use hypothesis (ii) to get

$$\mathsf{E}[V_p(\tau_1, x(\tau_1))] \leqslant \rho(\mathsf{E}[\alpha_1 \circ h(\tau_1, x(\tau_1))]) \leqslant \rho \circ \alpha_2 \circ h^\circ(t_0, x_0). \tag{3.12}$$

Consider now the interval $[\tau_1, \tau_2[$. From hypothesis (iii) we have

$$\mathsf{E}\big[V_{\sigma(\tau_1)}(\tau_2, x(\tau_2))\big] \leqslant \mathsf{E}\big[V_{\sigma(\tau_1)}(\tau_1, x(\tau_1))\big]\,.$$

Combining with hypothesis (ii) and applying (3.12) with $p = \sigma(\tau_2)$, we get

$$\mathsf{E}[\alpha_1 \circ h(\tau_2, x(\tau_2))] \leqslant \rho \circ \alpha_2 \circ h^\circ(t_0, x_0). \tag{3.13}$$

For all $p \in \mathcal{P}$, we have

$$\mathsf{E}[V_p(\tau_2, x(\tau_2))] \leqslant \mathsf{E}[\alpha_2 \circ h^\circ(\tau_2, x(\tau_2))]\,,$$

so by hypothesis (iv) and Jensen's inequality we get

$$\mathsf{E}[V_p(\tau_2, x(\tau_2))] \leqslant \rho(\mathsf{E}[\alpha_1 \circ h(\tau_2, x(\tau_2))]).$$

Now (3.12) gives

$$\mathsf{E}[V_p(\tau_2, x(\tau_2))] \leqslant \rho \circ \rho \circ \alpha_2 \circ h^\circ(t_0, x_0).$$

It is not difficult to see that the worst-case situation for maximum possible overshoot of the function $\mathsf{E}[V_\sigma]$ occurs when the switching signal σ visits every element of the set $\mathcal{P}$ without repetition until $\mathcal{P}$ is exhausted. Let $\tau_{j^\star}$ be the first switching instant after all the subsystems that participate in the dynamics have become active at least once since initialization at $t = t_0$. From the above computations, it is easy to see that, with

$$\rho^j := \underbrace{\rho \circ \ldots \circ \rho}_{j \text{ times}},$$

the estimate

$$\xi(\tau_{j^\star}) \leqslant \rho^{\mathrm{N}-1} \circ \alpha_2 \circ h^\circ(t_0, x_0),$$

is valid. Define the function

$$\gamma(\cdot) := \max\left\{\alpha_2(\cdot), \rho \circ \alpha_2(\cdot), \ldots, \rho^{\mathrm{N}-1} \circ \alpha_2(\cdot)\right\}.$$

From the above arguments and (3.8), it follows that $\xi(t) \leqslant \gamma \circ h^\circ(t_0, x_0)$. In view of the definition of ξ_0 and hypothesis (ii), this leads to

$$\xi(t) \leqslant \gamma \circ \alpha_1^{-1}(\xi_0) \qquad \forall\, t \geqslant t_0. \tag{3.14}$$

It remains to prove uniform global asymptotic convergence of Σ.

We distinguish two cases.

Case 1. *Switching stops in finite time.* Since σ eventually becomes constant from the κth switching instant, it follows that there are no impulses after $t = \tau_\kappa$. Therefore, the system (3.9) becomes an autonomous scalar ordinary differential equation after $t = \tau_\kappa$, with negative right-hand side for nonzero $\xi(\tau_\kappa)$. It follows that $\xi(t)$ monotonically decreases to 0 for all $t \geqslant \tau_\kappa$. From (3.14) uniform Lyapunov stability of Σ follows. Thus, (3.9) is GUAS. Theorem 3.8 now guarantees that (3.2) is (h°, h)-GUAS-M.

Case 2. *Switching continues indefinitely.* Consider the restatement of the inequality (3.8) with $\xi(\tau_i)$ as in (3.9)

$$\begin{aligned}\xi(\tau_j) - \xi(\tau_i) &\leqslant -\mathsf{E}[U \circ h^\circ(\tau_i, x(\tau_i))] = -\mathsf{E}\left[U \circ \alpha_2^{-1} \circ \alpha_2 \circ h^\circ(\tau_i, x(\tau_i))\right] \\ &\leqslant -U \circ \alpha_2^{-1}\left(\mathsf{E}[\alpha_2 \circ h^\circ(\tau_i, x(\tau_i))]\right),\end{aligned}$$

where we have utilized Jensen's inequality and convexity of $U \circ \alpha_2^{-1}$ in hypothesis (i). The infinite sequence $\{\xi(\tau_i)\}_{\{i \geqslant 0 | \sigma(\tau_i) = p\}}$ is monotonically nonincreasing and therefore must attain a limit, say, $c \geqslant 0$. Taking limits as $i \uparrow \infty$ on both sides of (3.8), we have

$$c - c \leqslant - \lim_{\substack{i \uparrow \infty \\ \sigma(\tau_i) = p}} U \circ \alpha_2^{-1}\left(\mathsf{E}[\alpha_2 \circ h^\circ(\tau_i, x(\tau_i))]\right),$$

which leads to

$$\lim_{\substack{i \uparrow \infty \\ \sigma(\tau_i) = p}} \mathsf{E}[\alpha_2 \circ h^\circ(\tau_i, x(\tau_i))] = 0.$$

Considering hypothesis (ii) and the reset equation in (3.9), we have

$$\lim_{\substack{i \uparrow \infty \\ \sigma(\tau_i) = p}} \xi(\tau_i) = \lim_{\substack{i \uparrow \infty, \\ \sigma(\tau_i) = p}} \mathsf{E}\left[V_{\sigma(\tau_i)}(\tau_i, x(\tau_i))\right] = 0. \tag{3.15}$$

Combining (3.14) and (3.15), we conclude that (3.9) is GUAS.

By Theorem 3.8, there exists a function $\beta \in \mathcal{KL}$ such that

$$\mathsf{E}[h(t, x(t))] \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0,$$

and we conclude that (3.2) is (h°, h)-GUAS-M. □

Remark 3.12. We note that in the above proof, Σ makes explicit use of state information of (3.2)—in the notation of (2.6), we use

$$\psi_i\left(\xi\left(\tau_i^-\right), y(x(\tau_i), \sigma(\tau_i))\right) = \mathsf{E}\left[V_{\sigma(\tau_i)}(\tau_i, x(\tau_i))\right]$$

in the reset equation. However, $\xi(\tau_i^-)$ is not utilized; cf. Remark 2.10.

Remark 3.13. For Euclidean norms, the function $\rho \in \mathcal{K}_\infty$ always exists if the function $\alpha_2 \circ \alpha_1^{-1}$ is concave (cf. Remark 2.11). Also, if α_1, α_2, ρ, and U are quadratic, as is typically the case for linear systems, hypothesis (i) is always satisfied.

Remark 3.14. It readily follows that for autonomous switched stochastic systems and h°, h specialized to Euclidean norms, Corollary 3.11 gives global asymptotic stability in the mean. Corollary 3.9 then implies global asymptotic stability in probability, which is derived in [8] without the aid of the comparison framework.

3.4. Stochastic stability under average dwell-time switching. In this subsection we investigate conditions on the average dwell-time τ_a of a switching signal such that (3.2) has the GUAS-M$_q$ property. We no longer retain the assumption that $\mathcal{P}$ is finite. We specialize to Euclidean norms for simplicity and propose the following result, which may be regarded as a stochastic counterpart of Theorem 2.12. A generalization of Theorem 3.15 to two measures and nonautonomous stochastic switched systems is readily done.

Let us consider the autonomous stochastic switched system

$$\mathrm{d}x = f_\sigma(x)\mathrm{d}t + G_\sigma(x)\mathrm{d}w, \qquad x(t_0) = x_0, \quad t \geqslant t_0, \tag{3.16}$$

where $x \in \mathbb{R}^n$, $f_p \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}^n]$, $G_p \in \mathrm{C}[\mathbb{R}^n, \mathbb{R}^{n\times m}]$, $f_p(0) = 0$, $G_p(0) = 0$ for every $p \in \mathcal{P}$, w is an m-dimensional Wiener process on the probability space $\mathbf{\Omega}$. We assume that f_p and G_p are smooth enough to ensure existence and uniqueness of the corresponding solution processes for every $p \in \mathcal{P}$; see, e.g., [36] for precise conditions.

We need the definition of average dwell-time from (2.17) for the following result.

THEOREM 3.15. *Consider the switched system* (3.2)*. Suppose that there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}^1[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$ for each $p \in \mathcal{P}$, and a positive number $\lambda_\circ$ such that*

$$\alpha_1(|x|^q) \leqslant V_p(x) \leqslant \alpha_2(|x|^q) \qquad \forall\, x \in \mathbb{R}^n \tag{3.17}$$

with α_1 convex, $q \geqslant 1$, and

$$\mathscr{L}_p V_p(x) \leqslant -\lambda_\circ V_p(x) \qquad \forall\, x \in \mathbb{R}^n. \tag{3.18}$$

Suppose also that there exists a positive constant μ such that for each $t \in \mathbb{R}_{\geqslant 0}$,

$$V_{p_1}(x) \leqslant \mu V_{p_2}(x) \qquad \forall\, x \in \mathbb{R}^n, \quad \forall\, p_1, p_2 \in \mathcal{P}. \tag{3.19}$$

Then (3.2) *is* GUAS-M$_q$ *for every switching signal σ with average dwell-time $\tau_a > \frac{\ln \mu}{\lambda_\circ}$.*

For a single system, (3.18) is a condition that implies global exponential stability in the qth mean [14], under an additional condition of $\alpha_i(r) = k_i r$, $i = 1, 2$. Theorem 3.15 is particularly simple for autonomous linear stochastic switched systems as we now show.

Consider a stochastic switched autonomous linear system

$$\mathrm{d}x = A_\sigma x\mathrm{d}t + B_\sigma x\mathrm{d}w_t, \qquad x(0) = x_0, \quad t \geqslant 0, \tag{3.20}$$

where $\sigma(t) \in \mathcal{P}$, $x \in \mathbb{R}^n$, $A_p, B_p \in \mathbb{R}^{n\times n}$, w is a scalar normalized Wiener process. For quadratic Lyapunov functions $V_p(x) = x^{\mathrm{T}} P_p x$, where P_p, $p \in \mathcal{P}$ is a positive definite symmetric matrix, with the aid of (3.3) the condition (3.18) simplifies to

$$A_p^{\mathrm{T}} P_p + P_p A_p + B_p P_p B_p^{\mathrm{T}} + \lambda_\circ P_p \leqslant 0, \tag{3.21}$$

which is a linear matrix inequality. To get the quadratic Lyapunov functions V_p, it is necessary to solve (3.21) for each $p \in \mathcal{P}$; see [5] for related methods of solution of such linear matrix inequalities, and see [14] for a discussion on quadratic Lyapunov functions in stability analysis of stochastic linear systems. Note that (3.19) is automatically satisfied if we can find such $V_p = x^{\mathrm{T}} P_p x$, with P_p satisfying (3.21).

We provide two different proofs of Theorem 3.15 to illustrate the versatility of our comparison framework.

An impulsive differential equation as a comparison system.

Proof. Consider an impulsive differential system of the type (2.6) with

$$\psi_i\left(\xi\left(\tau_i^-\right), y(x(\tau_i), \sigma(\tau_i))\right) = \mu\xi(\tau_i^-), \quad i \geqslant 1, \quad \mu > 0,$$

as the reset equation, and

$$\phi(t, \xi) = -\lambda_\circ \xi, \quad \lambda_\circ > 0.$$

Note that concavity of ϕ with respect to ξ is trivially satisfied. The complete system stands as

$$\Sigma' : \quad \begin{cases} \dot{\xi} = -\lambda_\circ \xi, & t \neq \tau_i, \\ \xi(\tau_i) = \mu\xi(\tau_i^-), & \mu > 0, \end{cases} \qquad \xi(t_0) = V_{\sigma(t_0)}(x_0), \quad i \geqslant 1, \quad t \geqslant t_0. \tag{3.22}$$

This system Σ' is the same as (2.21). From (3.17) it follows that hypothesis (ii) of Theorem 3.8 is satisfied with $h^\circ(t, x) = h(t, x) = |x|^q$. Further, from (3.18) and (3.19) together with the initial condition in (3.22), it follows that hypotheses (iii) and (iv) of Theorem 3.8 are satisfied, respectively.

Intuitively,

- the minimum rate of decay of the expected values of Lyapunov functions corresponding to each active subsystem is captured by the vector field of Σ'; and
- the maximum jump in the values of two Lyapunov functions corresponding to two consecutively active subsystems is captured by the reset equation of Σ' (since from (3.19) it follows that $\mathsf{E}[V_{p_1}(x(\tau_i))] \leqslant \mu\mathsf{E}\big[V_{p_2}(x(\tau_i^-))\big]$ for every $i \geqslant 1$, as in section 2.4, $\mu \geqslant 1$).

As in section 2.4, with $\tau_a > \frac{\ln \mu}{\lambda_\circ}$ the GUAS property of Σ' follows, which verifies hypothesis (v) of Theorem 3.8. In view of Remark 3.5, by Theorem 3.8 we conclude that (3.2) is GUAS-M$_q$. □

An ordinary differential equation as a comparison system.

Proof. Consider the following scalar autonomous system as a candidate comparison system:

$$\Sigma'' : \quad \dot{\xi} = \left(\frac{\ln \mu}{\tau_a} - \lambda_\circ\right)\xi, \qquad \xi(t_0) = \mu^{N_\circ} \mathrm{e}^{\lambda_\circ t_0} V_{\sigma(t_0)}(x_0), \quad t \geqslant t_0, \tag{3.23}$$

Note that concavity of the vector field of Σ'' with respect to ξ is trivially satisfied. Let the average dwell-time of σ be τ_a, and let $\nu := N_\sigma(T, t_0)$. Considering the least rate of decay of the expected values of Lyapunov functions corresponding to each active subsystem given by (3.18), we have for an arbitrary $T \geqslant 0$,

$$\mathsf{E}\big[V_{\sigma(\tau_i)}(x(\tau_{i+1}^-))\big] \leqslant \mathsf{E}\big[V_{\sigma(\tau_i)}(x(\tau_i))\big]\, \mathrm{e}^{-\lambda_\circ(\tau_{i+1} - \tau_i)}, \quad 0 \leqslant i \leqslant \nu,$$

and

$$\mathsf{E}\big[V_{\sigma(\tau_\nu)}(x(T))\big] \leqslant \mathsf{E}\big[V_{\sigma(\tau_\nu)}(x(\tau_\nu))\big]\, \mathrm{e}^{-\lambda_\circ(T - \tau_\nu)}.$$

Combining with the reset equation, we have

$$\mathsf{E}\big[V_{\sigma(\tau_\nu)}(x(T))\big] \leqslant \mu^\nu V_{\sigma(t_0)}(x_0) \mathrm{e}^{-\lambda_\circ(T - t_0)}.$$

The solution ξ of (3.23) is identical to (2.25). From (3.17) it follows that hypothesis (ii) of Theorem 3.8 is satisfied with $h^\circ(t,x) = h(t,x) = |x|^q$. Further, from (3.18) and (3.19) together with the initial condition in (3.23), it follows that hypotheses (iii) and (iv) of Theorem 3.8 are satisfied, respectively.

Intuitively,

- the initial condition of Σ'' captures the maximum possible overshoot in $\mathsf{E}[V_\sigma]$—this corresponds to the situation when all $N_\circ$ switches occur very close to $t = t_0$;
- $\xi(\cdot)$ forms an envelope of the sequence $\big\{\mathsf{E}\big[V_{\sigma(\tau_i)}(x(\tau_i))\big]\big\}_{i\geqslant 0}$ over the interval $[t_0, T]$.

As in section 2.4, $\lambda_\circ > \frac{\ln\mu}{\tau_a}$ ensures global uniform asymptotic stability of Σ''; this verifies Theorem 3.8 hypothesis (v). It follows that by Theorem 3.8, (3.2) is GUAS-M$_q$ for switching signals with $\tau_a > \frac{\ln\mu}{\lambda_\circ}$ in view of Remark 3.5. □

Remark 3.16. We note that in contrast to (3.9), the comparison systems (3.22) and (3.23) do not utilize state information in the form of the function y directly; cf. Remark 3.12.

Remark 3.17. From the solutions of (3.22) and (3.23) it is clear that for $\tau_a > \frac{\ln\mu}{\lambda_\circ}$, GUAS of Σ' and Σ'' are ensured, which in turn imply that (2.2) is GUAS-M$_q$. In other words, the GUAS-M$_q$ property of (3.2) is "uniform" over all switching signals with $\tau_a > \frac{\ln\mu}{\lambda_\circ}$. We therefore say that (3.2) is *globally "uniformly" asymptotically stable in the qth mean* over all such switching signals; see also Remark 2.15.

Remark 3.18. What we stated in Remark 2.17 for deterministic switched systems carries over to the stochastic case quite easily; it is not difficult to show that in the context of Theorem 3.8, under suitable hypotheses each subsystem may be exponentially unstable while the switched system (3.2) remains (h°, h)-GUAS-M$_q$.

Namely, consider the stochastic switched system (3.2), and suppose that σ is such that $N_\sigma(t, t_0)$ obeys (2.28) for some $N_\circ, \delta_a > 0$. Let there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}^1[\mathbb{R}^n, \mathbb{R}_{\geqslant 0}]$, $p \in \mathcal{P}$, and real numbers $\mu' \in]0,1[$ and $\lambda_\circ > 0$, such that

(i) the estimate (3.17) holds;
(ii) $\forall\, x \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$ and $\forall\, p \in \mathcal{P}$, we have $\mathscr{L}_p V_p(x) \leqslant \lambda_\circ V_p(x)$;
(iii) $\mathsf{E}\big[V_{\sigma(\tau_i)}(x(\tau_i))\big] \leqslant \mu' \mathsf{E}\Big[V_{\sigma(\tau_i^-)}(x(\tau_i^-))\Big] \quad \forall\, i \geqslant 1$;
(iv) $\delta_a < -\ln\mu'/\lambda_\circ$.

Then (3.2) is GUAS-M$_q$.

It may be verified that the comparison system $\widetilde{\Sigma}$ in (2.29) is a suitable comparison system under the above hypotheses, and the assertion follows from Theorem 3.8.

Remark 3.19. We mentioned in Remark 2.18 that for a system with state-dependent switching, in general we cannot conclude stability for more than one initial condition from the associated switched system (2.2). The situation is more complicated in the case of a hybrid system with continuous dynamics perturbed by a Wiener process; now the switching signals σ corresponding to different trajectories are different even for a fixed initial condition, making direct analysis of such a system difficult. The switched system (3.2) provides for simpler stability analysis for a fixed initial condition, but once again we cannot conclude stability of the hybrid system under variations in initial conditions from stability of the switched system. Theorem 3.8 is stated without any claim of uniformity with respect to initial conditions. However, in Theorem 3.15 we obtained uniform stability of a switched system perturbed by a Wiener process over the class of σ defined by a suitable average dwell-time, without taking into account an underlying hybrid system. This class of signals is potentially

useful for hybrid systems where variations in the initial condition preserve average dwell-time switching; see [39] for some results in this direction.

3.5. Remarks on other stability notions. In this subsection we study two notions of stochastic stability of switched systems, different from the ones considered so far, and utilize the framework of Theorem 3.8.

Stochastic practical stability. Stochastic practical stability is concerned with practical stability of systems perturbed by a Wiener process with respect to prespecified domains in the state space. We briefly study a representative notion of stochastic practical stability below; see [40] for a version of the definition for nonswitched systems and with $h^\circ(t,x) = h(t,x) = |x|^q$.

DEFINITION 3.20. *Let $h^\circ, h \in \Gamma$ and the pair (λ, A), $\lambda \in]0, A[$ be given. The stochastic switched system (3.2) is said to be (h°, h)*-uniformly practically stable in the mean with respect to (λ, A) *if for every $t_0 \in \mathbb{R}_{\geqslant 0}$, the property*

$$h^\circ(t_0, x_0) < \lambda \implies \mathsf{E}[h(t, x(t))] < A \qquad \forall\, t \geqslant t_0 \tag{3.24}$$

holds for all solution processes.

The following result provides sufficient conditions for (h°, h)-uniform practical stability of (3.2) in the mean with respect to a given pair (λ, A).

PROPOSITION 3.21. *Consider the stochastic switched system (3.2) with a given σ, $h^\circ, h \in \Gamma$, and let the pair (λ, A), $\lambda \in]0, A[$ be given. Suppose that there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, $V_p \in \mathrm{C}^{1,2}\big[\mathcal{B}(h, A) \cap \big((\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n) \setminus \mathcal{B}(h^\circ, \lambda)\big), \mathbb{R}_{\geqslant 0}\big]$ for each $p \in \mathcal{P}$, and a system Σ of the type (2.6), such that*

(i) *α_1 is convex, and ϕ is concave in the second argument;*

(ii) *the family $\{V_p \mid p \in \mathcal{P}\}$ is $\mathcal{P}$-uniformly h-positive definite and h°-decrescent in the sense of (2.5);*

(iii) *$\forall\, (t, x) \in \mathcal{B}(h, A) \cap \big((\mathbb{R}_{\geqslant 0} \times \mathbb{R}^n) \setminus \mathcal{B}(h^\circ, \lambda)\big)$ and $\forall\, p \in \mathcal{P}$, we have $\mathscr{L}_p V_p(t, x) \leqslant \phi(t, V_p(t, x))$;*

(iv) *$\forall\, (t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, there exists $\xi_0 \in \mathbb{R}_{\geqslant 0}$ such that $\mathsf{E}\big[V_{\sigma(\tau_i)}(\tau_i, x(\tau_i))\big] \leqslant \xi(\tau_i)$ for all $i \geqslant 0$, where $x(t)$ and $\xi(t)$ are the solution process of (3.2) and the solution of Σ, respectively, for these initial conditions;*

(v) *Σ is uniformly practically stable with respect to $(\alpha_2(\lambda), \alpha_1(A))$.*

Then (3.2) is (h°, h)-uniformly practically stable in the mean with respect to (λ, A).

Proof. In view of Remark 2.20, the uniform practical stability of Σ with respect to $(\alpha_2(\lambda), \alpha_1(A))$ implies that for every $t_0 \in \mathbb{R}_{\geqslant 0}$,

$$\xi_0 < \alpha_2(\lambda) \implies \xi(t) < \alpha_1(A) \qquad \forall\, t \geqslant t_0. \tag{3.25}$$

Select and arbitrary x_0 such that $h^\circ(t_0, x_0) < \lambda$, and pick $\xi_0 = \alpha_2 \circ h^\circ(t_0, x_0)$ and an arbitrary nonnegative integer ℓ. From hypothesis (iv) and Lemma 3.7 at time $\tau = \tau_\ell$, we have

$$\mathsf{E}\big[V_{\sigma(\tau_\ell)}(\tau_\ell, x(\tau_\ell))\big] \leqslant \xi(t) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[. \tag{3.26}$$

Combining (3.26) with (3.25) over the time interval $[\tau_\ell, \tau_{\ell+1}[$, and using hypothesis (iii), we have

$$\mathsf{E}[\alpha_1 \circ h(t, x(t))] < \alpha_1(A) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[. \tag{3.27}$$

Since α_1 is convex by hypothesis (i), by Jensen's inequality in (3.27), we reach

$$\alpha_1(\mathsf{E}[h(t, x(t))]) < \alpha_1(A) \qquad \forall\, t \in [\tau_\ell, \tau_{\ell+1}[. \tag{3.28}$$

The inequality (3.28), together with the arbitrariness of ℓ, indicate that (3.24) holds. The (h°, h)-uniform practical stability in the mean of (3.2) with respect to (λ, A) follows. □

Strong global uniform asymptotic stability in probability. We present a stronger version of global asymptotic stability in probability below; see, e.g., [33] for a version of the definition for nonswitched systems and Euclidean norms.

DEFINITION 3.22. *Let $h^\circ, h \in \Gamma$. The stochastic switched system* (3.2) *is said to be* (h°, h)-strongly globally uniformly asymptotically stable in probability ((h°, h)-SGUAS-P) *if for every $\eta \in]0,1[$, there exists a function $\beta \in \mathcal{KL}$ such that for every $(t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$, the inequality*

$$\mathsf{P}\left[\sup_{t \geqslant t_0} h(t, x(t)) \geqslant \beta(h^\circ(t_0, x_0), t - t_0)\right] < \eta \tag{3.29}$$

holds.

LEMMA 3.23. *The (h°, h)-*SGUAS-P *property is equivalent to the simultaneous verification of the following two properties:*

(SP1) *for every $\eta' \in]0,1[$, there exists a function $\delta \in \mathcal{K}_\infty$ such that for every $\varepsilon > 0$ and $t_0 \in \mathbb{R}_{\geqslant 0}$, we have*

$$h^\circ(t_0, x_0) < \delta(\varepsilon) \implies \mathsf{P}\left[\sup_{t \geqslant t_0} h(t, x(t)) \geqslant \varepsilon\right] < \eta'; \tag{3.30}$$

(SP2) *for every $\eta'' \in]0,1[$, $r, \varepsilon' > 0$, there exists a number $\widetilde{T}(r, \varepsilon') > 0$ such that for every $t_0 \in \mathbb{R}_{\geqslant 0}$, we have*

$$h^\circ(t_0, x_0) < r \implies \mathsf{P}\left[\sup_{t \geqslant t_0 + \widetilde{T}(r, \varepsilon')} h(t, x(t)) \geqslant \varepsilon'\right] < \eta''. \tag{3.31}$$

A proof of this result is provided in Appendix B.

It is clear that the (h°, h)-SGUAS-P property is stronger than the (h°, h)-GUAS-P property. However, the same hypotheses as those of Corollary 3.9 ensure this stronger property, as we prove below. In fact, the hypotheses can be slightly weaker—we can do away with the convexity assumption of α_1 in Theorem 3.8.

THEOREM 3.24. *Suppose the hypotheses of Theorem* 3.8 *hold true, with α_1 not necessarily convex. Then* (3.2) *is* (h°, h)-SGUAS-P.

Proof. In view of Lemma 3.23 it suffices to prove (SP1)–(SP2). We first prove (SP1). Let $\eta' \in]0,1[$ and $\varepsilon > 0$ be given. Since Σ is GUAS by hypothesis (v), in view of Remark 2.6 there exists a function $\delta_\xi \in \mathcal{K}_\infty$ such that for every $t_0 \in \mathbb{R}_{\geqslant 0}$, we have

$$\xi_0 < \delta_\xi(\eta' \alpha_1(\varepsilon)) \implies \xi(t) < \eta' \alpha_1(\varepsilon) \qquad \forall\, t \geqslant t_0.$$

Let $\delta(\cdot) := \alpha_2^{-1} \circ \delta_\xi(\eta' \alpha_1(\cdot))$, where we have suppressed the dependence of δ on η', which is implied. Choose an arbitrary x_0 such that $h^\circ(t_0, x_0) < \delta(\varepsilon)$ and let $\xi_0 := \alpha_2 \circ h^\circ(t_0, x_0)$. Then from hypotheses (iii), (iv) and Lemma 3.7, it follows that

$$\mathsf{E}\big[V_{\sigma(t)}(t, x(t))\big] \leqslant \xi(t) < \eta' \alpha_1(\varepsilon) \qquad \forall\, t \geqslant t_0. \tag{3.32}$$

Claim 1. We have $\mathsf{P}[\sup_{t \geqslant t_0} h(t, x(t)) \geqslant \varepsilon] < \eta'$.

Indeed, let τ_ε be the first exit time of $x(t)$ from $\mathcal{B}(h,\varepsilon)$, i.e., $\tau_\varepsilon := \inf\{t \geqslant t_0 \mid h(t,x(t)) \geqslant \varepsilon\} \leqslant \infty$. Therefore, from (3.32), we have

$$\mathsf{E}\big[V_{\sigma(\tau_\varepsilon\wedge t)}(\tau_\varepsilon \wedge t, x(\tau_\varepsilon \wedge t))\big] < \eta'\alpha_1(\varepsilon) \qquad \forall\, t \geqslant t_0. \tag{3.33}$$

Fix an arbitrary $t' > t_0$. We have from (3.33),

$$\eta'\alpha_1(\varepsilon) > \mathsf{E}\big[V_{\sigma(\tau_\varepsilon\wedge t')}(\tau_\varepsilon \wedge t', x(\tau_\varepsilon \wedge t'))\big] \geqslant \mathsf{E}\big[\mathbf{1}_{\{\tau_\varepsilon \leqslant t'\}} V_{\sigma(\tau_\varepsilon)}(\tau_\varepsilon, x(\tau_\varepsilon))\big], \tag{3.34}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. From hypothesis (ii) and the definition of τ_ε, it follows that

$$\mathsf{E}\big[\mathbf{1}_{\{\tau_\varepsilon \leqslant t'\}} V_{\sigma(\tau_\varepsilon)}(\tau_\varepsilon, x(\tau_\varepsilon))\big] \geqslant \mathsf{E}\big[\mathbf{1}_{\{\tau_\varepsilon \leqslant t'\}} \alpha_1 \circ h(\tau_\varepsilon, x(\tau_\varepsilon))\big] = \mathsf{P}[\tau_\varepsilon \leqslant t']\, \alpha_1(\varepsilon). \tag{3.35}$$

Combining (3.34) and (3.35) we get $\eta'\alpha_1(\varepsilon) > \alpha_1(\varepsilon)\mathsf{P}[\tau_\varepsilon \leqslant t']$, and considering the definition of τ_ε, this leads to

$$\mathsf{P}\left[\sup_{t\in[t_0,t']} h(t,x(t)) \geqslant \varepsilon\right] < \eta'.$$

Since $t' > t_0$ is arbitrary, we have

$$\mathsf{P}\left[\sup_{t\geqslant t_0} h(t,x(t)) \geqslant \varepsilon\right] < \eta',$$

whence Claim 1 is verified. The (SP1) property (3.30) of (3.2) follows.

We now sketch the proof of (SP2), which is very similar to the proof of (SP1). Let $\eta'' \in]0,1[$ and $r, \varepsilon' > 0$ be given. Since Σ is GUAS by hypothesis (v), in view of Remark 2.6 there exists a number $T(\alpha_2(r), \eta''\alpha_1(\varepsilon')) \geqslant 0$ such that for every $t_0 \in \mathbb{R}_{\geqslant 0}$ we have

$$\xi_0 < \alpha_2(r) \implies \xi(t) < \eta''\alpha_1(\varepsilon') \qquad \forall\, t \geqslant t_0 + T(\alpha_2(r), \eta''\alpha_1(\varepsilon')).$$

Choose x_0 such that $h^\circ(t_0,x_0) < r$ and let $\xi_0 := \alpha_2 \circ h^\circ(t_0,x_0)$. Then from hypotheses (iii), (iv) and Lemma 3.7, it follows that

$$\mathsf{E}\big[V_{\sigma(t)}(t,x(t))\big] \leqslant \xi(t) < \eta''\alpha_1(\varepsilon') \qquad \forall\, t \geqslant t_0 + T(\alpha_2(r), \eta''\alpha_1(\varepsilon')). \tag{3.36}$$

Claim 2. We have $\mathsf{P}\left[\sup_{t\geqslant t_0 + T(\alpha_2(r),\eta''\alpha_1(\varepsilon'))} h(t,x(t)) \geqslant \varepsilon'\right] < \eta''$.

Indeed, defining $\tau_{\varepsilon'} := \inf\{t \geqslant t_0 + T(\alpha_2(r), \eta''\alpha_1(\varepsilon')) \mid h(t,x(t)) \geqslant \varepsilon'\} \leqslant \infty$, and following the steps of the above proof of (SP1) with τ'_ε replacing τ_ε, for an arbitrary $t'' > t_0 + T(\alpha_2(r), \eta''\alpha_1(\varepsilon'))$, we obtain

$$\eta''\alpha_1(\varepsilon') > \mathsf{E}\big[V_{\sigma(\tau_{\varepsilon'}\wedge t'')}(\tau_{\varepsilon'} \wedge t'', x(\tau_{\varepsilon'} \wedge t''))\big] \geqslant \mathsf{E}\big[\mathbf{1}_{\{\tau_{\varepsilon'} \leqslant t''\}} V_{\sigma(\tau_{\varepsilon'})}(\tau_{\varepsilon'}, x(\tau_{\varepsilon'}))\big].$$

This leads to $\eta''\alpha_1(\varepsilon') > \mathsf{P}[\tau_{\varepsilon'} \leqslant t'']\, \alpha_1(\varepsilon')$. Since t'' is arbitrary, by the definition of $\tau_{\varepsilon'}$ we get

$$\mathsf{P}\left[\sup_{t\geqslant t_0 + \widetilde{T}(r,\varepsilon')} h(t,x(t)) \geqslant \varepsilon'\right] < \eta'',$$

where $\widetilde{T}(r,\varepsilon') := T(\alpha_2(r), \eta''\alpha_1(\varepsilon'))$, suppressing the dependence on η'', which is implied. This verifies Claim 2, and hence the (SP2) property (3.31) of (3.2).

We conclude that (3.2) is (h°, h)-SGUAS-P. □

Since (h°, h)-SGUAS-P implies (h°, h)-GUAS-P, we obtain the (h°, h)-GUAS-P property of (3.2) without the necessity of α_1 being convex; cf. Remark 3.10. See also [23] for a proof of the GUAS-P property of a single stochastic system without the convexity assumption on α_1.

4. Conclusion. We have established a general framework for stability analysis of deterministic and stochastic switched systems. In section 2 we have unified representative existing results on deterministic switched systems and provided illustrations of how we can improve upon the scope of applicability of Lyapunov's second method to switched systems. In section 3 we have established new results on stability of stochastic switched systems. We have carried out analysis in terms of two measures and demonstrated how our framework applies to various stability notions. To conclude, we make the following comments with an eye toward possible future work.

In this paper we have considered disturbances in the form of a Wiener process affecting the states of a switched system, and we have not paid attention to the mechanism of switching signal generation. A common way to generate a switching signal, which has received a lot of attention in recent literature, is via a Markov chain with state space $\mathcal{P}$; see, e.g., [2, 4, 21]. In situations where both Markovian switching and a Wiener process are present, we have to keep in mind two different probability spaces in general: one that generates the disturbances, and the other which governs the Markovian switching. Some results on stability analysis of piecewise deterministic systems with Markovian switching are under development by the authors and will be presented separately.

Although we have utilized only scalar Lyapunov functions in this paper, there can be a parallel development utilizing *vector Lyapunov functions*. For details on this method, see, e.g., [32] and also [29] for an extensive development and discussion on vector Lyapunov functions used in conjunction with analysis using two measures.

A suitably modified version of the comparison framework can be used for systems with inputs also; details for the case of input-to-state stability of switched systems will be reported elsewhere.

Appendix A. Equivalence of definitions of (h°, h)-GUAS.

Proof of the claim in Remark 2.3. *Sufficiency.* Assume that the (S1) property holds. With a fixed $\varepsilon > 0$ we have

$$h^\circ(t_0, x_0) \leqslant \delta(\varepsilon) \implies h(t, x(t)) \leqslant \varepsilon \qquad \forall\, t \geqslant t_0,$$

and the above implication holds uniformly over t_0. Defining the function $\varphi \in \mathcal{K}_\infty$ such that $\varphi(\cdot) := \delta^{-1}(\cdot)$, it follows that

$$h(t, x(t)) \leqslant \varphi \circ h^\circ(t_0, x_0) \qquad \forall\, t \geqslant t_0, \tag{A.1}$$

since $\varepsilon > 0$ is arbitrary.

Now we assume that the (S2) property holds.

A straightforward generalization of [31, Lemma 3.1] leads to the existence of a family of mappings $\{T_r\}_{r>0}$ with the properties

- for every fixed $r > 0$, $T_r \in \mathrm{C}[\mathbb{R}_{>0}, \mathbb{R}_{>0}]$ is surjective and strictly decreasing, and
- for every fixed $\varepsilon > 0$, $T_r(\varepsilon)$ is strictly increasing with r and $\lim_{r\uparrow\infty} T_r(\varepsilon) = \infty$,

such that $h^\circ(t_0, x_0) < r$ implies $h(t, x(t)) < \varepsilon \quad \forall\, t \geqslant t_0 + T_r(\varepsilon)$.

Let $\psi_r(\cdot) := T_r^{-1}(\cdot),\ r \in]0, \infty[$. Then $\psi_r \in \mathrm{C}[\mathbb{R}_{>0}, \mathbb{R}_{>0}]$ is surjective and strictly decreasing. We write $\psi_r(0) = \infty$, considering $\lim_{t \downarrow 0} \psi_r(t) = \infty$.

Claim. For $h^\circ(t_0, x_0) < r$ we have $h(t, x(t)) \leqslant \psi_r(t)$ for $t \geqslant t_0$.

Proof. It follows from the definition of T_r that for an arbitrary $r, \varepsilon > 0$,

$$h^\circ(t_0, x_0) < r \implies h(t, x(t)) < \varepsilon \qquad \forall\, t \geqslant t_0 + T_r(\varepsilon).$$

Since $t - t_0 = T_r \circ \psi_r(t - t_0)$, for $t > t_0$, we have $h(t, x(t)) < \psi_r(t - t_0)$. Combining this and $\psi_r(0) = \infty$, the validity of the claim follows.

For arbitrary $s, t \in \mathbb{R}_{\geqslant 0}$, let

$$\overline{\psi}(s, t) := \min \left\{ \varphi(s), \inf_{r \in\,]s, \infty[} \psi_r(t) \right\}.$$

By definition of φ and the above claim, for every $(t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$,

$$h(t, x(t)) \leqslant \overline{\psi}(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0.$$

The function $\overline{\psi}$ need not be of class $\mathcal{KL}$, so we majorize it as follows. By definition, $\overline{\psi}(\cdot, t)$ is a nondecreasing function and for some fixed s, $\overline{\psi}(s, t) \to 0$ as $t \uparrow \infty$. Define

$$\widetilde{\psi}(s, t) := \int_s^{s+1} \overline{\psi}(v, t) \mathrm{d}v + \frac{s}{(1 + s)(1 + t - t_0)}.$$

It is straightforward to prove that

- $\widetilde{\psi}(s, t)$ is increasing with s for every $t \in \mathbb{R}_{\geqslant 0}$,
- $\widetilde{\psi}(s, t)$ decreases to 0 as $t \uparrow \infty$ for every fixed $s \in \mathbb{R}_{\geqslant 0}$, and
- $\widetilde{\psi}(s, t) \geqslant \overline{\psi}(s, t)$.

Then defining $\beta(s, t) := \sqrt{\varphi(s) \widetilde{\psi}(s, t)}$, it follows that for an arbitrary $(t_0, x_0) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$,

$$h(t, x(t)) \leqslant \sqrt{\big(\varphi \circ h^\circ(t_0, x_0)\big)\, \big(\overline{\psi}(h^\circ(t_0, x_0), t - t_0)\big)} \leqslant \beta(h^\circ(t_0, x_0), t - t_0) \qquad \forall\, t \geqslant t_0.$$

The proof of sufficiency is complete.

Necessity.

- To see (S1), consider the inequality (2.4) at the initial condition. We recover a class $\mathcal{K}_\infty$ function $\beta(\cdot, 0) =: \gamma(\cdot)$ relating h° and h as $h(t, x) \leqslant \gamma \circ h^\circ(t, x)$ for all $(t, x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}^n$. Now for $\varepsilon > 0$, consider $\delta(\cdot) = \gamma^{-1}(\cdot)$.
- To see (S2), consider $r, \varepsilon > 0$ given and the existence of T follows from the property of the class $\mathcal{KL}$ function β.

The proof of necessity is complete. □

Appendix B. Equivalence of definitions of (h°, h)-SGUAS-P.

Proof of Lemma 3.23. First we show that (h°, h)-SGUAS-P implies (SP1)–(SP2).

We note that by the (h°, h)-SGUAS-P property, there exists a function $\beta \in \mathcal{KL}$ such that (3.29) holds. Let $\eta', \eta'' \in]0, 1[$, $r, \varepsilon, \varepsilon' > 0$ and $t_0 \in \mathbb{R}_{\geqslant 0}$ be given. We choose $\eta = \eta' \wedge \eta''$. From the deterministic equivalence claimed in Remark 2.3, we have two numbers $\delta(\varepsilon) > 0$ and $T(r, \varepsilon') \geqslant 0$ such that from the condition $\sup_{t \geqslant t_0} h(t, x(t)) < \beta(h^\circ(t_0, x_0), t - t_0)$ we get

(i) $h^\circ(t_0, x_0) < \delta(\varepsilon) \implies \sup_{t \geqslant t_0} h(t, x(t)) < \varepsilon$, and

(ii) $h^\circ(t_0, x_0) < r \implies \sup_{t \geqslant t_0 + T(r, \varepsilon')} h(t, x(t)) < \varepsilon'$.

Choose $x_0 \in \mathbb{R}^n$ such that $h^\circ(t_0, x_0) < \delta(\varepsilon) \wedge r$. Define the sets

$$\begin{aligned} \Omega'_t &:= \big\{h(t, x(t)) < \varepsilon\big\}, \\ \Omega''_t &:= \big\{h(t, x(t)) < \varepsilon'\big\} \cup \big\{t < t_0 + T(r, \varepsilon')\big\}. \end{aligned} \tag{B.1}$$

Note that the set $\big\{t < t_0 + T(r, \varepsilon')\big\}$ merely ensures that the probability measure of the set $\big\{h(t, x(t)) < \varepsilon'\big\}$ is meaningful only after $t \geqslant t_0 + T(r, \varepsilon')$, which is what we need. In the light of (i)–(ii) and the definitions in (B.1), we know that (3.29) implies that $\mathsf{P}[\cap_{t \geqslant t_0} (\Omega'_t \cap \Omega''_t)] \geqslant 1 - \eta$. But by our choice of η, this means that $\mathsf{P}[\cap_{t \geqslant t_0} (\Omega'_t \cap \Omega''_t)] \geqslant (1-\eta') \vee (1-\eta'')$. Therefore, $\mathsf{P}[\cap_{t \geqslant t_0} \Omega'_t] \geqslant 1-\eta'$ and $\mathsf{P}[\cap_{t \geqslant t_0} \Omega''_t] \geqslant 1 - \eta''$, which are the properties (SP1) and (SP2) in (3.30) and (3.31), respectively.

Now we show that (SP1)–(SP2) implies (h°, h)-SGUAS-P.

Let $\eta \in]0, 1[$, $r, \varepsilon, \varepsilon' > 0$ and $t_0 \in \mathbb{R}_{\geqslant 0}$ be given. We choose $\eta' = \eta'' = \eta/2$. From (SP1)–(SP2), we get two numbers $\delta(\varepsilon) > 0$ and $T(r, \varepsilon') \geqslant 0$. Choose x_0 such that $h^\circ(t_0, x_0) < \delta(\varepsilon) \wedge r$. Using the definitions of Ω'_t and Ω''_t in (B.1), from (3.30) and (3.31) we have $\mathsf{P}[\cap_{t \geqslant t_0} \Omega'_t] \geqslant 1 - \eta/2$, and $\mathsf{P}[\cap_{t \geqslant t_0} \Omega''_t] \geqslant 1 - \eta/2$. We know that for $A, B \subset \Omega$ such that $\mathsf{P}[A] \geqslant 1 - \eta/2$ and $\mathsf{P}[B] \geqslant 1 - \eta/2$, we have $\mathsf{P}[A \cap B] \geqslant 1 - \eta$. (This follows from the simple observation that $1 \geqslant \mathsf{P}[A \cup B] = \mathsf{P}[A] + \mathsf{P}[B] - \mathsf{P}[A \cap B]$.) It follows that $\mathsf{P}[\cap_{t \geqslant t_0} (\Omega'_t \cap \Omega''_t)] \geqslant 1 - \eta$, which is just

$$\mathsf{P}\left[\left\{\sup_{t \geqslant t_0} h(t, x(t)) < \varepsilon\right\} \bigcap \left\{\sup_{t \geqslant t_0 + T(r, \varepsilon')} h(t, x(t)) < \varepsilon'\right\}\right] \geqslant 1 - \eta.$$

In view of Remark 2.3 and (i)–(ii) above, since $r, \varepsilon, \varepsilon' > 0$ were arbitrary, it follows that there exists a function $\beta \in \mathcal{KL}$ such that $\mathsf{P}\big[\sup_{t \geqslant t_0} h(t, x(t)) < \beta(h^\circ(t_0, x_0), t - t_0)\big] \geqslant 1 - \eta$, which is the (h°, h)-SGUAS-P property (3.29). □

We remark that with little alteration of the above proof we obtain the equivalence of the definitions of (h°, h)-GUAS-P claimed in Remark 3.4. Indeed, for arbitrary fixed $t > t_0$, we need to eliminate the intersections of the sets Ω'_t and Ω''_t over $t \geqslant t_0$ in the equations above to establish the equivalence for time t. Since t is an arbitrary choice, we can conclude the validity of the result for every $t \geqslant t_0$.

Acknowledgments. We are indebted to S. P. Meyn for sharing with us his vision and discernment on stochastic stability. We thank A. D. Teel and V. Lakshmikantham for providing valuable insights on stability analysis in terms of two measures at various junctures.

REFERENCES

[1] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Krieger, Melbourne, FL, 1992.

[2] S. Battilotti and A. D. Santis, *Dwell time controllers for stochastic systems with switching Markov chain*, Automatica, 41 (2005), pp. 923–934.

[3] S. P. Bhat and D. S. Bernstein, *Finite-time stability of continuous autonomous systems*, SIAM J. Control Optim., 38 (2000), pp. 751–766.

[4] P. Bolzern, P. Colaneri, and G. D. Nicolao, *On almost sure stability of discrete-time Markov jump linear systems*, in Proceedings of the 43rd Conference on Decision and Control, 2004, pp. 3204–3208.

[5] S. Boyd, L. E. Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[6] M. S. Branicky, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.

[7] R. W. Brockett, *Lecture Notes on Stochastic Control*, Harvard University, Cambridge, MA, 1995.
[8] D. Chatterjee and D. Liberzon, *On stability of stochastic switched systems*, in Proceedings of 43rd Conference on Decision and Control, vol. 4, 2004, pp. 4125–4127.
[9] D. Chatterjee, *Stability Analysis of Deterministic and Stochastic Switched Systems via a Comparison Principle and Multiple Lyapunov Functions*, Master's thesis, University of Illinois, Urbana, IL, 2004.
[10] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Appl. Math. 5, 2nd ed., SIAM, Philadelphia, 1990.
[11] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.
[12] M. K. Ghosh, A. Arapostathis, and S. I. Marcus, *Ergodic control of switching diffusions*, SIAM J. Control Optim., 35 (1997), pp. 1952–1988.
[13] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
[14] R. Z. Haśminskii, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Groningen, The Netherlands, 1980.
[15] J. P. Hespanha, D. Liberzon, and A. R. Teel, *On input-to-state stability of impulsive systems*, in Proceedings of the 44th Conference on Decision and Control, 2005, pp. 3992–3997.
[16] J. P. Hespanha and A. S. Morse, *Stability of switched systems with average dwell-time*, in Proceedings of the 38th IEEE Conference on Decision and Control, vol. 3, 1999, pp. 2655–2660.
[17] J. P. Hespanha, *A model for stochastic hybrid systems with application to communication networks*, Nonlinear Analysis Special Issue on Hybrid Systems, 62 (2005), pp. 1353–1383.
[18] J. P. Hespanha, *Uniform stability of switched linear systems: Extensions of LaSalle's invariance principle*, IEEE Trans. Automat. Control, 49 (2004), pp. 470–482.
[19] L. Hou, A. N. Michel, and H. Ye, *Stability analysis of switched systems*, in Proceedings of the 35th IEEE Conference on Decision and Control, vol. 2, 1996, pp. 1208–1212.
[20] J. Hu, J. Lygeros, and S. Sastry, *Towards a theory of stochastic hybrid systems*, Lecture Notes in Comput. Sci. 1790, Springer-Verlag, 2000, pp. 160–173.
[21] Y. Ji and H. J. Chizeck, *Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[22] H. K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 2002.
[23] M. Krstić and H. Deng, *Stabilization of Nonlinear Uncertain Systems*, Springer-Verlag, Berlin, 1998.
[24] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice–Hall, Englewood Cliffs, 1985.
[25] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[26] G. S. Ladde and V. Lakshmikantham, *Random Differential Inequalities*, Academic Press, New York, 1980.
[27] V. Lakshmikantham, S. Leela, and A. A. Martynyuk, *Practical Stability of Nonlinear Systems*, World Scientific, Singapore, 1990.
[28] V. Lakshmikantham and S. Leela, *Differential and Integral Inequalities: Theory and Application*, Vol. 1, Academic Press, New York, 1969.
[29] V. Lakshmikantham and X. Liu, *Stability Analysis in Terms of Two Measures*, World Scientific, Singapore, 1995.
[30] D. Liberzon, *Switching in Systems and Control*, Birkhäuser, Boston, 2003.
[31] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.
[32] V. M. Matrosov, *Vector Lyapunov functions method in the analysis of dynamical properties of nonlinear differential equations*, Trends in Theory and Practice of Nonlinear Differential Equations, Lecture Notes in Pure and Appl. Math. 90, Marcel Dekker, New York, 1984, pp. 357–374.
[33] Y. N. Merenkov, *Stability-like properties of stochastic differential equations*, Differential Equations, 39 (2003), pp. 1703–1712.
[34] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, Berlin, 1993.
[35] A. A. Movchan, *Stability of processes with respect to two metrics*, Prikl. Mat. Mekh., 24 (1960), pp. 988–1001.
[36] B. K. Øksendal, *Stochastic Differential Equations*, 5th ed., Springer-Verlag, Berlin, 1998.
[37] P. Peleties and R. A. DeCarlo, *Asymptotic stability of m-switched systems using Lyapunov-like functions*, in Proceedings of the American Control Conference, 1991, pp. 1679–1684.

[38] S. Pettersson and B. Lennartson, *Stability and robustness for hybrid systems*, in Proceedings of the 35th IEEE Conference on Decision and Control, vol. 2, 1996, pp. 1202–1207.
[39] M. Prandini, J. P. Hespanha, and M. Campi, *Hysteresis-based switching control of stochastic linear systems*, in Proceedings of the 2003 European Control Conference, 2003.
[40] S. Sathananthan and L. H. Keel, *Optimal practical stabilization and controllability of systems with Markovian jumps*, Nonlinear Anal., 54 (2003), pp. 1011–1027.
[41] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed., Springer-Verlag, Berlin, 1998.
[42] J. Sun, Y. Zhang, and Q. Wu, *Less conservative conditions for asymptotic stability of impulsive control systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 829–831.
[43] A. D. Teel and L. Praly, *A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions*, ESAIM. Control Optim. Calculus Variations, 5 (2000), pp. 313–367.
[44] A. van der Schaft and H. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Springer-Verlag, Berlin, 1999.
[45] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, 2nd ed., Springer-Verlag, Berlin, 1985.
[46] M. Zakai, *A Lyapunov criterion for the existence of stationary probability distributions for systems perturbed by noise*, SIAM J. Control, 7 (1969), pp. 390–397.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 207–225

RECEDING HORIZON CONTROL WITH INCOMPLETE OBSERVATIONS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. To overcome the difficulties related to the computational requirements for solving the optimality systems for optimal control problems on long time horizons, receding horizon techniques provide an important alternative. Rather than finding the optimal solution, a suboptimal control is obtained which achieves the design objective with significantly less computational effort. Moreover, the obtained control can be interpreted as a state feedback control. In this work we continue our analysis of receding horizon strategies, considering the situation when only partial state observations are available. The receding horizon strategy is combined with a state estimator framework. A linear quadratic Gaussian design based on a linearization procedure is proposed and its asymptotic performance is analyzed for systems with nonlinear dynamics. Numerical examples validate the proposed methodology.

Key words. receding horizon control, incomplete observations, dynamic compensator

AMS subject classifications. 49N35, 93C41

DOI. 10.1137/S0363012903437988

1. Introduction.

We consider the optimal control problem of minimizing the performance index subject to a nonlinear control system:

$$\min \int_0^{T_\infty} f^0(x(t), u(t))\, dt, \tag{1.1}$$

where

$$\frac{d}{dt}x(t) = f(x(t)) + Bu(t) \quad \text{for} \quad t > 0, \tag{1.2}$$

$$x(0) = x_0, \tag{1.3}$$

for some $T_\infty \in (0, \infty]$ and $x_0 \in R^n$. In order to present our approach without the technical difficulties associated to infinite dimensional control systems, we assume here that the state space is finite dimensional. We refer to $x(\cdot)$ and $u(\cdot)$ as state and control functions with $x(t) \in R^n$ and $u(t) \in R^m$. Under appropriate conditions, (1.1)–(1.3) admit a solution which satisfies the minimum principle

$$\begin{cases} \frac{d}{dt}x(t) = H_p(x(t), u(t), p(t)),\ x(0) = x_0, \\ \frac{d}{dt}p(t) = -H_x(x(t), u(t), p(t)),\ p(T_\infty) = 0, \\ u(t) = \arg\min_{u \in R^m}\ H(x(t), u, p(t)), \end{cases} \tag{1.4}$$

*Received by the editors November 10, 2003; accepted for publication (in revised form) August 30, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sicon/45-1/43798.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (kito@math.ncsu.edu).
‡Institut für Mathematik, Karl-Franzens-Universität Graz, A-8010 Graz, Austria (karl.kunisch@uni-graz.at). This author was supported by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03 "Optimierung und Kontrolle."

where H is the Hamiltonian defined by $H(x,u,p) = f^0(x,u) + p \cdot (f(x) + Bu)$. The coupled system of two-point boundary value problems with initial condition for the primal equation and terminal condition for the adjoint equation represents a significant challenge for numerical computations in case the dimension n of the state or the time horizon T_∞ are large. It has therefore been the focus of many research efforts. An alternative approach consists in constructing the feedback solution based on Bellman's dynamic programming principle. Again, due to computational costs, this is tractable only for very limited examples.

One of the possibilities to overcome these difficulties is given by time-domain decomposition based on receding horizon formulations [ABQRW, GPM]. Receding horizon techniques have proved to be effective numerically both for optimal control problems governed by ordinary (e.g., [CA, JYH, K, MM, PND, SMR]) and by partial differential equations, e.g., in the form of the instantaneous control technique for problems in fluid mechanics [B, CHK, CTMC, HV].

To briefly explain the strategy let $0 = T_0 < T_1 \cdots < T_\infty$ describe a grid on $[0, T_\infty)$ and let $T \geq \max\{T_{i+1} - T_i : i = 0, \dots\}$. If $T > T_{i+1} - T_i$, we have overlapping domains. The receding horizon optimal control problem involves the successive finite horizon optimal control problems on $[T_i, T_i + T]$,

$$\min \int_{T_i}^{T_i+T} f^0(x(t), u(t))\, dt + G(x(T_i + T)), \tag{1.5}$$

subject to

$$\frac{d}{dt}x(t) = f(x(t)) + Bu(t),\ t \in [T_i, T_i + T], \tag{1.6}$$

$$x(T_i) = x^*(T_i) \text{ if } i \geq 1 \text{ and } x(0) = x_0 \text{ for } i = 0, \tag{1.7}$$

where x^* is the solution to the auxiliary problem on $[T_{i-1}, T_{i-1} + T]$.

The solution on $[0, T_\infty)$ is obtained by concatenation of the solutions on $[T_i, T_{i+1}]$ for $i = 0, \dots$. If the terminal cost G is chosen as a control Liapunov function, then the asymptotic stability and the performance estimate of the receding horizon synthesis are established in [IK1] in case (1.2) is posed in a finite dimensional state space, and in [IK2] for infinite dimensional state spaces. For example, if G is selected as the optimal value function for (1.1)–(1.3) on the infinite horizon, then the optimal pair (x^*, u^*) for the receding horizon control (1.5)–(1.8) is optimal for the original problem (1.1)–(1.3) on the infinite horizon. In general, G plays the role of a look-ahead term in the sense that it stabilizes the receding horizon synthesis and provides a good suboptimal control law for the given performance (1.1). Since it is impractical to assume that the optimal value functional is known we constructed in [IK1] and [IK2] control Liapunov functions for diverse classes of problems which are based on Liapunov or Riccati equations. For the numerical example of section 5 we select G as the solution to an associated Riccati equation.

If $x(T_i)$ is observed, then the receding horizon control technique is a state feedback method since the control on $[T_i, T_{i+1}]$ is determined as a function of the state $x^*(T_i)$. If the optimal pair $(x^*(t), u^*(t)),\ t \in [T_i, T_i + T]$ is shifted by $-T_i$ it satisfies the two-point boundary value problem (1.4) on the interval $[0, T]$ with the terminal condition $p(T) = G_x(x(T))$ and initial condition $x(0) = x^*(T_i)$.

In this paper we address the state estimator problem for the receding horizon technique. We shall also allow for additive noise in the system dynamics as well as

in the observation process and we admit uncertainty in the initial condition. The nonlinear time-independent control system with additive unmodeled disturbance is given by

$$\begin{cases} \frac{d}{dt}x(t) = f(x(t)) + Bu(t) + d(t) \ \text{ for } \ t > 0, \\ x(0) = x_0, \end{cases} \tag{1.8}$$

where $d(t)$ is an unknown disturbance process. The observation process providing partial observations $y(t) \in R^p$ of the state $x(t)$ is assumed to be of the form

$$y(t) = Cx(t) + n(t), \tag{1.9}$$

where $C \in \mathbb{R}^{p\times n}$ and $n(t)$ is a measurement noise process. The output feedback law will utilize the open loop optimal control u^* with associated optimal state x^* on the interval $[T_i, T_{i+1}]$ computed from (1.5) and (1.6). The initial condition for the computation of x^* is taken to be

$$x(0) = x_0 + \eta_0,$$

where η_0 denotes the uncertainty in the initial condition for $i = 0$, and then it is chosen as the state of the estimator, which will be introduced below, at time T_i, if $i \geq 1$. This can be considered as a preprocessing step for the construction of the feedback law, and in view of the fact that receding time horizon T is considered to be small compared to T_∞ it is considered to be comparatively cheap.

Once x^* is computed on $[T_i, T_i + T]$, the output feedback law is chosen to be of the form

$$u(t) = u^*(t) - B^T\Pi(w(t) - x^*(t)), \quad t \in [T_i, T_{i+1}), \tag{1.10}$$

where w denotes the state of the compensator. The construction of the feedback gain $B^T\Pi$ will be specified below. Suggested from linear quadratic Gaussian (LQG) design, the compensator is based on (1.10) together with the state estimator dynamics of the form

$$\frac{dw}{dt} = f'(\bar{x})(w(t) - x^*(t)) + f(x^*(t)) + Bu(t) + \Sigma C^T(y(t) - Cw(t)), \tag{1.11}$$

$$w(T_i) = w(T_i^-) \text{ for } i = 1, 2, \ldots, \text{ and } w(0) = x_0 + \eta_0, \tag{1.12}$$

for the state estimator $w(t)$, where $w(T_i^-)$ denotes the value of the estimator on $[T_{i-1}, T_{i-1} + T]$ at T_i and $f'(\bar{x})$ is the Jacobian of f at a reference state $\bar{x}$. This reference state $\bar{x}$ is selected on the basis of the optimal pair (x^*, u^*) on $[T_i, T_i + T]$, for example,

$$\bar{x} = \frac{1}{T}\int_{T_i}^{T_i+T} x^*(t)\,dt \quad \text{or} \quad \bar{x} = x^*(T_i + T). \tag{1.13}$$

The feedback synthesis (1.10)–(1.12) performs tracking of (1.8) to the optimal pair (x^*, u^*) on $[T_i, T_i+T]$ under the uncertainty of the initial condition, observation noise and an additive disturbance in the system dynamics. As described in section 2, our tracking method is based on the linearization of the nonlinearity f at the reference

state $\bar{x}$. The control gain $B^T\Pi$ and the filter gain ΣC^T will be constructed by two LQG Riccati equations, which we specify in section 2.

For our stability analysis of the output feedback synthesis (1.11)–(1.13) it is not necessary that (x^*, u^*) is assumed to be the optimal pair for the receding horizon control on $[T_i, T_i + T]$. It is only required that x^* is a stable trajectory of (1.2) corresponding to a given control u^*. However, if G is selected as a control Liapunov function, then the global stability and performance estimate of the receding control are established in [IK1]. Thus, we assume (x^*, u^*) is the optimal pair for the receding horizon control on each horizon $[T_i, T_i+T]$ and the corresponding (1.11)–(1.13) defines our proposed output feedback synthesis for (1.1)–(1.3).

The stability analysis and performance estimates for the proposed procedure are given in section 2. In section 3 the asymptotic behavior of the overall closed-loop system (1.8), with feedback control given by (1.10) and (1.11), is discussed. Our analysis uses the Liapunov stability arguments. In section 4 we briefly address a modification of the proposed method based on an H^∞-synthesis. Numerical examples that illustrate the feasibility of combining the receding horizon strategy with state estimator dynamics are given in section 5, by means of stabilization and tracking problems for the Burgers equation.

The controlled Burgers equation is an infinite dimensional system which, after discretization, is of the form considered in (1.2). In fact, while our analysis is carried out for finite dimensional systems, the proposed concepts can be extended to infinite dimensional systems. The details, however, require further research.

2. LQG design. In this section we describe the construction of the feedback and filters gains for (1.10) and (1.11). Subsequently we establish the stability and the performance estimate for the compensator dynamics (1.10), (1.11) based on the LQG design on a single time horizon $[0, T]$. The iterative procedure on the sequence of time horizons $[T_i, T_{i+1}]$ will be considered in section 3. The Jacobian of f at $x \in \mathbb{R}^n$ will be denoted by $A(x) = f'(x)$ and in particular we set

$$A = f'(\bar{x}).$$

We assume that

$$(A, B, C) \text{ is stabilizable and detectable.} \tag{2.1}$$

Consider the controlled linear equation, which results from linearizing f at $x^*(t)$ and subsequently replacing $x^*(t)$ by $\bar{x}$:

$$\begin{cases} \frac{d}{dt}(\hat{x}(t) - x^*(t)) = A(\hat{x}(t) - x^*(t)) + B(u(t) - u^*(t)) + d(t), \\ \hat{x}(0) = x_0 + \eta_0, \end{cases} \tag{2.2}$$

where $d(t)$ is the disturbance process. We shall further comment on the choice of the linearized equation in Remark 2.2. Let $Q \in \mathbb{R}^{n\times n}$ denote a positive definite matrix and consider the tracking problem to the pair (x^*, u^*):

$$\begin{cases} \min \frac{1}{2}\int_0^\infty ((\hat{x}(t) - x^*(t))^T Q(\hat{x}(t) - x^*(t)) + |u(t) - u^*(t)|^2)dt \\ \text{subject to (2.2) over } \ u \in L^2(0, \infty; \mathbb{R}^m). \end{cases} \tag{2.3}$$

Throughout this paper we use $|x|^2 = (x, x)$ to denote the square of the Euclidean norm of a vector x and $\|A\|$ for the subordinate matrix norm. Here (x, y) stands

for the dot product of vectors x, y which will at times also be denotes by $x^T y$. Let $V\colon \mathbb{R}^n \to \mathbb{R}^+$ denote the value functional associated to (2.3). It can be shown that

$$V(x) = \frac{1}{2}(x - x^*)^T \Pi (x - x^*) + v^T (x - x^*) + c,$$

where c is a constant, and the symmetric positive definite matrix $\Pi \in R^{n\times n}$ and $v \in R^n$ satisfy

$$\begin{cases} A^T \Pi + \Pi A - \Pi B B^T \Pi + Q = 0, \\ (A - B B^T \Pi)^T v + \Pi d = 0, \end{cases} \tag{2.4}$$

where we assume that the disturbance is constant in time. The optimal feedback loop control for (2.3) is given by

$$u = u^*(t) - B^T(\Pi(\hat{x}(t) - x^*(t)) + v). \tag{2.5}$$

Since we consider unknown, unmodeled disturbance $d = d(t)$ we do not include the feedforward input v in the feedback form (1.10).

Turning to the estimator we use the (Kalman) filter gain ΣC^T based on the linear system (A, C), where $\Sigma \in \mathbb{R}^{n\times n}$ is the positive definite solution to

$$A\Sigma + \Sigma A^T - \Sigma C^T C \Sigma + R = 0, \tag{2.6}$$

with $R \in R^{n\times n}$ a positive definite matrix. This results in the following equations for the compensator

$$\frac{dw}{dt} = A(\bar{x})(w(t) - x^*(t)) + f(x^*(t)) + Bu(t) + \Sigma C^T (y(t) - Cw(t)), \tag{2.7}$$

$$w(0) = x_0 + \eta_0 \tag{2.8}$$

and the associated feedback law

$$u(t) = u^*(t) - B^T \Pi (w(t) - x^*(t)). \tag{2.9}$$

Note that (2.1) guarantees the existence of Π and Σ with the specified properties. Moreover the spectra of $A - B B^T \Pi$ and $A - \Sigma C^T C$ are strictly contained in the left half of the complex plane. We henceforth assume that

$$d \in L^1_{loc}(0, \infty; R^n), \text{ and } n \in L^1_{loc}(0, \infty; R^p).$$

We further assume the existence of solutions x and w to (1.8) and (2.7), (2.8) on $[0, T]$ for every $u \in L^1(0, T; R^m)$, where y is given by (1.9).

PROPOSITION 2.1. *Let $W(x) = \frac{1}{2} x^T \Pi x$. For $t \in [0, T]$ we have*

$$\begin{aligned} \frac{d}{dt} W(x(t) - x^*(t)) = &-\frac{1}{2}(|B^T \Pi (x(t) - x^*(t))|^2 + (Q(x(t) - x^*(t)), x(t) - x^*(t))) \\ &+ (d(t) + r(x(t), x^*(t)), \Pi(x(t) - x^*(t))) + (B^T \Pi (x(t) - w(t)), B^T \Pi (x(t) - x^*(t))), \end{aligned}$$

where

$$r(x, x^*) = f(x) - f(x^*) - A(\bar{x})(x - x^*),$$

x is the solution to (1.8) *with u given in* (2.9)*, and x^* is the solution to* (1.6) *with the optimal open loop control u^*.*

Proof. Using (1.6) and (1.8) we have

$$\frac{d}{dt}(x(t)-x^*(t)) = A(x(t)-x^*(t)) - BB^T\Pi(x(t)-x^*(t))$$

$$+ BB^T\Pi(x(t)-w(t)) + d(t) + r(x(t),x^*(t)).$$

Thus,

$$\frac{d}{dt}W(x(t)-x^*(t)) = ((A-BB^T\Pi)(x(t)-x^*(t)) + BB^T\Pi(x(t)-w(t))$$

$$+ d(t) + r(x(t),x^*(t)), \Pi(x(t)-x^*(t))$$

and the proposition follows from the fact that

$$((A-BB^T\Pi)x,\Pi x) = -\frac{1}{2}\,(|B^T\Pi x|^2 + (Qx,x)). \qquad \square$$

PROPOSITION 2.2. *Let $\tilde{W}(z) = \frac{1}{2}\, z^T\Sigma^{-1}z$. For $t\in[0,T]$ we have*

$$\frac{d}{dt}\tilde{W}(x(t)-w(t)) = -\frac{1}{2}(|C(x(t)-w(t))|^2 + (R\Sigma^{-1}(x(t)-w(t)),\Sigma^{-1}(x(t)-w(t)))$$

$$+ (r(x(t),x^*(t)) + d(t), \Sigma^{-1}(x(t)-w(t))) - (n(t), C(x(t)-w(t)),$$

where x and w are the solutions to (1.8) *and* (2.7) *for some $u\in L^2(0,T;R^m)$.*

Proof. From (1.8) and (2.7) we have

$$\frac{d}{dt}(x(t)-w(t)) = A(x(t)-w(t)) + r(x(t),x^*(t)) + d(t) - \Sigma\, C^T(C(x(t)-w(t)) + n(t)),$$

and hence

$$\frac{d}{dt}\tilde{W}(x(t)-w(t)) = ((A-\Sigma C^TC)(x(t)-w(t))$$

$$+ r(x(t),x^*) + d(t) - \Sigma\, C^Tn(t), \Sigma^{-1}(x(t)-w(t))).$$

Thus the proposition follows from the fact that

$$((A-\Sigma C^TC)z,\Sigma^{-1}z) = -\frac{1}{2}\,(|Cz|^2 + (R\Sigma^{-1}z,\Sigma^{-1}z)). \qquad \square$$

To quantify the performance of the compensator we set

$$E(t) = (x(t)-x^*(t),\Pi(x(t)-x^*(t)))^{1/2} + (x(t)-w(t),\Sigma^{-1}(x(t)-w(t)))^{1/2}.$$

Note that $E(t) = (W(x(t)-x^*(t)) + \tilde{W}(x(t)-w(t)))^{1/2}$. We further introduce positive constants $\alpha_1,\ \alpha_2$ and $\beta_1,\ \beta_2$ such that

$$\alpha_1\, I \le \Pi \le \alpha_2\, I, \qquad \beta_1\, I \le \Sigma^{-1} \le \beta_2\, I, \tag{2.10}$$

where I is the identity matrix on R^n. We shall require the following assumptions: there exists $\omega > 0$ such that

$$\begin{aligned}\frac{1}{2}(|B^T\Pi x|^2 + (Qx,x) + |Cz|^2 + (R\Sigma^{-1}z, \Sigma^{-1}z))\\ - (B^T\Pi\Sigma\,\Sigma^{-1}z, B^T\Pi x) \geq \omega\,(W(x) + \tilde{W}(z)) \text{ for all } x,z \in R^n,\end{aligned} \tag{2.11}$$

and

$$E(0) \leq \frac{\delta}{2} \text{ and } |x^*(t) - \bar{x}| \leq \delta \text{ on } [0,T] \text{ for some } \delta > 0. \tag{2.12}$$

Turning to the effect of the nonlinearity we observe that

$$r(x,x^*) = f(x) - f(x^*) - A(x^*)(x - x^*) + (A(x^*) - A(\bar{x}))(x - x^*).$$

We assume that there exist a constant L such that

$$|r(x,x^*)| \leq L\,(|x^* - \bar{x}| + |x - x^*|)\,|x - x^*| \text{ for all } x, x^* \in \mathcal{S}, \tag{2.13}$$

where $\mathcal{S} = \{x \in R^n : |x - \bar{x}| \leq \delta(1 + \frac{1}{\sqrt{\alpha_1}})\}$. We shall further assume that

$$\tilde{\omega} = \omega - L\delta\,\frac{\sqrt{\alpha_1} + 1}{\alpha_1}\,\frac{\sqrt{\alpha_2} + \sqrt{\alpha_2 + \beta_2}}{2} > 0. \tag{2.14}$$

This requirement is trivially satisfied if f is linear. Let us define

$$\rho(t) = \sqrt{\alpha_2 + \beta_2}\,|d(t)| + \frac{1}{\sqrt{\beta_1}}\,\|C\||n(t)|.$$

We require the following smallness condition on the noise processes:

$$\int_0^t e^{-\tilde{\omega}(t-s)}\rho(s)\,ds < \frac{\delta}{2} \quad \text{for all } t \in [0,T]. \tag{2.15}$$

THEOREM 2.1. *If* (2.12) *and the stability conditions* (2.11), (2.13), (2.14) *and the smallness condition on the noise processes* (2.15) *are satisfied, then*

$$E(t) \leq e^{-\tilde{\omega}t}\,E(0) + \int_0^t e^{-\tilde{\omega}(t-s)}\rho(s)\,ds \tag{2.16}$$

holds for all $t \in [0,T]$.

Proof. Due to (2.12) there exists τ such that

$$E(\tau) \leq \delta \text{ on } [0,\tau].$$

Note that $\sqrt{\alpha_1}|x(t) - x^*(t)| \leq E(t)$ and therefore $x(t) \in \mathcal{S}$ for $t \in [0,\tau]$. Set $X(t) = |x(t) - x^*(t)|_\Pi$ and $Y(t) = |x(t) - w(t)|_{\Sigma^{-1}}$. Suppressing the dependence on t we obtain by (2.12), (2.13) that

$$\begin{aligned}&(r(x,x^*), \Pi(x - x^*)) + (r(x,x^*), \Sigma^{-1}(x - w))\\ &\leq L\delta\,\frac{\sqrt{\alpha_1} + 1}{\alpha_1}(\sqrt{\alpha_2}X^2 + \sqrt{\beta_2}XY)\\ &\leq L\delta\,\frac{\sqrt{\alpha_1} + 1}{\alpha_1}\,\frac{\sqrt{\alpha_2} + \sqrt{\alpha_2 + \beta_2}}{2}\,(X^2 + Y^2) \text{ for } t \in [0,\tau].\end{aligned}$$

Since

$$E(t)\frac{d}{dt}E(t) = \frac{d}{dt}(W(x(t)-x^*(t)) + \tilde{W}(x(t)-w(t))),$$

we find by Proposition 2.1, 2.2, with u given in (2.9), and (2.11), (2.13), (2.14) that

$$\frac{d}{dt}(W(x-x^*) + \tilde{W}(x-w))$$

$$\leq -\omega(W(x-x^*) + \tilde{W}(x-w)) + (d + r(x,x^*), \Pi(x-x^*) + \Sigma^{-1}(x-w)) - (n, C(x-w))$$

$$\leq -\tilde{\omega}(W(x-x^*) + \tilde{W}(x-w)) + (d, \Pi(x-x^*)) + (d, \Sigma^{-1}(x-w)) - (n, C(x-w))$$

$$\leq -\tilde{\omega}(W(x-x^*) + \tilde{W}(x-w)) + \rho(t)(W(x-x^*) + \tilde{W}(x-w))^{1/2}.$$

Consequently

$$\frac{d}{dt}E(t) \leq -\tilde{\omega}E(t) + \rho(t).$$

This implies (2.16) on $[0,\tau]$.

By (2.15) and continuity of $t \to \int_0^t e^{-\omega(t-s)}\rho(s)ds$ there exists $\alpha \in (0,1)$ such that

$$\int_0^t e^{-\tilde{\omega}(t-s)}\rho(s)ds \leq \frac{\alpha\delta}{2} \text{ for all } t \in [0,T],$$

and consequently

$$E(t) < \frac{\delta}{2}(1+\alpha) \text{ for all } t \in [0,\tau]. \tag{2.17}$$

A continuation argument implies that (2.16) and (2.17) hold on $[0,T]$. In fact, if (2.17) holds on $[0,T]$, then (2.16) can be continued from $[0,\tau]$ to $[0,T]$. Assuming that (2.17) is not valid on $[0,T]$, let $\tilde{\tau}$ denote the smallest value in $(0,T)$ such that $E(\tilde{\tau}) = \frac{\delta}{2}(1+\alpha)$. Then repeating the argument leading to (2.16) it can be shown that there exists $\epsilon > 0$ such that (2.16) holds in $[0,\tilde{\tau}+\epsilon]$. This implies $E(\tilde{\tau}+\epsilon) < \frac{\delta}{2}(1+\alpha)$, which is a contradiction. Thus (2.17) holds on $[0,T]$. □

Remark 2.1. Concerning assumption (2.12), we note that the indefinite term $-(B^T\Pi\Sigma\Sigma^{-1}, B^T\Pi x)$ can be estimated by

$$-(B^T\Pi\Sigma\Sigma^{-1}z, B^T\Pi x) \geq -\frac{1}{2\alpha}|B^T\Pi\Sigma\Sigma^{-1}z|^2 - \frac{\alpha}{2}|B^T\Pi x|^2$$

and the terms on the right-hand side can be combined with (Qx,x) and $(R\Sigma^{-1}z, \Sigma^{-1}z)$, and $\alpha > 0$ appropriately chosen, to check for (2.12). However, since Π and Σ depend nonlinearly on Q and R, respectively, we must check the validity of (2.12) for any given system. We have done so numerically for our test examples.

Remark 2.2. The linear equation (2.2) as a basis for the construction of the compensator is well suited for our purposes. Other choices are briefly indicated.

1. Linearizing (1.8) at $\bar{x}$ results in

$$\frac{d}{dt}(\hat{x}(t)-\bar{x}) = f(\bar{x}) + A(\hat{x}(t)-\bar{x}) + B(u(t)-u^*(t)) + d(t).$$

The resulting Riccati synthesis is of the form (2.4) with d replaced by $d+f(\bar{x})$. Since d is unknown it cannot be used for the construction of the estimator, and $f(\bar{x})$ remains as a known bias term. It necessitates to modify W to be $W(x) = \frac{1}{2}x^T\Pi x + x^T v$, with the bias v changing from one horizon to the next.

2. It is also possible to employ the time varying linearization

$$\begin{cases} \frac{d}{dt}(\hat{x}(t) - x^*(t)) = A(t)(\hat{x}(t) - x^*(t)) + B(u - u^*(t)) + \tilde{n}(t), \\ \\ \hat{x} = x_0, \end{cases} \tag{2.18}$$

where $A(t) = A(x^*(t))$ and use the corresponding time-varying gains $B^T\Pi(t)$ and $\Sigma(t)C^T$ determined by

$$\frac{d}{dt}\Pi(t) + A(t)^T\Pi(t) + \Pi(t)A(t) - \Pi(t)BB^T\Pi(t) + Q = 0$$

and

$$\frac{d}{dt}\Sigma(t) - A(t)\Sigma(t) - \Sigma(t)A(t)^T + \Sigma(t)BB^T\Sigma(t) - R = 0.$$

One can adapt our analysis and establish an error estimate analogous to (2.16) assuming that $\omega > 0$ and L are independent of $t \in [0, T]$.

3. Asymptotic performance of a closed-loop system. We will apply Theorem 2.1 repeatedly on the intervals $[T_i, T_i + T]$. Let us briefly recall the procedure. The open loop solution x^* to (1.5)–(1.7) is computed on $[T_i, T_{i+1}]$, and based on it $\bar{x}$ is determined; see (1.13). To refer to a specific horizon we henceforth use $\bar{x}_{T_i}$ for $\bar{x}$. This determines $A(\bar{x}_{T_i})$ and allows us to compute $\Pi = \Pi(\bar{x}_{T_i})$ and $\Sigma = \Sigma(\bar{x}_{T_i})$ as solutions to the corresponding Riccati equations. The compensator can then be defined on the basis of (1.10)–(1.12). To simplify the following discussion we assume that $T_{i+1} - T_i = T$ for all i. We shall assume that

$$E(0) \le \frac{\delta}{2} \text{ and } |x^*(t) - \bar{x}| \le \delta \text{ on } [T_i, T_{i+1}] \text{ for all } i = 0, \dots \tag{3.1}$$

and that (2.10) and (2.11) hold uniformly on all horizons $[T_i, T_{i+1}]$. In view of the continuity of $x \to A(x)$, we have continuity of $x \to \Pi(x)$ and $x \to \Sigma(x)$. Due to these properties and the fact that the open loop control x^* is typically guaranteed to be continuous and bounded on $[0, \infty)$ (see, e.g., [IK1]), assumptions (2.10), (2.11), and (3.1) are natural ones. As $\bar{x}_{T_i}$ changes from one horizon to the next, so does $\mathcal{S} = \mathcal{S}_{T_i}$ in (2.13). We assume that (2.13) holds uniformly for all horizons as well and that

$$\int_0^t e^{-\tilde{\omega}(t-s)}\rho(s + T_i)\, ds < \frac{\delta}{2} \text{ for } t \in [0, T] \text{ and all } i = 0, 1 \dots. \tag{3.2}$$

Finally we require that

$$\int_0^T e^{-\tilde{\omega}(T-t)}\, \rho(T_i + t)\, dt \le \frac{\delta}{2}(1 - e^{-\tilde{\omega}T}) \text{ for } i = 0, 1 \dots. \tag{3.3}$$

These assumptions imply that

$$E(t) \le e^{-\tilde{\omega}\,(t-T_i)}E(T_i) + \int_0^{t-T_i} e^{-\tilde{\omega}(t-T_i-s)}\, \rho(T_i + s)\, ds \text{ for } t \in [T_i, T_{i+1}] \tag{3.4}$$

for every i, provided that $E(T_i) \leq \frac{\delta}{2}$. Note that (3.3), (3.4) imply

$$E(T_{i+1}) \leq e^{-\tilde{\omega}\,T}\,E(T_i) + \frac{\delta}{2}(1 - e^{-\tilde{\omega}T}) \leq \frac{\delta}{2},$$

and hence by induction $E(T_i) \leq \frac{\delta}{2}$ for all i.

THEOREM 3.1. *If* (3.1)–(3.3) *hold and* (2.10), (2.11), *and* (2.13) *are satisfied uniformly, then*

$$E(t) \leq e^{-\tilde{\omega}\,(t-T_i)}E(T_i) + \int_0^{t-T_i} e^{-\tilde{\omega}(t-T_i-s)}\,\rho(T_i+s)\,ds$$

for $t \in [T_i, T_{i+1}]$ *and all* i. *In particular this implies*

$$E(t) \leq e^{-\tilde{\omega}t}E(0) + \int_0^t e^{-\tilde{\omega}(t-s)}\rho(s)\,ds \text{ for all } t > 0.$$

Note that on each horizon $[T_i, T_{i+1}]$ we must have $|x^*(t) - \bar{x}| \leq \delta$ on $[0, T]$. This may necessitate to take T smaller than $T_{i+1} - T_i$ to ensure that $|x^*(t) - \bar{x}| \leq \delta$ on $[T_i, T_i + T]$. In this case, we can further partition the interval $[T_i, T_{i+1}]$ into subintervals and use consecutive linearization on each subinterval so that the condition is satisfied. The extreme case of this procedure results in the time-varying synthesis as in Remark 2.2.

Concerning the condition $|x^*(t) - \bar{x}| \leq \delta$ we can use an alternative approach motivated by the H^∞ Riccati equation. This will be discussed in the next section.

4. H^∞ Riccati synthesis. In this section we present a modification of the approach proposed in section 1 motivated by H^∞ synthesis. We assume that there exists an attenuation bound $\gamma > 0$ such that

$$A^T\Pi + \Pi\,A - \Pi\left(B\,B^T - \frac{1}{\gamma}I\right)\Pi + \frac{1}{\gamma}I + Q = 0 \tag{4.1}$$

has a positive definite solution Π and

$$A\Sigma + \Sigma A^T - \Sigma\left(C^TC - \frac{1}{\gamma}I\right)\Sigma + \frac{1}{\gamma}I + R = 0 \tag{4.2}$$

has a positive definite solution Σ, where again $A = f'(\bar{x})$. These Riccati equations are similar to those used in the equivalence between H^∞ controllers and linear quadratic zero-some differential games where in our case u and w are the two players [BB]. In the following proposition x and w denote the solutions to (1.8) and (2.7)with u given in (2.9) and x^* is the solution to (1.6) with the optimal open loop control u^*.

PROPOSITION 4.1. *For* $t \in [0, T]$ *we have*

$$\begin{aligned}
&\frac{d}{dt}W(x - x^*)\\
&= -\tfrac{1}{2}(|B^T\Pi(x - x^*)|^2 + (Q(x - x^*), x - x^*) + \tfrac{1}{\gamma}\,|\Pi(x - x^*)|^2 + \tfrac{1}{\gamma}|x - x^*|^2)\\
&\qquad + (r(x, x^*) + d, \Pi(x - x^*)) + (B^T\Pi(x - w), B^T\Pi(x - x^*))
\end{aligned}$$

and

$$\frac{d}{dt}\tilde{W}(x-w)$$
$$= -\frac{1}{2}(|C(x-w)|^2 + (R\Sigma^{-1}(x-w), \Sigma^{-1}(x-w)) + \frac{1}{\gamma}|x-w|^2 + \frac{1}{\gamma}|\Sigma^{-1}(x-w)|^2)$$
$$+(r(x,x^*)+d, \Sigma^{-1}(x-w)) - (n, C(x-w)),$$

where

$$r(x,x^*) = f(x) - f(x^*) - A(\bar{x})(x-x^*),$$

and the dependence of the variables on t is suppressed.

Proof. The proposition follows from the proofs of Propositions 2.1 and 2.2 observing that

$$((A - BB^T\Pi)x, \Pi x) = -\frac{1}{2}\left(|B^T\Pi x|^2 + (Qx,x) + \frac{1}{\gamma}|x|^2 + \frac{1}{\gamma}|\Pi x|^2\right),$$

and

$$((A - \Sigma C^T C)z, \Sigma^{-1}z) = -\frac{1}{2}\left(|Cz|^2 + (R\Sigma^{-1}z, \Sigma^{-1}z) + \frac{1}{\gamma}|z|^2 + \frac{1}{\gamma}|\Sigma^{-1}z|^2\right). \qquad \square$$

Since

$$r(x(t), x^*(t)) = \left(\int_0^1 A(x^*(t) + \theta(x(t) - x^*(t)))\, d\theta - A\right)(x(t) - x^*(t)),$$

we can observe that the assumption

$$\left\|\int_0^1 A(x^*(t) + \theta(x(t) - x^*(t)))\, d\theta - A\right\| \leq \frac{1}{\sqrt{2}\gamma} \text{ for all } t \in [0,T] \tag{4.3}$$

implies that

$$(r(x(t), x^*(t)), \Pi(x(t) - x^*(t))) \leq \frac{1}{2\gamma}\left(|\Pi(x(t) - x^*(t))|^2 + \frac{1}{2}\,|x(t) - x^*(t)|^2\right), \tag{4.4}$$

and

$$(r(x(t), x^*(t)), \Sigma^{-1}(x(t) - x^*(t))) \leq \frac{1}{2\gamma}\left(|\Sigma^{-1}(x(t) - w(t))|^2 + \frac{1}{2}\,|x(t) - x^*(t)|^2\right) \tag{4.5}$$

Using (2.11) and (4.4), (4.5) we find

$$\frac{d}{dt}(W(x-x^*)+\tilde{W}(x-w))$$

$$\leq -\omega(W(x-x^*)+\tilde{W}(x-w))+(d+r(x,x^*),\Pi(x-x^*)+\Sigma^{-1}(x-w))-(n,C(x-w))$$

$$-\frac{1}{2\gamma}(|\Pi(x-x^*)|^2+|x-x^*|^2+|x-w|^2+|\Sigma^{-1}(x-w)|^2)$$

$$\leq -\omega(W(x-x^*)+\tilde{W}(x-w))+\rho(t)(W(x-x^*)+\tilde{W}(x-w))^{1/2}$$

$$+\frac{1}{2\gamma}(|\Pi(x-x^*)|^2+|x-x^*|^2+|\Sigma^{-1}(x-w)|^2)$$

$$-\frac{1}{2\gamma}(|\Pi(x-x^*)|^2+|x-x^*|^2+|x-w|^2+|\Sigma^{-1}(x-w)|^2),$$

where $\rho(t)$ was defined below (2.14). Hence it follows that

$$\frac{d}{dt}(W(x-x^*)+\tilde{W}(x-w)) \leq -\omega(W(x-x^*)+\tilde{W}(x-w))+\rho(t)((W(x-x^*)+\tilde{W}(x-w)))^{\frac{1}{2}},$$

and therefore

$$\frac{d}{dt}E(t) \leq -\omega E(t)+\rho(t).$$

We can summarize the above developments in the following result.

THEOREM 4.1. *If* (4.3) *and* (4.2) *admit solutions for* $\gamma > 0$ *and* (2.10), (2.11), (4.3) *hold on* $[0,T]$, *then*

$$E(t) \leq e^{-\omega t}E(0)+\int_0^t e^{-\omega(t-s)}\,\rho(s)\,ds \text{ for } t\in[0,T]. \tag{4.6}$$

If the assumptions hold uniformly on all intervals, then (4.6) *holds for all* $t\in[0,\infty)$.

5. Numerical examples. We validate the proposed approach by means of a class of optimal control problems for the Burgers equation

$$\min\ \frac{1}{2}\int_0^{T_\infty}|y(t)-z(t)|^2_{L^2(\Omega)}dt\ +\ \frac{\sigma}{2}\int_0^{T_\infty}|u(t)|^2_{L^2(\tilde{\Omega})}dt \tag{5.1}$$

subject to $u\in L^2(\tilde{\Omega})$ and

$$\begin{cases} \frac{d}{dt}\,y(t,x)=\nu\,y_xx(t,x)-y\cdot y_x(t,x)+Bu(t,x)+d(t,x), & t>0,\\ y(t,0)=y(t,1)=0, & t>0,\\ y(0,x)=\varphi(x), & x\in(0,1), \end{cases} \tag{5.2}$$

where $y(t)=y(t,x)$, $x\in\Omega=(0,1)$, and ν and σ are positive constants. Here the initial condition $\varphi\in L^2(\Omega)$, the control domain $\tilde{\Omega}$ is a subset of Ω, and B is the extension-by-zero-operator from $\tilde{\Omega}$ to Ω. Finally d represents noise to the system and the observation data are supposed to be of the form

$$\hat{y}(t)=C_{\hat{\Omega}}\,y(t)+n(t), \tag{5.3}$$

where C is the restriction operator from Ω to $\hat{\Omega}$ with $\hat{\Omega}$ a subset of Ω, and n stands for observation noise. We further need to specify the operators G for the terminal weight in the receding horizon cost (1.5), the tracking weight Q in (2.3) and the operator R in the Kalman filter equation. We shall set

$$R = r\,I \quad \text{and} \quad Q = q\,I. \tag{5.4}$$

After several tests G was taken to be 0 for stabilization problems ($z = 0$) and as a scalar multiple of $\frac{1}{2}(x, \Pi x)$, where Π is the nonnegative solution of the algebraic Riccati equation (2.4) with $A = -\nu\,\partial_{xx}$ for tracking problems with $z \neq 0$. In this way G does not depend on the specific receding horizon level and can be computed during the initialization phase. Choosing G instead as $G(x) = \alpha\,|x|^2$ resulted in longer computing times and less favorable tracking properties.

The spatial discretization was done by the standard Galerkin scheme applied to (5.2) based on the basis of linear finite elements with meshsize dx. The ordinary differential equations resulting from (5.2) and (2.2), (2.7) were solved with an implicit Euler scheme with stepsize dt, while resolving the nonlinearities by Newton's method. The resulting linear systems were solved by inexact GMRES iterations. Unless specified otherwise we took

$$\nu = .01, \quad dt = .05, \quad dx = .025,$$

and for the receding horizon, $T = .5$. Further, unless quoted otherwise we chose $\Omega = \tilde{\Omega} = \hat{\Omega}$, $q = 10^{-5}$, and $r = 10^3$. For Example 1 we took $\sigma = .0175$ and for Example 2 we calculated with $\sigma = 10^{-3}$. The MATLAB-routine CARE was used to solve the algebraic Riccati equations. Below, $\hat{J}$ denotes the tracking part of the cost in (5.1). Noise was simulated by choosing uniformly distributed random numbers in the interval $[-\delta, \delta]$. It was added to the spatial-temporal grid points for either d or n representing disturbance to the equation and noise in the observation, respectively. Analogously, noise in the initial data φ was simulated by adding random numbers in $[-\delta, \delta]$ to the values of φ at the spatial grid points. The initial condition for (5.2) on receding horizon intervals with $i \geq 2$ was chosen as the state of the estimator at time T_i.

EXAMPLE 5.1 (stabilization). *Here we choose*

$$\varphi(x) = \begin{cases} 5 \sin 2\pi x & \text{on } (0, \frac{1}{2}], \\ 0 & \text{on } (\frac{1}{2}, 1), \end{cases}$$

and $z = 0$. Further $T^\infty = 5$, $T = 0.5$, $T_i = iT$ so that we have 10 receding horizon intervals. The uncontrolled solution is depicted in Figure 5.1. For Figures 5.2–5.4 we chose $n = 0$ (no observation noise) and simulated disturbances to the system with noise level $\delta = 10$ for d. The result for Figure 5.2 shows an open-loop solution, i.e., first we compute the open-loop optimal control u^* to (5.1)–(5.2), with $d = 0$, and then use the optimal open-loop control u^* against the noisy Burgers equation with $d = 10$. Figure 5.3 is for the full observation case, obtained with the state estimator procedure introduced in section 1, with $\hat{\Omega} = \tilde{\Omega} = \Omega$. For Figure 5.4 we test a partial observation case, i.e., we chose the control domain $\tilde{\Omega} = (0, \frac{1}{2})$ and the observation domain $\hat{\Omega} = (\frac{1}{2}, 1)$. Thus control and observation domain are non-overlapping. It should be noted that the hyperbolic nature of the Burgers equation is such that information moves from smaller to larger x-values. Figures 5.5 and 5.6 show

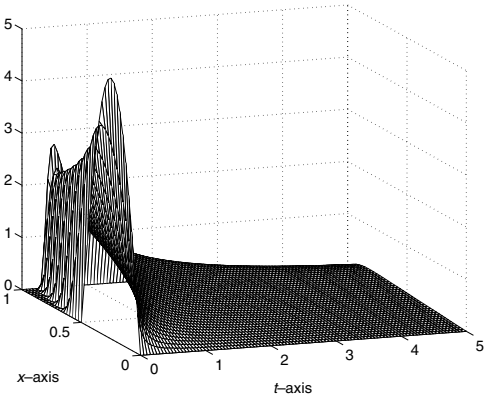


FIG. 5.1. *Uncontrolled solution, no noise.*

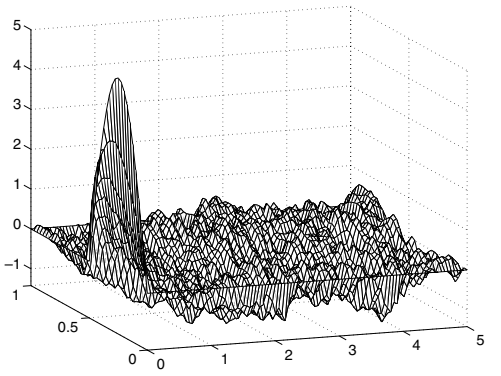


FIG. 5.2. *Open-loop controlled solution.*

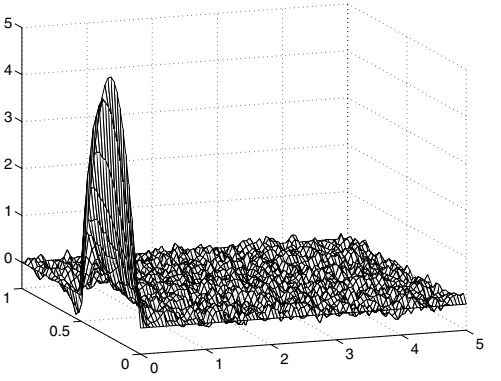


FIG. 5.3. *Estimator closed loop.*

results for the noisy observation case and we take again $\tilde{\Omega} = (0, \frac{1}{2})$ *and* $\hat{\Omega} = (\frac{1}{2}, 1)$*, i.e., the control and observation domains, which are as those used for Figure* 5.4*. Figure* 5.5 *disturbance to the equation was set* $d = 0$*, while now the observation data were taken to be noisy with noise level* $\delta = 10$ *for* n*. For Figure* 5.6 *noise was added to the initial condition* ($\delta = 5$) *and to the observation data* ($\delta = 10$).

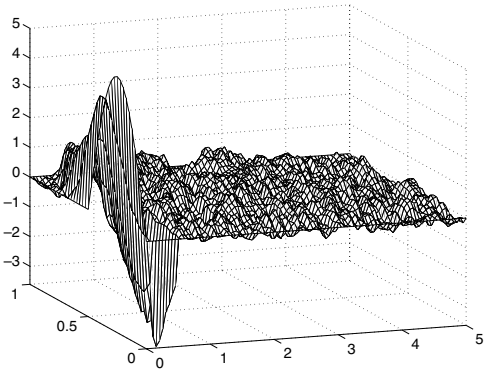


FIG. 5.4. *Partial control and observation domain.*

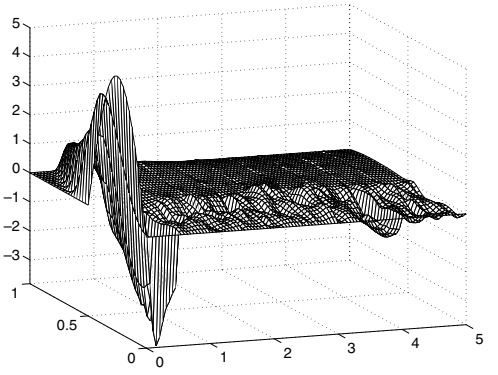


FIG. 5.5. *Noise in observation.*

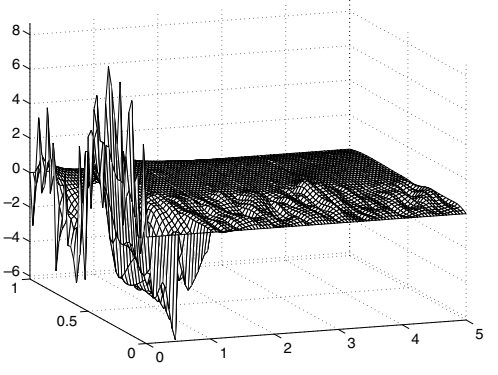


FIG. 5.6. *Noise in observation and initial condition.*

The run-times depend slightly on the random numbers that enter in d *or* n*. Typical times for Figures* 5.3–5.6 *are* 16 *to* 17 *units, whereas the run-time for Figure* 5.2*, which uses the open-loop optimal control on the complete interval* $[0, T^{\infty}]$*, is* 181 *units. The stabilization value for Figure* 5.2 *is* $\hat{J} = .540$*, whereas it is only* $\hat{J} = .146$ *for Figure* 5.3*. In Table* 1 *we give the stabilization values* $\hat{J}_{-}$ *for the results of Figures* 5.2–5.6 *on* $[0.5, 5]$*, i.e., for the receding horizon intervals* 2–10*.*

TABLE 1
$\hat{J}_-$.

Fig 5.2	Fig 5.3	Fig 5.4	Fig 5.5	Fig 5.6
.17-.5	.0008-.0009	.070-.114	.056-.065	.1-.5

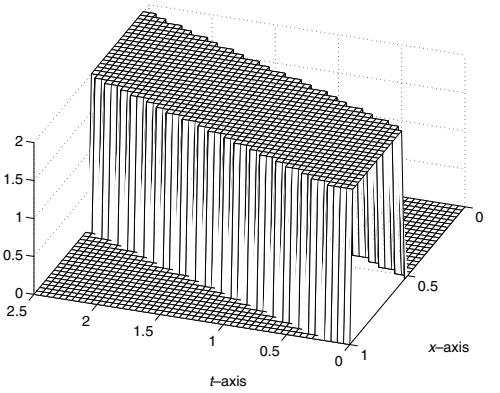


FIG. 5.7. *Desired state.*

We ran all tests with five randomly chosen initializations of the random number generator. In Table 1 we report the range of obtained values for $\hat{J}_-$. We can conclude that the receding horizon strategy combined with the dynamic observer introduced in section 1 attenuates noise significantly better than the open-loop controlled system; compare column 1 in Table 1 to columns 2–4. The computing times for the receding horizon strategy are much shorter than for the open-loop optimal control on the complete time interval. Finally the strategy can cope with partial observations.

EXAMPLE 5.2 (tracking). *Here we set the initial condition as*

$$\varphi(x) = \begin{cases} 1 & \text{on } (0, \frac{1}{2}], \\ 0 & \text{on } (\frac{1}{2}, 1), \end{cases}$$

and the target z as the characteristic function of the set $\{(t,x)\colon 2.5-5x < t < 5-5x$, $x \in (0,1)\}$; see Figure 5.7. *Further $T^\infty = 2.5$, $T = 0.5$, $T_i = iT$ so that we have five receding horizon intervals. Note that the uncontrolled solution of the Burgers equation would transport the jump at $x = \frac{1}{2}$ toward increasing x as t increases, while decreasing its height. Also, the desired state z moves into the direction which is opposite to the one of the characteristics of the Burgers equation. Thus this example can be considered as a challenging one. For the results presented here we took the control domain as $\tilde{\Omega} = \Omega$. The noise levels are* 5 *for disturbance to the equation and* 10 *for observation noise and disturbance to the initial condition. If noise is not mentioned, then it is set equal to* 0. *The random number generator is initialized at the same seed, independently for disturbances and observations.*

In Figures 5.8 *and* 5.9 *we show the open-loop and the full-observation ($\hat{\Omega} = \Omega$) closed-loop results, each with disturbances to the equation. Figures* 5.10 *and* 5.11 *show partial observation results with the observation domain $\hat{\Omega} = (1/4, 1)$ for nonzero disturbance d and nonzero observation noise n, respectively. Figure* 5.12 *provides the result for the case with the partial observation on $\hat{\Omega} = (1/4, 1)$, $d = 0$, and noise*

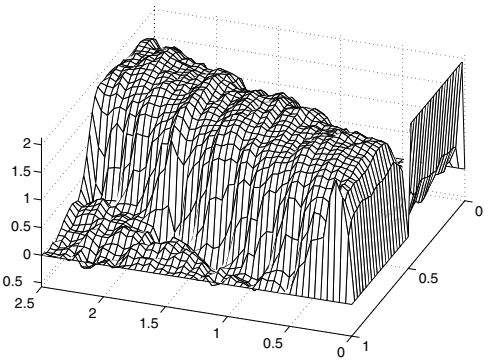


FIG. 5.8. *Open-loop controlled solution.*

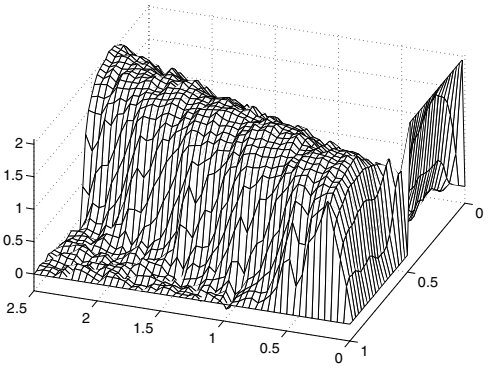


FIG. 5.9. *Estimator closed loop.*

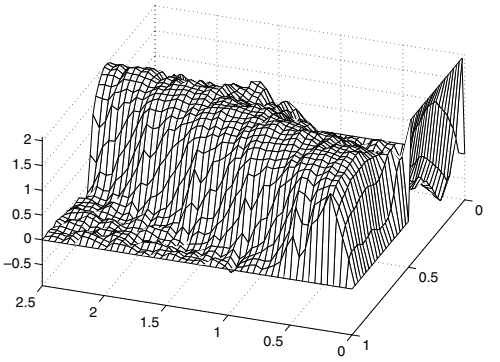


FIG. 5.10. *Estimator closed loop, partial obseration.*

entering to the observations as well as the initial condition. The tracking costs for Figures 5.8–5.12 *are given in Table* 2.

For Figure 5.8 *we again computed the open-loop optimal control from* $[0, T^{\infty}]$ *and applied it to the perturbed Burgers equation. The computing time for Figure* 5.8 *is* 56 *units, whereas for the other figures it is* 16.3 *to* 20 *units. Thus the ratio of the computing times for the open-loop control on* $[0, T^{\infty}]$ *and the receding horizon*

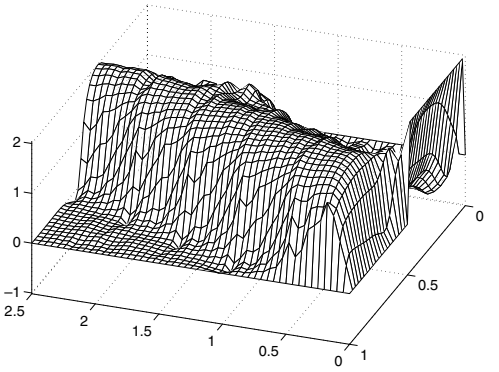


FIG. 5.11. *Estimator, noise in data.*

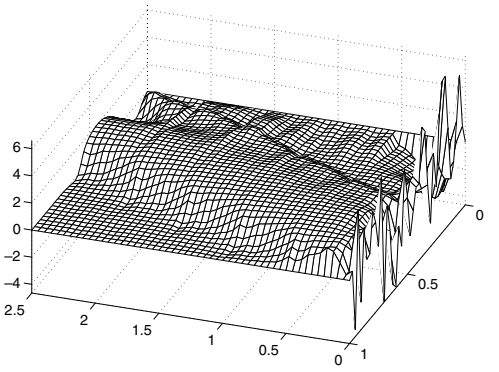


FIG. 5.12. *Noise in data and initial condition.*

TABLE 2
$\hat{J}$.

Fig 5.8	Fig 5.9	Fig 5.10	Fig 5.11	Fig 5.12
.179	.0816	.0814	.0812	.0887

controls is less favorable than for Example 5.1. *This is because the resulting two-point boundary value problems are less well-conditioned as a consequence of the terminal weight* $G \neq 0$.

We also tested with smaller observation domain $\hat{\Omega}$ *and obtained comparable results. Reducing the control domain* $\tilde{\Omega}$ *results in a significant increase of* $\hat{J}$ *due to the challenging nature of the problem. For example, for the settings of Figure* 5.10, *if* $\tilde{\Omega} = (0, .8)$, *then* $\hat{J} = .490$.

6. Conclusion. The receding horizon technique is a sequential time-domain decomposition method to reduce the computational requirements for solving the optimality systems on long time horizons. The obtained controls on each receding horizon can be interpreted as a state feedback control. In this paper we developed a compensator design in the context of receding horizon control when only partial state observations are available. The receding horizon synthesis was combined with the state estimator dynamics. An LQG design based on a linearization procedure was

used and its asymptotic performance was analyzed for nonlinear systems with disturbances and observation noise. Numerical examples for control problems for the Burgers equation validate the proposed methodology.

Acknowledgment. We would like to thank Prof. Stefan Volkwein for providing us with the code of his open loop optimal control routine for the Burgers equation.

REFERENCES

[ABQRW] F. Allgöwer, T. Badgwell, J. Qin, J. Rawlings, and S. Wright, *Nonlinear predictive control and moving horizon estimation—An introductory overview*, in Advances in Control, P. Frank, ed., Springer, London, 1999, pp. 391–449.

[BB] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems, A Dynamic Game Approach*, Birkhäuser, Boston, 1995.

[B] T. R. Bewley, *Flow control: New challenges for a new Renaissance*, Prog. Aerospace Sci., 37 (2001), pp. 21–58.

[CA] H. Chen and F. Allgöwer, *A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability*, Automatica, 34 (1998), pp. 1205–1217.

[CHK] H. Choi, M. Hinze, and K. Kunisch, *Instantaneous control of backward facing step flow*, Appl. Numer. Math., 31 (1999), pp. 133–158.

[CTMC] H. Choi, R. Temam, P. Moin, and J. Kim, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[GPM] C. E. Garcia, D. M. Prett, and M. Morari, *Model predictive control: Theory and practice—a survey*, Automatica, 25 (1989), pp. 335–348.

[HV] M. Hinze and S. Volkwein, *Analysis of instantaneous control for the Burgers equation*, Nonlinear Anal., 50 (2002), pp. 1–26.

[IK1] K. Ito and K. Kunisch, *On asymptotic properties of receding horizon optimal control*, SIAM J. Control. Optim., 40 (2002), pp. 1585–1610.

[IK2] K. Ito and K. Kunisch, *Receding horizon optimal control for infinite dimensional systems*, ESIAM Control Optim. Calc. Var., 8 (2002), pp. 741–760.

[JYH] A. Jadababaie, J. Yu, and J. Hauser, *Unconstrained Receding Horizon Control of Nonlinear Systems*, Trans. Automat. Control, 46 (2001), pp. 776–783.

[K] D. L. Kleinman, *An easy way to stabilize a linear constant system*, IEEE Trans. Automat. Control, 15 (1970), pp. 692–712.

[MM] D. Q. Mayne and H. Michalska, *Receding horizon control of nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 814–824.

[PND] J. A. Primbs, V. Nevistič, and J. C. Doyle, *A Receding Horizon Generalization of Pointwise Min-Norm Controllers*, Trans. Automat. Control, 45 (2000), pp. 898–909.

[SMR] P. Scokaert, D. Q. Mayne, and J. B. Rawlings, *Suboptimal predictive control (Feasibility implies stability)*, IEEE Trans. Automat. Control, 44 (1999), pp. 648–654.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 226–245

COMMON POLYNOMIAL LYAPUNOV FUNCTIONS FOR LINEAR SWITCHED SYSTEMS*

PAOLO MASON†, UGO BOSCAIN†, AND YACINE CHITOUR‡

Abstract. In this paper, we consider linear switched systems $\dot{x}(t) = A_{u(t)}x(t)$, $x \in \mathbb{R}^n$, $u \in U$, $\{A_u : \ u \in U\}$ compact, and the problem of asymptotic stability for arbitrary switching functions, uniform with respect to switching (UAS). Given a UAS system, it is always possible to build a common polynomial Lyapunov function. Our main result is that the degree of that common polynomial Lyapunov function is not uniformly bounded over all the UAS systems. This result answers a question raised by Dayawansa and Martin. A generalization to a class of piecewise-polynomial Lyapunov functions is given.

Key words. switched systems, stability, polynomial Lyapunov functions, uncertain systems

AMS subject classifications. 93D20, 37N35

DOI. 10.1137/040613147

1. Introduction.

In recent years, the problem of stability and stabilizability of switched systems has attracted increasing attention (see, for instance, [1, 4, 5, 8, 11, 13, 14, 20]), and still many questions remain unsolved.

In this paper, we address the problem of existence of common polynomial Lyapunov functions for linear switched systems.

By a switched system, we mean a family of continuous-time dynamical systems and a rule that determines at each time which dynamical system is responsible of the time evolution. More precisely, let $\{f_u : \ u \in U\}$ (where U is a subset of $\mathbb{R}^m$, $m \in \mathbb{N}$) be a finite or infinite set of sufficiently regular vector fields on a manifold M, and consider the family of dynamical systems:

$$\dot{x} = f_u(x), \quad x \in M. \tag{1}$$

The rule is given by assigning the so-called switching function, i.e., a function $u(.) : [0, \infty[\to U \subset \mathbb{R}^m$. Here, we consider the situation in which the switching function is not known a priori and represents some phenomenon (e.g., a disturbance) that is not possible to control. Therefore, the dynamics defined in (1) also fit into the framework of uncertain systems (cf., for instance, [9]).

In what follows, we use the notation $u \in U$ to label a fixed individual system and $u(.)$ to indicate the switching function.

These kinds of systems are sometimes called n-modal systems, dynamical polysystems, polysystems, or input systems. The term "switched system" is often reserved for situations in which the switching function $u(.)$ is piecewise continuous or the set U is finite. For the purpose of this paper, we only require $u(.)$ to be a measurable function and U to be a measurable set.

For a discussion of various issues related to switched systems, see [5, 13, 14].

A typical problem for switched systems goes as follows. Assume that for every $u \in U$, the dynamical system $\dot{x} = f_u(x)$ satisfies a given property (P). Then one can

*Received by the editors August 9, 2004; accepted for publication (in revised form) September 12, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sicon/45-1/61314.html
†SISSA-ISAS, Via Beirut 2-4, 34014 Trieste, Italy (mason@sissa.it, boscain@sissa.it).
‡Université Paris XI, Départment de Mathématiques, F-91405 Orsay, France (chitour@lss.supelec.fr).

investigate conditions under which property (P) still holds for $\dot{x} = f_{u(t)}(x)$, where $u(.)$ is an arbitrary switching function.

In [1, 8, 11, 12], the case of linear switched systems was considered:

$$\dot{x}(t) = A_{u(t)}x(t), \quad x \in \mathbb{R}^n, \quad A_u \in \mathbb{R}^{n\times n}, \tag{2}$$

where n is a positive integer and $u(.) : [0, \infty[\to U$ is a (measurable) switching function. For these systems, the problem of asymptotic stability of the origin, uniformly with respect to switching functions, was investigated.

Next, we set $\mathbf{A} := \{A_u \ : \ u \in \ U\}$ and, to simplify the notation, we still call *switching function* the measurable matrix-valued map $A(.) := A_{u(.)}$. In this way, the switching system (2) reads

$$\dot{x}(t) = A(t)x(t) \text{ with } x \in \mathbb{R}^n \text{ and } A(.) : [0, \infty[\to \mathbf{A} \text{ a measurable map.} \tag{3}$$

In the following, we assume that

(HO) the set $\mathbf{A}$ is a compact subset of the set of $n \times n$ real matrices.

Moreover, the set of switching functions, denoted by $\boldsymbol{\mathcal{A}}$, is the set of measurable functions $A(.) : [0, \infty[\to \mathbf{A}$. With our assumptions, for every switching function $A(.)$ and initial condition $x_0 \in \mathbb{R}^n$, the corresponding (Carathéodory) solution of (3) is defined for every $t \geq 0$. We use $\phi_t^{A(.)}(x_0)$ to denote the flow of (3) at time $t \geq 0$ corresponding to the switching function $A(.)$ and starting from x_0.

Let us recall the usual notions of stability used for the system (3).

DEFINITION 1. *Consider the switched system* (3). *We say that the origin is*

(S) *stable if for every* $A(.) \in \boldsymbol{\mathcal{A}}$ *and* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that (s.t.)* $\|\phi_t^{A(.)}(x_0)\| \leq \varepsilon$ *for every* $t \geq 0$, $\|x_0\| \leq \delta$;

(US) *uniformly stable if it is stable with* δ *not depending on* $A(.)$;

(U) *unstable if it is not stable (i.e., if there exists* $A(.) \in \boldsymbol{\mathcal{A}}$ *s.t. the system* $\dot{x}(t) = A(t)x(t)$ *is unstable as a linear time-varying system);*

(AS) *asymptotically stable, if it is stable and attractive (i.e., there exists* $\delta' > 0$ *so that we have* $\lim_{t\to\infty} \|\phi_t^{A(.)}(x_0)\| = 0$ *for every* $A(.) \in \boldsymbol{\mathcal{A}}$ *and* $x_0 \in \mathbb{R}^n$ *with* $\|x_0\| \leq \delta'$);

(UAS) *uniformly asymptotically stable if it is uniformly stable and if for every* $\varepsilon' > 0$ *and* $\delta' > 0$ *there exists* $T > 0$ *such that, for every switching function* $A(.) \in \boldsymbol{\mathcal{A}}$, $t \geq T$, *and* $\|x_0\| \leq \delta'$, *we have* $\|\phi_t^{A(.)}(x_0)\| \leq \varepsilon'$;

(GUES) *globally uniformly exponentially stable if there exist positive constants* M, λ *such that:* $\|\phi_t^{A(.)}(x_0)\| \leq M\ e^{-\lambda t}\|x_0\|$ *for every* $x_0 \in \mathbb{R}^n$, $t > 0$, $A(.) \in \boldsymbol{\mathcal{A}}$.

Because the dynamics are linear in the state variable, the local and global notions of stability are equivalent. More precisely, it was proved in [3] that, for system (3) subject to HO, the three notions AS, UAS, GUES and the notion of attractivity are all equivalent (see also the bibliographical note in [14]). In addition, if the system is unstable, then there exists a switching function $A(.) \in \boldsymbol{\mathcal{A}}$ and an initial condition x_0 such that $\lim_{t\to\infty} \|\phi_t^{A(.)}(x_0)\| \to \infty$. In the following, we just refer to the notions of stability, instability, and GUES.

Remark 1. Since for the stability issue, a system of type (3), subject to HO, is uniquely determined by a compact set $\mathbf{A}$ of $n \times n$ real matrices, we identify $\mathbf{A}$ with the corresponding system for the rest of the paper. For instance, when we say that $\mathbf{A}$ is GUES, we mean that the corresponding system of type (3) is GUES.

We will often consider the problem of determining whether a system, belonging to a certain class C of systems of type (3) subject to HO, is GUES or not. Notice

that fixing such a class of systems means to fix a set of compact subsets of $\mathbb{R}^{n\times n}$, i.e., C can be identified with a subset of $\{\mathbf{A} \subset \mathbb{R}^{n\times n} : \ \mathbf{A} \text{ compact}\}$.

For a system (3) subject to HO, it is well known that the GUES property is a consequence of the existence of a common Lyapunov function.

DEFINITION 2. *A common Lyapunov function $V : \mathbb{R}^n \longrightarrow \mathbb{R}^+$, for a switched system (S) of type* (3)*, is a continuous function such that V is positive definite (i.e., $V(x) > 0\ \forall x \neq 0$, $V(0) = 0$) and V is strictly decreasing along nonconstant trajectories of (S).*

Vice versa, it is known that, given a GUES system of the type (3) subject to (HO), it is always possible to build a $\mathcal{C}^\infty$ common Lyapunov function (see, for instance, [11, 15, 16, 17] and the bibliographical note in [14]).

Anyway, the problem of finding a Lyapunov function or proving the nonexistence of a Lyapunov function is in general a difficult task. Sometimes, it is even easier to prove directly that a system is GUES or unstable. An example is provided below by bidimensional switched systems.

1.1. Single-input bidimensional switched systems. Consider a bidimensional system with *single input* of the type

$$\dot{x}(t) = u(t)Ax(t) + (1 - u(t))Bx(t), \tag{4}$$

where $x \in \mathbb{R}^2$, A and B are two 2×2 real Hurwitz matrices, and $u(.)$ is a measurable function defined on $\mathbb{R}^+$ and taking values in U equal either to $[0, 1]$ or $\{0, 1\}$. In what follows, we call Ξ the class of bidimensional systems of the above form. This class is parameterized by couples of 2×2 real Hurwitz matrices.

Remark 2. Whether systems of type (4) are GUES or not is independent on the specific choice $U = [0, 1]$ or $U = \{0, 1\}$. In fact, this is a particular instance of a more general result stating that the stability properties of systems (3) subject to HO depend only on the convex hull of the set $\mathbf{A}$; see Proposition 1, Remark 6, and Appendix B.

In [19], the authors provide a necessary and sufficient condition on the pair (A, B) to share a *quadratic* Lyapunov function, but it is known (see for instance [11, 18, 10, 21]) that there exist GUES linear bidimensional systems not admitting a quadratic Lyapunov function.

In [11], Dayawansa and Martin posed the problem of finding the minimal degree of a *polynomial* Lyapunov function. More precisely, the problem posed by Dayawansa and Martin is the following.

Problem P. Define $\Xi_{\text{GUES}} \subset \Xi$ as the set of GUES systems of the type (4). Find the minimal integer m such that every system of Ξ_{GUES} admits a polynomial Lyapunov function of degree less than or equal to m.

Remark 3. In the problem posed by Dayawansa and Martin, it is implicitly assumed that a GUES system always admits a polynomial common Lyapunov function. This fact was first proved by Molchanov and Pyatnitskii in [15, 16] under the assumption that the set $\mathbf{A}$ is of the form $\mathbf{A}= \{(a_{ij})_{i,j=1,\dots,n} \ : \ a_{ij}^- \leq a_{ij} \leq a_{ij}^+\}$. In [17] the authors state the result, with no further details, under the more general hypothesis $\mathbf{A}$ just compact. In the case in which the convex hull of $\mathbf{A}$ is finitely generated, the existence of a polynomial common Lyapunov function for GUES systems was proved by Blanchini and Miani [6, 7] in the context of uncertain systems.

Since the proofs of Molchanov, Pyatnitskii, Blanchini, and Miani need nontrivial intermediate results, in this paper we provide a self-contained proof, based on a more

direct argument, for a set **A** satisfying the weaker hypothesis that its convex hull is compact (see Theorem 1 and Remark 6).

As for the GUES issue, it was completely resolved in [8], where a necessary and sufficient condition for a system of type (4) to be GUES was found directly, without looking for a Lyapunov function. (See section 3 and Appendix A for more details.) This is a typical example in which it is easier to directly study the stability rather than look for a Lyapunov function.

1.2. Sets of functions sufficient to check GUES. The concept of Lyapunov function is useful for practical purposes when one can prove that, for a certain class of systems, if a Lyapunov function exists, then it is possible to find one of a certain type and possibly as simple as possible (e.g., polynomial with a bound on the degree, piecewise quadratic, etc.).

More precisely, consider a class C of systems of type (3) in $\mathbb{R}^n$ subject to HO, in the sense of Remark 1. One would like to find a class of functions $\mathcal{S}_C$, identified by a finite number of parameters, which is sufficient to check GUES for systems belonging to C i.e., if a system of C admits a Lyapunov function, then it admits one in $\mathcal{S}_C$. Once such a class of functions is identified, then in order to verify GUES, one could use numerical algorithms to check (by varying the parameters) whether a Lyapunov function exists (in which case the system is GUES) or not (meaning that the system is not GUES).

For instance, a remarkable result for a given class C of systems in $\mathbb{R}^n$ could be the following.

Claim. There exists a positive integer m (depending on n) such that whenever a system of C admits a Lyapunov function, then it admits one that is polynomial of degree less than or equal to m. In other words, the class of polynomials of degree at most m is sufficient to check GUES for the class C.

If this result were true, one could use a numerical algorithm to check, among all polynomial of degree m (varying the coefficients), if there is one that is a Lyapunov function. Unfortunately, *this claim is not true*, even for the simplest nontrivial case of class of systems in $\mathbb{R}^2$, namely, systems of type Ξ (cf. (4)).

The next definition formalizes the idea of class of functions sufficient to check GUES.

DEFINITION 3. *We say that a subset $\mathcal{S}$ of $\mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$ is finitely (or q-finitely) parameterized if there exist $\Omega \subset \mathbb{R}^q$ for a positive integer q and a bijective map $\Psi : \Omega \subseteq \mathbb{R}^q \to \mathcal{S} \subset \mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$. A subset $\mathcal{S}$ of $\mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$ is said to be sufficient to check GUES for a class C of systems of type* (3) *in $\mathbb{R}^n$ (SSF) if every GUES system of C admits a Lyapunov function in $\mathcal{S}$. A subset $\mathcal{S}$ of $\mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$ is said to be a (q-parameters) finite set of functions sufficient to check GUES for a class C of systems of type* (3) *in $\mathbb{R}^n$ (finite-SSF) if $\mathcal{S}$ is q-finitely parameterized and is an SSF for C.*

If a subset $\mathcal{S}$ of $\mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$ is not finitely parameterizable but is an SSF for a class of systems C, we call $\mathcal{S}$ an ∞-SSF.

Remark 4. In [7], a concept similar to those introduced in the previous definition was provided and it was called "universal class of Lyapunov functions."

Using the previous definitions, the results and the problem formulated in [11] can be rephrased in the following way:

R1 for systems (3) subject to HO, the set $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ is an ∞-SSF;

R2 for linear bidimensional systems of the class Ξ (cf. (4)), the set of quadratic functions is not an SSF;

P let $\mathcal{P}^m$ be the set of polynomial functions of two variables with degree at most m. What is the minimal m such that $\mathcal{P}^m$ is a finite-SSF for the linear bidimensional systems of the class Ξ?

Remark 5. Notice that to check numerically the existence of a Lyapunov function using the concept of finite-SSF, one needs some regularity properties of the functions of the family, with respect to the parameters (at least continuity). This discussion is outside the purpose of this paper.

1.3. Main results. We first provide a proof that the implicit assumption of Dayawansa and Martin (i.e., that a linear GUES switched system always admits a polynomial Lyapunov function; cf. Remark 3) is correct for a set **A** just compact.

THEOREM 1. *If the origin is a GUES equilibrium for the switched system* (3) *subject to HO, then there exists a polynomial Lyapunov function.*

The above result can be stated equivalently as follows.

THEOREM 1BIS. *The set of polynomials from $\mathbb{R}^n$ to $\mathbb{R}$ is an ∞-SSF for linear switched systems* (3) *subject to HO.*

The proof of Theorem 1 is given in section 2 and the starting point is the construction of a homogeneous and convex Lyapunov function W, following the corresponding argument of [11]. The main idea is then to seek for a (homogeneous) polynomial $\tilde{W}$ whose level sets approximate, in some suitable sense, those of W and, finally, to show that $\tilde{W}$ is also a Lyapunov function. A similar argument was used in [7], where the authors used an intermediate approximation with polyhedral Lyapunov functions.

Remark 6. In Appendix B, it is proved that the GUES property of (3) depends only on the convex hull of **A** (proof of Proposition 1). The claim of Appendix B can be applied to a compact set **A** being the convex hull of a set not necessarily compact. For this reason Proposition 1 and Theorem 1 hold under the weaker hypothesis:

(HO-weak) the set **A** is a measurable subset of the set of $n \times n$ real matrices, whose convex hull is compact.

The core of the paper consists of showing that problem P does not have a solution, i.e., the minimum degree of a polynomial Lyapunov function cannot be uniformly bounded over the set of all GUES systems of the form (4). More precisely, we have the following.

THEOREM 2. *Let $\Xi_{\mathrm{GUES}} \subset \Xi$ be the set of all GUES systems of the type* (4). *If (A, B) is a pair of 2×2 real matrices giving rise to a system of Ξ_{GUES}, let $m(A, B)$ be the minimum value of the degree of any polynomial LF associated to that system. Then $m(A, B)$ cannot be bounded uniformly over Ξ_{GUES}.*

The above result can be stated equivalently as follows.

THEOREM 2BIS. *Let $\mathcal{P}^m$ the set of polynomial functions from $\mathbb{R}^2$ to $\mathbb{R}$ of degree at most m. Then $\mathcal{P}^m$ is not a finite-SSF for Ξ.*

The proof, given in section 4, is based on ideas developed in [8], where necessary and sufficient conditions for GUES of systems (4) are provided. We build a sequence of GUES systems corresponding to a sequence of pairs of matrices (A_i, B_i), $i \geq 1$. The sequence of systems is chosen in such a way that the limit system is uniformly stable but not attractive. In particular, that limit system admits a nontrivial periodic trajectory whose support Γ is a $\mathcal{C}^1$ but not a $\mathcal{C}^2$ submanifold of the plane. To each GUES system of the sequence, one considers any polynomial LF $V_{\bar{A}_i, \bar{B}_i}$ whose degree is at most m. We prove that a subsequence of $(V_{\bar{A}_i, \bar{B}_i})$ converges to a nonzero polynomial function V (of degree at most m) which admits Γ as a level set. Since Γ is not analytic, a contradiction is reached.

Remark 7. The result given by Theorem 2 generalizes to dimensions higher than 2 as follows. Let (A_i, B_i), $i \geq 1$, be a sequence of 2×2 matrices such that (i) the corresponding systems of type (4) are GUES, (ii) the limit is uniformly stable but not attractive. As explained above, for this sequence of systems it is not possible to build a sequence of polynomial Lyapunov function of uniformly bounded degree. Consider now the sequence of systems in $\mathbb{R}^n$, $n \geq 2$, of the form $\dot{\bar{x}} = u\bar{A}_i\bar{x} + (1-u)\bar{B}_i\bar{x}$ corresponding to the matrices

$$\bar{A}_i = \left(\begin{array}{c|ccc} A_i & 0 & \cdots & \vdots \\ \hline & -1 & \cdots & \vdots \\ 0 & 0 & -1 & \vdots \\ & \vdots & \ddots & \vdots \\ & 0 & \cdots & -1 \end{array}\right), \quad \bar{B}_i = \left(\begin{array}{c|ccc} B_i & 0 & \cdots & \vdots \\ \hline & -1 & \cdots & \vdots \\ 0 & 0 & -1 & \vdots \\ & \vdots & \ddots & \vdots \\ & 0 & \cdots & -1 \end{array}\right). \tag{5}$$

Each system of the sequence is GUES but the limit system is not (it is just uniformly stable). Now, if $V_{\bar{A}_i,\bar{B}_i}$, $i \geq 1$, are the corresponding polynomial Lyapunov functions, then they cannot be polynomials of uniformly bounded degree since this is not true for the restriction of $V_{\bar{A}_i,\bar{B}_i}$ to the first two variables.

Remark 8 (extension to piecewise polynomial functions). Another class of functions commonly used to check GUES is that of piecewise quadratic functions or more generally piecewise polynomial functions (PPF). Here, by a PPF, we mean a continuous function $V \in \mathcal{C}^0(\mathbb{R}^n, \mathbb{R})$ together with a finite number q of cones K_j, $1 \leq j \leq q$, based at zero and partitioning $\mathbb{R}^n$ so that V is a polynomial function of degree d_j on K_j, $1 \leq j \leq q$. We refer to $m(V) := \max\{q, d_1, \ldots, d_q\}$ as the *total degree* of V.

It is tempting to state a version of problem P by replacing polynomial functions of degree at most m with PPFs of total degree at most m.

Again, the PPF version of problem P does not have a solution for $n = 2$, i.e., the minimum total degree of a piecewise polynomial Lyapunov function cannot be uniformly bounded over the set of all GUES system of the form (4). The argument is a simple extension of the proof of Theorem 2 and it is briefly mentioned in Remark 14.

The last results of the paper concern the existence and the characterization of a finite-SSF for systems of the type (3) subject to HO.

Let us define the *convex semicone* generated by a set $D \subset \mathbb{R}^n$ as the set of points λx with $\lambda > 0$ and $x \in co(D)$, where $co(D)$ denotes the convex hull of the set D. With this definition, the point $x = 0$ does not belong to the convex semicone generated by a set D if $0 \notin co(D)$.

First, we prove the following. (See Appendix B for the argument.)

PROPOSITION 1. *For every compact subset* $\mathbf{A}$ *of* $\mathbb{R}^{n\times n}$ *(i.e., verifying HO), let* $S_{\mathbf{A}}$ *be the system of the type* (3) *associated to* $\mathbf{A}$. *Then, for* $\mathbf{A}$ *and* $\mathbf{A}'$ *verifying HO and generating the same convex semicone,* $S_{\mathbf{A}}$ *is GUES (resp., uniformly stable) if and only if* $S_{\mathbf{A}'}$ *is GUES (resp., uniformly stable).*

Based on converse Lyapunov theorems, one can deduce some trivial existence results for finite-SSFs. For instance, consider a class C of systems of type (3) in $\mathbb{R}^n$ subject to HO and satisfying the following property: for every $\mathbf{A} \in C$, the convex hull of $\mathbf{A}$ is generated by at most k matrices $n \times n$, where k is a positive integer. One can build a finite-SSF for the class C as follows.

First, it follows directly from Definition 3 that we may simply assume that C is made of GUES systems. Thanks to Proposition 1, the class C can be parameterized by k-tuples of $n\times n$ matrices, defined up to their norm. In this way, a $k(n^2-1)$-parameters

finite-SSF is provided for the class C. For instance, the class Ξ of two-dimensional systems of type (4) admits a 6-parameter finite-SSF.

The above construction is not explicit and therefore is not useful to check GUES. Similarly to Lyapunov functions, it is then clear that the real challenge for finite-SSFs concerns their *explicit* characterization. For classes of systems of type (3) in $\mathbb{R}^n$ with $n \geq 3$, that issue is completely open in general. In dimension two *we provide an explicit* 5*-parameter finite-SSF for* Ξ, using the necessary and sufficient conditions for GUES given in [8]. This is the content of section 5.

Clearly, Ξ can be parameterized by the pairs (A, B) of 2×2 real Hurwitz matrices, where both A and B are defined up to their norm. The construction of the explicit finite-SSF goes as follows. As done previously, we may assume that the pair (A, B) gives rise to a GUES system. By taking advantage of the complete characterization of GUES systems of the class Ξ given in [8], one can explicitly associate to every GUES pair (A, B) a Lyapunov function as explained next. We start by defining, from (A, B), a pair $(\tilde{A}, \tilde{B})$ giving rise to a system of Ξ which is uniformly stable but not attractive. Such a system admits a closed trajectory whose support Γ is a simple Jordan closed curve (cf. sections 3, 4). We then construct a homogeneous positive definite function V whose level set 1 is Γ. We finally show that V is a Lyapunov function for (A, B). Since the set of $(\tilde{A}, \tilde{B})$ built from the GUES pairs (A, B) can be parameterized by using five parameters, we end up with a 5-parameter finite-SSF.

1.4. Structure of the paper. Section 2 contains the proof of Theorem 1. In section 3, we recall the main ideas from [8] needed for the rest of the paper. For completeness, we provide in Appendix A the full statement of the main result of [8]. The proof of Theorem 2 is given in section 4 and the explicit construction of a 5-parameter finite-SSF for systems of type (4) is provided in section 5. Finally, in Appendix B, we prove Proposition 1.

2. Existence of common polynomial Lyapunov functions. In this section, we prove Theorem 1. The starting point of the argument follows the first part of the proof of an analogous result in [11].

We define the function $V : \mathbb{R}^n \to \mathbb{R}^+$ by

$$V(x) = \sup_{A(.)\in \boldsymbol{\mathcal{A}}} \int_0^{+\infty} \|\phi_t^{A(.)}(x)\|^2 \mathrm{d}t.$$

The function V is well defined since there exist positive constants C, μ such that for all $t \geq 0$ and $x \in \mathbb{R}^n$,

$$\|\phi_t^{A(.)}(x)\| \leq C\, e^{-\mu t}\|x\|.$$

Note that V is homogeneous of degree 2 and continuous. In addition, we next show that V is strictly convex. That fact will be crucial later in the argument. Fix $x, y \in \mathbb{R}^n$ and $x \neq y$. Let $A(.)$ be a switching function. The function $x \mapsto \|\phi_t^{A(.)}(x)\|^2$ is strictly convex. Moreover, for every $\lambda \in]0, 1[$, by compactness of $\mathbf{A}$, the expression

$$\lambda\, \|\phi_t^{A(.)}(x)\|^2 + (1-\lambda)\, \|\phi_t^{A(.)}(y)\|^2 - \|\phi_t^{A(.)}(\lambda x + (1-\lambda)y)\|^2$$

is nonnegative for every $t \geq 0$ and is bounded from below by a positive constant on some interval $[0, \bar{t}]$, uniformly with respect to $A(.)$.

Therefore, dividing the integration interval into the two intervals $[0, \bar{t}]$ and $[\bar{t}, +\infty]$ and taking a maximizing sequence of switching functions for $V(\lambda\, x + (1-\lambda)\, y)$, we

have

$$V(\lambda x + (1-\lambda) y) < \lambda V(x) + (1-\lambda) V(y) \quad \forall \lambda \in]0,1[, \quad \forall x, y \in \mathbb{R}^n.$$

It is shown in [11] that V is a LF. Nevertheless, we need to consider at least $\mathcal{C}^1$ Lyapunov functions; therefore we define

$$\tilde{V}(x) = \int_{SO(n)} f(R)\, V(Rx)\, \mathrm{d}R$$

where $f : SO(n) \longrightarrow [0, +\infty[$ is a smooth function with support on a small neighborhood of the identity matrix and $\int_{SO(n)} f(R)\, \mathrm{d}R = 1$.

In [11], it is also shown that $\tilde{V}$ is a smooth Lyapunov function except at the origin. Moreover, since V is homogeneous of degree 2 and strictly convex, it follows that $\tilde{V}$ also satisfies such properties.

We consider now the function $W(x) = \sqrt{\tilde{V}(x)}$, which is a continuous, positively homogeneous Lyapunov function. Therefore, $W^{-1}(1)$ is a compact set. Using the fact that the set $\{x : W(x) < 1\}$ is strictly convex, we construct a polynomial Lyapunov function $\tilde{W}$ by approximating the level sets of W. For this purpose, we need the following preliminary result, which describes a continuity property of the function $\nabla W(y) \cdot Dx$ with respect to x, y, D.

LEMMA 1. *Let us set*

$$M := \min_{x \in W^{-1}(1)} \min_{D \in \mathbf{A}} \big[-\nabla W(x) \cdot Dx \big].$$

Then, for every $\varepsilon \in (0, M)$, there exists $\delta \in (0,1)$ such that for every $x, y \in W^{-1}(1)$ with $\nabla W(y) \cdot x > 1 - \delta$ and every $D \in \mathbf{A}$, one has

$$\nabla W(y) \cdot Dx < -\varepsilon.$$

Proof of the Lemma. First, notice that M is well defined since it is the infimum of a continuous function over a compact set. Moreover, $M > 0$ because W is a Lyapunov function.

Since, by homogeneity, $\nabla W(y) \cdot y = W(y) = 1$, we have

$$\nabla W(y) \cdot x = 1 - \nabla W(y) \cdot (y - x),$$

and then the hypothesis is equivalent to $\nabla W(y) \cdot (y - x) < \delta$.

Reasoning by contradiction, assume that there exists a sequence (x_j, y_j, D_j) such that $\nabla W(y_j) \cdot D_j x_j \geq -\varepsilon$ and $\nabla W(y_j) \cdot (y_j - x_j)$ converges to 0 as j goes to infinity. By compactness, we can find a subsequence of (x_j, y_j, D_j) converging to $(\bar{x}, \bar{y}, \bar{D})$ and therefore, by continuity, $\nabla W(\bar{y}) \cdot \bar{D}\bar{x} \geq -\varepsilon$ and $\nabla W(\bar{y}) \cdot (\bar{y} - \bar{x}) = 0$.

Therefore $\bar{y} - \bar{x}$ belongs to the tangent space at $\bar{y}$ of the strictly convex set $W^{-1}([0,1])$. Since $\bar{x}$ also belongs to the boundary of that set, it must be $\bar{y} = \bar{x}$. It implies $\nabla W(\bar{y}) \cdot \bar{D}\bar{x} = \nabla W(\bar{x}) \cdot \bar{D}\bar{x} \leq -M$ and we reach a contradiction. ∎

Remark 9. Taking $-x$ instead of x, one obtains that for every $x, y \in W^{-1}(1)$ and every $D \in \mathbf{A}$, then $\nabla W(y) \cdot x < -1 + \delta \Longrightarrow \nabla W(y) \cdot Dx > \varepsilon$.

To conclude the proof of the theorem, we take $\delta \in (0,1)$ corresponding to some ε as in the lemma above, and for every $y \in W^{-1}(1)$ we consider the open sets $B_y = \big\{x \in \mathbb{R}^n : \nabla W(y) \cdot x > 1 - \delta/2\big\}$. Since $y \in B_y$, we have that $\{B_y\}_{y \in W^{-1}(1)}$ is an open covering of the compact set $W^{-1}(1)$, and therefore we can find $y_1, \ldots, y_N$ points of $W^{-1}(1)$ such that the union of B_{y_k}, $k = 1, \ldots, N$, covers $W^{-1}(1)$.

Let us define

$$\tilde{W}(x) := \sum_{k=1}^{N} (\nabla W(y_k) \cdot x)^{2p}.$$

We claim that for an integer p large enough, $\tilde{W}$ is a polynomial Lyapunov function. For $D \in \mathbf{A}$ and $x \in \mathbb{R}^n$, $x \neq 0$, we have

$$\nabla \tilde{W}(x) \cdot Dx = 2p \sum_{k=1}^{N} \big(\nabla W(y_k) \cdot x\big)^{2p-1} \nabla W(y_k) \cdot Dx, \tag{6}$$

and we want to show that $\nabla \tilde{W}(x) \cdot Dx < 0$. By homogeneity, it is enough to do it for $x \in W^{-1}(1)$. Set

$$K := \max_{x,y \in W^{-1}(1),\, D \in \mathbf{A}} \nabla W(y) \cdot Dx.$$

If, for some index k in $\{1, \dots, N\}$, one has $|\nabla W(y_k) \cdot x| \leq 1 - \delta$, then

$$|\big(\nabla W(y_k) \cdot x\big)^{2p-1} \nabla W(y_k) \cdot Dx| \leq (1-\delta)^{2p-1} K.$$

Otherwise, if the inequalities $1 - \delta/2 \geq |\nabla W(y_k) \cdot x| > 1 - \delta$ hold, then, by the previous lemma and remark, one has that the corresponding term in the summation must be negative.

Finally, since by the definition of the points y_k, there exist at least two distinct indices k_1 and k_2 such that $x \in B_{y_{k_1}}$ and $-x \in B_{y_{k_2}}$ we have that

$$\big(\nabla W(y_{k_i}) \cdot x\big)^{2p-1} \nabla W(y_{k_i}) \cdot Dx < -(1-\delta/2)^{2p-1} \varepsilon.$$

Summing up, we deduce that

$$\begin{aligned} \nabla \tilde{W}(x) \cdot Dx &< 2p\Big(-2(1-\delta/2)^{2p-1}\varepsilon + (N-2)(1-\delta)^{2p-1} K \Big) \\ &= -4p(1-\delta/2)^{2p-1}\varepsilon \left(1 - \frac{K(N-2)}{2\varepsilon} \left(\frac{1-\delta}{1-\delta/2} \right)^{2p-1} \right). \end{aligned}$$

For p large enough, the right-hand side of previous expression is negative, uniformly with respect to $D \in \mathbf{A}$ and $x \in W^{-1}(1)$. The theorem is proved.

Remark 10. One can also check that the level set $\tilde{W}^{-1}(1)$ approximates, as p tends to $+\infty$, the corresponding level set of the function $\max_{k=1,\dots,N} |\nabla W(y_k) \cdot x|$ (which is a polytope) and, therefore, the latter is a Lyapunov function as well (cf. [7]).

3. Necessary and sufficient conditions for GUES of bidimensional systems. Consider the following property:

($\mathcal{P}$) the bidimensional switched system given by

$$\dot{x}(t) = u(t)Ax(t) + (1-u(t))Bx(t), \text{ where } u(.) : [0, \infty[\to [0,1], \tag{7}$$

is GUES at the origin.

In this section, together with Appendix A, we recall the main ideas from [8], to get a necessary and sufficient condition on A and B under which $(\mathcal{P})$ holds, or under which we have at least uniform stability. The full statement of the theorem is reported in Appendix A.1.

Remark 11. Recall that by Proposition 1 (proved in Appendix B), the necessary and sufficient condition for stability of the system (7) are the same if we assume $u(.)$ taking values in $\{0,1\}$ or in $[0,1]$, or if we multiply A and B by two arbitrary positive constants.

Set $M(u) := uA + (1-u)B$, $u \in [0,1]$. In the class of constant functions the asymptotic stability of the origin of the system (7) occurs if and only if the matrix $M(u)$ has eigenvalues with strictly negative real part for each $u \in [0,1]$. So this is a necessary condition for GUES. On the other hand it is known that if $[A,B]=0$, then the system (7) is GUES. So, in what follows, we always assume the following conditions:

H1. Let λ_1, λ_2 (resp., λ_3, λ_4) be the eigenvalues of A (resp., B). Then $\mathrm{Re}(\lambda_1)$, $\mathrm{Re}(\lambda_2)$, $\mathrm{Re}(\lambda_3)$, $\mathrm{Re}(\lambda_4) < 0$.

H2. $[A,B] \neq 0$ (that implies that neither A nor B is proportional to the identity).

For simplicity we will also assume the following:

H3. A and B are diagonalizable in $\mathbb{C}$. (Notice that if H2 and H3 hold, then $\lambda_1 \neq \lambda_2$, $\lambda_3 \neq \lambda_4$.)

H4. Let $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{C}P^1$ (resp., $\mathbf{V}_3, \mathbf{V}_4 \in \mathbb{C}P^1$) be the eigenvectors of A (resp., B). Then $\mathbf{V}_i \neq \mathbf{V}_j$ for $i \in \{1,2\}$, $j \in \{3,4\}$. (Notice that, from H2 and H3, the V_i are uniquely defined, $\mathbf{V}_1 \neq \mathbf{V}_2$ and $\mathbf{V}_3 \neq \mathbf{V}_4$, and H4 can be violated only when both A and B have real eigenvalues.)

All the other cases in which H1 and H2 hold are the following:

- A or B are not diagonalizable. This case (in which $(\mathcal{P})$ can be true or false) can be treated with techniques entirely similar to the ones of [8].
- A or B are diagonalizable, but one eigenvector of A coincides with one eigenvector of B. In this case, using arguments similar to those of [8], it possible to conclude that $(\mathcal{P})$ is true.

We will call, respectively, (CC) the case where both matrices have nonreal eigenvalues, (RR) the case where both matrices have real eigenvalues, and (RC) the case where one matrix has real eigenvalues and the other nonreal eigenvalues.

Theorem 3, reported in Appendix A.1, gives necessary and sufficient conditions for the stability of the system (7) in terms of three (coordinates invariant) parameters given in Definition 4. The first two parameters, ρ_A and ρ_B, depend on the eigenvalues of A and B, respectively, and the third parameter $\mathcal{K}$ depends on $Tr(AB)$, which is a Killing-type pseudoscalar product in the space of 2×2 matrices. As explained in [8], the parameter $\mathcal{K}$ contains the interrelation between the two systems $\dot{x} = Ax$ and $\dot{x} = Bx$, and it has a precise geometric meaning. It is in $1:1$ correspondence with the cross ratio of the four points in the projective line $\mathbb{C}P^1$ that corresponds to the four eigenvectors of A and B.

Definition 4. *Let A and B be two 2×2 real matrices and suppose that H1, H2, H3, and H4 hold. Moreover, choose the labels $(_1)$ and $(_2)$ (resp., $(_3)$ and $(_4)$) so that $|\lambda_2| > |\lambda_1|$ (resp., $|\lambda_4| > |\lambda_3|$) if they are real or $\mathrm{Im}(\lambda_2) < 0$ (resp., $\mathrm{Im}(\lambda_4) < 0$) if they are complex. Define*

$$\rho_A := -i\frac{\lambda_1+\lambda_2}{\lambda_1-\lambda_2}; \qquad \rho_B := -i\frac{\lambda_3+\lambda_4}{\lambda_3-\lambda_4}; \qquad \mathcal{K} := 2\frac{Tr(AB)-\frac{1}{2}Tr(A)Tr(B)}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_4)}.$$

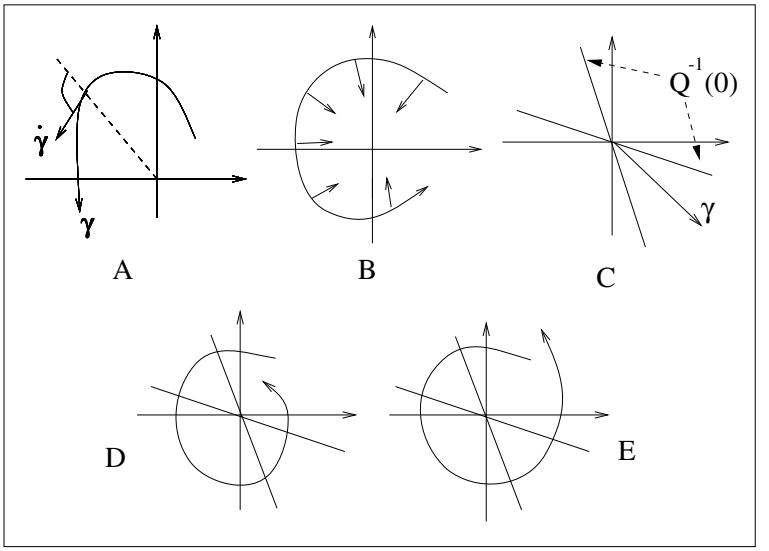


FIG. 1. *Proof of the stability conditions.*

Moreover, define the following function of $\rho_A, \rho_B, \mathcal{K}$*:*

$$\mathcal{D} := \mathcal{K}^2 + 2\rho_A\rho_B\mathcal{K} - (1 + \rho_A^2 + \rho_B^2). \tag{8}$$

Notice that ρ_A is a positive real number if and only if A has nonreal eigenvalues and $\rho_A \in i\mathbb{R}$, $\rho_A/i > 1$ if and only if A has real eigenvalues. The same holds for B. Moreover, $\mathcal{D} \in \mathbb{R}$.

Remark 12. Under hypotheses H1 to H4, using a suitable 3-parameter changes of coordinates, it is always possible to put the matrices A and B, up the their norm (cf. Remark 11), in the normal forms given in Appendix A.2, where $\rho_A, \rho_B, \mathcal{K}$ appear explicitly. (See [8] for more details.)

The parameter $\mathcal{K}$ contains important information about the matrices A and B. They are stated in the following proposition, which can be easily proved using the normal forms given in Appendix A.2.

PROPOSITION 2. *Let* A *and* B *be as in Definition* 4. *Then* (i) *if* A *and* B *have both complex eigenvalues, then* $\mathcal{K} \in \mathbb{R}$ *and* $|\mathcal{K}| > 1$*;* (ii) *if* A *and* B *have both real eigenvalues, then* $\mathcal{K} \in \mathbb{R} \setminus \{\pm 1\}$*;* (iii) A *and* B *have one complex and the other real eigenvalues if and only if* $\mathcal{K} \in i\mathbb{R}$.

Theorem 3, stated in Appendix A.1, is the main result of [8] and gives necessary and sufficient conditions for $(\mathcal{P})$ holding true. We next describe the main idea of the proof. All details can be found in [8].

DEFINITION 5. *The worst trajectory* γ_{x_0} *is the trajectory (based at* x_0*) having the following property. At each time* t, $\dot\gamma_{x_0}(t)$ *forms the smallest angle (in absolute value) with the (exiting) radial direction (Figure* 1A*).*

Then the system (7) is GUES if and only if, for each $x_0 \in \mathbb{R}^2$, the worst trajectory γ_{x_0} tends to the origin. The worst trajectory is constructed as follows. We study the locus $Q^{-1}(0)$ (where $Q(x) := \det(Ax, Bx)$) where the two vector fields Ax and Bx are collinear. The quantity $\mathcal{D}$, defined in Definition 4, is proportional to the discriminant of the quadratic form Q. We have several cases:

- If $Q^{-1}(0)$ contains only the origin, then, in the (CC) and (RC) case, one vector field points always on the same side of the other and the worst trajectory is a trajectory of a fixed vector field (either Ax or Bx). In that case, the system is GUES (case (CC.1) and (RC.1) of Theorem 3); see Figure 1, case B. The situation is similar in case (RR.1) (the worst trajectory tends to the origin).
- If $Q^{-1}(0)$ does not contain only the origin then it is the union of two lines passing through the origin (since Q is a quadratic form). If at each point of $Q^{-1}(0)$, the two vector fields have opposite direction, then there exists a trajectory going to infinity corresponding to a constant switching function (see Figure 1, case C). This correspond to cases (CC.2.1), (RC.2.1) and (RR.2.1) of Theorem 3. In that situation, there exists $u \in [0,1]$ such that the matrix $M(u) := uA+(1-u)B$, $u \in [0,1]$ admits an eigenvalue with positive real part. If at each point of $Q^{-1}(0)$, the two vector fields have the same direction, then the system is GUES if and only if the worst trajectory turns around the origin and after one turn the distance from the origin is decreased (see Figure 1, cases D and E). The quantities $\rho_{CC}, \rho_{RC}, \rho_{RR}$ defined in Theorem 3 (for the three cases (CC), (RC), (RR), resp.) represent the distance from the origin of the worst trajectory (that at time zero is at distance 1), after one half turn. This correspond to cases (CC.2.2), (RC.2.2), and (RR.2.2) of Theorem 3.
- Finally (CC.3), (RC.3), and (RR.3) are the degenerate cases in which the two straight lines coincide.

4. Nonexistence of a uniform bound on the minimal degree of polynomial Lyapunov functions. In this section, we prove Theorem 2. The starting point of the argument is to consider a pair of matrices A and B having both nonreal eigenvalues ((CC) case) and satisfying

$$\mathcal{D} > 0, \quad \mathcal{K} > 1, \quad \rho_{CC} = 1. \tag{9}$$

Such a pair exists. Indeed, Figure 2 translates graphically the contents of Theorem 3 for a fixed $\mathcal{K} > 1$, in the region of the (ρ_A, ρ_B)-plane where $\rho_A, \rho_B > 0$. The open shadowed region corresponds to values of the parameters ρ_A, ρ_B for which the system is GUES. We denote by S^+ the open subset of the shadowed region where $\mathcal{D} > 0$. The curve C represents the limit case where $\rho_{CC} = 1$. To each internal point of that curve, it is associated a system verifying (9), since $\mathcal{D} > 0$. A system corresponding to such a limit case is not asymptotically stable but just stable. Moreover, the worst trajectory (see Definition 5, and its construction in section 3) is a periodic curve, whose support is of class $\mathcal{C}^1$ but not of class $\mathcal{C}^2$ (recall that the switchings occur on $Q^{-1}(0)$, i.e., when the linear vector fields corresponding to A and B are parallel).

Fix a point $(\rho_A, \rho_B) \in C$ corresponding to (A, B). Since C is a subset of the boundary of the open set S^+ in the space of parameters (see Figure 2), there exists a sequence of points $(\rho_{A_k}, \rho_{B_k}) \in S^+$, for $k \geq 1$, converging to (ρ_A, ρ_B). This exactly means that there exists a sequence of GUES pairs (A_k, B_k), $k \geq 1$, such that (A_k, B_k) tends to (A, B) as k goes to ∞.

Let $x = (x_1, x_2)$. For every $k \geq 1$, consider a polynomial Lyapunov function V_k of degree at most m_k, i.e., $V_k = \sum_{1 \leq i+j \leq m_k} a_{ij}^{(k)} x_1^i x_2^j$. Arguing by contradiction, we assume that the sequence (m_k) is bounded by a positive integer m. Up to multiplication by a constant, we can choose $\sum_{1 \leq i+j \leq m_k} |a_{ij}^{(k)}| = 1$. By compactness, there exists a subsequence of (V_k) (still denoted by (V_k)) which converges (uniformly on compact subsets of $\mathbb{R}^2$) to some nonzero polynomial V with degree at most m. Note that $V(0) = 0$ since the V_k are Lyapunov functions.

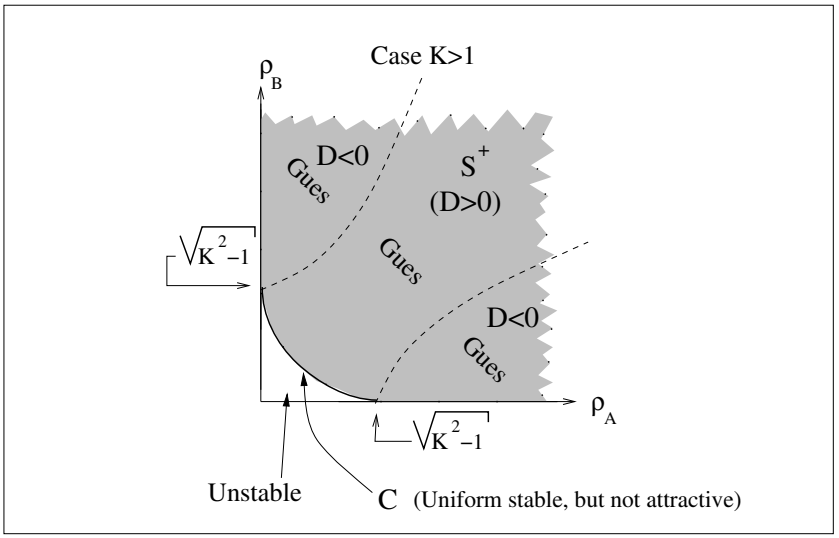


FIG. 2. *GUES property in the space of parameters and explicit construction of a* 5*-parameter SSF for systems of type* (4). S^+ *is the region in the* (ρ_A, ρ_B)*-plane in which the system is GUES and* $\mathcal{D} > 0$.

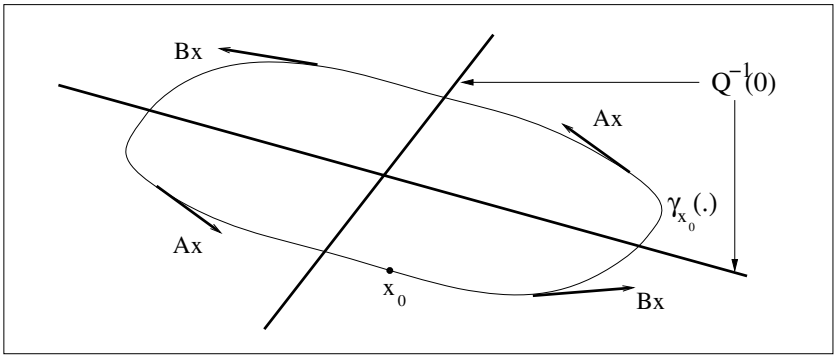


FIG. 3. *The worst trajectory.*

Fix $x_0 \in \mathbb{R}^2$, $x_0 \neq 0$. Let $T > 0$ be the period of the worst trajectory γ_{x_0} corresponding to the pair (A, B) and starting at x_0. Note that T is independent of x_0. The curve $\gamma_{x_0} : [0, T] \to \mathbb{R}^2$ can be seen as the concatenation of at most five arcs of integral curves of $\dot{x} = A\,x$ and $\dot{x} = B\,x$ (see Figure 3) and satisfies the Cauchy problem:

$$\begin{cases} \dot{x} = C(t)\,x, \\ x(0) = x_0, \end{cases}$$

where $C(t)$ is equal to A or B on subintervals of $[0, T]$.

For $k \geq 1$, consider the Cauchy problem:

$$\begin{cases} \dot{x} = C_k(t)\,x, \\ x(0) = x_0, \end{cases}$$

where $C_k(t) = A_k$ if $C(t) = A$ and $C_k(t) = B_k$ if $C(t) = B$. Then, γ_k is a trajectory of the switched system of the type (4) associated to (A_k, B_k). Since, the right-hand side of the previous equation is Lipschitz continuous in x and piecewise continuous in t, then the solutions γ_k converge uniformly to γ_{x_0} on $[0, T]$.

We next show that V remains constant on γ_{x_0}. For $k \geq 1$ and $t \in [0, T]$, one has

$$\|V_k \circ \gamma_k(t) - V \circ \gamma_{x_0}(t)\| \leq \|V_k \circ \gamma_k(t) - V \circ \gamma_k(t)\| + \|V \circ \gamma_k(t) - V \circ \gamma_{x_0}(t)\|.$$

By uniform convergence of V_k to V and of γ_k to γ_{x_0}, and by continuity of V, we deduce that $V_k \circ \gamma_k(t)$ converges to $V \circ \gamma_{x_0}(t)$ for every fixed t.

Since for every $k \geq 1$, V_k is a Lyapunov function for the switched system of the type (4) associated to (A_k, B_k), then $V_k \circ \gamma_k$ is a decreasing function and, hence, $V \circ \gamma_{x_0}$ is nonincreasing. Moreover, $V \circ \gamma_{x_0}(T) = V \circ \gamma_{x_0}(0)$. Therefore, $V \circ \gamma_{x_0}$ must be constant. It implies that there exists $t_1 > 0$ such that either $V(e^{At}x_0)$ or $V(e^{Bt}x_0)$ is constant on $[0, t_1]$. With no loss of generality, assume the first alternative. Since the map $t \mapsto V(e^{At}x_0)$ is real analytic, it follows that $V(e^{At}x_0)$ is constant over the whole real line. By letting t go to $+\infty$, since $e^{At}x_0 \to 0$, we deduce that $V(x_0) = V(0) = 0$. Since x_0 is an arbitrary non zero point of $\mathbb{R}^2$, we get that $V \equiv 0$, which is not possible.

Remark 13. The construction of the sequence (A_i, B_i) with unbounded degree for polynomial Lyapunov function was performed for matrices having both non real eigenvalues (that corresponds to the (CC) case). The same construction can be reproduced for the (RC) and (RR) cases.

Remark 14. For the PPF case (see Remark 8), the above argument can be easily modified to get that the minimum total degree of a piecewise polynomial Lyapunov function cannot be uniformly bounded over the set of all GUES systems of the form Ξ (cf. (4)). Indeed, let V_k be the sequence of PPFs taking the value $V_k^l(x) = \sum_{1 \leq i+j \leq m} a_{ijl}^{(k)} x_1^i x_2^j$ in the cone K_k^l, for $1 \leq l \leq m$. Here, to simplify the notation, we assume without loss of generality that for each element of the sequence, the number of cones and the degree of $V_k^l(x)$ is always m.

Each cone can be identified by a couple of angles with the x_1-direction. Therefore to each function V_k we can associate a m-uple of angles $(\alpha_1^k, \ldots, \alpha_m^k)$ such that the cone K_k^l coincides with the region between the lines corresponding to α_k^l and α_k^{l+1}. In particular, up to subsequences, we can assume that the numbers α_k^l converge to α^l. Similarly to the case above, we can normalize the coefficients of the Lyapunov functions V_k by $\sum_{l=1}^m \sum_{1 \leq i+j \leq m} |a_{ijl}^{(k)}| = 1$ and consider a subsequence of the coefficients converging to a_{ijl}. Then, if we define V as the PPF such that $V(x) = V^l(x) = \sum_{1 \leq i+j \leq m} a_{ijl} x_1^i x_2^j$ on the cone K^l defined by the angles α^l and α^{l+1}, it is easy to verify that $V_k(x)$ converges uniformly on compact subsets of $\mathbb{R}^2$ to $V(x)$. We can conclude the proof as before showing that $V^l(x_0) = V(0) = 0$ for arbitrary x_0, which leads to a contradiction.

5. Explicit construction of a finite-SSF for systems of type (4). In this section, we provide a 5-parameter finite-SSF for the class Ξ of bidimensional systems of type (4). Recall that for what concern the stability issue, Ξ can be parameterized by the 6-parameter family provided by the pairs (A, B) (of 2×2 matrices) defined up to their norm.

As explained in the introduction, it is enough to construct a Lyapunov function for a pair (A, B) giving rise to a GUES system of Ξ. We treat only the (CC.2.2) case since, in all the other cases, the construction is entirely similar.

In the (CC) case, after a 3-parameter change of coordinates, the normal form for the pair (A, B) is given by (cf. Appendix A.2)

$$A = \begin{pmatrix} -\rho_A & -1/E \\ E & -\rho_A \end{pmatrix}, \qquad B = \begin{pmatrix} -\rho_B & -1 \\ 1 & -\rho_B \end{pmatrix}, \qquad E > 0.$$

Moreover, in the (CC.2.2) case, we have $\mathcal{K} > 1$, $\mathcal{D} > 0$, and $\rho_{CC} < 1$, where $\mathcal{K} := 1/2(E + 1/E)$, $\mathcal{D}$ and ρ_{CC} being respectively defined in (4) and (11). Recall that for fixed $\mathcal{K} > 1$ (i.e., fixed $E > 1$), Figure 2 describes, in the (ρ_A, ρ_B)-plane, the status of each point with respect to the GUES issue.

We now associate, to every GUES pair (A, B), a pair $(\tilde{A}, \tilde{B})$ corresponding to a system of the type (4) uniformly stable but not attractive. Consider in Figure 2 the line segment joining the point $(0, 0)$ to (ρ_A, ρ_B) in the S^+ region. That segment intersects the curve C in a point $(\tilde{\rho}_A, \tilde{\rho}_B)$. That results from the Jordan separation theorem and the fact that C connects the points $(\sqrt{\mathcal{K}^2 - 1}, 0)$ and $(0, \sqrt{\mathcal{K}^2 - 1})$. Therefore, there exists a $\zeta \in (0, 1)$ such that, for the system given by

$$\tilde{A} = \begin{pmatrix} -\tilde{\rho}_A & -1/E \\ E & -\tilde{\rho}_A \end{pmatrix}, \qquad \tilde{B} = \begin{pmatrix} -\tilde{\rho}_B & -1 \\ 1 & -\tilde{\rho}_B \end{pmatrix}, \qquad \tilde{\rho}_A = \rho_A \zeta, \qquad \tilde{\rho}_B = \rho_B \zeta,$$

the worst trajectories γ_{x_0} are closed curves, i.e., $\rho_{CC} = 1$.

Moreover, one can easily compute

$$\det(\tilde{A}x, Ax) = \rho_A(1 - \zeta)\left({x_1}^2 E + \frac{{x_2}^2}{E}\right) > 0,$$
$$\det(\tilde{B}x, Bx) = \rho_B(1 - \zeta)|x|^2 > 0.$$

Therefore, the vector fields Ax, Bx point inside the area delimited by a fixed worst trajectory (that is, closed curve) of the modified switched system and so, passing to angular coordinates, the function

$$V(r, \alpha) = \frac{r}{\tilde{r}(\alpha)}, \tag{10}$$

where $\tilde{r}(\alpha)$ is a parameterization of the fixed worst trajectory, is a Lyapunov function for the system defined by (A, B).

Hence, we have provided a 5-parameter SSF in the (CC.2.2) case. The five parameters are $\mathcal{K}$, the ratio ρ_B/ρ_A, and the three parameters involved in the change of coordinates to get the normal forms (14).

Remark 15. Notice that in the cases (CC.1) and (CC.2.1) (cf. section 3 and Theorem 3 in Appendix A.1), one can choose as SSF the set of quadratic polynomials, which actually is parameterized by two parameters.

Remark 16. Let us come back to the general system (3), subject to HO. Notice that the question of finding the smallest m such that there exists an m-parameters finite-SSF, for a certain class C of systems, has no real meaning if one does not require suitable conditions on the map Ψ in Definition 3. Indeed, it is always possible to build a countable SSF for the class of systems of type (3) in $\mathbb{R}^n$ subject to HO.

Appendix A. Stability conditions for bidimensional systems.

A.1. Statement of the stability conditions.

THEOREM 3. *Let A and B be two real matrices such that* H1, H2, H3, *and* H4, *given in section* 3, *hold and define $\rho_A, \rho_B, \mathcal{K}, \mathcal{D}$ as in Definition* 4. *We have the following stability conditions.*

Case (CC) *If A and B have both complex eigenvalues, then*

Case (CC.1). *if $\mathcal{D} < 0$, then $(\mathcal{P})$ is true;*

Case (CC.2). *if $\mathcal{D} > 0$, then*

Case (CC.2.1). *if $\mathcal{K} < -1$, then $(\mathcal{P})$ is false;*

Case (CC.2.2). *if $\mathcal{K} > 1$, then $(\mathcal{P})$ is true if and only if it holds the following condition:*

$$\rho_{CC} := \exp\left[-\rho_A \arctan\left(\frac{-\rho_A\mathcal{K}+\rho_B}{\sqrt{\mathcal{D}}}\right) - \rho_B \arctan\left(\frac{\rho_A-\rho_B\mathcal{K}}{\sqrt{\mathcal{D}}}\right) - \frac{\pi}{2}(\rho_A+\rho_B)\right] \times \sqrt{\frac{(\rho_A\rho_B+\mathcal{K})+\sqrt{\mathcal{D}}}{(\rho_A\rho_B+\mathcal{K})-\sqrt{\mathcal{D}}}} < 1. \tag{11}$$

Case (CC.3). *If $\mathcal{D} = 0$, then $(\mathcal{P})$ holds true or false whether $\mathcal{K} > 1$ or $\mathcal{K} < -1$.*

Case (RC). *If A and B have one of them complex and the other real eigenvalues, define $\chi := \rho_A\mathcal{K} - \rho_B$, where ρ_A and ρ_B are chosen in such a way $\rho_A \in i\mathbb{R}$, $\rho_B \in \mathbb{R}$. Then*

Case (RC.1). *if $\mathcal{D} > 0$, then $(\mathcal{P})$ is true;*

Case (RC.2). *if $\mathcal{D} < 0$, then $\chi \neq 0$ and we have:*

Case (RC.2.1). *if $\chi > 0$, then (P) is false. Moreover, in this case $\mathcal{K}/i < 0$;*

Case (RC.2.2). *if $\chi < 0$, then*

Case (RC2.2.A). *if $\mathcal{K}/i \leq 0$, then $(\mathcal{P})$ is true;*

Case (RC2.2.B). *if $\mathcal{K}/i > 0$, then $(\mathcal{P})$ is true if and only if it holds the following condition:*

$$\rho_{RC} := \left(\frac{m^+}{m^-}\right)^{-\frac{1}{2}(\rho_A/i-1)} e^{-\rho_B\bar{t}} \times \left(\sqrt{1-\mathcal{K}^2}\, m^- \sin\bar{t} - \left(\cos\bar{t} - \frac{\mathcal{K}}{i}\sin\bar{t}\right)\right) < 1, \tag{12}$$

where

$$m^{\pm} := \frac{-\chi \pm \sqrt{-\mathcal{D}}}{(-\rho_A/i-1)\mathcal{K}/i},$$

$$\bar{t} = \arccos\frac{-\rho_A/i + \rho_B\mathcal{K}/i}{\sqrt{(1-\mathcal{K}^2)(1+\rho_B^2)}}.$$

Case (RC.3). *If $\mathcal{D} = 0$, then $(\mathcal{P})$ holds true whether $\chi < 0$ or $\chi > 0$.*

Case (RR). *If A and B have both real eigenvalues, then*

Case (RR.1). *if $\mathcal{D} < 0$, then $(\mathcal{P})$ is true; moreover we have $|\mathcal{K}| > 1$;*

Case (RR.2). *if $\mathcal{D} > 0$, then $\mathcal{K} \neq -\rho_A\rho_B$ (notice that $-\rho_A\rho_B > 1$) and*

Case (RR.2.1). *if $\mathcal{K} > -\rho_A\rho_B$, then (P) is false;*

Case (RR.2.2). *if $\mathcal{K} < -\rho_A\rho_B$, then*

Case (RR.2.2.A). *if $\mathcal{K} > -1$, then (P) is true;*

Case (RR.2.2.B). *if $\mathcal{K} < -1$, then (P) is true if and only if the following condition holds:*

$$\rho_{RR} := -f^{sym}(\rho_A, \rho_B, \mathcal{K}) f^{asym}(\rho_A, \rho_B, \mathcal{K}) \times f^{asym}(\rho_B, \rho_A, \mathcal{K}) < 1, \tag{13}$$

where

$$f^{sym}(\rho_A, \rho_B, \mathcal{K}) := \frac{1 + \rho_A/i + \rho_B/i + \mathcal{K} - \sqrt{\mathcal{D}}}{1 + \rho_A/i + \rho_B/i + \mathcal{K} + \sqrt{\mathcal{D}}};$$

$$f^{aym}(\rho_A, \rho_B, \mathcal{K}) := \left(\frac{\rho_B/i - \mathcal{K}\rho_A/i - \sqrt{\mathcal{D}}}{\rho_B/i - \mathcal{K}\rho_A/i + \sqrt{\mathcal{D}}} \right)^{\frac{1}{2}(\rho_A/i - 1)}.$$

Case (RR.3). *If $\mathcal{D} = 0$, then $(\mathcal{P})$ holds true or false whether $\mathcal{K} < -\rho_A\rho_B$ or $\mathcal{K} > -\rho_A\rho_B$.*

Finally, if $(\mathcal{P})$ is not true, then in case CC.2.2 *with $\rho_{CC} = 1$, case* (RC.2.2.B)*, with $\rho_{RC} = 1$, case* (RR.2.2.B)*, with $\rho_{RR} = 1$, case* (CC.3) *with $\mathcal{K} < -1$, case* (RC.3) *with $\chi > 0$ and case* (RR.3) *with $\mathcal{K} > -\rho_A\rho_B$, the origin is just stable. In the other cases, the system is unstable.*

Remark 17. Formula (13) is a corrected version of Formula (6), p. 93, of [8] and it is proved in Appendix A.3.

A.2. Normal forms of 2×2 matrices.

PROPOSITION 3. *Let A, B be two 2×2 real matrices satisfying conditions* H1*,* H2*,* H3*, and* H4 *given in section* 3*. In the case in which one of the two matrices has real and the other nonreal eigenvalues (i.e., the* (RC) *case), assume that A is the one having real eigenvalues. Then there exists a* 3*-parameter change of coordinates and two constant $\alpha_A, \alpha_B > 0$ such that the matrices A/α_A and B/α_B (still denoted below by A and B) are in the following normal forms*

Case in which A and B have both nonreal eigenvalues ((CC) case):

$$A = \begin{pmatrix} -\rho_A & -1/E \\ E & -\rho_A \end{pmatrix}, \qquad B = \begin{pmatrix} -\rho_B & -1 \\ 1 & -\rho_B \end{pmatrix}, \tag{14}$$

where $\rho_A, \rho_B > 0$, $|E| > 1$. In this case, $\mathcal{K} = \frac{1}{2}(E + \frac{1}{E})$. Moreover, the eigenvalues of A and B are, respectively, $-\rho_A \pm i$ and $-\rho_B \pm i$.

Case in which A has real and B nonreal eigenvalues ((RC) case):

$$A = \begin{pmatrix} -\rho_A/i + 1 & 0 \\ 0 & -\rho_A/i - 1 \end{pmatrix}, \tag{15}$$

$$B = \begin{pmatrix} -\rho_B - \mathcal{K}/i & -\sqrt{1 - \mathcal{K}^2} \\ \sqrt{1 - \mathcal{K}^2} & -\rho_B + \mathcal{K}/i \end{pmatrix}, \tag{16}$$

where $\rho_B > 0$, $\rho_A/i > 1$, $\mathcal{K} \in i\mathbb{R}$. In this case, the eigenvalues of A and B are, respectively, $-\rho_A/i \pm 1$ and $-\rho_B \pm i$.

Case in which A and B have both real eigenvalues ((RR) case):

$$A = \begin{pmatrix} -\rho_A/i + 1 & 0 \\ 0 & -\rho_A/i - 1 \end{pmatrix}, \tag{17}$$

$$B = \begin{pmatrix} \mathcal{K} - \rho_B/i & 1 - \mathcal{K} \\ 1 + \mathcal{K} & -\mathcal{K} - \rho_B/i \end{pmatrix}, \tag{18}$$

where $\rho_A/i, \rho_B/i > 1$ and $\mathcal{K} \in \mathbb{R} \setminus \{\pm 1\}$. In this case, the eigenvalues of A and B are, respectively, $-\rho_A/i \pm 1$ and $-\rho_B/i \pm 1$.

A.3. Proof of formula (12). In this paragraph, we prove formula (12), i.e., in the (RC.2.2.B) case, we determine an inequality defining the set of parameters ρ_A, ρ_B, $\mathcal{K}$ such that the property $(\mathcal{P})$, stated in section 3, holds.

Thanks to Proposition 3 (see also [8, Appendix B, p. 110]), we can find a coordinate transformation such that (up to a rescaling of the matrices) A and B are given by equations (15), (16). In the case (RC.2.2.B), we have $\mathcal{D} := \mathcal{K}^2 + 2\rho_A\rho_B\mathcal{K} - (1 + \rho_A^2 + \rho_B^2) < 0$, $\chi := \rho_A\mathcal{K} - \rho_B < 0$, $\mathcal{K}/i > 0$. Moreover, the set $Q^{-1}(0)$ is the union of two lines passing from the origin and, at each point of $Q^{-1}(0)$, the two vector fields point in the same direction. One easily checks that the slope of the two lines defining $Q^{-1}(0)$ is

$$m^{\pm} = \frac{-\chi \pm \sqrt{-\mathcal{D}}}{(-\rho_A/i - 1)\sqrt{1-\mathcal{K}^2}}.$$

Notice that in our case we have $m^{\pm} < 0$ and $m^+ < m^-$.

In this case, the worst trajectories are concatenations of arcs of integral curves of the vector fields Ax, Bx and rotate counterclockwise around the origin. More precisely, they are integral curves of Ax from the line $x_2 = m^+x_1$ to the line $x_2 = m^-x_1$ and integral curves of Bx otherwise.

Therefore, starting from the point $\binom{1}{m^+}$ (with the field Ax), we follow the worst trajectory until it touches again the line $x_2 = m^+x_1$. Property $(\mathcal{P})$ is then satisfied if and only if $\rho_{RC} < 1$, where ρ_{RC} is the absolute value of the first coordinate of the final point.

One can easily compute that the first switching time is $t_1 = \frac{1}{2}\log\frac{m^+}{m^-}$, which is positive since $\frac{m^+}{m^-} > 1$. Moreover, the integral curve of Bx starting from the point $\binom{1}{m^-}$ is

$$e^{-\rho_B t}\begin{pmatrix} -\sqrt{1-\mathcal{K}^2}\, m^- \sin t + (\cos t - \frac{\mathcal{K}}{i}\sin t) \\ \sqrt{1-\mathcal{K}^2}\sin t + m^-(\cos t + \frac{\mathcal{K}}{i}\sin t)\end{pmatrix},$$

and, setting the ratio between the second coordinate and the first one equal to m^+, one obtains that the second switching time is $t_2 = \arccos\frac{-\rho_A/i + \rho_B\mathcal{K}/i}{\sqrt{(1-\mathcal{K}^2)(1+\rho_B^2)}}$. Notice that t_2 is well defined if and only if $\mathcal{D} < 0$ (condition which is satisfied in our case). Moreover, t_2 is positive and less than π. Finally, the inequality we was seeking for is

$$\rho_{RC} = \left(\frac{m^+}{m^-}\right)^{-\frac{1}{2}(\rho_A/i-1)} e^{-\rho_B t_2}\left(\sqrt{1-\mathcal{K}^2}\, m^- \sin t_2 - \left(\cos t_2 - \frac{\mathcal{K}}{i}\sin t_2\right)\right) < 1\,.$$

Appendix B. The stability properties of (3) depend only on the convex hull of A. We provide here the proof of Proposition 1. First, let us show the following.

Claim. Consider the switched system (3), under HO, and let $\mathbf{A}'$ be a measurable subset of $\mathbf{A}$ such that the convex hull of $\mathbf{A}'$ contains $\mathbf{A}$. Then the following two conditions are equivalent:

(i) the system is GUES (resp., uniformly stable), with $A(.)$ measurable, taking values in $\mathbf{A}$;

(ii) the system is GUES (resp., uniformly stable), with $A(.)$ measurable, taking values in $\mathbf{A}'$.

Proof of the Claim. Let $\boldsymbol{\mathcal{A}}$ (resp., $\boldsymbol{\mathcal{A}}'$) be the set of measurable functions $A(.) : [0,\infty[\to \mathbf{A}$ (resp., $A(.) : [0,\infty[\to \mathbf{A}'$). Since $\mathbf{A}'$ is contained in $\mathbf{A}$, then the implication (i) $\Rightarrow$ (ii) is obvious.

Let us prove the other implication (which is strictly related to the classical approximability theorems in control theory). We start considering uniform stability. By contradiction, assume that we can find $\epsilon > 0$ satisfying the following. There exists a sequence of points (x_l) tending to zero and a sequence of controls $A_l(.) \in \boldsymbol{\mathcal{A}}$ such that the corresponding trajectory γ_l starting at x_l exits the interior of the ball of radius ϵ for some time t_l. Using classical approximability results (see, for instance, [2]), the trajectory γ_l can be approximated in the L^∞-norm on $[0, t_l]$ by a trajectory γ'_l corresponding to a switching function $A'_l(.) \in \boldsymbol{\mathcal{A}}'$ and starting at x_l. Hence γ'_l exits the interior of the ball of radius $\epsilon/2$ at time t_l. We reached a contradiction.

Now we want to prove that GUES holds in the case $A(.) \in \boldsymbol{\mathcal{A}}'$ implies GUES holds in the case $A(.) \in \boldsymbol{\mathcal{A}}$. Since $\mathbf{A}$ is compact, we know (see Definition 1 and below) that attractivity and GUES are equivalent for the corresponding switched system. Therefore, proceeding by contradiction, we can assume that there is a trajectory $\gamma(.)$ of the switched system corresponding to $A(.) \in \boldsymbol{\mathcal{A}}$ not converging to zero. That means that there exist $\epsilon > 0$ and a sequence t_n of times tending to infinity such that $|\gamma(t_n)| > \epsilon$. As before, we can approximate $\gamma(.)$ on the interval $[0, t_n]$ with a trajectory $\gamma_n(.)$ corresponding to controls taking values in $\mathbf{A}'$, in such a way that $|\gamma_n(t_n)| > \epsilon/2$. But this is impossible since we have assumed GUES for the switched system with $A(.) \in \boldsymbol{\mathcal{A}}'$. □

Notice that one can provide an alternative argument for the GUES part of Proposition 1, by using Lyapunov functions.

Then one immediately extends to semicones, observing that the stability properties of the system (3), subject to the compactness hypothesis HO, depend only on the shape of the trajectories and not on the way in which they are parameterized.

Acknowledgments. The authors are grateful to Franco Blanchini for bringing to their attention the references [15, 16] and for many suggestions.

REFERENCES

[1] A. A. Agrachev and D. Liberzon, *Lie-algebraic stability criteria for switched systems*, SIAM J. Control Optim., 40 (2001), pp. 253–269.

[2] A. A. Agrachev and Yu. L. Sachkov, *Control Theory from the Geometric Viewpoint*, Encyclopedia of Math. Sci. 87, Springer-Verlag, Berlin, 2004.

[3] D. Angeli, *A note on stability of arbitrarily switched homogeneous systems*, Systems Control Lett., to appear.

[4] D. Angeli and D. Liberzon, *A note on uniform global asymptotic stability of nonlinear switched systems in triangular form*, in Proceedings of the 14th International Symposium on Mathematical Theory of Networks and Systems, 2000.

[5] A. Bacciotti and F. Ceragioli, *Closed loop stabilization of planar bilinear switched systems*, Internat J. Control, to appear.

[6] F. Blanchini, *Nonquadratic Lyapunov function for robust control*, Automatica J. IFAC, 31 (1995), pp. 451–461.

[7] F. Blanchini and S. Miani, *A new class of universal Lyapunov functions for the control of uncertain linear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 641–647.

[8] U. Boscain, *Stability of planar switched systems: The linear single input case*, SIAM J. Control Optim., 41 (2002), pp. 89–112.

[9] Z. Bubnicki, *Analysis and Decision Making in Uncertain Systems*, Comm. Control Engrg. Ser., Springer-Verlag, London, 2004.

[10] G. Chesi, A. Garulli, A. Tesi, and A. Vicino, *Homogeneous Lyapunov functions for systems with structured uncertainties*, Automatica J. IFAC, 39 (2003), pp. 1027–1035.

[11] W. P. Dayawansa and C. F. Martin, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.

[12] D. Liberzon, J. P. Hespanha, and A. S. Morse, *Stability of switched systems: A lie-algebraic condition*, Systems Control Lett., 37 (1999), pp. 117–122.

[13] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems Magazine, 19 (1999), pp. 59–70.

[14] D. Liberzon, *Switching in Systems and Control*, Systems Control Found. Appl., Birkhäuser, Boston, 2003.

[15] A. P. Molchanov and E. S. Pyatnitskii, *Lyapunov functions that specify necessary and sufficient conditions of absolute stability of nonlinear nonstationary control systems. Part* I, Autom. Remote Control, 47 (1986), pp. 344–354.

[16] A. P. Molchanov and E. S. Pyatnitskii, *Lyapunov functions that specify necessary and sufficient conditions of absolute stability of nonlinear nonstationary control systems. Part* II, Autom. Remote Control, 47 (1986), pp. 443–451.

[17] A. P. Molchanov and E. S. Pyatnitskii, *Criteria of asymptotic stability of differential and difference inclusions encountered in control theory*, Systems Control Lett., 13 (1989), pp. 59–64.

[18] A. Olas, *On robustness of systems with structured uncertainties*, in Proceedings of the 4th Workshop on Control Mechanics University of Southern California, 1991.

[19] R. Shorten and K. Narendra, *Necessary and sufficient conditions for the existence of a common quadratic Lyapunov function for two stable second order linear time-invariant systems*, in Proceedings of the American Control Conference, 1999, pp. 1410–1414.

[20] X. Xu and P. J. Antsaklis, *Stabilization of second order LTI switched systems*, Internat. J. Control, 73 (2000), pp. 1261–1279.

[21] A. L. Zelentsovsky, *Nonquadratic Lyapunov functions for robust stability analysis of linear uncertain systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 135–138.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 246–272

AN EVOLUTION EQUATION APPROACH TO NONAUTONOMOUS LINEAR SYSTEMS WITH STATE, INPUT, AND OUTPUT DELAYS*

SAID HADD†

Abstract. We propose a new approach which brings nonautonomous linear systems with state, input, and output delays in the line with the standard theory of nonautonomous linear systems. To this purpose, we establish, using the concept of Lebesgue extensions, a new variation of constants formula for nonhomogenous delay equations. From this we deduce another new one for nonautonomous linear systems with state and input delays. Inspiriting from this formula we show that a given delay system determines a nonautonomous absolutely regular linear system with the same input and output spaces as of the delay system. Our abstract results will be applied to investigate the existence and the uniqueness of the solutions for a class of nonautonomous neutral equations in Banach spaces.

Key words. nonautonomous delay equations, delay systems, nonautonomous regular systems, Lebesgue extension, representation, neutral equations

AMS subject classifications. 34K35, 34K17, 34K40, 47D06, 93C05, 93C23

DOI. 10.1137/040612178

1. Introduction.

The delay systems (also called hereditary or systems with aftereffects) represent a class of infinite-dimensional systems largely used to describe propagation phenomena or population dynamics. Roughly speaking, the reaction of real-world systems to exogenous signals is never "instantaneous" and it needs some time, time which can be translated into a mathematical language by some delay terms. A distinguished feature of this class of systems is that their evolution rate is described by differential equations which include information on the past history. Into a mathematical framework, such systems may be described in several ways, and we mention, for example, differential equations on abstract spaces, over rings of operators or functional differential equations. We note that the delay effects on the stability and control of dynamical systems (delays in the state and/or in the input) are problems of recurring interest since the delay presence may induce complex behaviors (oscillations, instability, bad performances) for the (closed-loop) schemes.

The purpose of the current paper is to introduce an evolution equation approach which brings nonautonomous linear systems with delays in the state, control, and observation variables in line with the standard theory of nonautonomous systems developed recently by Schnaubelt [26], a nonautonomous version of the known abstract linear systems theory introduced mainly by Salamon [25] and Weiss [32]. We note that the results in [26] extend those obtained in the works by Hinrichsen and Pritchard [14], Jacob [16, 17], and Jacob, Dragan, and Pritchard [18].

In this paper we give conditions on delay operators so as to reformulate the delay systems as absolutely regular nonautonomous systems. In particular, this serves to establish a new representation of the state, as well as of the output function for delay systems in terms of some extensions of the delay operators. We show how this approach leads also to introduce a new definition of the solution of input delay equations.

*Received by the editors July 24, 2004; accepted for publication (in revised form) September 15, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sicon/45-1/61217.html
†Department of Mathematics, Cadi Ayyad University, Faculty of Sciences Semlalia, B.P.: 2390, 40000 Marrakesh, Morocco (hadd@ucam.ac.ma).

A quite complete theory of representation for autonomous delay systems exists. We quote the classical work by Ichikawa [15] and Pritchard and Salamon [23] as well as the early papers by Delfour [7], Vinter and Kwong [29] and the books by Bensoussan et al. [3, Chap. 4] and Curtain and Zwart [6, Chap. 4.2]. In [23], the authors showed that autonomous delay systems are represented as Pritchard–Salamon systems (see, e.g., [5]), a strict subclass of that of Salamon and Weiss. Likewise we note that in [10, 11] we introduced a large class of autonomous delay systems which can be reformulated as Salamon–Weiss systems.

In the nonautonomous case, we have already started the investigation of delay systems; see [12]. There, we studied the robustness of the (absolute) regularity of a nonautonomous system under perturbation by appropriate state and input delay operators.

In the current paper a representation of nonautonomous linear systems with state, input, and output delay is considered. Our approach is based on a study of absolutely regular systems.

Before summarizing precisely our results, we first fix some notation which will be used throughout the paper. The state space X, the control space U, and the observation space Y are supposed to be Banach spaces, and $r > 0$ is a real number. For a Banach space Z, if $z : [-r, \infty) \to Z$, then its history is the function $z_t : [-r, 0] \to Z$ defined by $z_t(\theta) = z(t + \theta),\ t \geq 0$. We denote by $C([-r, 0], Z)$, the space of Z-valued continuous functions on $[-r, 0]$, $L^2([-r, 0], Z)$ the customary Lebesgue space of square integrable Z-valued functions on $[-r, 0]$, $W^{1,2}([-r, 0], Z)$ the Sobolev space, and $L^2_{loc}(J, Z)$ is the space of all functions $\phi : J \to Z$ locally 2-integrable endowed with the usual Fréchet topology, where $J \subset \mathbb{R}$ is a closed interval. Hereafter, the function $f \in L^2_{loc}([s, \infty), X)$ for $s \geq 0$ will denote the nonhomogenous term for evolution equations. If $(S_Z(t))_{t\geq 0}$ is the left shift semigroup on $L^2([-r, 0], Z)$ (see section 3), then $\mathcal{S}_Z := (\mathcal{S}_Z(t, s))_{t\geq s}$ is a strongly continuous evolution family on this space, where we set $\mathcal{S}_Z(t, s) = S_Z(t - s),\ t \geq s$.

Let us consider the nonhomogenous (state) delay equation

$$\dot{x}(t) = A(t)x(t) + L(t)x_t + f(t), \quad x(s) = \eta, \quad x_s = \varphi, \quad t \geq s \geq 0. \tag{1.1}$$

If $A(t)$ are linear operators on X which satisfy Aquistapace and Terreni assumptions (see [2] and the comments on evolution equations in section 4), then generate a strongly continuous evolution family $T := (T(t, s))_{t\geq s}$ on X; see [1]. Moreover, $L(t)$ are bounded linear operators from $C([-r, 0], X)$ into X, expressed in terms of a kernel $\ell(\cdot, \cdot) : [-r, 0] \times \mathbb{R}_+ \to \mathcal{L}(X),\ t \geq 0$, satisfying the assumption (H); see section 3. In [12], it is shown that (H) implies that $L(t)$ are nonautonomous observation operators for $\mathcal{S}_X$. We denote by $\tilde{L}(t)$ the *Lebesgue extensions* of $L(t)$ with respect to $\mathcal{S}_X$ (see (2.5)). We shall see that for initial conditions $\eta \in X$ and $\varphi \in L^2([-r, 0], X)$, the equation (1.1) has a solution $x : [s - r, \infty) \to X$ satisfying $x_t \in D(\tilde{L}(t))$ for almost everywhere (a.e.) $t \geq s \geq 0,\ \tilde{L}(\cdot)x_\cdot \in L^2_{loc}([s, \infty), X)$, and

$$x(t) = T(t, s)\eta + \int_s^t T(t, \tau)[\tilde{L}(\tau)x_\tau + f(\tau)]\, d\tau, \quad t \geq s. \tag{1.2}$$

What seems surprising is that if $x : [s - r, \infty) \to X$ is the solution of (1.1) corresponding to the initial conditions $\eta \in X$ and $\varphi \in C([-r, 0], X)$ such that $\eta = \varphi(0)$, then the formula (1.2) coincides with the usual one in the C-setting (see, e.g., [9]). In fact, as we shall see in section 4, in this case we have $\tilde{L}(t)x_t = L(t)x_t$ for almost every $t \geq s$. This motivates us to call (1.2) the *mild solution* of (1.1). In what follows we

will indicate the importance of the formula (1.2) when we study the solution of input delay equations.

Let us consider a control function $u \in L^2([s-r,\infty),U)$ such that $u_s = \zeta$ for initial time $s \geq 0$. We now consider the following delay system:

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + L(t)x_t + K(t)u_t, \quad t \geq s \geq 0, \\ &x(s) = \eta, \quad x(t) = \varphi, \quad u(t) = \zeta, \quad t \leq s, \end{aligned} \tag{1.3}$$

$$y(t) = G(t)x_t + D(t)u_t. \tag{1.4}$$

Here the delay operators $K(t) : C([-r,0],U) \to X$, $G(t) : C([-r,0],X) \to Y$ and $D(t) : C([-r,0],U) \to Y$ are defined by kernels $k(\cdot,\cdot)$, $g(\cdot,\cdot)$, and $d(\cdot,\cdot)$ satisfying the condition (H).

If one regards (1.3) as a nonhomogenous equation of the form (1.1), then to use the abstract framework cited above it is necessary that the X-valued function $f(t) := K(t)u_t$, $t \geq s \geq 0$, be locally 2-integrable. This is not clear since $K(t)$ are not bounded on $L^2([-r,0],U)$. However, for *smooth input,* i.e., $u \in W^{1,2}_{loc}([s-r,\infty),U)$ with $u(s) = 0$, we have proved, in fact, that f is well defined and belongs to $L^2_{loc}([s,\infty),X)$ (see section 3). Thus, due to the aforementioned result, the input delay equation (1.3) has a mild solution satisfying (1.2).

Analyzing in profile what happens when the input u is not smooth, one has to make sense to "$K(t)u_t$" for a.e. $t \geq s \geq 0$. Observe that $v(t,\cdot) := u_t$ is the state trajectory of the boundary control system $\frac{\partial}{\partial t}v(t,\cdot) = \frac{\partial}{\partial \theta}v(t,\cdot)$, $v(t,0) = u(t)$ for $t \geq s \geq 0$. Now, due to the rich existing theory of (absolutely) regular nonautonomous systems (see section 2), one considers $f(t) = K(t)u_t$, $t \geq s$, as the observation equation associated with this boundary control system, where we denote the obtained system by (BS). In section 3, we have proved that the hypothesis (H) on $K(t)$ implies that (BS) determines an absolutely regular nonautonomous system on the state space $L^2([-r,0],U)$, the control space U, and the observation space X. Now due to Theorem 2.3, we have $u_t \in D(\tilde{K}(t))$ for a.e. $t \geq s$ and $\tilde{f}(t) := \tilde{K}(t)u_t$ (as we shall see, this is the extension of f) belongs to $L^2_{loc}([s,\infty),X)$, where $\tilde{K}(t)$ are the Lebesgue extensions of $K(t)$ with respect to the evolution family $\mathcal{S}_U$. We thus apply formula (1.2) to $\tilde{f}$ and we get a new solution $x : [s-r,\infty) \to X$ satisfying $x_t \in D(\tilde{L}(t))$ for a.e. $t \geq s$, $\tilde{L}(\cdot)x_\cdot \in L^2_{loc}([s,\infty),X)$ and

$$x(t) = T(t,s)\eta + \int_s^t T(t,\tau)[\tilde{L}(\tau)x_\tau + \tilde{K}(\tau)u_\tau]\,d\tau, \quad t \geq s. \tag{1.5}$$

If we restrict ourselves to the initial condition $\varphi \in C([-r,0],X)$ such that $\eta = \varphi(0)$ and the smooth input u, then we have established that $\tilde{L}(t)x_t = L(t)x_t$ and $\tilde{K}(t)u_t = K(t)u_t$ for almost every $t \geq s$. This suggests that we call the function x satisfying (1.5) a *generalized solution* of (1.3). By density, every input u can approximated by a sequence of smooth inputs u^n. Now if x^n is the mild solution of (1.3) corresponding to u^n, then we shall see that the generalized solution of (1.3) corresponding to u satisfies $x^n \to x$ in $L^2_{loc}([s-r,\infty),X)$.

However, the formula for the generalized (and then the mild) solution still contains a delay term. In order to obtain a delay-free representation for the input delay equation (1.3), we have transformed (1.3) into a nonautonomous control system without delay $(\mathcal{T}_{L,K},\Phi_{L,K})$ on the new state space $\mathcal{X} := X \times L^2([-r,0],X) \times L^2([-r,0],U)$ and the control space U, by adding the history of the control to the dynamics of the system. Moreover, the state trajectory and the control function of the new system

are $z(t) = (x(t), x_t, u_t)$ and $u(t)$ for $t \geq s \geq 0$, respectively, where x is the generalized solution of (1.3). We thus see that there is a natural connection between the state z and the delayed observation equation (1.4). This suggests to regard y as the observation equation associated with the new control system. Typically, we have transformed the delay system (1.3)–(1.4) into an absolutely regular nonautonomous system $\Sigma_{L,K}$ on the state space $\mathcal{X}$, the control space U, and the observation space Y. This transformation allows us to prove that the output equation (1.4) is represented as $y(t) = \tilde{G}(t)x_t + \tilde{D}(t)u_t$ for almost every $t \geq s$, where $\tilde{G}(t)$ (resp., $\tilde{D}(t)$) are the Lebesgue extensions of $G(t)$ (resp., $D(t)$) with respect to $\mathcal{S}_X$ (resp., $\mathcal{S}_U$).

If in (1.3)–(1.4) we take $L(t) = 0$, $G(t) = \delta_0$, $X = Y = U$ and invoking the feedback law '$u(t) = y(t)$', then the delay system corresponds to the usual nonautonomous neutral equations (see, e.g., [13, Chap. 9]). Our representation theory allows us to introduce a kind of solutions for such neutral equations (see Definition 6.1). On the other hand, we show that the condition (H) implies that $\Delta(t) = I_X$ are admissible feedback operators for $\Sigma_{0,K}$, which give sense to the aforementioned feedback law. In view of our representation theorem (see Theorem 5.8) for nonautonomous delay systems and the feedback theory for nonautonomous systems (see section 2) we prove that the nonautonomous neutral equation has a unique (generalized) solution (see Theorem 6.3).

Section 2 is devoted to preliminaries from (absolutely) regular nonautonomous systems. Section 3 is on the study of the absolute regularity of a boundary nonautonomous system closely related to the left shift semigroup. In section 4 we investigate mild and generalized solutions of nonhomogenous delay evolution equations. In section 5 we prove that the delay system (1.3)–(1.4) determines an absolutely regular nonautonomous system. The last section is concerned with the study of a class of parabolic nonautonomous neutral equations.

2. Background on regular nonautonomous systems. Since our results are based on the theory of (absolutely) regular nonautonomous systems, we review the relevant theory from Schnaubelt [26]. We note that the results in [26] extend those obtained in the works by Hinrichsen and Pritchard [14], Jacob [16, 17], and Jacob, Dragan, and Pritchard [18]. We equally mention that the paper [26] can be regarded as the nonautonomous version of the Weiss works [30, 31, 32] and the Staffans book [28, Chaps. 5 and 7] on well-posed linear systems.

A *(strongly continuous) evolution family* on X is a set $T = (T(t,s))_{t \geq s \geq 0} \subset \mathcal{L}(X)$ (the space of bounded linear operators on X) such that

(i) $T(t,s) = T(t,\rho)T(\rho,s)$, $T(s,s) = Id$,

(ii) for each $x \in X$ the map $(t,s) \mapsto T(t,s)x$ is continuous for all $t \geq s \geq 0$, and

(iii) $\|T(t,s)\| \leq Me^{\omega(t-s)}$

for all $t \geq \rho \geq s \geq 0$ and constants $M \geq 1$ and $\omega \in \mathbb{R}$.

This notion of evolution families arises naturally from the theory of evolution equations which are well posed (see, e.g., [4, Chap. 2], [8, sect. VI.9], [22], [27]). That is, the evolution family arises as the solution operator of the well-posed nonautonomous evolution equation $\dot{x}(t) = A(t)x(t)$, $x(s) = \eta \in X$, for $t \geq s \geq 0$, where $A(t)$ is (in general) an unbounded linear operator for every fixed t. We mention that generally, the function $T(t,s)\eta$, as a function of t, is not differential. However, the differentiability of such a function is guaranteed if $A(t)$ satisfy additional conditions. (See section 4 for some comments and further references.) Moreover, we are concerned here with the notion of evolution families rather than that of evolution equations involving concrete differential operators.

For an evolution family T, we set

$$(\mathbb{K}_s^T f)(t) := \int_s^t T(t,\tau) f(\tau)\, d\tau$$

for all $t \geq s \geq 0$ and $f \in L^2_{loc}([s,\infty), X)$.

The pair $(T,\Phi) := (T, \{\Phi(t,s) : t \geq s \geq 0\})$ is called a *nonautonomous control system* (on X, U) if $\Phi(t,s) : L^2_{loc}([s,\infty), U) \to X,\ t \geq s \geq 0$, are linear operators such that

$$\begin{aligned} &\Phi(t,s)u = \Phi(t,\rho)(u \mid [\rho,\infty)) + T(t,\rho)\Phi(\rho,s)u, \quad t \geq \rho \geq s \geq 0, \\ &\|\Phi(t,s)u\|_X \leq \beta \|u\|_{L^2([s,t],U)}, \quad 0 \leq t-s \leq t_0, \end{aligned} \tag{2.1}$$

for $u \in L^2_{loc}([s,\infty), U),\ t_0 > 0$, and constant $\beta = \beta(t_0) > 0$.

Let $s \geq 0,\ u \in L^2_{loc}([s,\infty), U)$. Then the function $x : [s,\infty) \to X$ defined by $x(s) = \eta$ and

$$x(t) = T(t,s)\eta + \Phi(t,s)u, \quad t \geq s, \tag{2.2}$$

is called the *state trajectory* of the control system (T,Φ).

Let $\Psi(s) : X \longrightarrow L^2_{loc}([s,\infty), Y),\ s \geq 0$, be linear operators satisfying

$$\Psi(s)\eta = \Psi(t)T(t,s)\eta \quad \text{on} \quad [t,\infty) \quad \text{and} \quad \int_s^{s+t_0} \|(\Psi(s)\eta)(t)\|^2\, dt \leq \gamma^2 \|\eta\|^2 \tag{2.3}$$

for $t \geq s \geq 0$, $\eta \in X$, $t_0 > 0$, and constant $\gamma = \gamma(t_0) > 0$. In this case we say that $(T,\Psi) := (T, \{\Psi(s),\ s \geq 0\})$ is a *nonautonomous observation system* (on X, Y).

For linear operators $C(s) : D(C(s)) \subseteq X \to Y,\ s \geq 0$, we define the set

$$\begin{aligned} \mathcal{D}_s(C(\cdot)) := \{f \in L^2_{loc}([s,\infty), X) : f(t) \in D(C(t)) \ \text{ for a.e. } \ t \geq s, \\ C(\cdot)f(\cdot) \in L^2_{loc}([s,\infty), Y)\}. \end{aligned}$$

Let $C(s) : D(C(s)) \subseteq X \longrightarrow Y,\ s \geq 0$, be densely defined linear operators such that $T(\cdot,s)\eta \in D_s(C(\cdot))$ and

$$\int_s^{s+t_0} \|C(t)T(t,s)\eta\|^2\, dt \leq \gamma^2 \|\eta\|^2$$

for $s \geq 0$, $t_0 > 0$, $\eta \in D(C(s))$, and a constant $\gamma = \gamma(t_0) > 0$. Then we say that $C(s),\ s \geq 0$, are *admissible observation operators* for T. Note that the admissibility of $C(t)$ for T implies that the mapping

$$\Psi(s) : D(C(s)) \to L^2_{loc}([s,\infty), Y), \quad \Psi(s)\eta := C(\cdot)T(\cdot,s)\eta, \quad s \geq 0, \tag{2.4}$$

possess unique extensions (again denoted by $\Psi(s)$) to linear continuous operators from X to $L^2_{loc}([s,\infty), Y)$ which yield a nonautonomous observation system; see [26, Lem. 2.5].

Conversely, let (T,Ψ) be a nonautonomous observation system and $s \geq 0$. We define the operators

$$\begin{aligned} D(\tilde{C}(s)) &:= \Big\{\eta \in X : \lim_{\tau \searrow 0} \frac{1}{\tau} \int_s^{s+\tau} (\Psi(s)\eta)(\sigma)\, d\sigma \quad \text{exists in Y}\Big\}, \\ \tilde{C}(s)\eta &:= \lim_{\tau \searrow 0} \frac{1}{\tau} \int_s^{s+\tau} (\Psi(s)\eta)(\sigma)\, d\sigma. \end{aligned} \tag{2.5}$$

We say that $\tilde{C}(\cdot)$ *represent* (T,Ψ) or that $\tilde{C}(t)$ are the *Lebesgue extensions* of $C(t)$ with respect to T if $\Psi(s)$ is given by (2.4).

The following result, due to Schnaubelt [26, Thm. 2.7], can be regarded as a nonautonomous version of the well-known Weiss representation theorem [30].

THEOREM 2.1. *Let (T,Ψ) be a nonautonomous observation system on X,Y. Let $\tilde{C}(t)$, $t\geq 0$, be defined as in (2.5). Then $T(\cdot,s)\eta\in\mathcal{D}_s(\tilde{C}(\cdot))$ and $\Psi(s)\eta=\tilde{C}(\cdot)T(\cdot,s)\eta$ for $s\geq 0$ and $\eta\in X$.*

The following proposition will be used constantly throughout this paper (see [26, Prop. 2.11] and its proof).

PROPOSITION 2.2. *Let (T,Ψ) be a nonautonomous observation system represented by $\tilde{C}(t)$. Then $\mathbb{K}_s^T f\in\mathcal{D}_s(\tilde{C}(\cdot))$ and*

$$\|\tilde{C}(\cdot)\mathbb{K}_s^T f\|_{L^2([s,s+t_0],Y)}\leq c\ t_0^{\frac{1}{2}}\ \|f\|_{L^2([s,s+t_0],X)}$$

for $s\geq 0$, $0<t_0\leq t_1$, $f\in L^2_{loc}(\mathbb{R}_+,X)$, and a constant $c=c(t_1)>0$.

Let (T,Φ) and (T,Ψ) be nonautonomous control and observation systems. If there are linear operators $\mathbb{F}(s):L^2_{loc}([s,\infty),U)\to L^2_{loc}([s,\infty),Y)$ satisfying

$$\mathbb{F}(s)u=\Psi(t)\Phi(t,s)u+\mathbb{F}(t)(u\mid[t,\infty)) \quad \text{on } [t,\infty), \tag{2.6}$$

$$\|\mathbb{F}(s)u\|_{L^2([s,s+t_0],Y)}\leq\kappa\|u\|_{L^2([s,s+t_0],U)} \tag{2.7}$$

for $u\in L^2_{loc}([s,\infty),U)$, $t\geq s\geq 0$, $t_0>0$, and a constant $\kappa=\kappa(t_0)>0$, then $\Sigma=(T,\Phi,\Psi,\mathbb{F})$ is called a *well-posed nonautonomous system* (on X,U,Y) with *input-output operators* $\mathbb{F}(s)$. We note that $\mathbb{F}(s)u=0$ on $[s,t]$ and $\mathbb{F}(s)u=\mathbb{F}(t)(u|[t,\infty))$ on $[t,\infty)$ is u vanishes on $[s,t]$. Then one defines the following restrictions:

$$\mathbb{F}(s)|[s,t]=:\mathbb{F}(t,s):L^2([s,t],U)\to L^2([s,t],Y),\qquad t\geq s\geq 0.$$

Let $\Sigma=(T,\Phi,\Psi,\mathbb{F})$ be a well-posed nonautonomous system with state trajectory $x:[s,\infty)\to X$ (see (2.2)) and control function $u\in L^2([s,\infty),U)$ for initial time $s\geq 0$. In the rest of this section we assume that Ψ is given by nonautonomous admissible observation operators $C(t)$ as in (2.4), so we denote by $\tilde{C}(t)$ the Lebesgue extensions of $C(t)$ with respect to T. We now consider the system

$$\text{(CS)}\quad\begin{cases}x(t)=T(t,s)\eta+\Phi(t,s)u, & t\geq s\geq 0,\\ y(t)=C(t)x(t).\end{cases}$$

We say that y is the *observation function* of (CS) (or Σ). When $\eta=0$, the operators $\mathbb{F}(s)$ relate the input u to the output y (i.e., $u\to y$). Since $C(t)$ are unbounded operators, then $y(t)$ has a sense only if $\mathbb{F}(t)$ satisfy additional regularity properties.

A well-posed nonautonomous linear system $(T,\Phi,\Psi,\mathbb{F})$ is called *regular* (with feedthrough $D=0$) if

$$Y-\lim_{\tau\searrow 0}\frac{1}{\tau}\int_t^{t+\tau}(\mathbb{F}(t)u_z)(\sigma)\,d\sigma=0$$

and *absolutely regular* if

$$\lim_{\tau\searrow 0}\frac{1}{\tau}\int_t^{t+\tau}\|(\mathbb{F}(t)u_z)(\sigma)\|_Y^2\,d\sigma=0$$

for $t\geq 0$ and $z\in U$, where $u_z(s):=z$ for $s\geq 0$. We note that the absolute regularity implies the regularity of Σ.

The following result is a consequence of Theorem 2.1 and [26, Thm. 3.11].

THEOREM 2.3. *Let $\Sigma = (T, \Phi, \Psi, \mathbb{F})$ be a regular nonautonomous system, with state trajectory x, control function u, and observation function y. The following assertions hold:*

(i) $\Phi(\cdot, s)u \in \mathcal{D}_s(\tilde{C}(\cdot))$ *for* $s \geq 0$.
(ii) $x(t) \in D(\tilde{C}(t))$ *for a.e.* $t \geq s \geq 0$.
(iii) $y = \Psi(s)\eta + \mathbb{F}(s)u$ *for* $s \geq 0$.
(iv) $y(t) = \tilde{C}(t)x(t)$ *for a.e.* $t \geq s \geq 0$.

We end this section by recalling the concept of admissible feedback operators for nonautonomous linear systems.

DEFINITION 2.4. *Let $\Sigma = (T, \Phi, \Psi, \mathbb{F})$ be a well-posed nonautonomous system. We call $\Delta(\cdot) \in L^\infty(\mathbb{R}_+, \mathcal{L}_s(Y, U))$ (the space of essentially bounded and strongly measurable operator functions) an* admissible feedback *for Σ if there exists $t_0 > 0$ such that the operators $I_Y - \mathbb{F}(s + t_0, s)\Delta(\cdot)$, $s \geq 0$, have uniformly bounded inverses on $L^2([s, s + t_0], Y)$.*

If Σ is absolutely regular and $\Delta(t)$ are admissible feedback operators for Σ, then the closed-loop system Σ^Δ for Σ and $\Delta(\cdot)$ exists, and it is also absolutely regular. Moreover, we have several formulas relating the open-loop and closed-loop systems. To put the formulas in a concise form, we define the operators $\Psi(t, s)x := (\Psi_s x)_{|[s,t]}$ and

$$\Sigma(t, s) := \begin{pmatrix} T(t, s) & \Phi(t, s) \\ \Psi(t, s) & \mathbb{F}(t, s) \end{pmatrix} : \ X \times L^p([s, t], U) \longrightarrow X \times L^p([s, t], Y) \tag{2.8}$$

for $t \geq s \geq 0$. Then it holds

$$\Sigma^\Delta(t, s) - \Sigma(t, s) = \Sigma(t, s) \begin{pmatrix} 0 & 0 \\ 0 & \Delta(\cdot) \end{pmatrix} \Sigma^\Delta(t, s) = \Sigma^\Delta(t, s) \begin{pmatrix} 0 & 0 \\ 0 & \Delta(\cdot) \end{pmatrix} \Sigma(t, s). \tag{2.9}$$

These facts are shown in Theorem 4.4 and Proposition 5.1 of [26], where one can find further results on the relations ship between Σ and Σ^Δ.

3. A boundary control nonautonomous system. Let us first introduce some basic facts about bounded variation functions. For a Banach space Z, we shall denote by $BV([-r, 0], Z)$ the space of all functions $\nu : [-r, 0] \to Z$ of bounded variation, i.e., the total variation of ν

$$Var(\nu)_{-r}^0 := \sup \left\{ \sum_{j=1}^N \|\nu(\theta_j) - \nu(\theta_{j-1})\| : 0 = \theta_0 > \theta_1 \cdots > \theta_N = -r, \ N \in \mathbb{N} \right\}$$

is finite. Elements of $BV([-r, 0], Z)$ are normalized throughout this paper by the requirements $\nu(-r) = 0$ and that $\nu(\cdot)$ is left-continuous on $[-r, 0]$. Hence, extending $\nu \in BV([-r, 0], Z)$ by 0 to $(-\infty, 0]$, $\nu(\cdot)$ can also be considered as an element of $BV((-\infty, 0], Z)$. (This space is defined as $BV([-r, 0], Z)$.)

Throughout E and F are supposed to be Banach spaces. We shall study a boundary control system closely related to the left shift semigroup $S_E := (S_E(t))_{t \geq 0}$ on $L^2([-r, 0], E)$, where

$$(S_E(t)\psi)(\theta) := \mathbb{I}_{[-r,0]}(t + \theta)\psi(t + \theta), \quad \psi \in L^2([-r, 0], E), \ t \geq 0, \ \theta \in [-r, 0].$$

Here the symbol $\mathbb{I}_J$ denotes the constant function equal to one in the interval $J \subset \mathbb{R}$ and zero otherwise. It can be verified that the generator of S_E is given by

$$Q_E\psi := \psi' \quad \text{for} \quad \psi \in D(Q_E) := \{\psi \in W^{1,2}([-r, 0], E) : \psi(0) = 0\}.$$

Let $\mathcal{S}_E := (\mathcal{S}_E(t,s))_{t\geq s\geq 0}$ be the family defined by $\mathcal{S}_E(t,s) := S_E(t-s)$ for $t \geq s \geq 0$. Then $\mathcal{S}_E$ is a strongly continuous evolution family on $L^2([-r,0],E)$.

Observe that for $s \geq 0$, $\zeta \in L^2([-r,0],E)$ and $u \in L^2_{loc}([s-r,\infty),E)$ such that $u_s = \zeta$ we have

$$u_t = \mathcal{S}_E(t,s)\zeta + \Phi(t,s)u, \qquad t \geq s, \tag{3.1}$$

where $\Phi(t,s) : L^2([s,\infty),E) \to L^2([-r,0],E)$ are the linear operators defined by

$$(\Phi(t,s)u)(\theta) := \begin{cases} u(t+\theta), & \theta > s-t, \\ 0, & \theta \leq s-t, \end{cases} \tag{3.2}$$

for $t \geq s \geq 0,\ u \in L^2([s,\infty),E)$, and $\theta \in [-r,0]$. Clearly, these operators satisfy

$$\|\Phi(t,s)u\|_{L^2([-r,0],E)} = \|u\|_{L^2([s,t],E)}$$

for all $t \geq s \geq 0$ and $u \in L^2_{loc}([s,\infty),E)$. Moreover, one can see that $(\mathcal{S}_E,\Phi)$ satisfies the first identity in (2.1). Thus we have the following lemma.

LEMMA 3.1. *Let $\Phi(t,s)$, $t \geq s \geq 0$, be the operators defined by* (3.2). *Then $(\mathcal{S}_E,\Phi)$ is a nonautonomous control system on $L^2([-r,0],E),E$. Let now $u(t)$ and $v(t,\theta)$ for $t \geq s \geq 0,\ \theta \in [-r,0]$, be the control function and the state trajectory of this system, respectively, such that $v(s,\cdot) = \zeta$ is the initial state. Then $v(t,\cdot) = u_t$ for $t \geq s$.*

Proof. The proof follows immediately by (3.1). □

We note that the system $(\mathcal{S}_E,\Phi)$ is represented by the boundary control equation

$$\begin{aligned} &\frac{\partial}{\partial t}v(t,\theta) = \frac{\partial}{\partial \theta}v(t,\theta), \quad v(s,\theta) = \zeta(\theta), \quad t \geq s \geq 0, \ -r \leq \theta \leq 0, \\ &v(t,0) = u(t), \end{aligned} \tag{3.3}$$

which has the input segment $u_t,\ t \geq s$, as the state trajectory. Next, we couple (3.3) with the following observation equation:

$$y(t) = P(t)u_t, \qquad t \geq s \geq 0. \tag{3.4}$$

Here the operators $P(t) : C([-r,0],E) \to F$ are defined by

$$P(t)\psi = \int_{-r}^{0} dp(t,\theta)\psi(\theta), \qquad \psi \in C([-r,0],E), \quad t \geq 0, \tag{3.5}$$

where the kernel $p(\cdot,\cdot) : \mathbb{R}_+ \times [-r,0] \to \mathcal{L}(E,F)$ satisfies the following assumption:

(**H**) The function $\mathbb{R}_+ \times [-r,0] \ni (t,\theta) \mapsto p(t,\theta) \in \mathcal{L}(E,F)$ is strongly measurable in θ and strongly continuous in t such that $p(t,\cdot) \in BV([-r,0],\mathcal{L}(E,F))$ with total variation $\mu_p(t,\cdot) := Var(p(t,\cdot))_0^{-r} \leq c_p < +\infty$ for all $t \geq 0$ and $\sup_t \mu_p(t,[-\sigma,0]) \to 0$ as $\sigma \to 0$. Moreover,

$$\int_0^{\alpha} \|p(s+t,\theta'-t) - p(s+t,\theta-t)\|\,dt \leq c'_p|\theta'-\theta|,$$

where $c'_p = c'_p(\alpha_0),\ 0 < \alpha \leq \alpha_0,\ \theta',\theta \in [-r,0]$, and $s \geq 0$.

We already showed (see [12]) that there are examples of $p(\cdot,\cdot)$ for which condition (H) holds. For the sake of completeness we recall the following example.

EXAMPLE 3.2. *Let $q(\cdot,\cdot):\mathbb{R}_+\times[-r,0]\to\mathcal{L}(E,F)$ be strongly continuous in $t\geq 0$ such that $q(t,\cdot):[-r,0]\to\mathcal{L}(E,F)$ is Lipschitzien, $q(t,-r)=0$, and $\|q(t,\cdot)\|_{Lip}\leq c$ for all $t\geq 0$, where $\|\cdot\|_{Lip}$ is the Lipschitzien norm. We set*

$$p(t,\theta):=q(t,\theta)\varrho(\theta),\quad (t,\theta)\in\mathbb{R}_+\times[-r,0],$$

where $\varrho(\cdot)\in BV([-r,0],\mathcal{L}(E,F))$ is such that $Var(\varrho)_{-\tau}^{0}\to 0$ as $\tau\to 0$. Then the kernel $p(\cdot,\cdot)$ satisfies (H).

Since $W^{1,2}([-r,0],E)$ is continuously embedded in $C([-r,0],E)$, then $P(t)\in\mathcal{L}(W^{1,2}([-r,0],E),E)$. Next, we set

$$C_0([-r,0],E):=\{\psi\in C([-r,0],E):\psi(0)=0\}.$$

The following result shows that the linear operators with kernels satisfying the assumption (H) are nonautonomous admissible observation operators for the evolution family $\mathcal{S}_E$. For the proof see [12] (see also the proof of Lemma 3.5 below).

LEMMA 3.3. *Assume that the kernel $p(\cdot,\cdot)$ satisfies* (H). *Then*

$$\int_s^{s+\alpha}\|P(t)\mathcal{S}_X(t,s)\psi\|^2\,dt\leq\gamma\,\|\psi\|_2^2$$

for $0<\alpha\leq r$, $s\geq 0$, $\psi\in C_0([-r,0],E)$, and a constant $\gamma>0$.

Now Lemma 3.3 implies that the maps

$$\Psi(s):C_0([-r,0],E)\to L^2_{loc}([s,\infty),F),\quad \psi\mapsto P(\cdot)\mathcal{S}_E(\cdot,s)\psi,\quad s\geq 0, \tag{3.6}$$

admit unique continuous extensions to $L^2([-r,0],E)$ (which we denote again by $\Psi(s)$). Moreover one can see that $(\mathcal{S}_E,\Psi)$ is a nonautonomous observation system on $L^2([-r,0],E)$, F. Throughout we denote by $\tilde{P}(t)$ the Lebesgue extensions of $P(t)$ with respect to $\mathcal{S}_E$.

Now we are interested in studying the absolute regularity of the boundary linear system formed by (3.3) and (3.4). Before that, we shall give some preliminaries.

LEMMA 3.4. *Let $v:[s-r,\infty)\longrightarrow E$. The following three properties are equivalent:*

(i) $v\in W^{1,2}_{loc}([s-r,\infty),E)$,
(ii) $v_\cdot\in C^1([s,\infty),L^2([-r,0],E))$,
(iii) $v_\cdot\in C([s,\infty),W^{1,2}([-r,0],E))$.

Proof. The proof of this lemma is similar to that given in [3, p. 226], where $E=\mathbb{R}^n$, $n\geq 1$, was considered. □

We will frequently use the following dense spaces in $L^2_{loc}([s,\infty),E)$:

$$\mathcal{C}_s([s,\infty),X):=\{u\in C([s,\infty),E):u(s)=0\},$$
$$W^{1,2}_{s,loc}([s,\infty),E):=\{u\in W^{1,2}_{loc}([s,\infty),E):u(s)=0\}.$$

LEMMA 3.5. *Assume that the kernel $p(\cdot,\cdot)$ satisfies* (H) *and let $s\geq 0$. Then*

(a) *$\Phi(\cdot,s)u\in C([s,\infty),W^{1,2}([-r,0],E))$ for all $u\in W^{1,2}_{s,loc}([s,\infty),E)$.*
(b) *For $u\in\mathcal{C}_s([s,\infty),E)$ we have $\Phi(t,s)u\in C([-r,0],E)$ and*

$$\int_s^{s+\alpha}\|P(t)\Phi(t,s)u\|^2\,dt\leq\kappa\,\|u\|^2_{L^2([s,s+\alpha],E)} \tag{3.7}$$

for $t\geq s$, $0<\alpha<r$, a constant $\kappa>0$.

Proof. We show (a). Let $u \in W^{1,2}_{s,loc}([s,\infty),E)$ and take $\zeta \in D(Q_E)$. Now we define the function $v : [s-r,\infty) \longrightarrow E$ by

$$v(t) = \begin{cases} u(t), & t \geq s, \\ \zeta(t-s), & s-r \leq t < s. \end{cases}$$

Then we have $v \in W^{1,2}_{loc}([s-r,\infty),X)$ and $v_t = \mathcal{S}_E(t,s)\zeta + \Phi(t,s)u$ for $t \geq s$. So that $\Phi(\cdot,s)u \in C([s,\infty),W^{1,2}([-r,0],X))$, due to Lemma 3.4(iii).

We show (b), so we will proceed as in [12]. To this purpose, let $u \in \mathcal{C}_s([s,\infty),E)$ and $0 < \alpha < r$. Clearly, by (3.2), we have $\Phi(t,s)u \in C([-r,0],E)$ for $t \geq s$. Now we have

$$\int_s^{s+\alpha} \|P(t)\Phi(t,s)u\|^2\,dt = \int_0^\alpha \Big\| \int_{-t}^0 dp(t+s,\theta)u(t+s+\theta)\Big\|^2 dt. \tag{3.8}$$

Now we define $\theta_m := -mr/n,\ m = 0,1,\dots,n$, a subdivision of the interval $[-r,0]$. For a fixed $t \in [0,\alpha]$, $D_{\sigma,n} = \{-\theta_0 - t,\dots,-\theta_n - t\}$ is a subdivision of the interval $[-t,-t+\alpha]$, since we have extended $p(t,.)$ by $p(t,0)$, the extra terms on $[0,-t+r]$ are all equal to 0. So, for each fixed σ, we have

$$\int_{-t}^0 dp(s+t,\theta)u(t+s+\theta) = \lim_{n\to\infty} \sum_{m=0}^n \Delta_m(t,s)u(s-\theta_m), \tag{3.9}$$

where we set $\Delta_m(t,s) := p(t+s,-\theta_m - t) - p(t+s,-\theta_{m-1}-t)$. Observe that (H) implies $\sum_{m=0}^n \|\Delta_m(t,s)\| \leq c_p$. Now, due to (3.8), (3.9), Fatou's lemma, and the Cauchy–Schwartz inequality, we get

$$\begin{aligned}
&\int_s^{s+\alpha} \|P(t)\Phi(t,s)u\|^2\,dt \\
&\quad\leq \liminf_{n\to\infty} \int_0^\alpha \left(\sum_{m=0}^n \|\Delta_m(t,s)\|^{\frac12} \|\Delta_m(t,s)\|^{\frac12} \|u(s-\theta_m)\| \right)^2 dt \\
&\quad\leq c_p \liminf_{n\to\infty} \int_0^\alpha \sum_{m=0}^n \|\Delta_m(t,s)\| \|u(s-\theta_m)\|^2\,dt \\
&\quad\leq c_p \liminf_{n\to\infty} \sum_{m=0}^n \|u(s-\theta_m)\|^2 \int_0^\alpha \|\Delta_m(t,s)\|\,dt \\
&\quad\leq c_p\ c_p' \liminf_{n\to\infty} \sum_{m=0}^n \|u(s-\theta_m)\|^2\ |\theta_m - \theta_{m-1}| \\
&\quad\leq c_p c_p' \|u\|^2_{L^2([s,s+\alpha],E)}. \qquad (3.10)
\end{aligned}$$

Thus the lemma follows. □

According to Lemma 3.5 we can now define the following linear operators:

$$\mathbb{F}(s) : \mathcal{C}_s([s,\infty),E) \to L^2_{loc}([s,\infty),F), \quad \mathbb{F}(s)u = P(\cdot)\Phi(\cdot,s)u, \quad s \geq 0. \tag{3.11}$$

Further, by density and (3.7), we extend $\mathbb{F}(s)$ to bounded operators (denoted again by $\mathbb{F}(s)$) from $L^2_{loc}([s,\infty),E)$ to $L^2_{loc}([s,\infty),E)$. Moreover, we have the following main result of this section.

THEOREM 3.6. *Assume that $p(\cdot,\cdot)$ satisfies* (H). *Then $\Sigma = (\mathcal{S}_E, \Phi, \Psi, \mathbb{F})$ is an absolutely regular nonautonomous system on $L^2([-r,0],E), E, F$. Furthermore, if u is the control function of Σ, then the input segment u_t is the state trajectory of Σ which satisfies $u_\bullet \in \mathcal{D}_s(\tilde{P}(\cdot))$ for $s \geq 0$. Moreover, the output function of Σ satisfies the representation $y(t) = \tilde{P}(t)u_t$ for almost every $t \geq s \geq 0$.*

Proof. Let us first show that $\mathbb{F}(s)$ satisfy (2.6). For this let $0 \leq s \leq t \leq \rho$, $u^0 \in \mathcal{C}([s,\infty),E)$, and $u^1 \in \mathcal{C}([t,\infty),E)$ such that $u^0(s) = u^0(t) = u^1(t) = 0$. We now define the function

$$u(\tau) = \begin{cases} u^1(\tau), & t \leq \tau \leq \rho, \\ u^0(\tau), & s \leq \tau \leq t, \end{cases}$$

so that $u \in \mathcal{C}_s([s,\infty),E)$. Now Lemma 3.5 implies that $\Phi(t,s)u^0 \in C_0([-r,0],E)$ and $\Phi(\rho,t)u^1 \in C([-r,0],E)$. Now, by (2.1), (3.6), and (3.11), we obtain

$$\begin{aligned} (\mathbb{F}(s)u)(\rho) &= P(\rho)\Phi(\rho,s)u \\ &= P(\rho)\mathcal{S}_E(\rho,t)\Phi(t,s)u^0 + P(\rho)\Phi(\rho,t)u^1 \\ &= (\Psi(t)\Phi(t,s)u^0)(\rho) + (\mathbb{F}(t)u^1)(\rho). \end{aligned}$$

Then, by density, Σ is a well-posed nonautonomous system on $L^2([-r,0],E), E, F$. Next, we show that Σ is an absolutely regular nonautonomous system. For this we set $u_z(s) = z$ for $z \in E$ and $s \geq 0$. Moreover, let $\varepsilon_n \in C([s,\infty))$ such that $0 \leq \varepsilon_n \leq 1$, $\varepsilon_n(s) = 0$, and $\varepsilon_n(t) = 1$ for $t \geq s + 1/n$, $n \in \mathbb{N}\backslash\{0\}$. We now set $v^n = \varepsilon_n u_z$ for $n \in \mathbb{N}\backslash\{0\}$, so we have $v^n \in \mathcal{C}_s([s,\infty),U)$ and $v^n \to u_z$ in $L^2_{loc}([s,\infty),E)$. Let $0 < t < \inf\{1,r\}$ and $s \geq 0$. Due to (3.11) we get

$$\begin{aligned} \frac{1}{t}\int_s^{s+t} \|(\mathbb{F}(s)u_z)(\tau)\|^2\,d\tau &= \lim_{n\to\infty}\frac{1}{t}\int_s^{s+t}\|(\mathbb{F}(s)v^n)(\tau)\|^2\,d\tau \\ &= \lim_{n\to\infty}\frac{1}{t}\int_s^{s+t}\|P(\tau)\Phi(\tau,s)v^n\|^2\,d\tau \\ &\leq \frac{1}{t}\int_s^{s+t}\left(\int_{s-\tau}^0 d\mu_p(\tau,\theta)\right)^2 d\tau\,\|z\|^2 \\ &\leq \frac{1}{t}\int_s^{s+t}\Big(\mu_p(\tau,[-t,0])\Big)^2 d\tau\,\|z\|^2 \\ &\leq \|z\|^2\left(\sup_{s\leq\tau\leq s+r}\mu_p(\tau,[-t,0])\right)^2. \end{aligned}$$

Consequently, we have the absolute regularity of Σ, due to (H). Finally, the rest of the proof follows by Lemma 3.1 and Theorem 2.3(ii) and (iv). □

COROLLARY 3.7. *Assume that $p(\cdot,\cdot)$ satisfies* (H) *and let $\Sigma = (\mathcal{S}_E, \Phi, \Psi, \mathbb{F})$ be the absolutely regular nonautonomous system obtained in Theorem* 3.6. *If the initial condition $\zeta \in D(Q_E)$ and the control function $u \in W^{1,2}_{s,loc}([s,\infty),E)$, then the output function of Σ satisfies $y(t) = P(t)u_t$ for $t \geq s \geq 0$.*

Proof. Let $\zeta \in D(Q_E)$ and $u \in W^{1,2}_{s,loc}([s,\infty),E)$. Then, by Theorem 2.3(iii), (3.1), (3.6), and (3.11), we get

$$y(t) = (\Psi(s)\zeta)(t) + (\mathbb{F}(s)u)(t) = P(t)(\mathcal{S}_E(t,s)\zeta + \Phi(t,s)u) = P(t)u_t$$

for almost every $t \geq s \geq 0$. □

4. Nonhomogenous delay equations. Let us introduce some basic facts on the existence and regularity of the evolution equations,

$$\text{(CP)} \quad \begin{cases} \dot{x}(t) = A(t)x(t) + f(t), & t > s \geq 0, \\ x(s) = \eta, \end{cases}$$

for linear operators $A(t)$, $t \geq s \geq 0$, on a Banach space X, where $\eta \in X$ and $f \in L^2_{loc}([s,\infty), X)$. The homogenous problem with $f = 0$ is denoted by $(\text{CP})_0$.

Among various types of solutions x of (CP) we consider the following one.

DEFINITION 4.1. *A classical solution of* (CP) *is a function* $x \in W^{1,2}_{loc}([s,\infty), X)$ *such that* $x(t) \in D(A(t))$ *for* $t > s$ *and* (CP) *holds.*

Contrary to the autonomous case, the solvability of (CP) in the classical sense requires additional conditions on $A(t)$ (see, e.g., [2], [20], [24], [27] and the references therein). Here in the current paper we restrict ourselves to a class of parabolic equations, namely, those introduced by Aquistapace and Terreni [2]. We assume that

(AT1) $A(t)$, $t \geq 0$, are linear operators on a Banach space X with dense domains $D(A(t))$ and there are constants $\beta \geq 0$, $w \in \mathbb{R}$, and $\phi \in (\pi/2, \pi)$ such that $\lambda \in \rho(A(t))$ and $\|R(\lambda, A(t))\| \leq \beta(1+|\lambda - w|)^{-1}$ for $\lambda \in \Sigma(\phi, w)$ and $t \geq s \geq 0$.

(AT2) There are constants $\kappa \geq 0$ and $\mu, \nu \in (0,1]$ with $\mu + \nu > 1$ such that

$$|\lambda|^\nu \|A_w(t)R(\lambda, A_w(t))(A_w(t)^{-1} - A_w(t')^{-1})\| \leq \kappa |t - t'|^\mu$$

for $A_w(t) := A(t) - w$, $t, t' \geq s \geq 0$, and $|\arg \lambda| \leq \phi$.

Here we set $\Sigma(\phi, w) = \{w\} \cup \{\lambda \in \mathbb{C} \backslash \{w\} : |\arg(\lambda - w)| \leq \phi\}$. If (AT1) and (AT2) hold, then we say that (AT) is satisfied and (CP) is a *parabolic equation.*

Let us recall (see, e.g., Tanabe [24]) that if $D(A(t))$ do not depend on t, it suffices to suppose that $A(\cdot) : [0, t_0] \to \mathcal{L}(Z, X)$ is Hölder continuous, where $Y := D(A(0))$ is endowed with the graph norm.

THEOREM 4.2. *Assume that* (AT) *holds. Then there is a strongly continuous evolution family* $T := (T(t,s))_{t \geq s \geq 0}$ *on* X *such that* $T(t,s)X \subseteq D(A(t))$ *for* $t > s$.

If $A(t)$ satisfy (AT), the $(\text{CP})_0$ has a unique classical solution $x = T(\cdot, s)\eta$ for $\eta \in X$. We say that T solves $(\text{CP})_0$ or $A(t)$ generate T.

Let (AT) be satisfied. According to [1, Props. 3.2, 5.1], if x is a classical solution of (CP), then it is given by

$$x(t) = T(t,s)\eta + \int_s^t T(t,\sigma) f(\sigma)\, d\sigma, \qquad t \geq s \geq 0. \tag{4.1}$$

A function $x : [s,\infty) \to X$ satisfying (4.1) is called the *mild solution* of (CP).

Finally, we mention the important case of the *hyperbolic equations* studied by Kato [19], that is, if $A(t)$ are the generators of contraction semigroups satisfying $D(A(t)) \equiv Z$ and $A(\cdot)\eta \in C^1(\mathbb{R}_+, X)$ for $\eta \in X$. Then $(\text{CP})_0$ is well posed. (See also [20] for a more general version.)

The perturbation problem of $(\text{CP})_0$ by nonautonomous bounded perturbations is studied in [22, p. 129]. We note that the well-posedness of $(\text{CP})_0$ is not persisted even by bounded perturbations $B(\cdot) \in C_b(\mathbb{R}_+, \mathcal{L}(X))$ (see [21, Ex. 4.1.7]).

In what follows, we are concerned with the notion of perturbed evolution families rather than that of perturbed evolution equations involving concrete differential operators.

PROPOSITION 4.3. *Let* $B(t) : D(B(t)) \to X$ *be nonautonomous admissible observation operators for the strongly continuous evolution family* T *on* X. *Let* $\tilde{B}(t)$ *be the Lebesgue extensions of* $B(t)$ *with respect to* T. *The following assertions hold:*

(i) *There is a unique evolution family V on X such that $V(\cdot,s)\eta \in \mathcal{D}_s(\tilde{B}(\cdot))$, and*

$$\begin{aligned} &\|\tilde{B}(\cdot)V(\cdot,s)x\|_{L^2([s,s+t_0],X)} \le \gamma\,\|\eta\|, \\ &V(\cdot,s)\eta = T(\cdot,s)\eta + \mathbb{K}_s^T\tilde{B}(\cdot)V(\cdot,s)\eta \end{aligned} \tag{4.2}$$

for $s \ge 0,\ t_0 > 0,\ \eta \in X$ and a constant $\gamma = \gamma(t_0) > 0$.

(ii) *Let $C(t) : D(C(t)) \to Y$ be nonautonomous admissible observation operators for T. For any $s \ge 0,\ \eta \in X$, and $f \in L^2_{loc}([s,\infty),X)$ we have*

$$V(\cdot,s)\eta,\ \mathbb{K}_s^V f \in \mathcal{D}_s(\tilde{C}(\cdot)), \tag{4.3}$$

$$\|\tilde{C}(\cdot)\mathbb{K}_s^V f\|_{L^2([s,s+t_0],Y)} \le c t_0^{1/2}\|f\|_{L^2([s,s+t_0],X)} \tag{4.4}$$

for some constants $t_0 > 0$ and $c = c(t_0) > 0$ independent of f and s.

Proof. The assertion (i) can be obtained by [26, Rem. 4.6(a)]. Now we prove (ii). Let $s \ge 0,\ \eta \in X$, and $f \in L^2_{loc}([s,\infty),X)$. The fact that $V(\cdot,s)\eta \in \mathcal{D}_s(\tilde{C}(\cdot))$ is due to Theorem 2.1, Proposition 2.2, and (4.2). Due to (i) above, the function $[s,s+t_0]\times[s,s+t_0] \ni (t,\sigma) \mapsto \tilde{B}(\tau)V(\tau,\sigma)f(\sigma)$ is measurable for f in the subspaces

$$\mathfrak{F}_s = \operatorname{span}\{\xi(\cdot)V(\cdot,\sigma)x : x \in X,\ \sigma \ge s,\ \xi \in C_c(\mathbb{R}_+),\ \xi(t) = 0 \text{ for } s \le t \le \sigma\}.$$

Thus, by density it is measurable for $f \in L^2_{loc}([s,\infty),X)$ (the density of $\mathfrak{F}_s$ is proved in [4, Thm. 3.12]), where $C_c(\mathbb{R}_+)$ is the space of continuous functions with compact support in $[0,\infty)$. Fubini–Tonelli's theorem and (4.2) imply that

$$\begin{aligned} \int_s^{s+t_0}\int_s^t \|\tilde{B}(t)V(t,\sigma)f(\sigma)\|^2 d\sigma\,dt &= \int_s^{s+t_0}\int_\sigma^{s+t_0} \|\tilde{B}(t)V(t,\sigma)f(\sigma)\|^2 dt\,d\sigma \\ &\le \gamma^2\|f\|^2_{L^p([s,s+t_0],X)}. \end{aligned} \tag{4.5}$$

Now we define the function $g_s(t) := \int_s^t \tilde{B}(t)V(t,\sigma)f(\sigma)\,d\sigma$ for $t \ge s \ge 0$. Using Hölder's inequality, Fubini's theorem, and (4.5) we get

$$\begin{aligned} \int_s^{s+t_0} \|g_s(\tau)\|^2\,d\tau &\le t_0\int_s^{s+t_0}\int_s^\tau \|\tilde{B}(\tau)V(\tau,\sigma)f(\sigma)\|^2\,d\sigma\,dt \\ &\le t_0\gamma^2\|f\|^2_{L^2([s,s+t_0],X)}. \end{aligned} \tag{4.6}$$

On the other hand, by (4.2) and Fubini's theorem, we obtain

$$\begin{aligned} (\mathbb{K}_s^V f)(t) &= (\mathbb{K}_s^T f)(t) + \int_s^t\int_\sigma^t T(t,\tau)\tilde{B}(\tau)V(\tau,\sigma)f(\sigma)\,d\tau\,d\sigma \\ &= (\mathbb{K}_s^T f)(t) + \int_s^t T(t,\tau)\int_s^\tau \tilde{B}(\tau)V(\tau,\sigma)f(\sigma)\,d\sigma\,d\tau \\ &= (\mathbb{K}_s^T(f+g_s))(t) \end{aligned}$$

for $t \ge s \ge 0$. Thus, the rest of the proof now follows from (4.6) and Proposition 2.2. □

Let V be the evolution family obtained in Proposition 4.3. Let $s \ge 0$ and $f \in L^2_{loc}([s,\infty),X)$. We are looking for functions $x : [s,\infty) \to X$ satisfying

$$x(t) = V(t,s)\eta + \int_s^t V(t,\sigma)f(\sigma)\,d\sigma, \qquad t \ge s. \tag{4.7}$$

We need the following lemma.

LEMMA 4.4. *Let T be a strongly continuous evolution family on X. Let $B(t) : D(B(t)) \subseteq X \to X$ be nonautonomous admissible observation operators for T. Let $\tilde{B}(t)$ be the Lebesgue extensions of $B(t)$ with respect to T. Then*

$$\tilde{B}(t)(\mathbb{K}_s^V f)(t) = \int_s^t \tilde{B}(t)V(t,\sigma)f(\sigma)\,d\sigma$$

for a.e. $t \geq s$ and all $f \in \mathcal{C}_s([s,\infty),X)$, where V is the strongly continuous evolution family satisfying (4.2).

Proof. Due to Proposition 4.3(i) one defines $\Psi'(s)\eta = \tilde{B}(\cdot)V(\cdot,s)\eta$ for $\eta \in X$ and $s \geq 0$. Then (V,Ψ') is a nonautonomous observation system. Now we define

$$\tilde{B}_n(s)\eta = n\int_s^{s+\frac{1}{n}} (\Psi'(s)\eta)(\sigma)\,d\sigma$$

for $\eta \in X$, $s \geq 0$, and $n \in \mathbb{N}\backslash\{0\}$, so we have $B_n(s) \in \mathcal{L}(X)$. Next we take $f \in \mathcal{C}_s([s,\infty),X)$ for $s \geq 0$. The function $\rho \ni [0,t_0] \mapsto B_n(\cdot)V(\cdot,\rho)f(\rho) - \Psi'(\rho)f(\rho)$ is measurable, due to [26, Lem. 2.8]. On the other hand, for $t_0 > 0$ and $s \geq 0$, we have

$$\begin{aligned}
&\left\| B_n(\cdot)\mathbb{K}_s^V f - \int_s^{\cdot} [\Psi'(\rho)f(\rho)](\cdot)\,d\rho \right\|^2_{L^2([s,s+t_0],X)} \\
&\qquad \leq \int_s^{s+t_0} \left(\int_s^{\sigma} \|B_n(\sigma)V(\sigma,\rho)f(\rho) - [\Psi'(\rho)f(\rho)](\sigma)\| d\rho \right)^2 d\sigma \\
&\qquad \leq t_0^{\frac{1}{2}} \int_s^{s+t_0} \int_\rho^{\rho+t_0} \|B_n(\sigma)V(\sigma,\rho)f(\rho) - [\Psi'(\rho)f(\rho)](\sigma)\|^2 d\sigma d\rho \\
&\qquad := t_0^{\frac{1}{2}} \int_s^{s+t_0} h_{n,f}(\rho)\,d\rho, \qquad (4.8)
\end{aligned}$$

due to Hölder's inequality and Fubini's theorem, where we set

$$h_{n,f}(\rho) := \int_\rho^{\rho+t_0} \|B_n(\sigma)V(\sigma,\rho)f(\rho) - [\Psi'(\rho)f(\rho)](\sigma)\|^2 d\sigma, \quad s+t_0 \geq \rho \geq s,\ n \geq 1.$$

It can be verified (see the proof of [26, Lem. 2.9]) that $h_{n,f}(\rho) \leq c\|f(\rho)\|$ for $s \leq \rho \leq s+t_0$, $n \geq 1$, and a constant $c = c(t_0) > 0$, so $h_{n,f} \in L^2_{loc}([s,\infty))$. Now, for $f \in \mathfrak{F}_s$ one can see that $h_{n,f}(\rho) \to 0$ as $n \to \infty$ for $s \leq \rho \leq s+t_0$, due to [26, Lem. 2.9]. We now approximate f by $f_m \in \mathfrak{F}_s$ in $L^2_{loc}([s,\infty),X)$, so we have $f_{\phi(m)}(\rho) \to f(\rho)$ in X as $m \to \infty$ for a.e. $\rho \geq s$. Then

$$\begin{aligned}
h_{n,f}(\rho) &\leq \|B_n(\cdot)V(\cdot,\rho)(f(\rho) - f_{\phi(m)}(\rho))\|^2_{L^2([\rho,\rho+t_0],X)} \\
&\qquad + h_{n,f_{\phi(m)}} + \|\Psi'(\rho)(f(\rho) - f_{\phi(m)}(\rho))\|^2_{L^2([\rho,\rho+t_0])} \\
&\leq \delta\,\|f(\rho) - f_{\phi(m)}(\rho)\|^2_X + h_{n,f_{\phi(m)}}
\end{aligned}$$

for a.e. $s \leq \rho \leq s+t_0$ and a constant $\delta = \delta(t_0) > 0$, due to the proof of [26, Lem. 2.9]. Then

$$\lim_{n\to\infty} h_{n,f}(\rho) \leq \delta\,\|f(\rho) - f_{\phi(m)}(\rho)\|^2_X$$

for a.e. $s \leq \rho \leq s+t_0$. Now by letting $m \to \infty$ in the estimate above we obtain that $h_{n,f}(\rho) \to 0$ as $n \to \infty$. Lebesgue's dominated convergence theorem and (4.8) now

imply

$$\int_s^t [\Psi'(\sigma)f(\sigma)](t)\,d\sigma = \lim_{n\to\infty} B_{\phi(n)}(t)\mathbb{K}_s^V f = \tilde{B}(t)(\mathbb{K}^V f)(t)$$

for a.e. $t \geq s$, due to [26, Lem. 2.10]. □

THEOREM 4.5. *Assume that* (AT) *holds. Let T be the strongly continuous evolution family generated by $A(t)$. Let $B(t) : D(B(t)) \to X$ be nonautonomous admissible observation operators for T. Let $\tilde{B}(t)$ the Lebesgue extensions of $B(t)$ with respect to T. Finally, let $f \in L^2_{loc}([s,\infty), X)$, $s \geq 0$. Then the function x given by* (4.7) *satisfies $x \in \mathcal{D}_s(\tilde{B}(\cdot))$ and*

$$x(t) = T(t,s)\eta + \int_s^t T(t,\sigma)\Big[\tilde{B}(\sigma)x(\sigma) + f(\sigma)\Big]\,d\sigma, \qquad t \geq 0. \tag{4.9}$$

Proof. Let $x : [s,\infty) \to X$ be the function given by (4.7) corresponding to $\eta \in X$ and $f \in L^2_{loc}([s,\infty), X)$. Then $x \in \mathcal{D}_s(\tilde{B}(\cdot))$, due to (4.3). We now approximate f by a sequence f^n in $\mathcal{C}_s([s,\infty), X)$ and let $x^n : [s,\infty) \to X$ be the solution of (4.7) corresponding to η and f^n. Then, due to Hölder's inequality and (4.7), we obtain $x^n \to x$ in $L^2_{loc}([s,\infty), X)$. Moreover, by (4.2), (4.7), Fubini's theorem, and Lemma 4.4, we have

$$\begin{aligned}
x^n(t) &= T(t,s)\eta + \int_s^t T(t,\sigma)f(\sigma)\,d\sigma + \int_s^t T(t,\sigma)\tilde{B}(\sigma)V(\sigma,s)\eta\,d\sigma \\
&\qquad + \int_s^t \int_\sigma^t T(t,\tau)\tilde{B}(\tau)V(\tau,\sigma)f^n(\sigma)\,d\tau\,d\sigma \\
&= T(t,s)\eta + \int_s^t T(t,\sigma)f^n(\sigma)\,d\sigma + \int_s^t T(t,\sigma)\tilde{B}(\sigma)V(\sigma,s)\eta\,d\sigma \\
&\qquad + \int_s^t T(t,\tau)\int_s^\tau \tilde{B}(\tau)V(\tau,\sigma)f^n(\sigma)\,d\sigma\,d\tau \\
&= T(t,s)\eta + \int_s^t T(t,\sigma)\Big[\tilde{B}(\sigma)x^n(\sigma) + f^n(\sigma)\Big]\,d\sigma \qquad (4.10)
\end{aligned}$$

for $t \geq s \geq 0$. Let us denote by $v(t)$ the right-hand side of (4.9), so using again Hölder's inequality and (4.10) we get $x^n \to v$. Thus $x = v$ by uniqueness and continuity. □

In the remainder of this section we study mild solutions of the nonhomogenous delay equation (1.1). We assume that the state delay operators $L(t) : C([-r,0], X) \to X$ are defined by

$$L(t)f = \int_{-r}^0 d\ell(t,\theta)f(\theta), \quad f \in C([-r,0], X),\ t \geq 0, \tag{4.11}$$

where the kernel $\ell(\cdot,\cdot)$ satisfies (H).

At first we assume that the initial conditions $\eta \in X$ and $\varphi \in C([-r,0], X)$ are such that $\eta = \varphi(0)$. In this case, a *mild solution* of (1.1) is a function $x \in C([s-r,\infty), X)$ such that

$$\begin{aligned}
x(t) &= T(t,s)\eta + \int_s^t T(t,\sigma)\Big[L(\sigma)x_\sigma + f(\sigma)\Big]\,d\sigma, \quad t \geq s \geq 0, \\
x(s+\theta) &= \varphi(\theta), \quad -r \leq \theta \leq 0.
\end{aligned} \tag{4.12}$$

Next, we want to solve (1.1) for the initial condition $\binom{\eta}{\varphi}$ in the Banach space $\mathcal{X}_0 := X \times L^2([-r,0],X)$ endowed with the norm $\|\binom{\eta}{\varphi}\| = \|x\| + \|\varphi\|_2$. To this purpose, we consider the family $\mathcal{T} := (\mathcal{T}(t,s))_{t\geq 0}$ of bounded linear operators on $\mathcal{X}_0$ defined by

$$\mathcal{T}(t,s) := \begin{pmatrix} T(t,s) & 0 \\ T_{t,s} & \mathcal{S}_X(t,s) \end{pmatrix}, \quad t \geq s \geq 0, \tag{4.13}$$

where $(T_{t,s}\eta)(\theta) := T(t+\theta,s)\eta$ for $t+\theta \geq s$ and zero if not, for $\eta \in X$, $t \geq s \geq 0$, $\theta \in [-r,0]$. Clearly, $\mathcal{T}$ is a strongly continuous evolution family on $\mathcal{X}_0$.

In the sequel we need the subspace

$$\mathfrak{D} := \Big\{ \binom{\eta}{\varphi} \in X \times C([-r,0],X) : \varphi(0) = \eta \Big\},$$

which is dense in $\mathcal{X}_0$. Moreover, we consider the linear operators

$$\mathfrak{B}(t) := \begin{pmatrix} 0 & L(t) \\ 0 & 0 \end{pmatrix}, \qquad D(\mathfrak{B}(t)) := \mathfrak{D}, \quad t \geq 0.$$

As shown in [12], Lemma 3.3 implies that $\mathfrak{B}(t)$ are nonautonomous admissible observation operators for $\mathcal{T}$. Throughout we denote by $\tilde{\mathfrak{B}}(t)$ the Lebesgue extensions of $\mathfrak{B}(t)$ with respect to $\mathcal{T}$. So, due to Proposition 4.3, there exists a unique strongly continuous evolution family $\mathcal{T}_L := (\mathcal{T}_L(t,s))_{t\geq s\geq 0}$ on $\mathcal{X}_0$ satisfying

$$\mathcal{T}_L(\cdot,s)\binom{\eta}{\varphi} \in \mathcal{D}_s(\tilde{\mathfrak{B}}(\cdot)), \tag{4.14}$$

$$\mathcal{T}_L(\cdot,s)\binom{\eta}{\varphi} = \mathcal{T}(\cdot,s)\binom{\eta}{\varphi} + \mathbb{K}_s^{\mathcal{T}}\tilde{\mathfrak{B}}(\cdot)\mathcal{T}_L(\cdot,s)\binom{\eta}{\varphi} \tag{4.15}$$

for $\binom{\eta}{\varphi} \in \mathcal{X}_0$ and $s \geq 0$. We note that $\mathcal{T}_L$ solves the homogenous delay equation associated to (1.1) (i.e., when $f = 0$); see [12] for more details.

Next, we are looking for functions $w : [s,\infty) \to \mathcal{X}_0$ satisfying

$$w(t) = \mathcal{T}_L(t,s)\binom{\eta}{\varphi} + \int_s^t \mathcal{T}_L(t,\sigma)\binom{f(\sigma)}{0}\,d\sigma \tag{4.16}$$

for $t \geq s \geq 0$, $\binom{\eta}{\varphi} \in \mathcal{X}_0$, and $f \in L^2_{loc}([s,\infty),X)$.

Due to Theorem 4.5, the function w given by (4.16) satisfies $w \in \mathcal{D}_s(\tilde{\mathfrak{B}}(\cdot))$ and

$$w(t) = \mathcal{T}(t)\binom{\eta}{\varphi} + \int_s^t \mathcal{T}(t,\sigma)\Big[\tilde{\mathfrak{B}}(\sigma)w(\sigma) + \binom{f(\sigma)}{0}\Big]\,d\sigma \tag{4.17}$$

for $t \geq s \geq 0$ and $\binom{\eta}{\varphi} \in \mathcal{X}_0$.

In view of (4.13) and (4.17), the function w can be computed explicitly if we know the expressions of the Lebesgue extensions $\tilde{\mathfrak{B}}(t)$. In what follows, we will see that the expression of $\tilde{\mathfrak{B}}(t)$ depends on $\tilde{L}(t)$, the Lebesgue extension of $L(t)$ with respect to $\mathcal{S}_X$. Before showing this we have to prove the following lemma.

LEMMA 4.6. *Assume that the kernel $\ell(\cdot,\cdot)$ satisfies* (H). *Let $\eta \in X$ and set $\psi(\theta) := \eta$ for all $\theta \in [-r,0]$. Then $\psi \in D(\tilde{L}(s))$ and $\tilde{L}(s)\psi = L(s)\psi$ for all $s \geq 0$.*

Proof. Let $s \geq 0$, $\eta \in X$, and let $(\mathcal{S}_X, \Psi_L)$ be the nonautonomous observation system associated to $L(t)$. Let the sequence $\alpha_n \in C([-r,0])$ such that $0 \leq \alpha_n \leq 1$, $\alpha_n(0) = 0$ and $\alpha_n(\theta) = 1$ for $-r \leq \theta \leq -\frac{r}{n}$. Now we set $\psi^n = \alpha_n\psi$, so that $\psi^n \to \psi$

in $L^2([-r,0],X)$ as $n\to\infty$, $\psi^n\in C_0([-r,0],X)$ and $\Psi_L(s)\psi^n = L(\cdot)\mathcal{S}_X(\cdot,s)\psi^n$. Next, we take $0<\frac{r}{n}\le t\le\frac{r}{2}$ for $n\ge 3$. Then

$$\begin{aligned}\frac{1}{t}\int_s^{s+t}(\Psi_L(s)\psi)(\tau)\,d\tau - L(s)\psi \\ = \frac{1}{t}\int_s^{s+t}[\Psi_L(s)(\psi-\psi^n)](\tau)\,d\tau + \frac{1}{t}\int_s^{s+t}L(\tau)(\mathcal{S}_X(\tau,s)\psi^n-\psi)\,d\tau \\ + \frac{1}{t}\int_s^{s+t}[L(\tau)\psi - L(s)\psi]\,d\tau. \qquad (4.18)\end{aligned}$$

First, by Hölder's inequality,

$$\Big\|\frac{1}{t}\int_s^{s+t}[\Psi_L(s)\psi-\psi^n](\tau)\,d\tau\Big\| \le \frac{\gamma}{\sqrt{t}}\|\psi^n-\psi\|_2 \underset{n\to\infty}{\longrightarrow} 0.$$

Second, by the strong continuity of $(L(\tau))_{\tau\ge s}$,

$$\lim_{t\searrow 0}\frac{1}{t}\int_s^{s+t}[L(\tau)\psi - L(s)\psi]\,d\tau = 0.$$

Further,

$$\begin{aligned}\frac{1}{t}\int_s^{s+t}L(\tau)[\mathcal{S}_X(\tau,s)\psi^n-\psi]\,d\tau &= \frac{1}{t}\int_0^t\int_{-r}^{-\sigma}d\ell(s+\sigma,\theta)[\psi^n(\sigma+\theta)-\psi(\theta)]]\,d\sigma \\ &= \frac{1}{t}\int_0^t\int_{-\frac{r}{n}-\sigma}^{-\sigma}d\ell(s+\sigma,\theta)[\psi^n(\sigma+\theta)-\psi(\theta)]\,d\sigma,\end{aligned}$$

which implies that

$$\begin{aligned}\Big\|\frac{1}{t}\int_s^{s+t}L(\tau)[\mathcal{S}_X(\tau,s)\psi^n-\psi]\,d\tau\Big\| &\le 2\|\eta\|\frac{1}{t}\int_0^t\mu_\ell\Big(s+\sigma,\Big[-\frac{r}{n}-\sigma,-\sigma\Big]\Big)\,d\sigma \\ &\le 2\|\eta\|\sup_{0\le\sigma\le r}\mu_\ell(s+\sigma,[-2t,0])\to 0;\ (\text{as } t\searrow 0),\end{aligned}$$

due to (H). Thus the lemma now follows by taking the limits, first $n\to\infty$ and second $t\searrow 0$, in both sides of (4.18). □

Let us define $\Psi_{\mathfrak{B}}(s)\binom{\eta}{\varphi} = \mathfrak{B}(\cdot)\mathcal{T}(\cdot,s)\binom{\eta}{\varphi}$ for $\binom{\eta}{\varphi}\in\mathfrak{D}$ and $s\ge 0$. Then $(\mathcal{T},\Psi_{\mathfrak{B}})$ is a nonautonomous observation system. Moreover, by density of $\mathfrak{D}$ in $\mathcal{X}_0$, Hölder's inequality, and (4.13) one can see that

$$\begin{aligned}\frac{1}{t}\int_s^{s+t}\Big[\Psi_{\mathfrak{B}}(s)\binom{\eta}{\varphi}\Big](\tau)\,d\tau = \frac{1}{t}\int_s^{s+t}\binom{(\Psi_L(s)\varphi)(\tau)}{0}d\tau - \frac{1}{t}\int_s^{s+t}\binom{(\Psi_L(s)\psi)(\tau)}{0}d\tau \\ + \frac{1}{t}\int_s^{s+t}[\Psi_{\mathfrak{B}}(s)\binom{\eta}{\psi}](\tau)\,d\tau \qquad (4.19)\end{aligned}$$

for all $\binom{\eta}{\varphi}\in\mathcal{X}_0$, where we set $\psi(\theta)=\eta$.

LEMMA 4.7. *Let* (AT) *be holds and the kernel* $\ell(\cdot,\cdot)$ *satisfies* (H). *Then*

$$D(\tilde{\mathfrak{B}}(s)) = X\times D(\tilde{L}(s)) \quad \textit{and} \quad \tilde{\mathfrak{B}}(s) = \begin{pmatrix}0 & \tilde{L}(s)\\ 0 & 0\end{pmatrix}, \qquad s\ge 0.$$

Proof. The proof follows immediately by Lemma 4.6, (4.19), and the strong continuity of $\mathfrak{B}(t)$. □

The following theorem follows by (4.17) and Lemma 4.7.

THEOREM 4.8. *Assume that* (AT) *holds and* $\ell(\cdot,\cdot)$ *satisfies* (H). *Then the function* $w:[s,\infty)\to\mathcal{X}_0$ *given by* (4.16) *satisfies*

$$\begin{pmatrix} x(t)\\ x_t\end{pmatrix}=\mathcal{T}_L(t,s)\begin{pmatrix}\eta\\ \varphi\end{pmatrix}+\int_s^t\mathcal{T}_L(t,\sigma)\begin{pmatrix} f(\sigma)\\ 0\end{pmatrix}d\sigma \tag{4.20}$$

for $t\ge s\ge 0$ *and* $\begin{pmatrix}\eta\\ \varphi\end{pmatrix}\in\mathcal{X}_0$, *where the function* $x:[s-r,\infty)\to X$ *satisfies* $x_\cdot\in\mathcal{D}_s(\tilde{L}(\cdot))$ *and* (1.2).

The following result shows the relationship between the formulas (1.2) and (4.12).

PROPOSITION 4.9. *Assume that* (AT) *holds and* $\ell(\cdot,\cdot)$ *satisfies* (H). *Let the initial condition* $\begin{pmatrix}\eta\\ \varphi\end{pmatrix}\in\mathfrak{D}$. *Then the formulas given by* (1.2) *and* (4.12) *coincide.*

Proof. We first take $h\in C([-r,0],X)$ and $s\ge 0$. Due to the strong continuity of the family $(\mathfrak{B}(t))_{t\ge 0}$ one can see that $\tilde{\mathfrak{B}}(s)(h(0),h)^\top=\mathfrak{B}(s)(h(0),h)^\top$. Thus by Lemma 4.7, we obtain

$$L(s)h=\tilde{L}(s)h\quad\text{for}\quad h\in C([-r,0],X)\cap D(\tilde{L}(s)). \tag{4.21}$$

Let now the initial condition $\begin{pmatrix}\eta\\ \varphi\end{pmatrix}\in\mathfrak{D}$ and let x be the corresponding mild solution of (1.1). Then $(x(t),x_t)\in\mathfrak{D}$ for $t\ge s\ge 0$. Since by Theorem 4.8 we have $x_t\in D(\tilde{L}(t))$ for a.e. $t\ge s$, then $L(t)x_t=\tilde{L}(t)x_t$ for a.e. $t\ge 0$, due to (4.21). The proposition now follows by (1.2). □

In view of Theorem 4.8 and Proposition 4.9 we introduce the following definition.

DEFINITION 4.10. *Let* $s\ge 0$, $\begin{pmatrix}x\\ \varphi\end{pmatrix}\in\mathcal{X}_0$ *and* $f\in L^2_{loc}([s,\infty),X)$. *A continuous function* $x(\cdot):[s-r,\infty)\to X$ *is called a mild solution of* (1.1) *if* $x_\cdot\in\mathcal{D}_s(\tilde{L}(\cdot))$ *and satisfies* (1.2).

5. Nonautonomous linear systems with delays in the state, control, and observation variables. In this section we want to reformulate the delay system formed by (1.3) and (1.4) as an absolutely regular nonautonomous system.

Let us first prove the existence of a variation of constants formula for the delayed control equation (1.3). This formula will helps us to prove that (1.3) determines a nonautonomous control system.

As the kernel $k(\cdot,\cdot)$ satisfies (H), then $K(t)$ are nonautonomous admissible observation operators for $\mathcal{S}_U$. We then denote by $\tilde{K}(t)$ the Lebesgue extensions of $K(t)$ with respect to $\mathcal{S}_U$. Moreover, due to Theorem 3.6, $K(t)$ define an absolutely regular nonautonomous system on $L^2([-r,0],U),U,X$, which we denote throughout by $\Sigma_K:=(\mathcal{S}_U,\Phi,\Psi_K,\mathbb{F}_K)$. Note that the input segment u_t is the state trajectory of Σ_K. Then we have $u_\cdot\in\mathcal{D}_s(\tilde{K}(\cdot))$, due to Theorem 3.6.

The following proposition shows the existence of the mild solution of (1.3) in the case of smooth input u.

PROPOSITION 5.1. *Let* (AT) *be verified and assume that the kernels* $\ell(\cdot,\cdot)$ *and* $k(\cdot,\cdot)$ *satisfy* (H). *Let the initial condition* $\begin{pmatrix}\eta\\ \varphi\end{pmatrix}\in\mathcal{X}_0$. *If the input* u *is smooth, i.e.,* $u\in W^{1,2}_{loc}([s-r,\infty),U)$ *is such that* $u(s)=0$, *then* $u_t\in W^{1,2}([-r,0],U)$, $t\ge s\ge 0$, *and the function* $t\mapsto K(t)u_t$ *belongs to* $L^2_{loc}([s,\infty),U)$. *In particular, the equation* (1.3) *has a mild solution given by*

$$x(t)=T(t,s)\eta+\int_s^t T(t,\tau)[\tilde{L}(\tau)x_\tau+K(\tau)u_\tau]\,d\tau,\quad t\ge s. \tag{5.1}$$

If in addition $\binom{\eta}{\varphi} \in \mathfrak{D}$, *then this mild solution also satisfies*

$$\begin{aligned} x(t) &= T(t,s)\eta + \int_s^t T(t,\tau)\Big[L(\tau)x_\tau + K(\tau)u_\tau\Big]\,d\tau, \quad t \geq s \geq 0, \\ x(s+\theta) &= \varphi(\theta), \quad -r \leq \theta \leq 0. \end{aligned} \tag{5.2}$$

Proof. Let $\binom{\eta}{\varphi} \in \mathcal{X}_0$ and u be a smooth input, so $u_s = \zeta \in D(Q_U)$. Since $u_t = \mathcal{S}_U(t,s)\zeta + \Phi(t,s)u$ for $t \geq s \geq 0$, then by Lemma 3.5(a) we have $u_t \in W^{1,2}([-r,0],U)$, $t \geq s$. Now due to Corollary 3.7 we have $f(t) := K(t)u_t = \tilde{K}(t)u_t$ for a.e. $t \geq s$, which belongs to $L^2_{loc}([s,\infty),U)$. Thus the rest of the proof follows by Theorem 4.8 and Proposition 4.9. □

Let us now deal with with the solution of (1.3) for general input u.

THEOREM 5.2. *Let* (AT) *be verified and assume that the kernels* $\ell(\cdot,\cdot)$ *and* $k(\cdot,\cdot)$ *satisfy* (H). *Let the initial conditions* $\binom{\eta}{\varphi} \in \mathcal{X}_0$ *and* $\zeta \in L^2([-r,0],U)$. *For* $u \in L^2_{loc}[s-r,\infty),U)$ *there exists a squence of mild solutions* $x^n : [s-r,\infty) \to X$ *of* (1.3) *such that* $x^n \to x$ *in* $L^2_{loc}([s-r,\infty),X)$, *as* $n \to \infty$, *where* $x : [s-r,\infty) \to X$ *is the function satisfying* $x_\cdot \in \mathcal{D}_s(\tilde{L}(\cdot))$ *and*

$$\begin{aligned} x(t) &= T(t,s)\eta + \int_s^t T(t,\tau)[\tilde{L}(\tau)x_\tau + \tilde{K}(\tau)u_\tau]\,d\tau, \qquad t \geq s \geq 0. \\ x(t) &= \varphi(t-s), \qquad \text{a.e.} \quad t \leq s. \end{aligned} \tag{5.3}$$

Proof. Let $\binom{\eta}{\varphi} \in \mathcal{X}_0$ and $\zeta \in L^2([-r,0],U)$. Let us approximate the input $u \in L^2_{loc}[s-r,\infty),U)$ by a sequence of smooth inputs $u^n \in W^{1,2}_{loc}([s-r,\infty),U)$ such that $u^n(s) = 0$, so $\zeta^n := u^n_s \to \zeta$ in $L^2([-r,0],U)$ as $n \to \infty$. We set $f^n(t) := K(t)u^n_t$ for $t \geq s$ and let $x^n : [s-r,\infty) \to X$ with $x^n(s) = \eta$ and $x^n_s = \varphi$ be the mild solution of (1.3) corresponding to u^n and f^n, due to Proposition 5.1. By Theorem 4.8, the function $w^n(t) = (x^n(t), x^n_t)$ satisfies (4.20). Next, we set $f(t) := \tilde{K}(t)u_t$ for a.e. $t \geq s$, (which belongs to $L^2_{loc}([s,\infty),U)$) and let x be the function defined by (5.3). Thus, by Theorem 4.8, the function $w(t) = (x(t), x_t)^\top$ satisfies (4.20). We now claim that $x^n \to x$ in $L^2_{loc}([s,\infty),X)$ as $n \to \infty$. By using Theorem 2.3(iii), Theorem 3.6, and Corollary 3.7, we obtain $f^n \to f$ in $L^2_{loc}([s,\infty),U)$. Thus, Hölder's inequality and (4.20) imply that $w^n \to w$ in $L^2_{loc}([s,\infty),\mathcal{X}_0)$. In particular, $x^n \to x$ in $L^2_{loc}([s,\infty),X)$ and $x^{\phi(n)}_t \to x_t$ in $L^2([-r,0],X)$ for a.e. $t \in [s,s+t_0]$ as $n \to \infty$. □

Now let us introduce the following definition.

DEFINITION 5.3. *Assume that* (AT) *holds, and* $\ell(\cdot,\cdot)$ *and* $k(\cdot,\cdot)$ *satisfy* (H). *Let* (1.3) *with initial condition* $\binom{\eta}{\varphi} \in \mathcal{X}_0$, $\zeta \in L^2([-r,0],U)$ *and* $s \geq 0$. *Let the control function* $u \in L^2_{loc}([s,\infty),U)$ *such that* $u_s = \zeta$. *A function* $x : [s-r,\infty) \to X$ *is called the* generalized solution *of* (1.3) *if* $x_\cdot \in \mathcal{D}_s(\tilde{L}(\cdot))$ *and it satisfies* (5.3).

The following result follows immediately by Proposition 5.1, which motives Definition (5.3).

COROLLARY 5.4. *Assume that* (AT) *holds and the kernels* $\ell(\cdot,\cdot)$ *and* $k(\cdot,\cdot)$ *satisfy* (H). *Let the initial conditions* $s \geq 0$, $\binom{\eta}{\varphi} \in \mathfrak{D}$, *and* $\zeta \in D(Q_U)$ *and assume that* $u \in W^{1,2}_{loc}([s-r,\infty),U)$. *Then the generalized and mild solutions of* (1.3) *coincide.*

The following proposition is a consequence of Theorems 4.8 and 5.2, which will help us, later, to show that (1.3) determines a certain control system.

PROPOSITION 5.5. *Assume that* (AT) *holds and* $\ell(\cdot,\cdot)$ *and* $k(\cdot,\cdot)$ *satisfy* (H). *Let the initial conditions* $\binom{\eta}{\varphi} \in \mathcal{X}_0$, $\zeta \in L^2([-r,0],U)$, $s \geq 0$, *and* $u \in L^2_{loc}([s-r,\infty),U)$ *such that* $u_s = \zeta$. *Then* (1.3) *has a generalized solution* $x : [s-r,\infty) \to X$ *satisfying*

$x_\cdot \in \mathcal{D}_s(\tilde{L}(\cdot))$ and (1.5)*. Moreover, we have*

$$\begin{pmatrix} x(t) \\ x_t \end{pmatrix} = \mathcal{T}_L(t,s)\begin{pmatrix} \eta \\ \varphi \end{pmatrix} + \int_s^t \mathcal{T}_L(t,\sigma)\begin{pmatrix} \tilde{K}(\sigma)u_\sigma \\ 0 \end{pmatrix} d\sigma, \qquad t \geq s \geq 0. \tag{5.4}$$

In what follows, we endowed the Banach space $\mathcal{X} := \mathcal{X}_0 \times L^2([-r,0],U)$ with the norm $\|(\eta,\varphi,\zeta)\| = \|\eta\| + \|\varphi\|_2 + \|\zeta\|_2$. Further we define the linear operators

$$\mathcal{W}(t,s) : L^2([-r,0],U) \to \mathcal{X}_0, \quad \mathcal{W}(t,s)\zeta = \int_s^t \mathcal{T}_L(t,\sigma)\begin{pmatrix} (\Psi_K(s)\zeta)(\sigma) \\ 0 \end{pmatrix} d\sigma, \quad t \geq s \geq 0.$$

Using Hölder's inequality it can be verified that $\mathcal{W}(t,s)$, $t \geq s$, are bounded from $L^2([-r,0],U)$ to $\mathcal{X}_0$. Next, let us define the linear operators

$$\mathfrak{R}(t,s) : L^2_{loc}([s,\infty),U) \to \mathcal{X}_0, \quad \mathfrak{R}(t,s)u = \int_s^t \mathcal{T}_L(t,\sigma)\begin{pmatrix} (\mathbb{F}_K(s)u)(\sigma) \\ 0 \end{pmatrix} d\sigma, \quad t \geq s \geq 0.$$

Clearly, $\mathfrak{R}(t,s)$ are bounded linear operators from $L^2_{loc}([s,\infty,U)$ to $\mathcal{X}_0$.

Due to Theorem 2.3(iii) and Proposition 5.5 one can see that

$$(x(t), x_t, u_t)^\top = \mathcal{T}_{L,K}(t,s)(\eta,\varphi,\zeta) + \Phi_{L,K}(t,s)u \tag{5.5}$$

for $(\eta,\varphi,\zeta)^\top \in \mathcal{X}, t \geq s \geq 0$, and $u \in L^2_{loc}([s,\infty),U)$, where x is given by (1.5),

$$\mathcal{T}_{L,K}(t,s) := \left(\begin{array}{c|c} \mathcal{T}_L(t,s) & \mathcal{W}(t,s) \\ \hline 0 & \mathcal{S}_U(t,s) \end{array}\right), \tag{5.6}$$

and

$$\Phi_{L,K}(t,s)u := \begin{pmatrix} \mathfrak{R}(t,s)u \\ \Phi(t,s)u \end{pmatrix}. \tag{5.7}$$

It is not difficult to prove that $\mathcal{T}_{L,K} := (\mathcal{T}_{L,K}(t,s))_{t\geq s\geq 0}$ is a strongly continuous evolution family on $\mathcal{X}$. Moreover, using the regularity of the system Σ_K one can see that $(\mathcal{T}_{L,K}, \Phi_{L,K})$ is a nonautonomous control system on $\mathcal{X}, U$ (see [12] for a similar situation). Now due to (5.5) we have the following result.

PROPOSITION 5.6. *Assume that* (AT) *holds and $\ell(\cdot,\cdot)$ and $k(\cdot,\cdot)$ satisfy* (H). *Then $(\mathcal{T}_{L,K}, \Phi_{L,K})$ is a nonautonomous control system on $\mathcal{X}, U$, with control function $u \in L^2([s,\infty),U)$, $s \geq 0$, and state trajectory $z(t) = (x(t), x_t, u_t)$, $t \geq s \geq 0$, where x is given by* (5.3).

Next, we are interested in studying the output function given by (1.4). By assumption the delay operators $G(t)$ and $D(t)$ are given by the kernel $g(\cdot,\cdot)$ and $d(\cdot,\cdot)$ satisfying (H), so there are nonautonomous admissible observation operators for $\mathcal{S}_X$ and $\mathcal{S}_U$, respectively. Throughout we denote by $\tilde{G}(t)$ (resp., $\tilde{D}(t)$) the Lebesgue extensions of $G(t)$ (resp., $D(t)$) with respect to $\mathcal{S}_X$ (resp., $\mathcal{S}_U$). Further, by Theorem 3.6, the operators $D(t)$ are issued from an absolutely regular nonautonomous system which we denote by $\Sigma_D := (\mathcal{S}_U, \Phi, \Psi_D, \mathbb{F}_D)$.

According to Corollary 5.4 and Proposition 5.6, the control delay equation (1.3) determines the nonautonomous control system $(\mathcal{T}_{L,K}, \Phi_{L,K})$. It is then interesting to regard (1.4) as an observation equation for this latter.

Let us first consider the operators

$$\mathcal{G}(t) := [0\ G(t)] : \mathfrak{D} \longrightarrow Y, \qquad t \geq 0.$$

It is not difficult to show that $\mathcal{G}(t)$ are admissible observation operators for $\mathcal{T}$, so we denote by $\tilde{\mathcal{G}}(t)$ the Lebesgue extensions of $\mathcal{G}(t)$ with respect to $\mathcal{T}$. Arguing as in the proof of Lemma 4.7 we get

$$D(\tilde{\mathcal{G}}(t)) = X \times D(\tilde{G}(t)) \quad \text{and} \quad \tilde{\mathcal{G}}(t) = [0\ \tilde{G}(t)], \qquad t \geq 0. \tag{5.8}$$

Moreover, Proposition 4.3 and (4.15) imply that $\mathcal{T}_L(\cdot, s)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right) \in \mathcal{D}_s(\tilde{\mathfrak{B}}(\cdot)) \cap \mathcal{D}_s(\tilde{\mathcal{G}}(\cdot))$ for all $\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right) \in \mathcal{X}_0$.

We set $\Omega_t := X \times D(\tilde{G}(t)) \times D(\tilde{D}(t))$ for $t \geq s \geq 0$ and we define the operators

$$\mathfrak{C}(t) := [0\ \tilde{G}(t)\ \tilde{D}(t)] : \Omega_t \longrightarrow Y, \qquad t \geq s. \tag{5.9}$$

Further, we consider

$$\begin{gathered}\Psi_{L,K}(s) : \mathcal{X} \longrightarrow L^2_{loc}([s,\infty), Y),\\ \Psi_{L,K}(s)(\eta,\varphi,\zeta) := \mathfrak{C}(\cdot)\mathcal{T}_{L,K}(\cdot, s)(\eta,\varphi,\zeta), \quad s \geq 0.\end{gathered} \tag{5.10}$$

Due to Theorem 2.1 and Proposition 4.3, $\Psi_{L,K}(s)$ are well-defined, bounded linear operators from $\mathcal{X}$ to $L^2_{loc}([s,\infty), Y)$. Moreover, we have the following proposition.

Proposition 5.7. *Assume that* (AT) *holds and the kernels* $\ell(\cdot,\cdot)$, $k(\cdot,\cdot)$, $g(\cdot,\cdot)$, *and* $d(\cdot,\cdot)$ *satisfy* (H). *Let* $\Psi_{L,K}(t)$, $t \geq 0$, *be the operators defined by* (5.10). *Then* $(\mathcal{T}_{L,K}, \Psi_{L,K})$ *is a nonautonomous observation system on* $\mathcal{X}, Y$. *Let then* $\mathcal{C}(t)$ *be the operators representing this system as in* (2.5). *Then for* $\tau \geq s$,

$$\Omega_\tau \subset D(\mathcal{C}(\tau)) \quad \textit{and} \quad \mathcal{C}(\tau) = \mathfrak{C}(\tau) \quad \textit{on} \quad \Omega_\tau. \tag{5.11}$$

Proof. As we mentioned above, the operators $\Psi_{L,K}(s)$ are linear bounded from $\mathcal{X}$ to $L^2_{loc}([s,\infty), Y)$. Now, clearly we have

$$[\Psi_{L,K}(t)\mathcal{T}_{L,K}(t,s)(\eta,\varphi,\zeta)](\rho) = [\Psi_{L,K}(s)(\eta,\varphi,\zeta)](\rho)$$

for $(\eta,\varphi,\zeta)^\top \in \mathcal{X}$ and $\rho \geq t \geq s \geq 0$. This shows that $(\mathcal{T}_{L,K}, \Psi_{L,K})$ is a nonautonomous observation system on $\mathcal{X}, Y$. Let now $\mathcal{C}(t)$ be the operators representing this system as in (2.5). Let us take $\tau \geq s \geq 0$, $t \in (0,1]$ and $(\eta,\varphi,\zeta)^\top \in \Omega_\tau$. Then, by combining (4.15), (5.6), (5.8), and (5.10), we have

$$\begin{aligned}\Psi_{L,K}(\tau)(\eta,\varphi,\zeta) = \tilde{\mathcal{G}}(\cdot)\mathcal{T}(\cdot,\tau)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right) + \tilde{\mathcal{G}}(\cdot)\mathbb{K}^{\mathcal{T}}_\tau\tilde{\mathfrak{B}}(\cdot)\mathcal{T}_L(\cdot,\tau)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right)\\ + \tilde{\mathcal{G}}(\cdot)\mathbb{K}^{\mathcal{T}_L}_\tau\left(\begin{smallmatrix}\Psi_K(s)\zeta\\ 0\end{smallmatrix}\right) + \Psi_D(\tau)\zeta.\end{aligned}$$

Since $\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right) \in D(\tilde{\mathcal{G}}(\tau))$, then

$$\lim_{t\searrow 0}\frac{1}{t}\int_\tau^{\tau+t}\tilde{\mathcal{G}}(\sigma)\mathcal{T}(\sigma,\tau)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right)d\sigma = \tilde{\mathcal{G}}(\tau)\left(\begin{smallmatrix}x\\ \varphi\end{smallmatrix}\right).$$

On the other hand, due to (4.14), Hölder's inequality, and Proposition 2.2, we have

$$\left\| \frac{1}{t} \int_\tau^{\tau+t} \tilde{\mathcal{G}}(\sigma)[\mathbb{K}_\tau^{\mathcal{T}} \tilde{\mathfrak{B}}(\cdot)\mathcal{T}_L(\cdot,\sigma)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right)](\sigma)\, d\sigma \right\| \le c\|\tilde{\mathfrak{B}}(\cdot)\mathcal{T}_L(\cdot,\tau)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right)\|_{L^2([\tau,\tau+t],\mathcal{X}_0)} \underset{t\searrow 0}{\longrightarrow} 0.$$

Similarly, by Hölder's inequality and Proposition (4.3)(ii), we get

$$\lim_{t\searrow 0} \frac{1}{t} \int_\tau^{\tau+t} \tilde{\mathcal{G}}(\sigma)\left[\mathbb{K}_\tau^{\mathcal{T}_L}\left(\begin{smallmatrix}\Psi_K(\tau)\zeta\\ 0\end{smallmatrix}\right)\right](\sigma)\, d\sigma = 0.$$

Consequently,

$$\lim_{t\searrow 0} \frac{1}{t} \int_\tau^{\tau+t} [\Psi_{L,K}(\tau)(\eta,\varphi,\zeta)](\sigma) = \tilde{\mathcal{G}}(\tau)\left(\begin{smallmatrix}\eta\\ \varphi\end{smallmatrix}\right) + \tilde{D}(\tau)\zeta = \mathfrak{C}(\tau)(\eta,\varphi,\zeta),$$

so (5.11) is now verified. □

We note that the semigroup $\mathcal{T}_{L,K}$ corresponds to (1.3) in the case when $u(t) = 0$ for $t \ge s$. Thus Proposition 5.7 shows that the uncontrolled system associated with (1.3)–(1.4) is well posed.

Now we want to construct the operators which associate with every input function u the corresponding output function y defined by (1.4). We first remark, by Theorem 2.3(i), (5.7), Proposition 4.3(ii), and (5.8) that $\Phi_{L,K}(t,s)u \in \Omega_t$ for all $t \ge s \ge 0$ and $u \in L^2_{loc}([s,\infty),U)$. This allows us to introduce the following operators:

$$\mathbb{F}_{L,K}(s) : L^2_{loc}([s,\infty),U) \to L^2_{loc}([s,\infty),Y), \quad \mathbb{F}_{L,K}(s)u := \mathfrak{C}(\cdot)\Phi_{L,K}(\cdot,s)u, \quad s \ge 0.$$

Using Hölder's inequality, (5.7), and Proposition 4.3(ii) one can see that $\mathbb{F}_{L,K}(s)$ are linear bounded operators.

THEOREM 5.8. *Assume that* (AT) *holds and the kernels* $\ell(\cdot,\cdot)$, $k(\cdot,\cdot)$, $g(\cdot,\cdot)$, *and* $d(\cdot,\cdot)$ *satisfy* (H). *Let* $s \ge 0$, $(\eta,\varphi,\zeta)^\top \in \mathcal{X}$, *and* $u \in L^2_{loc}([s,\infty),U)$ *such that* $u_s = \zeta$. *Then the quadruple* $\Sigma_{L,K} := (\mathcal{T}_{L,K}, \Phi_{L,K}, \Psi_{L,K}, \mathbb{F}_{L,K})$ *is an absolutely regular nonautonomous system on* $\mathcal{X}, U, Y$ *having* $u(t)$ *as control function,* $z(t) = (x(t), x_t, u_t)$ *as state trajectory and output function* $y(t)$ *satisfying the representation*

$$y(t) = \tilde{G}(t)x_t + \tilde{D}(t)u_t \tag{5.12}$$

for almost every $t \ge s \ge 0$, *where* x *is the function given by* (1.5).

Proof. Let $0 \le s \le t \le \rho$ and $u \in L^2_{loc}([s,\infty),U)$ such that $u^0 := u_{|[s,t]}$ and $u^1 := u_{|[t,\rho]}$. Since $\Phi_{L,K}(t)$ satisfy (2.1), then by (5.10) we get

$$\begin{aligned}(\mathbb{F}_{L,K}(s)u)(\rho) &= \mathfrak{C}(\rho)[\Phi_{L,K}(\rho,t)u^1 + \mathcal{T}_{L,K}(\rho,t)\Phi_{L,K}(t,s)u^0] \\ &= (\mathbb{F}_{L,K}(t)u^1)(\rho) + (\Psi_{L,K}(t)\Phi(t,s)u^0)(\rho).\end{aligned}$$

Hence $\Sigma_{L,K}$ is a well-posed nonautonomous system on $\mathcal{X}, U, Y$. Now to prove the absolute regularity of $\Sigma_{L,K}$, let us consider the step $u_z(s) = z$, $s \ge 0$, $z \in U$, and $t \in (0,1]$. We have

$$\mathbb{F}_{L,K}(s)u_z = \tilde{\mathcal{G}}(\cdot)\mathbb{K}_s^{\mathcal{T}_L}\left(\begin{smallmatrix}\mathbb{F}_K(s)u_z\\ 0\end{smallmatrix}\right) + \mathbb{F}_D(s)u_z, \quad s \ge 0.$$

Then, by Proposition 4.3(ii), there exists a constant c independents of t such that

$$\frac{1}{t}\int_s^{s+t}\left\|\tilde{\mathcal{G}}(\tau)\left[\mathbb{K}_s^{\mathcal{T}_L}\begin{pmatrix}\mathbb{F}_K(s)u_z\\0\end{pmatrix}\right](\tau)\right\|^2 d\tau \le c\int_s^{s+t}\|[\mathbb{F}_K(s)u_z](\tau)\|^2\,d\tau \;\to\; 0 \text{ (as } t\searrow 0).$$

Thus our claim now follows by the absolute regularity of Σ_D (see Theorem 3.6). Now Proposition 5.6 shows that $u(t)$ and $z(t)=(x(t),x_t,u_t)$, $t\ge 0$, are the control function and the state trajectory of $\Sigma_{L,K}$, respectively, where $x(t)$ is given by (5.3). Due to Theorem 3.6 we have $z(t)\in\Omega_t$ for a.e. $t\ge s$. Let now $y(t)$ be the output function of $\Sigma_{L,K}$ and let $\mathcal{C}(t)$ be the operators representing the system $(\mathcal{T}_{L,K},\Psi_{L,K})$. Then, by (5.11), we have

$$y(t)=\mathcal{C}(t)z(t)=\mathfrak{C}(t)z(t)=\tilde{G}(t)x_t+\tilde{D}(t)u_t$$

for a.e. $t\ge s$. □

The following corollary shows, in fact, that the delay system (1.3)–(1.4) is completely determined by the absolutely regular nonautonomous system $\Sigma_{L,K}$ obtained in Theorem 5.8.

COROLLARY 5.9. *Assume that* (AT) *holds and the kernels* $\ell(\cdot,\cdot)$, $k(\cdot,\cdot)$, $g(\cdot,\cdot)$, *and* $d(\cdot,\cdot)$ *satisfy* (H). *Let the initial conditions* $s\ge 0$, $(\eta,\varphi,\zeta)^\top\in\mathfrak{D}\times D(Q_U)$. *Let the control function* $u\in W^{1,2}_{loc}([s-r,\infty),U)$ *such that* $u_s=\zeta$. *Let* $\Sigma_{L,K}$ *be the absolutely regular nonautonomous system obtained in Theorem* 5.8. *Then the state trajectory of this system is* $z(t)=(x(t),x_t,u_t)$ *for* $t\ge s$, *where* $x(t)$ *is the function given by* (5.2). *Further, the output function of* $\Sigma_{L,K}$ *satisfies* $y(t)=G(t)x_t+D(t)u_t$ *for a.e.* $t\ge s$.

Proof. The first statement of the corollary follows by Corollary 5.4 and Proposition 5.6. For the second, we know that $x_t\in C([-r,0],X)$, $t\ge s$, for initial condition $\binom{\eta}{\varphi}\in\mathfrak{D}$. Moreover, $x_t\in D(\tilde{G}(t))$ for a.e. $t\ge s$, due to Theorem 3.6. Thus $\tilde{G}(t)x_t=G(t)x_t$ for a.e. $t\ge s$, due to (4.21). Further, Corollary 3.7 implies that $\tilde{D}(t)u_t=D(t)u_t$ for a.e. $t\ge s$. Thus the last statement of the corollary follows. □

6. Application to a parabolic nonautonomous neutral equation. In this section we investigate a class of parabolic nonautonomous neutral equations on Banach space. Our abstract results obtained in section 5 allow as to introduce a new definition of the solutions of nonautonomous neutral equations and to prove the existence and uniqueness of those solutions.

Let $\mathcal{O}$ be a bounded domain in $\mathbb{R}^n$ such that the boundary $\partial\mathcal{O}$ is of class C^2. We set

$$A(t,\xi,\partial)=\sum_{i,j=1}^n a_{ij}(t,\xi)\partial_i\partial_j+\sum_{i=1}^n a_j(t,\xi)\partial_j+a_0(t,\xi),\qquad t\ge 0,\ \xi\in\mathcal{O},$$

where the functions $a_{ij},a_i,a_0:\mathbb{R}_+\times\overline{\mathcal{O}}$, $i,j=1,\dots,n$, are bounded and uniformly Hölder continuous and satisfying

$$\sum_{i,j=1}^n a_{ij}(t,\xi)\vartheta_i\vartheta_j\ge\gamma|\vartheta|^2$$

for a constant $\gamma>0$ and all $\vartheta\in\mathbb{R}^n$, $t\ge 0$, and $\xi\in\overline{\mathcal{O}}$.

We consider the nonautonomous neutral equation

$$\text{(nEq)}\quad \begin{cases} \dfrac{d}{dt}\Big[u(t,\xi) - \displaystyle\int_{-r}^{0} d\mathfrak{d}(t,\theta,\xi)u(t+\theta,\xi)\Big] \\ \qquad\qquad = A(t,\xi,\partial)\Big[u(t,\xi) - \displaystyle\int_{-r}^{0} d\mathfrak{d}(t,\theta,\xi)u(t+\theta,\xi)\Big] \\ \qquad\qquad\qquad + \displaystyle\int_{-r}^{0} dk(t,\theta,\xi)u(t+\theta,\xi), \qquad t \geq s \geq 0,\ \xi \in \overline{\mathcal{O}}, \\ \lim\limits_{t\to s}\Big[u(t,\xi) - \displaystyle\int_{-r}^{0} d\mathfrak{d}(t,\theta,\xi)u(t+\theta,\xi)\Big] = \eta(\xi), \qquad \xi \in \mathcal{O}, \\ u(t,\xi) = \zeta(t-s,\xi), \qquad \xi \in \mathcal{O},\ t \leq s, \\ u(t,\xi) = \displaystyle\int_{-r}^{0} d\mathfrak{d}(t,\theta,\xi)u(t+\theta,\xi), \qquad \xi \in \partial\mathcal{O},\ t \geq s, \end{cases}$$

for $\eta \in C_0(\mathcal{O})$ and $\zeta \in L^2([-r,0]\times\mathcal{O})$, where $\mathfrak{d}(\cdot,\cdot,\cdot), k(\cdot,\cdot,\cdot) : \mathbb{R}_+ \times [-r,0]\times\overline{\mathcal{O}} \to \mathbb{R}$ are some appropriate kernels.

Let us now introduce the auxiliary function

$$x(t,\xi) = u(t,\xi) - \int_{-r}^{0} d\mathfrak{d}(t,\theta,\xi)u(t+\theta,\xi), \quad t \geq s,\ \xi \in \overline{\mathcal{O}}.$$

Then (nEq) becomes

$$\text{(6.1)}\quad \begin{cases} \dfrac{d}{dt}x(t,\xi) = A(t,\xi,\partial)x(t,\xi) + \displaystyle\int_{-r}^{0} dk(t,\theta,\xi)u(t+\theta,\xi), & t \geq s \geq 0,\ \xi \in \overline{\mathcal{O}}, \\ x(s,\xi) = \eta(\xi), \quad u(t,\xi) = \zeta(t-s,\xi), & \xi \in \mathcal{O},\ t \leq s, \\ x(t,\xi) = 0, & t \geq s,\ \xi \in \partial\mathcal{O}, \\ u(t,\xi) = x(t,\xi) + \displaystyle\int_{-r}^{0} d\mathfrak{d}(t,\theta,\xi)u(t+\theta,\xi), & \xi \in \overline{\mathcal{O}},\ t \geq s. \end{cases}$$

On $X = L^2(\mathcal{O})$ we introduce the realizations $A(t)$ of the differential operator $A(t,\xi,\partial)$ with domains

$$\mathcal{D}(A(t)) = \Big\{h \in W^{2,2}(\mathcal{O}) : h = 0 \text{ on } \partial\mathcal{O}\Big\}.$$

In order to be in the setting of the previous section we set

$$D(t)u_t = \int_{-r}^{0} d\mathfrak{d}(t,\theta,\cdot)u(t+\theta,\cdot) \quad \text{and} \quad K(t)u_t = \int_{-r}^{0} dk(t,\theta,\cdot)u(t+\theta,\cdot).$$

On X we rewrite (6.1) in the following abstract form:

$$\text{(6.2)}\quad \begin{cases} \dfrac{d}{dt}x(t) = A(t)x(t) + K(t)u_t, & t \geq s \geq 0, \\ x(s) = \eta, \quad u_s = \zeta, & \\ u(t) = x(t) + D(t)u_t, & t \geq s. \end{cases}$$

It is known that $A(t)$ satisfy (AT) (see, e.g., [27]). Thus $A(t)$ generate the evolution family $T := (T(t,s))_{t\geq s\geq 0}$ on X, where

$$T(t,s)\eta = \int_{\mathcal{O}} \mathfrak{b}(t,s,\cdot,\xi)\eta(\xi)\,d\xi$$

for $t \geq s \geq 0$, $\eta \in X$, and a continuous kernel $\mathfrak{b}(t,s,\sigma,\xi)$, $t \geq s \geq 0$, $\sigma \in \overline{\mathcal{O}}$, $\xi \in \mathcal{O}$, satisfying the Gaussian estimate

$$|\mathfrak{b}(t,s,\sigma,\xi)| \leq M(t-s)^{-\frac{n}{2}} \exp\Big(-\frac{w|\sigma-\xi|^2}{t-s} + \tilde{w}(t-s)\Big)$$

for $0 < t-s \leq t_0$ and constants $M, w > 0$, and $\tilde{w} \in \mathbb{R}$ (see [26, sect. 6] and the references therein).

Next, we shall assume that the kernels $\mathfrak{d}(t,\cdot,\cdot,), k(t,\cdot,\cdot) \in BV([-r,0],\mathcal{L}(X))$, $t \geq 0$, satisfy the condition (H).

Observe that (6.2) corresponds to the delay system (1.3)–(1.4) with $L(t) = 0$, $G(t) = \delta_0$, $X = U = Y$, and the feedback law $y(t) = u(t)$. Thus to solve (nEq) it suffices to study the existence of the closed-loop system associated to (1.3)–(1.4) with respect to an appropriate feedback law.

The framework obtained in section 5 justifies the following definition of the solution of (nEq).

Definition 6.1. *A generalized solution of the initial value problem* (nEq) *is a pair* (x,u) *of functions* $x : [s,\infty) \to X$ *and* $u : [s-r,\infty) \to X$ *such that*

(i) x *is continuous,* $u_\bullet \in \mathcal{D}_s(\tilde{D}(\cdot)) \cap \mathcal{D}_s(\tilde{K}(\cdot))$,

(ii) $u_s = \zeta$ *a.e. on* $[-r,0]$,

(iii) *the pair* (x,u) *satisfies the following equations:*

$$x(t) = T(t,s)\eta + \int_s^t T(t,\tau)\tilde{K}(\tau)u_\tau \, d\tau, \tag{6.3}$$

$$x(t) = u(t) - \tilde{D}(t)u_t \tag{6.4}$$

for $\eta \in X$ *and a.e.* $t \geq s \geq 0$.

In the rest of this section, we are interested in studying the existence of the generalized solutions of the nonautonomous neutral equation (nEq). To this purpose, we denote by $\mathcal{T}_{0,K}$ and $\Phi_{0,K}$ the operators families given by (5.6) and (5.7), respectively, with $L(t) = 0$, $\mathcal{T}_0 = T$,

$$\mathcal{W}(t,s)\zeta = \int_s^t T(t,\sigma)(\Psi_K(s)\zeta)(\sigma)\, d\sigma \quad \text{and} \quad \mathfrak{R}(t,s)u = \int_s^t T(t,\sigma)(\mathbb{F}_K(s)u)(\sigma)\, d\sigma$$

for $t \geq s \geq 0$, $\zeta \in L^2([-r,0],X)$ and $u \in L^2_{loc}([s,\infty),X)$. Then, as in section 5, $(\mathcal{T}_{0,L}, \Phi_{0,L})$ is a nonautonomous control system on $\mathcal{X}_0, X$.

Let us now rewrite the operators $\mathfrak{C}(t)$ given by (5.9) as

$$\mathfrak{C}(t) := [I \ \tilde{D}(t)] : \Omega_t \longrightarrow X, \qquad t \geq s, \tag{6.5}$$

where $\Omega_t = X \times \mathcal{D}(\tilde{D}(t))$ for $t \geq s \geq 0$. So we denote by $\Psi_{0,K}(s)$, $s \geq 0$, the operators given by (5.10) with $L(t) = 0$, $Y = X$, so that $(\mathcal{T}_{0,K}, \Psi_{0,K})$ is a nonautonomous observation system on $\mathcal{X}_0, X$. Finally, we set

$$\mathbb{F}_{0,K}(s) : L^2_{loc}([s,\infty),X) \to L^2_{loc}([s,\infty),X), \quad \mathbb{F}_{0,K}(s)u := \mathfrak{C}(\cdot)\Phi_{0,K}(\cdot,s)u, \quad s \geq 0.$$

Lemma 6.2. *Let the kernels* $\mathfrak{d}(\cdot,\cdot,\cdot,)$ *and* $k(\cdot,\cdot,\cdot)$ *satisfy* (H) *and assume that* $\sup_t Var(\mathfrak{d}(t,\cdot,\cdot))^0_{-r}$ *is small enough. Then* $\Sigma_{0,K} := (\mathcal{T}_{0,K}, \Phi_{0,K}, \Psi_{0,K}, \mathbb{F}_{0,K})$ *is an absolutely regular nonautonomous system on* $\mathcal{X}_0, X, X$ *with admissible feedback operators* $\Delta(t) = I_X$, $t \geq 0$.

Proof. Theorem 5.8 shows that $\Sigma_{0,K} := (\mathcal{T}_{0,K}, \Phi_{0,K}, \Psi_{0,K}, \mathbb{F}_{0,K})$ is an absolutely regular nonautonomous system on $\mathcal{X}_0, X, X$. Now (5.7) and (6.5) imply

$$I_X - \mathbf{F}_K(s) = I_X - \mathbb{F}_D(s) - \mathfrak{R}(\cdot, s), \qquad s \geq 0, \tag{6.6}$$

where $\mathbb{F}_D(s)$ are the input-output operators of the absolutely regular nonautonomous system on $L^2([-r,0],X), X, X$ associated with the delay operators $D(t)$, $t \geq 0$, (see Theorem 3.6). We claim that $\Delta(t) = I_X$, $t \geq 0$, are admissible feedback operators for Σ_D. For this let $t_0 > 0$ and $v \in L^2_{loc}([s,\infty), X)$. Due to (3.10) we have

$$\int_s^{s+t_0} \|(\mathbb{F}_D(s)v)(t)\|^2 \, dt \leq c \sup_t Var(\mathfrak{d}(t,\cdot))^0_{-r} \|v\|_{L^2([s,s+t_0],X)}. \tag{6.7}$$

According to (6.7) and the fact that $\sup_t Var(\mathfrak{d}(t,\cdot))^0_{-r}$ is small enough, the operators $I_X - \mathbb{F}_D(s+t_0, s)$, $s \geq 0$, have uniformly bounded inverses on $L^2([s, s+t_0], X)$ for some $t_0 > 0$. Then, due to (6.6), $I_X - \mathbf{F}_K(s+t_1, s)$, $s \geq 0$, have uniformly bounded inverses on $L^2([s, s+t_1], X)$ for some $t_1 > 0$ as well. This ends the proof. ∎

THEOREM 6.3. *Let the kernels* $\mathfrak{d}(\cdot,\cdot,\cdot,)$ *and* $k(\cdot,\cdot,\cdot)$ *satisfy* (H) *and assume that* $\sup_t Var(\mathfrak{d}(t,\cdot,\cdot))^0_{-r}$ *is small enough. Then the nonautonomous neutral equation* (nEq) *has a unique generalized solution.*

Proof. We are looking for a pair of functions (x,u) satisfying Definition 6.1. For this we let v and y denote the control and the observation functions, respectively, of the absolutely regular nonautonomous system $\Sigma_{0,K}$. Due to Lemma 6.2, the closed-loop system $\Sigma^I_{0,K}$ associated with $\Sigma_{0,K}$ and $\Delta(t) = I_X$ exists. Now if we consider the feedback law $v = y + v_c$ for a suitable input v_c then, by (2.9) and Theorem 2.3(iii), the state trajectory of $\Sigma^I_{0,K}$ is given by $z(t) = \mathcal{T}_{0,K}(t,s)(\eta,\zeta) + \Phi_{0,K}(t,s)(v_c + y)$ for $t \geq s \geq 0$ and $(\eta,\zeta)^\top \in \mathcal{X}_0$. We now choose $v_c = 0$ and $u = y$ with $u_s = \zeta$. Then $z(t) = (x(t), u_t)$ for $t \geq s \geq 0$, where x satisfies (6.3) (see Theorem 5.8). Again by Theorem 5.8 we have $u(t) = y(t) = \mathfrak{C}(t)z(t) = x(t) + \tilde{D}(t)u_t$ for a.e. $t \geq s \geq 0$. Thus x satisfies (6.4). This ends the proof. ∎

Acknowledgments. I would like to thank the associate editor and the two referees whose detailed comments helped me to improve the organization and the content of the paper.

REFERENCES

[1] P. ACQUISTAPACE, *Evolution operators and strong solutions of abstract linear parabolic equations*, Differential Integral Equations, 1 (1988), pp. 433–457.

[2] P. ACQUISTAPACE AND B. TERRENI, *A unified approach to abstract linear nonautonomous parabolic equations*, Rend. Sem. Mat. Univ. Padova, 87 (1987), pp. 47–107.

[3] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite-Dimensional Systems, Vol.* I, Birkhäuser, Boston, 1992.

[4] C. CHICONE AND Y. LATUSHKIN, *Evolution Semigroups in Dynamical Systems and Differential Equations*, AMS, Providence, RI, 1999.

[5] R. F. CURTAIN, H. LOGEMANN, S. TOWNLEY, AND H. ZWART, *Well-posedness, stabilizability and admissibility for Pritchard-Salamon systems*, J. Math. Systems Estim. Control, 4 (1994), pp. 1–38.

[6] R. F. CURTAIN AND H. ZWART, *Introduction to Infinite-Dimensional Linear Systems*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.

[7] M. C. DELFOUR, *The linear quadratic optimal control problem with delays in state and control variables: A state space approach*, SIAM J. Control Optim., 24 (1986), pp. 835–883.

[8] K. J. ENGEL AND R. NAGEL, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.

[9] G. Gühring, F. Räbiger, and R. Schnaubelt, *A characteristic equation for nonautonomous partial functional differential equations*, J. Differential Equations, 181 (2002), pp. 439–462.
[10] S. Hadd and A. Idrissi, *Regular linear systems governed by systems with state, input and output delays*, IMA J. Control Information, 22 (2005), pp. 423–439.
[11] S. Hadd, A. Idrissi, and A. Rhandi, *The regular linear systems associated to the shift semigroups and application to control delay systems*, Math. Control Signals Systems, to appear.
[12] S. Hadd, A. Rhandi, and R. Schnaubelt, *Feedback theory for nonautonomous linear systems with input delays*, submitted.
[13] J. K. Hale and S. M. V. Lunel, *Introduction to Functional Differential Equations*, Applied Math. Sci., 99, Springer-Verlag, New York, 1993.
[14] D. Hinrichsen and A. J. Pritchard, *Robust stability of linear evolution operators on Banach spaces*, SIAM J. Control Optim., 32 (1994), pp. 1503–1541.
[15] A. Ichikawa, *Quadratic control of evolution equations with delays in control,* SIAM J. Control Optim., 20 (1982), pp. 645–668.
[16] B. Jacob, *Time-Varying Infinite Dimensional State-Space Systems*, Ph.D. thesis, Fachbereich Mathematik/Informatik, Universität Bremen, Bremen, Germany, 1995.
[17] B. Jacob, *Optimal control of time-varying well-posed linear systems on finite time horizon*, in Mathematical Theory of Networks and Systems, Proceedings of the MTNS-98, A. Beghi, I. Finesso, and G. Picci, eds., II Polografo, Padova, Italy, 1998, pp. 483–486.
[18] B. Jacob, V. Dragan, and A. J. Pritchard, *Robust stability of infinite dimensional time-varying systems with respect to non-linear perturbation,* Integral Equations Operator Theory, 22 (1995), pp. 440–462.
[19] T. Kato, *Integration of the equation of evolution in Banach space,* J. Math. Soc. Japan, 5 (1953), pp. 208–234.
[20] T. Kato, *Linear evolution equations of "hyperbolic" type*, J. Fac. Sci. Univ. Tokyo, 25 (1970), pp. 241–258.
[21] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, 1995.
[22] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci., 44, Springer-Verlag, New York, 1983.
[23] A. J. Pritchard and D. Salamon, *The linear quadratic control problem for retarded systems with delays in control and observation,* IMA J. Control Inform., 2 (1985), pp. 335–362.
[24] H. Tanabe, *Remarks on the equations in a Banach space*, Osaka Math. J., 112 (1960), pp. 145–166.
[25] D. Salamon, *Infinite-dimentional linear system with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[26] R. Schnaubelt, *Feedback for nonautonomous regular linear systems*, SIAM J. Control Optim., 41 (2002), pp. 1141–1165.
[27] R. Schnaubelt, *Asymptotic behaviour of parabolic nonautonomous evolution equations*, in Functional Analytic Methods for Evolution Equations 1815, M. Iannelli, R. Nagel, S. Piazzera, eds., Springer, New York, 2004, pp. 401–472.
[28] O. J. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, Cambridge, 2005.
[29] R. B. Vinter and R. H. Kwong, *The infinite time quadratic control problem for linear systems with state and control delays: An evolution equation approach*, SIAM J. Control Optim., 19 (1981), pp. 139–153.
[30] G. Weiss, *Admissible observation operators for linear semigoups*, Israel J. Math., 65 (1989), pp. 17–43.
[31] G. Weiss, *Admissibility of unbounded control operators,* SIAM J. Control Optim., 27 (1989), pp. 527–545.
[32] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 273–297

EXPONENTIAL STABILIZATION OF WELL-POSED SYSTEMS BY COLOCATED FEEDBACK*

RUTH F. CURTAIN† AND GEORGE WEISS‡

Abstract. We consider well-posed linear systems whose state trajectories satisfy $\dot{x} = Ax + Bu$, where u is the input and A is an essentially skew-adjoint and dissipative operator on the Hilbert space X. This means that the domains of A^* and A are equal and $A^* + A = -Q$, where $Q \geq 0$ is bounded on X. The control operator B is possibly unbounded, but admissible and the observation operator of the system is B^*. Such a description fits many wave and beam equations with colocated sensors and actuators, and it has been shown for many particular cases that the feedback $u = -\kappa y + v$, with $\kappa > 0$, stabilizes the system, strongly or even exponentially. Here, y is the output of the system and v is the new input. We show, by means of a counterexample, that if B is sufficiently unbounded, then such a feedback may be unsuitable: the closed-loop semigroup may even grow exponentially. (Our counterexample is a simple regular system with feedthrough operator zero.) However, we prove that if the original system is exactly controllable and observable and if κ is sufficiently small, then the closed-loop system is exponentially stable.

Key words. well-posed linear system, regular linear system, positive-real transfer function, output feedback, exact controllability and observability, skew-adjoint operator, colocated sensors and actuators, exponential stability

AMS subject classifications. 93C20, 93C25, 93D15, 93B05

DOI. 10.1137/040610489

1. Introduction and the main result.

In this paper, we consider the stabilization of a special class of well-posed linear systems, as described below. To specify our terminology and notation, we recall that for any well-posed linear system Σ with input space U, state space X and output space Y, all Hilbert spaces, the state trajectories $z \in C([0,\infty), X)$ are described by the differential equation

$$\dot{z}(t) = Az(t) + Bu(t), \tag{1.1}$$

where $u \in L^2_{loc}([0,\infty), U)$ is the input function. The operator $A : \mathcal{D}(A) \to X$ is the generator of a strongly continuous semigroup of operators $\mathbb{T}$ on X and the (possibly unbounded) operator B is an admissible control operator for $\mathbb{T}$. In general, the output function y is in $L^2_{loc}([0,\infty), Y)$. If $u = 0$ and $z(0) \in \mathcal{D}(A)$, then y is given by

$$y(t) = Cz(t) \qquad \forall\, t \geq 0,$$

where $C : \mathcal{D}(A) \to Y$ is an admissible observation operator for $\mathbb{T}$. B is called the control operator of Σ and C is called the observation operator of Σ. If $z(0) = 0$, then the input and output functions u and y are related by the formula

$$\hat{y}(s) = \mathbf{G}(s)\hat{u}(s), \tag{1.2}$$

*Received by the editors June 24, 2004; accepted for publication (in revised form) September 5, 2005; published electronically March 15, 2006. This research was supported by EPSRC (from the UK) under the portfolio partnership grant GR/S61256/01.
http://www.siam.org/journals/sicon/45-1/61048.html
†Department of Mathematics, University of Groningen, 9700 AV Groningen, The Netherlands (R.F.Curtain@rug.nl).
‡Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK (G.Weiss@imperial.ac.uk).

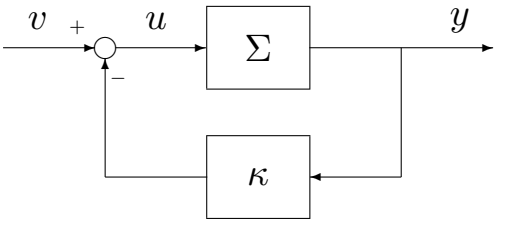


FIG. 1. *The open-loop system Σ with negative output feedback via κ. If the number $\kappa > 0$ is sufficiently small, then this is a new well-posed linear system Σ^κ, called the closed-loop system, which is input-output stable. If Σ is exactly controllable and exactly observable, then Σ^κ is exponentially stable.*

where a hat denotes the Laplace transform and $\mathbf{G}$ is the transfer function of Σ. The formula (1.2) holds for all $s \in \mathbb{C}$ with $\operatorname{Re} s$ sufficiently large. We refer to sections 3 and 4 for more details and references on admissibility and on well-posed linear systems. Now we specify the special class of systems studied in this paper.

Assumption ESAD. The operator A is *essentially skew-adjoint and dissipative*, which means that $\mathcal{D}(A) = \mathcal{D}(A^*)$ and there exists a $Q \in \mathcal{L}(X)$ with $Q \geq 0$ such that

$$Ax + A^*x = -Qx \qquad \forall\, x \in \mathcal{D}(A). \tag{1.3}$$

This implies that $\mathbb{T}$ is a contraction semigroup. Note that A is a bounded perturbation of the skew-adjoint operator $A + \frac{1}{2}Q$. Such a model is often used to describe the dynamics of oscillating systems, such as waves or flexible structures (often $Q = 0$, so that $\mathbb{T}$ is unitary); for a literature survey see section 2.

Assumption COL. $Y = U$ and $C = B^*$.

In the literature on the stabilization of flexible structures, a very popular way of implementing actuators and sensors is through colocated pairs, i.e., an actuator and sensor pair act at the same physical position. This often leads to assumption COL being satisfied, often with a finite-dimensional U.

Our aim is to show that for certain numbers $\kappa > 0$, the static output feedback law $u = -\kappa y + v$ stabilizes the system, where v is the new input function. The closed-loop system Σ^κ is shown as a block diagram in Figure 1 (see section 4 for some background on output feedback). The system Σ^κ is called *input-output stable* if its transfer function $\mathbf{G}^\kappa = \mathbf{G}(I + \kappa\mathbf{G})^{-1}$ is uniformly bounded on the open right half-plane where $\operatorname{Re} s > 0$. It is called *exponentially stable* if the growth bound of its semigroup is negative. Our main result is the following theorem.

THEOREM 1.1. *Suppose that Σ is a well-posed linear system which is exactly controllable, exactly observable and satisfies assumptions* ESAD *and* COL. *Then there exists a $\kappa_0 > 0$ (possibly $\kappa_0 = \infty$) such that for all $\kappa \in (0, \kappa_0)$, the feedback law $u = -\kappa y + v$ (where u and y are the input and the output of Σ) leads to a closed-loop system Σ^κ which is well-posed and exponentially stable.*

In fact, this result is only a corollary of Theorem 5.8, in which the assumptions are weaker than exact controllability and exact observability. We also give a formula for κ_0 based on the transfer function $\mathbf{G}$, see Theorem 5.8.

In all the published examples that we are aware of, the feedback $u = -\kappa y + v$ is stabilizing for all $\kappa > 0$, at least in the input-output sense, and often strongly or exponentially. In section 5 we give an example of a simple open-loop system Σ which fits into our framework, and it is regular with feedthrough operator zero (see section 4 for definitions), but for which the feedback $u = -\kappa y + v$ is only exponentially stabilizing for sufficiently small $\kappa > 0$. For too large a κ, the closed-loop semigroup $\mathbb{T}^\kappa$ will have a positive growth rate.

In section 6 we introduce the theoretical framework for discussing colocated feedback for systems described by second order differential equations in time, with a suitable version of Theorem 1.1. As an illustration, we outline the problem of stabilizing a Rayleigh beam with two sensors located in one point, which fits into the theory developed in this section. With the actuators designed such that $B = C^*$ and with proportional output feedback with not too high feedback gain, the closed-loop system is exponentially stable, for any position of the two sensors.

Exact controllability and exact observability are very restrictive conditions, especially if U is finite-dimensional, see, for example, the discussion in Rebarber and Weiss [29]. There is a rich literature dealing with various specific linear systems or classes of systems that satisfy ESAD and COL and that are approximately controllable and observable (or satisfy other related assumptions). In these papers, the main conclusion is usually the weak or strong stability of the closed-loop system (and various nonexponential decay rates of the energy). This area will be examined in a sequel to this paper, which will contain a unified theory of the strong stabilization of systems satisfying ESAD and COL.

2. Comments on the literature and a self-contained presentation of the finite-dimensional case. Many models of controlled flexible structures satisfy assumptions ESAD and COL. The feedback $u = -\kappa y + v$ is very simple to implement and it is often used in the stabilization of these structures. Our results may be regarded as an abstract unifying theory for colocated exponential stabilization of flexible structures. However, it must be pointed out that not all such examples in the literature satisfy all our assumptions. In particular, the open-loop system is not always well posed.

Much of the early work on the stabilization of flexible structures concerned finite-dimensional systems, see, for example, Benhabib et al. [9] and Joshi [18] and the references therein. They used the feedback $u = -\kappa B^* z + v$ because of its simplicity and its nice robustness properties. Indeed, it works for all systems with a positive-real transfer function, as we shall explain below (see also Desoer and Vidyasagar [12]). A valuable recent source for the finite-dimensional theory is Gawronski [13]. For a survey on the stabilization of finite-dimensional linear systems by static output feedback we refer to Syrmos et al. [39] (see also Zeheb and Hertz [54]).

We think that it will be instructive to give here a short self-contained presentation of the finite-dimensional version of our main result (Theorem 1.1), as well as of some related results, without any claims to novelty. Our infinite-dimensional arguments will go along the same lines, but with many more technicalities.

Recall that a square matrix-valued transfer function $\mathbf{G}$, analytic on the open right half-plane $\mathbb{C}_0$, is called *positive-real* if $\overline{\mathbf{G}(s)} = \mathbf{G}(\overline{s})$ and

$$\mathbf{G}(s)^* + \mathbf{G}(s) \geq 0 \qquad \forall\, s \in \mathbb{C}_0. \tag{2.1}$$

Positive-real transfer functions were introduced in electrical network theory, but they have strong connections with systems theory formulated in state space, see Anderson and Vongpanitlerd [3]. In particular, if the real square matrix A is dissipative, then for any real matrix B of appropriate dimensions, $\mathbf{G}(s) = B^*(sI - A)^{-1}B$ is positive-real. Such transfer functions often occur as models of flexible structures with colocated actuators and sensors, see [9, 13].

Now consider an arbitrary (but square) $m \times m$ matrix-valued transfer function $\mathbf{G}$. If the closed-loop system with transfer function $\mathbf{G}^\kappa$ is obtained from the open-loop system with transfer function $\mathbf{G}$ via the feedback $u = -\kappa y + v$, as in Figure 1, then

$\mathbf{G}^\kappa = \mathbf{G}(I + \kappa\mathbf{G})^{-1}$. The following lemma gives a simple sufficient condition for $\mathbf{G}^\kappa$ to be bounded on $\mathbb{C}_0$, a fact which is written as $\mathbf{G}^\kappa \in H^\infty$.

PROPOSITION 2.1. *Suppose that $cI + \mathbf{G}$ is positive-real for some $c \geq 0$, and denote $\kappa_0 = \frac{1}{c}$ (for $c = 0$, take $\kappa_0 = \infty$). Then for any $\kappa \in (0, \kappa_0)$, $\mathbf{G}^\kappa \in H^\infty$.*

Proof. Denoting $a = \frac{1}{\kappa}$ and $T(s) = aI + \frac{1}{2}(\mathbf{G}(s)^* + \mathbf{G}(s))$, we have $T(s) \geq (a - c)I > 0$. Hence, for every $\mathrm{v} \in \mathbb{C}^m$ with $\|\mathrm{v}\| = 1$ and for all $s \in \mathbb{C}_0$,

$$\|(aI + \mathbf{G}(s))\mathrm{v}\| \geq \mathrm{Re}\,\langle (aI + \mathbf{G}(s))\mathrm{v}, \mathrm{v}\rangle = \langle T(s)\mathrm{v}, \mathrm{v}\rangle \geq a - c,$$

hence

$$\left\|(aI + \mathbf{G}(s))^{-1}\right\| \leq \frac{1}{a - c} \tag{2.2}$$

so that $(aI + \mathbf{G})^{-1} \in H^\infty$. We rewrite $\mathbf{G}^\kappa$ in the form

$$\mathbf{G}^\kappa = a\left[I - a(aI + \mathbf{G})^{-1}\right],$$

which shows that $\mathbf{G}^\kappa \in H^\infty$ (a bound will be given below). □

PROPOSITION 2.2. *With the assumptions of Proposition 2.1, denote $c^\kappa = \frac{c}{1 - \kappa c}$ so that $c^\kappa \geq 0$. Then $c^\kappa I + \mathbf{G}^\kappa$ is positive-real.*

Proof. Notice that $\kappa c \in [0, 1)$. Introduce the transfer function

$$\mathbf{H} = \frac{\kappa}{1 - \kappa c}(cI + \mathbf{G})$$

so that clearly $\mathbf{H}$ is positive-real. It is readily verified that

$$c^\kappa I + \mathbf{G}^\kappa = \frac{1}{\kappa(1 - \kappa c)}\,\mathbf{H}(I + \mathbf{H})^{-1}. \tag{2.3}$$

This implies that for all $s \in \mathbb{C}_0$,

$$\begin{aligned}\kappa(1 - \kappa c)&\left[(c^\kappa I + \mathbf{G}^\kappa(s))^* + (c^\kappa I + \mathbf{G}^\kappa(s))\right] \\ &= (I + \mathbf{H}(s)^*)^{-1}\left[\mathbf{H}(s)^* + \mathbf{H}(s) + 2\mathbf{H}(s)^*\mathbf{H}(s)\right](I + \mathbf{H}(s))^{-1}.\end{aligned}$$

It is easy to see that the right-hand side of the above equation is nonnegative. Since clearly $\overline{\mathbf{G}^\kappa(s)} = \mathbf{G}^\kappa(\overline{s})$, it follows that $c^\kappa I + \mathbf{G}^\kappa$ is positive-real. □

LEMMA 2.3. *If H is an $m \times m$ matrix such that $\mathrm{Re}\,\langle H\mathrm{v}, \mathrm{v}\rangle \geq 0$ for all $\mathrm{v} \in \mathbb{C}^m$, then $I + H$ is invertible and $\|H(I + H)^{-1}\| \leq 1$.*

Proof. The invertibility is clear. Denote $T = H(I + H)^{-1}$; then it is easy to see that $H(I - T) = T$. If in $\mathrm{Re}\,\langle H\mathrm{v}, \mathrm{v}\rangle \geq 0$ we choose $\mathrm{v} = (I - T)z$, then we get

$$0 \leq \mathrm{Re}\,\langle Tz, (I - T)z\rangle = \mathrm{Re}\,\langle Tz, z\rangle - \|Tz\|^2$$

so that $\|Tz\|^2 \leq \mathrm{Re}\,\langle Tz, z\rangle$. Taking $\|z\| = 1$ and using the Cauchy inequality, we obtain $\|Tz\|^2 \leq \|Tz\|$, which implies that $\|Tz\| \leq 1$ for all z with $\|z\| = 1$. □

It is easy to see that, with the above notation, we also have $\|(I + H)^{-1}\| \leq 1$.

PROPOSITION 2.4. *With the assumptions of Proposition 2.1, we have*

$$\|c^\kappa I + \mathbf{G}^\kappa(s)\| \leq \frac{1}{\kappa(1 - \kappa c)} \qquad \forall\, s \in \mathbb{C}_0. \tag{2.4}$$

Proof. Since $\mathbf{H}$ is positive-real, the last lemma implies that

$$\|\mathbf{H}(s)(I+\mathbf{H}(s))^{-1}\| \le 1 \qquad \forall\, s \in \mathbb{C}_0.$$

Now (2.4) follows from this estimate and (2.3). □

The bound (2.4) is quite sharp: in the scalar case, as s approaches a pole of $\mathbf{G}$ on the imaginary axis, (2.4) tends to an equality. Note that in Propositions 2.1–2.4, $\mathbf{G}$ was not assumed to be rational.

Now consider a finite-dimensional system Σ described by the equations

$$\begin{cases} \dot z(t) &= Az(t)+Bu(t), \\ y(t) &= B^*z(t)+Du(t), \end{cases}$$

where A, B and D are real matrices of appropriate dimensions and $A^*+A = -Q \le 0$ so that assumptions ESAD and COL are satisfied. If the number κ is such that $I+\kappa D$ is invertible, then the feedback $u = -\kappa y + v$ leads to the closed-loop system Σ^κ described by the equations

$$\begin{cases} \dot z(t) &= \big(A - B\kappa(I+\kappa D)^{-1}B^*\big)\, z(t) + B(I+\kappa D)^{-1}v(t), \\ y(t) &= (I+\kappa D)^{-1}B^*z(t) + D(I+\kappa D)^{-1}v(t). \end{cases} \tag{2.5}$$

We are interested in conditions that guarantee that the matrix

$$A^\kappa = A - B\kappa(I+\kappa D)^{-1}B^*$$

appearing in (2.5) is stable, i.e., its eigenvalues are in the open left half-plane. The finite-dimensional version of Theorem 1.1 reads as follows.

PROPOSITION 2.5. *With the above notation, if (A,B) is controllable and (A,B^*) is observable, then there exists a $\kappa_0 > 0$ such that for all $\kappa \in (0,\kappa_0)$, A^κ is stable.*

Like Theorem 1.1, this is only a corollary of a stronger result (Proposition 2.6 below) which is a little more complicated to state. This stronger result is the finite-dimensional counterpart of Theorem 5.8. To state it, we introduce the notation

$$\kappa_0 = \frac{1}{c}, \quad \text{where} \quad c = \|E^+\|, \quad E = -\frac{1}{2}(D^*+D). \tag{2.6}$$

Here, E^+ denotes the positive part of E, i.e., $E^+ = EP^+$ where P^+ is the spectral projector corresponding to all the positive eigenvalues of E (hence $E^+ \ge E$ but $\|E^+\| \le \|E\|$). Note that if $E \le 0$, then $c = 0$, and then we put $\kappa_0 = \infty$.

PROPOSITION 2.6. *With the above notation, if (A,B) is stabilizable and (A,B^*) is detectable, then for all $\kappa \in (0,\kappa_0)$, A^κ is stable.*

Proof. The transfer function of Σ is $\mathbf{G}(s) = B^*(sI-A)^{-1}B + D$ and we have

$$\mathbf{G}(s)^* + \mathbf{G}(s) \ge D^* + D \qquad \forall\, s \in \mathbb{C}_0.$$

This implies that, denoting $E = -\frac{1}{2}(D^*+D)$, $\mathbf{G}+E$ is positive-real, and hence $\mathbf{G}+E^+$ is positive-real. Then for $c = \|E^+\|$, $cI+\mathbf{G}$ is positive-real. (A simpler but more restrictive choice would be $c = \|E\|$.) The transfer function of the closed-loop system Σ^κ from (2.5) is $\mathbf{G}^\kappa$. By Proposition 2.1, we have $\mathbf{G}^\kappa \in H^\infty$ for all $\kappa \in (0,\kappa_0)$. It is easy to show that Σ^κ is stabilizable and detectable (because these properties are preserved by output feedback). It is well known that if a finite-dimensional system is stabilizable, detectable and input-output stable, then it is stable. □

We now return to the discussion of the infinite-dimensional case. In this paper we replace the concept of positive-real transfer function with the more general concept of a positive transfer function. An analytic $\mathcal{L}(U)$-valued function on $\mathbb{C}_0$ is called a *positive transfer function* if (2.1) holds. Note that for simplicity we have dropped the condition concerning complex conjugates, since it is not needed in our arguments: in particular, it is not needed in Proposition 2.1. (Defining the complex conjugate of an operator is a bit awkward and not necessary.) In the finite-dimensional case, this slight generalization amounts to dropping the requirement that the system matrices should be real. Note that if the generator A is dissipative and $B \in \mathcal{L}(U, X)$ (i.e., B is bounded), then $\mathbf{G}(s) = B^*(sI - A)^{-1}B$ is a positive transfer function. This is not always true for unbounded B, as we shall show.

The first PDE examples fitting into our framework assumed that A is dissipative, B is bounded and the open-loop transfer function is $\mathbf{G}(s) = B^*(sI - A)^{-1}B$ (i.e., $D = 0$); see Bailey and Hubbard [5], Balakrishnan [6, 7], Russell [30], and Slemrod [33, 34]. In this case, $\mathbf{G}$ is positive and (by Proposition 2.1) the feedback $u = -\kappa y + v$ stabilizes in an input-output sense. Of course, the most desirable type of stability is exponential stability and for this, in the special case $A^* + A = 0$, we need the system to be exactly controllable (or equivalently, exactly observable). This is the setup studied in Haraux [15], Liu [24] and others.

Many examples of flexible beams, plates and hybrid structures with colocated control and observation have been shown to be exponentially stabilizable by static output feedback; see, for example, Chen [10, 11], Rebarber [27], Triggiani [40], Tucsnak and Weiss [41], Luo, Guo and Morgul [25], Guo and Luo [14], and Ammari and Tucsnak [2]. In many of these examples the approach is a classical Lyapunov one, with the key step being the appropriate PDE formulation so that the energy of the system can play the role of a Lyapunov functional. If one examines these examples carefully, one can recognize that they fit into our framework (the assumptions ESAD and COL are satisfied) and they use the feedback $u = -\kappa y + v$ for stabilization. There are also examples in the literature where the open-loop system is not well posed, but application of the static feedback results in a well-posed exponentially stable closed-loop system, see Rebarber [27], Rebarber and Townley [28], Weiss [49], Lasiecka and Triggiani [20]. These examples are not covered by the theory in this paper (except for some partial results in Remarks 5.3 and 5.4).

The recent paper [20] is interesting because it also gives an abstract framework for treating exponential stabilization by colocated feedback. The assumptions are ESAD (with $Q = 0$), COL, $0 \in \rho(A)$ and $A^{-\frac{1}{2}}B \in \mathcal{L}(U, X)$. The last two assumptions (and also the fact that $Q = 0$) are more restrictive than in our framework, but on the other hand, they do not require the open-loop system to be well posed. We note that our examples in sections 5 and 6 are not covered by the theory in [20], because they do not satisfy $A^{-\frac{1}{2}}B \in \mathcal{L}(U, X)$. The main thrust of [20] is to give examples of non-well-posed systems satisfying the assumptions mentioned earlier, for which $A - BB^*$ is exponentially stable. (See also the corrections to [20] in [21].)

As mentioned at the end of section 1, there is also a rich literature dealing with weak or strong stabilization by colocated feedback, which we shall discuss elsewhere.

3. Admissible control and observation operators. In this section we gather, for easy reference, some basic facts about admissible control and observation operators and various controllability and observability concepts. For proofs and for more details we refer to the literature.

We assume that X is a Hilbert space and $A : \mathcal{D}(A) \to X$ is the generator of a strongly continuous semigroup $\mathbb{T}$ on X. We define the Hilbert space X_1 as $\mathcal{D}(A)$ with the norm $\|z\|_1 = \|(\beta I - A)z\|$, where $\beta \in \rho(A)$ is fixed (this norm is equivalent to the graph norm). The Hilbert space X_{-1} is the completion of X with respect to the norm $\|z\|_{-1} = \|(\beta I - A)^{-1}z\|$. This space is isomorphic to $\mathcal{D}(A^*)^*$, the dual of $\mathcal{D}(A^*)$ with respect to the pivot space X, and we have the continuous embeddings

$$X_1 \subset X \subset X_{-1}. \tag{3.1}$$

$\mathbb{T}$ extends to a semigroup on X_{-1}, denoted by the same symbol. The generator of this extended semigroup is an extension of A, whose domain is X, so that $A : X \to X_{-1}$, see Weiss [42]. We denote by $\omega(\mathbb{T})$ the growth bound of $\mathbb{T}$. The semigroup $\mathbb{T}$ is called *exponentially stable* if $\omega(\mathbb{T}) < 0$.

We assume that U is a Hilbert space and $B \in \mathcal{L}(U, X_{-1})$ is an *admissible control operator* for $\mathbb{T}$, defined as in [42]. This means that if z is the solution of $\dot{z}(t) = Az(t) + Bu(t)$, as in (1.1), which is an equation in X_{-1}, with $z(0) = z_0 \in X$ and $u \in L^2([0,\infty), U)$, then $z(t) \in X$ for all $t \geq 0$. In this case, z is a continuous X-valued function of t. We have that for all $t \geq 0$,

$$z(t) = \mathbb{T}_t z_0 + \Phi_t u, \tag{3.2}$$

where $\Phi_t \in \mathcal{L}(L^2([0,\infty), U), X)$ is defined by

$$\Phi_t u = \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma) d\sigma. \tag{3.3}$$

The above integration is done in X_{-1}, but the result is in X. The Laplace transform of z is

$$\hat{z}(s) = (sI - A)^{-1} \left[z_0 + B\hat{u}(s) \right].$$

B is called *bounded* if $B \in \mathcal{L}(U, X)$ (and unbounded otherwise). If B is an admissible control operator for $\mathbb{T}$, then for every $\omega > \omega(\mathbb{T})$ there exists a positive constant δ such that

$$\|(sI - A)^{-1} B\|_{\mathcal{L}(U,X)} \leq \frac{\delta}{\sqrt{\operatorname{Re} s}} \qquad \forall \operatorname{Re} s > \omega. \tag{3.4}$$

If $\dim U < \infty$ and $\mathbb{T}$ is normal or contractive, then (3.4) implies the admissibility of B, see Jacob and Partington [16] and Weiss [47, 48]. Similarly, if $\mathbb{T}$ is left-invertible, then again (3.4) implies the admissibility of B, see [47]. If $\mathbb{T}$ is invertible and $\tilde{\mathbb{T}}$ is the inverse semigroup (i.e., $\tilde{\mathbb{T}}_t = (\mathbb{T}_t)^{-1}$), then B is admissible for $\tilde{\mathbb{T}}$ if and only if it is admissible for $\mathbb{T}$, see, for example, [42].

The *degree of unboundedness* of an operator $B \in \mathcal{L}(U, X_{-1})$, denoted $\alpha(B)$, is the infimum of those $\alpha \geq 0$ for which there exist positive constants δ, ω such that

$$\|(\lambda I - A)^{-1} B\|_{\mathcal{L}(U,X)} \leq \frac{\delta}{\lambda^{1-\alpha}} \qquad \forall \lambda \in (\omega, \infty). \tag{3.5}$$

It is clear from (3.4) that for any admissible $B \in \mathcal{L}(U, X_{-1})$ we have $\alpha(B) \leq \frac{1}{2}$, and if B is bounded then $\alpha(B) = 0$ (see [29] for further comments on $\alpha(B)$).

We assume that Y is another Hilbert space and $C \in \mathcal{L}(X_1, Y)$ is an *admissible observation operator* for $\mathbb{T}$, defined as in Weiss [43]. This means that for every $T > 0$ there exists a $K_T \geq 0$ such that

$$\int_0^T \|C\mathbb{T}_t z_0\|^2 dt \leq K_T^2 \|z_0\|^2 \qquad \forall\, z_0 \in \mathcal{D}(A). \tag{3.6}$$

C is called *bounded* if it can be extended such that $C \in \mathcal{L}(X, Y)$.

We regard $L^2_{loc}([0,\infty), Y)$ as a Fréchet space with the seminorms being the L^2 norms on the intervals $[0, n]$, $n \in \mathbb{N}$. Then the admissibility of C means that there is a continuous operator $\Psi : X \to L^2_{loc}([0,\infty), Y)$ such that

$$(\Psi z_0)(t) = C\mathbb{T}_t z_0 \qquad \forall\, z_0 \in \mathcal{D}(A). \tag{3.7}$$

The operator Ψ is completely determined by (3.7), because $\mathcal{D}(A)$ is dense in X. We introduce an extension of C, called the Λ-extension of C, defined by

$$C_\Lambda z_0 = \lim_{\lambda \to +\infty} C\lambda(\lambda I - A)^{-1} z_0, \tag{3.8}$$

whose domain $\mathcal{D}(C_\Lambda)$ consists of all $z_0 \in X$ for which the limit exists. We shall also use the *weak Λ-extension* of C, $C_{\Lambda w}$. It is defined as in (3.8), but replacing the strong limit by the weak limit. Thus, $C_{\Lambda w}$ is an extension of C_Λ to an even larger subspace of X, denoted by $\mathcal{D}(C_{\Lambda w})$. If we replace C by C_Λ, formula (3.7) becomes true for all $z_0 \in X$ and for almost every $t \geq 0$. If $y = \Psi z_0$, then its Laplace transform is

$$\hat{y}(s) = C(sI - A)^{-1} z_0. \tag{3.9}$$

The following duality result holds: if $\mathbb{T}$ is a semigroup on X with generator A, then $B \in \mathcal{L}(U, X_{-1})$ is an admissible control operator for $\mathbb{T}$ if and only if $B^* : \mathcal{D}(A^*) \to U$ is an admissible observation operator for the dual semigroup $\mathbb{T}^*$.

The dual version of (3.4) is as follows: if C is an admissible observation operator for $\mathbb{T}$, then for every $\omega > \omega(\mathbb{T})$ there exists a positive constant δ such that

$$\|C(sI - A)^{-1}\|_{\mathcal{L}(X,Y)} \leq \frac{\delta}{\sqrt{\operatorname{Re} s}} \qquad \forall\, \operatorname{Re} s > \omega. \tag{3.10}$$

The *degree of unboundedness* of C, denoted $\alpha(C)$, is defined similarly as $\alpha(B)$. We have $\alpha(C) = \alpha(C^*)$, where C^* is regarded as a control operator for $\mathbb{T}^*$.

DEFINITION 3.1. *Let A be the generator of a strongly continuous semigroup $\mathbb{T}$ on X and let $B \in \mathcal{L}(U, X_{-1})$ be an admissible control operator for $\mathbb{T}$. The pair (A, B) is exactly controllable in time $T > 0$, if for every $x_0 \in X$ there exists a $u \in L^2([0, T], U)$ such that*

$$\Phi_T u = \int_0^T \mathbb{T}_{T-\sigma} B u(\sigma) d\sigma = x_0.$$

(A, B) is exactly controllable if it is exactly controllable in some finite time $T > 0$.

It will be useful to note that if A is the generator of a strongly continuous group (i.e., $-A$ is also a semigroup generator), then (A, B) is exactly controllable in time T if and only if $(-A, B)$ is exactly controllable in time T.

We introduce observability concepts via duality. Suppose that A is the generator of the strongly continuous semigroup $\mathbb{T}$ on X and $C \in \mathcal{L}(X_1, Y)$ is an admissible

observation operator for $\mathbb{T}$. Of course, this is equivalent to C^* being an admissible control operator for the dual semigroup $\mathbb{T}^*$. We say that (A, C) is *exactly observable* (*in time* T) if (A^*, C^*) is exactly controllable (in time T).

For more details on exact controllability in an operator-theoretic setting we refer also to Avdonin and Ivanov [4], Guo and Luo [14], Jacob and Zwart [17], Miller [26], Rebarber and Weiss [29], Russell and Weiss [31] and the references therein. In the PDE setting, the relevant literature on controllability is overwhelming, and we mention the books of Lions [23], Lagnese and Lions [22] and Komornik [19] and the paper of Bardos, Lebeau and Rauch [8].

4. Some background on well-posed linear systems. In this section we collect some basic facts about well-posed and regular linear systems, their transfer functions, output feedback and closed-loop systems. Only one result (Proposition 4.1) is new, and of course its proof is included.

By a *well-posed linear system* we mean a linear time-invariant system such that on any finite time interval, the operator from the initial state and the input function to the final state and the output function is bounded. The input, state and output spaces are Hilbert spaces, and the input and output functions are of class $L^2_{\rm loc}$. To express this more clearly, let us denote by U the input space, by X the state space and by Y the output space of a well-posed linear system Σ. The input and output functions u and y are locally L^2 functions with values in U and in Y. The state trajectory z is an X-valued function. The boundedness property mentioned earlier means that for every $\tau > 0$ there is a $c_\tau \geq 0$ such that

$$\|z(\tau)\|^2 + \int_0^\tau \|y(t)\|^2 \, dt \leq c_\tau^2 \left(\|z(0)\|^2 + \int_0^\tau \|u(t)\|^2 \, dt \right) \tag{4.1}$$

(with c_τ independent of $z(0)$ and of u). For the detailed definition, background and examples we refer to Salamon [32], Staffans [35, 36, 37], Weiss [45, 46], Weiss and Rebarber [51] and Weiss, Staffans and Tucsnak [52].

We recall some necessary facts about well-posed linear systems. Let Σ be such a system, with input space U, state space X and output space Y. Then there are operators A, B, C satisfying the assumptions in the previous section, which are related to Σ in the following way. First of all, the state trajectories of Σ satisfy the equation (1.1), so that they are given by (3.2). $\mathbb{T}$ is called the *semigroup* of Σ, A is called its *semigroup generator*, the family $\Phi = (\Phi_t)_{t \geq 0}$ is called the *input maps* of Σ, and B is called the *control operator* of Σ. If u is the input function of Σ, z_0 is its initial state, and y is the corresponding output function, then

$$y = \Psi z_0 + \mathbb{F} u. \tag{4.2}$$

Here, Ψ is an operator as in (3.7), called the (extended) *output map* of Σ and C is called the *observation operator* of Σ. $\mathbb{F}$ is a continuous linear operator from $L^2_{loc}([0, \infty), U)$ to $L^2_{loc}([0, \infty), Y)$, called the (extended) *input-output map* of Σ. The formal definition of Σ in [44, 45] is via $\mathbb{T}, \Phi, \Psi$ and $\mathbb{F}$ so that $\Sigma = (\mathbb{T}, \Phi, \Psi, \mathbb{F})$.

The operator $\mathbb{F}$ appearing above is easiest to represent using Laplace transforms. To state this in precise terms, we introduce the notation $\mathbb{C}_\alpha$ for the open right half-plane consisting of those $s \in \mathbb{C}$ for which $\operatorname{Re} s > \alpha$. An operator-valued analytic function $\mathbf{G}$ is called *well posed* (or proper) if there exists $\alpha \in \mathbb{R}$ such that the domain of $\mathbf{G}$ contains $\mathbb{C}_\alpha$ and $\mathbf{G}$ is uniformly bounded on $\mathbb{C}_\alpha$. We do not distinguish between two well-posed functions if one is a restriction of the other (to a smaller domain in

$\mathbb{C}$). If $\mathbb{F}$ is the input-output map of a well-posed system, then there exists a unique $\mathcal{L}(U,Y)$-valued well-posed function $\mathbf{G}$, called the *transfer function* of Σ, which determines $\mathbb{F}$ as follows. If $u \in L^2([0,\infty),U)$ and $y = \mathbb{F}u$, then y has a Laplace transform $\hat{y}$ and, for $\operatorname{Re} s > \max\{\omega(\mathbb{T}),0\}$ (recall that $\omega(\mathbb{T})$ is the growth bound of $\mathbb{T}$),

$$\hat{y}(s) = \mathbf{G}(s)\hat{u}(s).$$

This determines $\mathbb{F}$, since $L^2([0,\infty),U)$ is dense in $L^2_{loc}([0,\infty),U)$. $\mathbf{G}$ is uniformly bounded on every half-plane $\mathbb{C}_\omega$ with $\omega > \omega(\mathbb{T})$. Because of the identification of well-posed functions mentioned earlier, by a transfer function we mean in fact an equivalence class of analytic functions. For all $s,\beta \in \mathbb{C}_{\omega(\mathbb{T})}$ we have

$$\mathbf{G}(s) - \mathbf{G}(\beta) = C\left[(sI-A)^{-1} - (\beta I - A)^{-1}\right]B. \tag{4.3}$$

This shows that $\mathbf{G}$ is determined by A, B and C up to an additive constant operator. We call (A,B,C) the *generating triple* of Σ. In the time domain, the output function y corresponding to the input function u and state trajectory z is given by

$$y(t) = C_\Lambda\left[z(t) - (\beta I - A)^{-1}Bu(t)\right] + \mathbf{G}(\beta)u(t), \tag{4.4}$$

valid for almost every $t \geq 0$, if $\beta \in \mathbb{C}_{\omega(\mathbb{T})}$. Thus, the system Σ is completely determined (via (1.1) and (4.4)) by its generating triple (A,B,C) and by the value of its transfer function at one point.

The well-posed linear system Σ is called *regular* if the limit

$$\lim_{\lambda \to +\infty} \mathbf{G}(\lambda)\mathrm{v} = D\mathrm{v} \tag{4.5}$$

exists for every $\mathrm{v} \in U$, where λ is real (see [44, 45]). In this case, the operator $D \in \mathcal{L}(U,Y)$ is called the *feedthrough operator* of Σ. Regularity is equivalent to the fact that the product $C_\Lambda(sI-A)^{-1}B$ makes sense, for some (hence, for every) $s \in \rho(A)$. In this case, the formula for $\mathbf{G}$ looks like the finite-dimensional one:

$$\mathbf{G}(s) = C_\Lambda(sI-A)^{-1}B + D. \tag{4.6}$$

Moreover, the function y from (4.2) satisfies, for almost every $t \geq 0$,

$$y(t) = C_\Lambda z(t) + Du(t), \tag{4.7}$$

where z is the state trajectory of the system (compare this with (4.4)). The operators A,B,C,D are called the *generating operators* of Σ, because they determine Σ via (1.1) and (4.7). In particular, if A is a generator, one of B and C is admissible and the other is bounded, then for any $D \in \mathcal{L}(U,Y)$, the operators A,B,C and D are the generating operators of a regular linear system.

The well-posed linear system Σ is called *weakly regular* if the limit in (4.5) exists in the weak sense, see [37]. In this case, the weak limit still defines an operator $D \in \mathcal{L}(U,Y)$, called the *feedthrough operator* of Σ. Weak regularity is equivalent to the fact that the product $C_{\Lambda w}(sI-A)^{-1}B$ makes sense, for some (hence, for every) $s \in \rho(A)$. In this case, the formulas (4.6) and (4.7) become valid after we replace C_Λ by $C_{\Lambda w}$. This slight generalization of regular systems (to weakly regular ones) is sometimes needed because the dual of a regular linear system is not regular, in general, but it is weakly regular. Note that if Y is finite-dimensional, then the regularity of Σ is equivalent to its weak regularity.

PROPOSITION 4.1. *Let Σ be a well-posed linear system with control operator B, observation operator C and transfer function $\mathbf{G}$. If*

$$\alpha(B) + \alpha(C) < 1, \tag{4.8}$$

then Σ is regular. In fact, $\lim_{\lambda\to+\infty} \mathbf{G}(\lambda)$ exists in the operator norm.

Proof. From (4.3) it follows that $\mathbf{G}'(s) = C(sI - A)^{-2}B$ for all $s \in \mathbb{C}_{\omega(\mathbb{T})}$. So for all $\lambda > 0$ sufficiently large, we have

$$\|\mathbf{G}'(\lambda)\|_{\mathcal{L}(U,Y)} \le \|C(\lambda I - A)^{-1}\|_{\mathcal{L}(X,Y)} \cdot \|(\lambda I - A)^{-1}B\|_{\mathcal{L}(U,X)}.$$

Hence, for every $\alpha_1 > \alpha(C)$ and $\alpha_2 > \alpha(B)$ we can find $\delta_1, \delta_2 > 0$ such that, for all $\lambda > 0$ sufficiently large,

$$\|\mathbf{G}'(\lambda)\|_{\mathcal{L}(U,Y)} \le \frac{\delta_1}{\lambda^{1-\alpha_1}} \cdot \frac{\delta_2}{\lambda^{1-\alpha_2}}.$$

If (4.8) holds, then α_1 and α_2 can be chosen such that $\alpha_1 + \alpha_2 = 1 - \varepsilon$, with $\varepsilon > 0$. Denoting $\delta = \delta_1\delta_2$, we obtain that for all sufficiently large $\lambda > 0$,

$$\|\mathbf{G}'(\lambda)\|_{\mathcal{L}(U,Y)} \le \frac{\delta}{\lambda^{1+\varepsilon}}, \qquad \varepsilon > 0.$$

By integration we obtain that for $\lambda_1 < \lambda_2$ sufficiently large,

$$\|\mathbf{G}(\lambda_1) - \mathbf{G}(\lambda_2)\| \le \int_{\lambda_1}^{\lambda_2} \|\mathbf{G}'(\lambda)\| d\lambda \;\le\; \frac{\delta}{\varepsilon} \left| \frac{1}{\lambda_1^{\varepsilon}} - \frac{1}{\lambda_2^{\varepsilon}} \right|$$

so that $\lim_{\lambda\to+\infty} \mathbf{G}(\lambda)$ exists (which is stronger than (4.5)). ∎

It is often useful to introduce the space

$$Z = X_1 + (\beta I - A)^{-1}BU, \tag{4.9}$$

where $\beta \in \rho(A)$ (the space Z does not depend on the choice of β). Z is a Hilbert space with the following factor space norm:

$$\|z\|_Z = \inf_{z=x+(\beta I-A)^{-1}Bv} \left(\|x\|_1^2 + \|v\|_U^2\right)^{\frac{1}{2}}.$$

It is easy to see that $X_1 \subset Z \subset X$ with continuous embeddings, but X_1 need not be dense in Z. It was shown in [37, section 3] that for any well-posed system, C can be extended to an operator $\overline{C} \in \mathcal{L}(Z;Y)$ (this extension may be nonunique). For every $s \in \rho(A)$, the product $\overline{C}(sI - A)^{-1}B$ exists. For every such extension $\overline{C}$ there exists a unique $D \in \mathcal{L}(U,Y)$ such that $\mathbf{G}(s) = \overline{C}(sI - A)^{-1}B + D$. (For regular systems, we may take $\overline{C} = C_\Lambda$ and then D becomes the feedthrough operator of the system.) If $u \in \mathcal{H}^1_{loc}(0,\infty;U)$ and $Az(0) + Bu(0) \in X$, then for all $t \ge 0$ we have $z(t) \in Z$, the equation $\dot{z}(t) = Az(t) + Bu(t)$ holds for every $t \ge 0$, the output function y is continuous, and it is given for all $t \ge 0$ by

$$y(t) = \overline{C}z(t) + Du(t). \tag{4.10}$$

Let Σ be a well-posed linear system with generating triple (A, B, C) and transfer function $\mathbf{G}$. An operator $K \in \mathcal{L}(Y,U)$ is called an *admissible feedback operator* for Σ

(or for $\mathbf{G}$) if $I - \mathbf{G}K$ has a well-posed inverse (equivalently, if $I - K\mathbf{G}$ has a well-posed inverse). If this is the case, then the system with output feedback $u = Ky + v$ (see Figure 1, but with $-K$ in place of κ) is well posed (its input is v, its state and output are the same as for Σ), see [46]. This new system is called the *closed-loop system* corresponding to Σ and K, and it is denoted by Σ^K. Its transfer function is $\mathbf{G}^K = \mathbf{G}(I - K\mathbf{G})^{-1} = (I - \mathbf{G}K)^{-1}\mathbf{G}$. We have that $-K$ is an admissible feedback operator for Σ^K and the corresponding closed-loop system is Σ. Let us denote by (A^K, B^K, C^K) the generating triple of Σ^K. Then for every $x_0 \in \mathcal{D}(A^K)$ and for every $z_0 \in \mathcal{D}(A)$,

$$A^K x_0 = \left(A + BKC^K\right) x_0, \qquad Az_0 = \left(A^K - B^K KC\right) z_0. \tag{4.11}$$

Note that $B^K \in \mathcal{L}(U, X_{-1}^K)$ and $C^K \in \mathcal{L}(X_1^K, Y)$, where X_{-1}^K and X_1^K are the counterparts of X_{-1} and X_1 for Σ^K. Any interconnection of finitely many well-posed linear systems can be thought of as a closed-loop system in the above sense.

The following invariance result is taken from [46, section 6].

PROPOSITION 4.2. *Let Σ be a well-posed linear system, let K be an admissible feedback operator for Σ, and let Σ^K be the corresponding closed-loop system. We denote by (A, B, C) the generating triple of Σ and by (A^K, B^K, C^K) the generating triple of Σ^K. Then the following holds:*

(a) *(A, B) is exactly controllable if and only if (A^K, B^K) has the same property,*

(b) *(A, C) is exactly observable if and only if (A^K, C^K) has the same property.*

We recall some concepts that are used in the optimal control literature (under different names), following the formulation in Weiss and Rebarber [51].

DEFINITION 4.3. *Let A be the generator of a strongly continuous semigroup $\mathbb{T}$ on X and suppose that $B \in \mathcal{L}(U, X_{-1})$ is an admissible control operator for $\mathbb{T}$. Then (A, B) is optimizable if for every $z_0 \in X$ there exists a $u \in L^2([0, \infty), U)$ such that the state trajectory z defined in (3.2) is in $L^2([0, \infty), X)$.*

Let $C \in \mathcal{L}(X_1, Y)$ be an admissible observation operator for $\mathbb{T}$. Then (A, C) is estimatable if (A^*, C^*) is optimizable.

Estimatability can be formulated also directly, without using adjoints, see [51]. Optimizability is one possible generalization of the concept of stabilizability from finite-dimensional systems to infinite-dimensional ones. Similarly, estimatability is one possible generalization of detectability. It is clear that exact controllability implies optimizability, and exact observability implies estimatability. It was shown in [51] that optimizability and estimatability are invariant under output feedback (just like exact controllability and exact observability).

PROPOSITION 4.4. *With the notation from Proposition 4.2, (A, B) is optimizable if and only if (A^K, B^K) is optimizable and (A, C) is estimatable if and only if (A^K, C^K) is estimatable.*

We quote from [51] the following characterization of exponential stability.

THEOREM 4.5. *A well-posed linear system is exponentially stable if and only if it is optimizable, estimatable and input-output stable.*

It follows from this theorem that a well-posed linear system is exponentially stable if it is exactly controllable, exactly observable, and input-output stable.

5. Positivity and exponential stabilization. In this section we prove a rather technical result (Theorem 5.2) about the positivity of the transfer function $\mathbf{G} + E$, where $\mathbf{G}$ is the transfer function of the system Σ satisfying ESAD and COL, and E is given by a certain limit, see (5.4). If B is not too unbounded (more precisely, if

$\alpha(B) < \frac{1}{2}$), then Σ is regular and

$$E = -\frac{1}{2}(D^* + D), \tag{5.1}$$

where D is the feedthrough operator of Σ. Formula (5.1) may hold even if $\alpha(B) = \frac{1}{2}$, but we show by a counterexample that (5.1) may fail if $\alpha(B) = \frac{1}{2}$. If $0 \in \rho(A)$ and $Q = 0$, then it turns out that E is related to the direct current (DC)-gain of the system, $\mathbf{G}(0)$:

$$E = -\frac{1}{2}\left[\mathbf{G}(0)^* + \mathbf{G}(0)\right],$$

see (5.5). For this formula to hold, Σ does not need to be regular.

We also prove a result (Proposition 5.7) about the stabilization of systems whose transfer function $\mathbf{G}$ is such that for some $c \geq 0$, $cI + \mathbf{G}$ is a positive transfer function. These two results imply our main result about exponential stabilization, Theorem 5.8 (which is stronger than Theorem 1.1). We consider the general class of well-posed linear systems, but we also explain some consequences for the smaller but simpler class of weakly regular linear systems.

Notation and assumptions. We consider a well-posed system Σ with input and output space U, state space X, semigroup $\mathbb{T}$, and transfer function $\mathbf{G}$. Σ satisfies the assumptions ESAD and COL from section 1, and the operators A, Q and B are as in section 1. The space X_{-1} is defined as in section 3.

It follows from ESAD that the growth bound of $\mathbb{T}$ satisfies $\omega(\mathbb{T}) \leq 0$ so that $\mathbf{G}$ is defined on $\mathbb{C}_0$. Assumption ESAD also implies that $\mathbb{T}$ is invertible, so that $\sigma(A)$ is contained in a vertical strip in the closed left half-plane. Thus, $\rho(A)$ includes a left half-plane. We now introduce a natural extension of $\mathbf{G}$ to $\rho(A)$.

LEMMA 5.1. *With the above notation and assumptions, there exists a unique extension of* $\mathbf{G}$ *to* $\rho(A)$ *which satisfies*

$$\mathbf{G}(s) - \mathbf{G}(\beta) = B^*\left[(sI - A)^{-1} - (\beta I - A)^{-1}\right]B \qquad \forall\, s, \beta \in \rho(A). \tag{5.2}$$

This extension is analytic, and it is bounded on any right half-plane $\mathbb{C}_\varepsilon$ *with* $\varepsilon > 0$, *as well as on some left half-plane.*

Note that if $\rho(A)$ is connected, then the above extension of $\mathbf{G}$ coincides with its analytic continuation to $\rho(A)$.

Proof. It is clear that the original $\mathbf{G}$ satisfies (5.2) on $\mathbb{C}_0$, because (5.2) is just (4.3) with $C = B^*$. This implies that $\mathbf{G}(s)$ can be defined for $s \in \rho(A)$ by (5.2) with $\beta \in \mathbb{C}_0$, and the definition is independent of the choice of β. It is also clear that the extended $\mathbf{G}$ is analytic, and it satisfies (5.2) if at least one of the numbers s, β is in $\mathbb{C}_0$. We have to check that (5.2) also holds if neither s nor β is in $\mathbb{C}_0$. To show this, we choose $z \in \mathbb{C}_0$ and we decompose

$$\mathbf{G}(s) - \mathbf{G}(\beta) = [\mathbf{G}(s) - \mathbf{G}(z)] + [\mathbf{G}(z) - \mathbf{G}(\beta)].$$

Now (5.2) follows from the fact that it holds for each of the two terms. It follows from the theory in section 4 that $\mathbf{G}$ is bounded on $\mathbb{C}_\varepsilon$ for any $\varepsilon > 0$, since $\omega(\mathbb{T}) \leq 0$.

Finally, we have to show that the extended $\mathbf{G}$ is bounded on some left half-plane. Choose $\mu > 0$ such that $\sigma(A)$ is contained in the vertical strip, where $-\mu \leq \operatorname{Re} s \leq 0$. Then for $\operatorname{Re} s > \mu$, (5.2) with the resolvent identity implies that

$$\mathbf{G}(s) - \mathbf{G}(-\overline{s}) = 2(\operatorname{Re} s)B^*(sI - A)^{-1}(\overline{s}I + A)^{-1}B.$$

B is an admissible control operator for the inverse semigroup $\tilde{\mathbb{T}}_t = (\mathbb{T}_t)^{-1}$ (see section 3, after (3.4)). It follows from (3.4) rewritten for $\tilde{\mathbb{T}}$ that for some $\delta^1, \omega^1 > 0$,

$$\|(sI + A)^{-1}B\|_{\mathcal{L}(U,X)} \leq \frac{\delta^1}{\sqrt{\operatorname{Re} s}} \qquad \forall \operatorname{Re} s > \omega^1.$$

Applying this estimate and (3.10), we obtain that for $\operatorname{Re} s$ sufficiently large,

$$\|\mathbf{G}(s) - \mathbf{G}(-\overline{s})\| \leq 2\delta\delta^1.$$

Since $\mathbf{G}$ is bounded on $\mathbb{C}_\varepsilon$ (for any $\varepsilon > 0$), this estimate shows that it is also bounded on some left half-plane. □

The last lemma can be derived also as a consequence of the theory of time-invertible systems in Staffans and Weiss [38].

THEOREM 5.2. *With the above notation and assumptions, there exist operators $E = E^* \in \mathcal{L}(U)$ such that $\mathbf{G} + E$ is a positive transfer function, i.e.,*

$$\mathbf{G}(s)^* + \mathbf{G}(s) + 2E \geq 0 \qquad \forall\, s \in \mathbb{C}_0. \tag{5.3}$$

One such operator E is given by

$$E = -\frac{1}{2} \lim_{\lambda \to +\infty} \left[\mathbf{G}(\lambda)^* + \mathbf{G}(-\lambda)\right]. \tag{5.4}$$

If $0 \in \rho(A)$, then the same E is also given by

$$E = -\frac{1}{2}\left[\mathbf{G}(0)^* + \mathbf{G}(0)\right] + B^*(A^*)^{-1}QA^{-1}B. \tag{5.5}$$

If Σ is weakly regular, with feedthrough operator D, then E is also given by

$$\langle E\mathrm{v}, \mathrm{v}\rangle = -\frac{1}{2}\langle (D^* + D)\mathrm{v}, \mathrm{v}\rangle + \lim_{\lambda \to +\infty} \lambda \|(\lambda I - A)^{-1}B\mathrm{v}\|^2 \qquad \forall\, \mathrm{v} \in U. \tag{5.6}$$

Note that the limit in (5.4) exists in the operator norm, even if Σ is not regular.

Proof. We shall use the following identity (a consequence of ESAD):

$$(\overline{s}I - A^*)^{-1} + (sI - A)^{-1} = (\overline{s}I - A^*)^{-1}\big[2(\operatorname{Re} s)I + Q\big](sI - A)^{-1}. \tag{5.7}$$

Using the formulas (4.3) and (5.7), we calculate

$$\begin{aligned}
&[\mathbf{G}(s)^* + \mathbf{G}(s)] - [\mathbf{G}(\beta)^* + \mathbf{G}(\beta)] \\
&= B^*[(sI - A)^{-1} - (\beta I - A)^{-1}]B + B^*[(\overline{s}I - A^*)^{-1} - (\overline{\beta}I - A^*)^{-1}]B \\
&= B^*[(\overline{s}I - A^*)^{-1} + (sI - A)^{-1}]B - B^*[(\overline{\beta}I - A^*)^{-1} + (\beta I - A)^{-1}]B \\
&= B^*(\overline{s}I - A^*)^{-1}[2(\operatorname{Re} s)I + Q](sI - A)^{-1}B \\
&\qquad -B^*(\overline{\beta}I - A^*)^{-1}[2(\operatorname{Re}\beta)I + Q](\beta I - A)^{-1}B,
\end{aligned}$$

for all $s, \beta \in \mathbb{C}_0$. Rearranging the above formula, we obtain

$$\begin{aligned}
&[\mathbf{G}(s)^* + \mathbf{G}(s)] - B^*(\overline{s}I - A^*)^{-1}[2(\operatorname{Re} s)I + Q](sI - A)^{-1}B \\
=&[\mathbf{G}(\beta)^* + \mathbf{G}(\beta)] - B^*(\overline{\beta}I - A^*)^{-1}[2(\operatorname{Re}\beta)I + Q](\beta I - A)^{-1}B.
\end{aligned}$$

Thus, both sides are equal to a bounded, self-adjoint operator on U, which depends neither on s nor on β. We denote this operator by $-2E$. Now since

$$B^*(\overline{s}I - A^*)^{-1}[2(\operatorname{Re} s)I + Q](sI - A)^{-1}B \geq 0 \qquad \forall\, s \in \mathbb{C}_0,$$

we deduce that (5.3) holds. For all $\lambda > 0$ we have

$$2E = -[\mathbf{G}(\lambda)^* + \mathbf{G}(\lambda)] + B^*(\lambda I - A^*)^{-1}[2\lambda I + Q](\lambda I - A)^{-1}B. \tag{5.8}$$

Since $-A^* = A + Q$, we have

$$(\lambda I - A^*)^{-1} = (\lambda I + A)^{-1} - (\lambda I - A^*)^{-1}Q(\lambda I + A)^{-1}.$$

Substituting this into (5.8), we obtain

$$\begin{aligned} 2E = &-[\mathbf{G}(\lambda)^* + \mathbf{G}(\lambda)] + 2\lambda B^*(\lambda I + A)^{-1}(\lambda I - A)^{-1}B \\ &+ B^*(\lambda I + A)^{-1}Q(\lambda I - A)^{-1}B - B^*(\lambda I - A^*)^{-1}Q(\lambda I + A)^{-1}[2\lambda I + Q](\lambda I - A)^{-1}B. \end{aligned}$$

We want to examine what happens in this rather long formula when $\lambda \to +\infty$. Since B^* is an admissible observation operator for the inverse semigroup generated by $-A$, by (3.10) $\|B^*(\lambda I + A)^{-1}\|$ decays like $1/\sqrt{\lambda}$ for large λ. Similarly, by (3.4), the factors $\|(\lambda I - A)^{-1}B\| = \|B^*(\lambda I - A^*)^{-1}\|$ decay like $1/\sqrt{\lambda}$ for large λ. The factor $(\lambda I + A)^{-1}[2\lambda I + Q]$ remains bounded as $\lambda \to +\infty$. It follows from these considerations that the last two terms in the long formula (the ones containing a factor Q) tend to zero as $\lambda \to +\infty$. Hence,

$$2E = \lim_{\lambda \to +\infty} \left[-\mathbf{G}(\lambda)^* - \mathbf{G}(\lambda) + 2\lambda B^*(\lambda I + A)^{-1}(\lambda I - A)^{-1}B \right].$$

On the other hand, we know from (5.2) and the resolvent identity that for $\lambda > 0$ sufficiently large, using the extension of $\mathbf{G}$ to $\rho(A)$, the following holds:

$$\mathbf{G}(\lambda) - \mathbf{G}(-\lambda) = 2\lambda B^*(\lambda I + A)^{-1}(\lambda I - A)^{-1}B.$$

Substituting this into the previous formula, we get (5.4).

To prove (5.5), assume that $0 \in \rho(A)$. Taking $\lambda \to 0$ in (5.8), we obtain (5.5).

To prove (5.6), suppose now that Σ is weakly regular with feedthrough operator D. Then, applying (5.8) to $\mathrm{v} \in U$, taking the scalar product with v and taking limits as $\lambda \to +\infty$, we obtain

$$\begin{aligned} 2\langle E\mathrm{v}, \mathrm{v}\rangle &= -\lim_{\lambda \to +\infty} \langle (\mathbf{G}(\lambda)^* + \mathbf{G}(\lambda))\mathrm{v}, \mathrm{v}\rangle + \lim_{\lambda \to +\infty} 2\lambda \|(\lambda I - A)^{-1}B\mathrm{v}\|^2 \\ &\quad + \lim_{\lambda \to +\infty} \langle Q(\lambda I - A)^{-1}B\mathrm{v}, (\lambda I - A)^{-1}B\mathrm{v}\rangle \\ &= -\langle (D + D^*)\mathrm{v}, \mathrm{v}\rangle + 2 \lim_{\lambda \to +\infty} \lambda \|(\lambda I - A)^{-1}B\mathrm{v}\|^2, \end{aligned}$$

since the inequality (3.4) shows that the limit containing Q is zero. □

Remark 5.3. It follows from the last theorem that $cI + \mathbf{G}$ is a positive transfer function for $c = \|E^+\|$, where E^+ is the positive part of the self-adjoint operator E from (5.4). Indeed, this follows from $cI \geq E^+ \geq E$ (for finite-dimensional systems this was in the proof of Proposition 2.6). Note that the proof of (5.3), (5.4) and (5.5) did not use the well posedness of Σ, but only the admissibility of B. Thus, if A satisfies ESAD, B is an admissible control operator for the semigroup generated by A, and $C = B^*$, then for any function $\mathbf{G}$ satisfying (4.3) there exists a $c \geq 0$ such that $cI + \mathbf{G}$ is a positive transfer function. This $\mathbf{G}$ need not be well posed.

Remark 5.4. In the case that $Q = 0$, the conclusions of the previous remark remain valid without assuming that B is admissible; we only need the fact that

$B \in \mathcal{L}(U, X_{-1})$. Thus, in this case, (5.3), (5.4) and (5.5) still hold without assuming well posedness or admissibility. Moreover, (5.4) and (5.5) can be replaced by

$$E = -\frac{1}{2}[\mathbf{G}(\lambda)^* + \mathbf{G}(-\lambda)] \qquad \forall\, \lambda > 0.$$

Recall that $\alpha(B)$ denotes the degree of unboundedness of B, introduced in (3.5). In the following proposition we discuss the consequences of B being less than maximally unbounded, i.e., $\alpha(B) < \frac{1}{2}$, in the context of Theorem 5.2.

PROPOSITION 5.5. *With the notation and assumptions of this section, suppose that* $\alpha(B) < \frac{1}{2}$. *Then* Σ *is regular and* E *from* (5.4) *is given by* (5.1). *Moreover, in this case,* E *given by* (5.1) *is the smallest self-adjoint operator in* $\mathcal{L}(U)$ *satisfying* (5.3).

By the minimality of E we mean the following: If $F = F^* \in \mathcal{L}(U)$ is such that $\mathbf{G} + F$ is a positive transfer function, then $E \leq F$.

Proof. First we note that the regularity of Σ follows from Proposition 4.1 (since $\alpha(B^*) \leq \frac{1}{2}$). The formula (5.1) follows from (5.6), where the second term is now zero. Let $F = F^*$ be such that $\mathbf{G} + F$ is a positive transfer function so that

$$2F + [\mathbf{G}(\lambda)^* + \mathbf{G}(\lambda)] \geq 0 \qquad \forall\, \lambda > 0.$$

Taking limits as $\lambda \to +\infty$, we obtain that $2F + D + D^* \geq 0$. From here, using (5.1) we obtain $2F - 2E \geq 0$, i.e., E is minimal. ∎

By Theorem 5.2 $\mathbf{G} + E$ is a positive transfer function. If Σ is regular and (5.1) holds, then this implies that $\mathbf{G} - D$ is also a positive transfer function. From (4.6) with $C = B^*$ we then obtain that $B_\Lambda^*(sI - A)^{-1}B$ is a positive transfer function.

The following example is a regular linear system satisfying ESAD and COL, with feedthrough operator $D = 0$, for which the operator E from (5.4) or (5.6) is nonzero. For this example, $B_\Lambda^*(sI - A)^{-1}B$ is *not* a positive transfer function and the feedback $u = -\kappa y + v$ is exponentially destabilizing if κ is too large.

Example 5.6. Consider the usual realization of a delay line of length h, $h > 0$, as given, e.g., on p. 831 of [45]. The state space of this system Σ_0 is $X = L^2[-h, 0]$, the semigroup is the left shift operator with zero entering from the right, with the generator

$$A_0 = \frac{\mathrm{d}}{\mathrm{d}x}, \qquad \mathcal{D}(A_0) = \{z_0 \in \mathcal{H}^1(-h, 0) \mid z_0(0) = 0\}.$$

The control operator is $B = \delta_0$ and the observation operator is $C = \delta_{-h}^*$, which means that $Cz_0 = z_0(-h)$ for $z_0 \in \mathcal{D}(A_0)$. The feedthrough operator is zero and the transfer function of this system is $\mathbf{G}_0(s) = e^{-hs}$. We call the input w and the output y. (For an isomorphic system described by the wave equation we refer to [53, section 7]. We close a positive unity feedback loop around this delay line, meaning that $w = y + u$, where u is the new input function. This leads to a new well-posed linear system Σ with the transfer function

$$\mathbf{G}(s) = \frac{e^{-hs}}{1 - e^{-hs}}. \tag{5.9}$$

The semigroup $\mathbb{T}$ of this new system is the periodic left shift semigroup on X, which is unitary. The generating operators of Σ can be computed directly, or using the formulas from [46, section 7]. The generator of $\mathbb{T}$ is again $A = \frac{\mathrm{d}}{\mathrm{d}x}$, but now the domain is

$$\mathcal{D}(A) = \{z_0 \in \mathcal{H}^1(-h, 0) \mid z_0(0) = z_0(-h)\}.$$

It is easy to see that this A is skew-adjoint. The operators B and C remain basically the same, but of course the new C is defined on the new $\mathcal{D}(A)$ and this results in $C = B^*$. Thus, the system Σ fits into the framework of this paper (it satisfies ESAD and COL). Moreover, Σ is regular and its feedthrough operator is zero. It is not difficult to check that B is maximally unbounded, i.e., $\alpha(B) = \frac{1}{2}$.

We remark that the space Z from (4.9) is now $Z = \mathcal{H}^1(-h, 0)$ and for every $z_0 \in Z$ we have $C_\Lambda z_0 = z_0(-h)$. It follows from what we said at (4.10) that if $u \in \mathcal{H}^1(0, \infty)$ and if z_0, the initial state of Σ satisfies $z_0 \in \mathcal{H}^1(-h, 0)$ and $z_0(0) = z_0(-h) + u(0)$ (equivalently, $Az_0 + Bu(0) \in X$), then for all $t \geq 0$ we have

$$z(\cdot, t) \in \mathcal{H}^1(-h, 0), \qquad z(0, t) = z(-h, t) + u(t),$$

and the output function y is given by $y(t) = z(-h, t)$.

The spectrum $\sigma(A)$ now consists of the poles of $\mathbf{G}$, so that $\rho(A)$ is a connected set. The extension of $\mathbf{G}$ to $\rho(A)$, as introduced in Lemma 5.1, is the analytic continuation of $\mathbf{G}$, still given by (5.9). Formula (5.4) gives $E = \frac{1}{2}$. It is not difficult to verify that

$$\mathbf{G}(i\omega)^* + \mathbf{G}(i\omega) = -1 \qquad \forall\, \omega \in \mathbb{R}$$

so that $\mathbf{G}$ is not positive-real, but of course $\mathbf{G} + E = \mathbf{G} + \frac{1}{2}$ is positive-real. Note that E is minimal in the sense of Proposition 5.5, even though $\alpha(B) = \frac{1}{2}$.

If we close a negative feedback loop around Σ by putting $u = -\kappa y + v$, where v is the new input, then we get the closed-loop transfer function

$$\mathbf{G}^\kappa(s) = \frac{e^{-hs}}{1 - (1 - \kappa)e^{-hs}}.$$

This transfer function is bounded on some right half-plane, since

$$|\mathbf{G}^\kappa(s)| \leq e^{-h\sigma}/|1 - (1 - \kappa)e^{-h\sigma}|, \qquad \sigma = \operatorname{Re} s,$$

and so the closed-loop system is well posed. $\mathbf{G}^\kappa$ is stable for $0 < \kappa < 2$, but for $\kappa \geq 2$ the transfer function has unstable poles. This shows that the closed-loop semigroup becomes unstable. Moreover, the larger κ becomes, the more unstable the closed-loop system becomes (its poles move to the right).

If a system with transfer function $\mathbf{G}$ is such that $cI + \mathbf{G}$ is positive, we show that sufficiently small output feedbacks stabilize the system in an input-output sense. If the system is optimizable and estimatable (for example, if it is exactly controllable and exactly observable), then we can conclude exponential stability. During the following proposition (and its proof) we suspend the standing notation and assumptions introduced at the beginning of this section. Thus, for example, the system Σ is not required to satisfy assumptions ESAD and COL.

PROPOSITION 5.7. *Suppose that Σ is a well-posed linear system with input and output space U. Assume that its transfer function $\mathbf{G}$ is such that for some $c \geq 0$, $cI + \mathbf{G}$ is a positive transfer function. Denote $\kappa_0 = \frac{1}{c}$ (for $c = 0$, take $\kappa_0 = \infty$). Then for every $\kappa \in (0, \kappa_0)$ the operator $K = -\kappa I$ is an admissible feedback operator for Σ, and the corresponding closed-loop system Σ^κ is input-output stable, i.e., $\mathbf{G}^\kappa = \mathbf{G}(I + \kappa\mathbf{G})^{-1} \in H^\infty(\mathcal{L}(U))$. Moreover, if Σ is optimizable and estimatable, then the closed-loop system is exponentially stable.*

Proof. Denote $a = \frac{1}{\kappa}$. The proof of the fact that for all $\mathrm{v} \in U$ with $\|\mathrm{v}\| = 1$,

$$\|(aI + \mathbf{G}(s))\mathrm{v}\| \geq a - c \qquad \forall\, s \in \mathbb{C}_0, \tag{5.10}$$

is exactly like in the proof of Proposition 2.1. However, since U may now be infinite-dimensional, boundedness from below is not enough to conclude the invertibility of $aI + \mathbf{G}(s)$. By a similar argument, we have that $\|(aI + \mathbf{G}(s))^*\mathrm{v}\| \geq a - c$, and this together with (5.10) implies the desired invertibility and also that (2.2) holds. From here we easily get $\mathbf{G}^\kappa \in H^\infty$, as in the proof of Proposition 2.1.

Now applying Theorem 4.5, we can conclude exponential stability if Σ^κ is optimizable and estimatable. If Σ is optimizable and estimatable, then Σ^κ inherits these properties, according to Proposition 4.4. □

The above proof appears to be short and simple, but it relies on Proposition 4.4 and Theorem 4.5, and the proof of the latter (in [51]) is rather involved.

THEOREM 5.8. *With the notation and assumptions from the beginning of this section, let $c = \|E^+\|$, where E^+ is the positive part of E from* (5.4) *and denote $\kappa_0 = \frac{1}{c}$ (for $c = 0$, take $\kappa_0 = \infty$). Then for every $\kappa \in (0, \kappa_0)$, $K = -\kappa I$ is an admissible feedback operator for Σ, leading to the closed-loop system Σ^κ which is input-output stable. If Σ is optimizable and estimatable, then Σ^κ is exponentially stable.*

Proof. According to Remark 5.3, $cI + \mathbf{G}$ is a positive transfer function for $c = \|E^+\|$. Now the theorem follows from Proposition 5.7 using this c. □

The following proposition may be useful for checking if a specific system fits into the framework used in this paper.

PROPOSITION 5.9. *Suppose that A satisfies* ESAD *and $B \in \mathcal{L}(U, X_{-1})$. Then B is an admissible control operator for $\mathbb{T}$ if and only if B^* is an admissible observation operator for $\mathbb{T}$. Moreover, we have $\alpha(B) = \alpha(B^*)$.*

Proof. Denote $A_0 = A + \frac{1}{2}Q$ so that A_0 is skew-adjoint. Take $s \in \mathbb{C}_0$. From

$$(sI - A_0)^{-1} - (sI - A)^{-1} = \frac{1}{2}(sI - A_0)^{-1}Q(sI - A)^{-1}, \tag{5.11}$$

we see that the norms $\|x\|_{-1} = \|(sI - A)^{-1}x\|$ and $\|x\|^0_{-1} = \|(sI - A_0)^{-1}x\|$ are equivalent. Hence, the space X_{-1} for A and for A_0 is the same. Multiplying the last identity with B from the right and taking norms, we can see that $\alpha(B)$ with respect to A and A_0 is the same. We denote by $\mathbb{S}$ the (unitary) semigroup generated by A_0, and by $\mathbb{S}^*$ its adjoint (or inverse) semigroup, generated by $-A_0$. It is easy to see that B is an admissible control operator for $\mathbb{T}$ if and only if it is for $\mathbb{S}$ (because the perturbation $\frac{1}{2}Q$ is bounded). Similarly, we can show that $\alpha(B^*)$ with respect to A and A_0 is the same and B^* is an admissible observation operator for $\mathbb{T}$ if and only if it is for $\mathbb{S}$. Thus, so far we have shown that for our purposes, there is no difference between A and A_0 (i.e., between $\mathbb{T}$ and $\mathbb{S}$).

It is easy to see that the admissibility of B for $\mathbb{S}$ and for $\mathbb{S}^*$ are equivalent (see the text after (3.4)). On the other hand, the admissibility of B for $\mathbb{S}^*$ is equivalent to the admissibility of B^* for $\mathbb{S}$ (see section 3). Hence, B is an admissible control operator for $\mathbb{S}$ if and only if B^* is an admissible observation operator for $\mathbb{S}$. It remains to prove that $\alpha(B) = \alpha(B^*)$. This follows from

$$\begin{aligned}\|(\lambda I - A_0)^{-1}B\|^2 &= \|B^*(\lambda I + A_0)^{-1}(\lambda I - A_0)^{-1}B\| \\ &= \|B^*(\lambda I - A_0)^{-1}(\lambda I + A_0)^{-1}B\| = \|B^*(\lambda I - A_0)^{-1}\|^2. \qquad \square\end{aligned}$$

6. A class of undamped second order systems. In this section we introduce a class of undamped second order systems satisfying ESAD and COL. For these systems we derive an explicit expression for E from (5.4) that shows clearly that E^+ is not always zero. Hence, as we have seen in section 5, the range of exponentially stabilizing feedback gains is bounded in general.

Let U_0, U_1 and H be Hilbert spaces and let $A_0 : \mathcal{D}(A_0) \to H$ be positive and boundedly invertible on H. Consider the system described by the following second order differential equation and two output equations:

$$\left\{ \begin{array}{rcll} \ddot{q} + A_0 q & = & C_0^* u_0 + A_0^{-1} C_1^* \dot{u}_1, & \\ y_0 & = & \overline{C_0}\dot{q}, & y_1 = \overline{C_1} q, \end{array} \right. \tag{6.1}$$

for sufficiently smooth signals u_0, u_1 and compatible initial conditions $q(0)$ and $\dot{q}(0)$. The input signals u_0, u_1 and the output signals y_0, y_1 are such that $u_0(t) \in U_0$, $u_1(t) \in U_1$, $y_0(t) \in U_0$, $y_1(t) \in U_1$. To formulate these equations in the form (1.1) and (4.10), which describe a well-posed linear system for sufficiently smooth u and compatible $z(0)$, we need to introduce various spaces and operators, and to make some assumptions. For every $\mu > 0$, we define $H_\mu = \mathcal{D}(A_0^\mu)$, with the norm $\|\varphi\|_\mu = \|A_0^\mu \varphi\|_H$, and we define $H_{-\mu} = H_\mu^*$ (duality with respect to the pivot space H). We denote $H_0 = H$ and $\|\varphi\|_0 = \|\varphi\|_H$. We assume that

$$C_0 \in \mathcal{L}(H_{\frac{1}{2}}, U_0), \qquad C_1 \in \mathcal{L}(H_1, U_1).$$

We identify U_0 and U_1 with their duals so that $C_0^* \in \mathcal{L}(U_0, H_{-\frac{1}{2}})$, $C_1^* \in \mathcal{L}(U_1, H_{-1})$. We assume that C_0 and C_1 have extensions $\overline{C_0}$ and $\overline{C_1}$ such that the operators

$$D_0 = \overline{C_0} A_0^{-1} C_1^* \in \mathcal{L}(U_1, U_0), \qquad D_1 = \overline{C_1} A_0^{-1} C_0^* \in \mathcal{L}(U_0, U_1)$$

exist. We introduce the input (and output) space U and the state space X:

$$U = U_0 \times U_1, \qquad X = H_{\frac{1}{2}} \times H. \tag{6.2}$$

Now the equations of the system Σ from (6.1) can be rewritten in the form

$$\left\{ \begin{array}{rcl} \dot{z} & = & Az + Bu, \\ y & = & \overline{C} z + Du, \end{array} \right. \tag{6.3}$$

where

$$z = \left[\begin{array}{c} q \\ w \end{array} \right] \in X, \quad u = \left[\begin{array}{c} u_0 \\ u_1 \end{array} \right] \in U, \quad y = \left[\begin{array}{c} y_0 \\ y_1 \end{array} \right] \in U, \tag{6.4}$$

$$A : \mathcal{D}(A) \to X, \quad A = \left[\begin{array}{cc} 0 & I \\ -A_0 & 0 \end{array} \right], \tag{6.5}$$

$$X_1 = \mathcal{D}(A) = H_1 \times H_{\frac{1}{2}}, \quad X_{-1} = H \times H_{-\frac{1}{2}}, \quad A^* = -A,$$

$$B = \left[\begin{array}{cc} 0 & A_0^{-1} C_1^* \\ C_0^* & 0 \end{array} \right] \in \mathcal{L}(U, X_{-1}), \quad C = B^* = \left[\begin{array}{cc} 0 & C_0 \\ C_1 & 0 \end{array} \right] \in \mathcal{L}(X_1, U), \tag{6.6}$$

$$\overline{C} = \left[\begin{array}{cc} 0 & \overline{C_0} \\ \overline{C_1} & 0 \end{array} \right], \qquad D = \left[\begin{array}{cc} 0 & D_0 \\ 0 & 0 \end{array} \right]. \tag{6.7}$$

We shall now prove that for sufficiently smooth u, (6.1) is equivalent to (6.3). We also show how the condition of "compatible initial conditions," $Az(0) + Bu(0) \in X$, can be expressed in terms of the functions and operators appearing in (6.1). The significance of this condition was explained in section 4 (around (4.10)).

PROPOSITION 6.1. *For $u \in \mathcal{H}^1_{loc}(0,\infty;U)$, the equations* (6.1) *(for a specific $t \geq 0$) are equivalent to* (6.3) *(for the same t) with z, u, y as in* (6.4) *and with*

$$w(t) = \dot{q}(t) - A_0^{-1}C_1^* u_1(t).$$

The first part of (6.1) *is regarded as an equation in $H_{-\frac{1}{2}}$, and the first part of* (6.3) *is regarded as an equation in X_{-1}. Moreover, for any $t \geq 0$, the conditions*

$$A_0 q(t) - C_0^* u_0(t) \in H, \qquad \dot{q}(t) \in H_{\frac{1}{2}} \tag{6.8}$$

are equivalent to $Az(t) + Bu(t) \in X$.

Proof. We rewrite the first part of (6.3):

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} q(t) \\ w(t) \end{bmatrix} = \begin{bmatrix} 0 & I \\ -A_0 & 0 \end{bmatrix}\begin{bmatrix} q(t) \\ w(t) \end{bmatrix} + \begin{bmatrix} 0 & A_0^{-1}C_1^* \\ C_0^* & 0 \end{bmatrix}\begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix},$$

or, equivalently,

$$\dot{q}(t) = w(t) + A_0^{-1}C_1^* u_1(t), \qquad \dot{w}(t) = -A_0 q(t) + C_0^* u_0(t).$$

By differentiating the first formula, we obtain the first equation in (6.1). To derive the last two equations in (6.1), we compute, starting from (6.3),

$$\begin{aligned}\begin{bmatrix} y_0(t) \\ y_1(t) \end{bmatrix} &= \begin{bmatrix} 0 & \overline{C_0} \\ \overline{C_1} & 0 \end{bmatrix}\begin{bmatrix} q(t) \\ w(t) \end{bmatrix} + \begin{bmatrix} 0 & \overline{C_0}A_0^{-1}C_1^* \\ 0 & 0 \end{bmatrix}\begin{bmatrix} u_0(t) \\ u_1(t) \end{bmatrix} \\ &= \begin{bmatrix} \overline{C_0}(\dot{q}(t) - A_0^{-1}C_1^* u_1(t)) + \overline{C_0}A_0^{-1}C_1^* u_1(t) \\ \overline{C_1}q(t) \end{bmatrix} = \begin{bmatrix} \overline{C_0}\dot{q}(t) \\ \overline{C_1}q(t) \end{bmatrix}.\end{aligned}$$

To obtain (6.3) from (6.1), we just reverse the order of the computations. It is easy to verify that the conditions (6.8) are equivalent to $Az(t) + Bu(t) \in X$. ∎

Clearly, Σ satisfies ESAD and COL, but it need not be well posed. The optimal control of systems of this type, but with $C_1 = 0$, has been studied in [49] without assuming well posedness. Here, *we do assume that our system Σ is well posed* (but not necessarily regular). Then, the equations (6.3) correspond to the representation of well-posed systems via (1.1) and (4.10).

As explained in section 4 (before (4.10)), the transfer function of Σ is given by $\mathbf{G}(s) = \overline{C}(sI - A)^{-1}B + D$, which is easy to compute in terms of A_0, C_0 and C_1: for $-s^2 \in \rho(A_0)$,

$$\mathbf{G}(s) = \begin{bmatrix} C_0 s \\ \overline{C_1} \end{bmatrix}(s^2 I + A_0)^{-1}\begin{bmatrix} C_0^* & A_0^{-1}C_1^* s \end{bmatrix}. \tag{6.9}$$

Note the curious fact that $\mathbf{G}$ does not depend on the extended operator $\overline{C_0}$. This extended operator appears in the representation (6.3) of Σ, but (6.9) shows that the system Σ is in fact independent of the choice of the extension $\overline{C_0}$.

In particular, we see from (6.9) that

$$\mathbf{G}(0) = \begin{bmatrix} 0 & 0 \\ D_1 & 0 \end{bmatrix}.$$

According to Theorem 5.2, we have $E = -\frac{1}{2}[\mathbf{G}(0)^* + \mathbf{G}(0)] + B^*(A^*)^{-1}QA^{-1}B$ so that (since $Q = 0$)

$$E = -\frac{1}{2}\begin{bmatrix} 0 & D_1^* \\ D_1 & 0 \end{bmatrix}. \tag{6.10}$$

Such systems, but with $C_1 = 0$, have been considered in [2, 14, 49].

PROPOSITION 6.2. *Suppose that the well-posed system* Σ *from* (6.3) *is exactly controllable (or equivalently, exactly observable). Let* $E \in \mathcal{L}(U)$ *be the operator from* (6.10)*, let* E^+ *be the positive part of* E*, and put* $c = \|E^+\|$, $\kappa_0 = \frac{1}{c}$. *Then for every* $\kappa \in (0, \kappa_0)$, $K = -\kappa I$ *is an admissible feedback operator for* Σ *and the resulting closed-loop system* Σ^κ *is exponentially stable.*

Proof. The equivalence of exact controllability and exact observability follows by duality, since $A^* = -A$ (see the comments after Definition 3.1). It is clear that exact controllability implies optimizability, and (by duality) exact observability implies estimatability. Thus, we can apply Theorem 5.8 to obtain this proposition. □

Example 6.3. We describe a well-posed system which fits into the above framework, so that proportional output feedback can exponentially stabilize it. Since the computations are rather long, they are the subject of a separate paper [50].

The physical system that we are modeling consists of a hinged elastic beam with two sensors: one measures the angular velocity of the beam at a point ξ and the other measures the bending (curvature) of the beam at the same point. These two measurements are advantageous because they make the open-loop system exactly observable. Our aim is to design the actuators and the feedback law in order to exponentially stabilize this system. Using Proposition 6.2, we shall design the actuators such that they are colocated, meaning that $B = C^*$, and then the open-loop system is described by equations of the form (6.1). Here, C_0 will be the operator corresponding to the measurement of the angular velocity at ξ, and C_1 will be the operator corresponding to the measurement of the bending at ξ. It turns out that the actuators cause a discontinuity of the bending exactly at ξ. As in the preceding theory, we are forced to use an extension of the operator C_1, which means that we have to decide if the corresponding sensor measures the left or the right limit of the bending at ξ, or a combination of the lateral limits.

We model the open-loop system as a homogenous Rayleigh beam situated along the interval $[0, \pi]$, with the two sensors located at $\xi \in (0, \pi)$. This is an extension of an example discussed in [1] and [52], and these papers contain further references to Rayleigh beam models. The equations describing the open-loop system are

$$\ddot{q}(x,t) - \alpha \frac{\partial^2 \ddot{q}}{\partial x^2}(x,t) + \frac{\partial^4 q}{\partial x^4}(x,t) = \mathcal{U}(x,t), \tag{6.11}$$

$$q(0,t) = q(\pi,t) = 0, \qquad \frac{\partial^2 q}{\partial x^2}(0,t) = \frac{\partial^2 q}{\partial x^2}(\pi,t) = 0, \quad t \geq 0, \tag{6.12}$$

$$y_0(t) = \frac{\partial \dot{q}}{\partial x}(\xi,t), \qquad y_1(t) = \frac{\partial^2 q}{\partial x^2}(\xi,t). \tag{6.13}$$

Here $q(x,t)$ represents the transverse displacement of the beam ($x \in [0,\pi]$ and $t \geq 0$), and $\alpha > 0$ is a constant, proportional to the moment of inertia of the cross section of the beam. In (6.11), $\mathcal{U}$ denotes the control terms which are to be designed.

In order to fit this system into the framework of (6.1), (6.2), we denote $H = \mathcal{H}_0^1(0,\pi)$, $V = \mathcal{H}^2(0,\pi) \cap \mathcal{H}_0^1(0,\pi)$. On H we define the inner product such that

$$\langle \varphi, \psi \rangle_H = \left\langle \left(I - \alpha \frac{\mathrm{d}^2}{\mathrm{d}x^2} \right) \varphi, \psi \right\rangle_{L^2} \qquad \forall\, \varphi, \psi \in V.$$

We introduce the operator $\mathcal{R} : L^2(0,\pi) \to V$ defined by

$$\mathcal{R} = \left(I - \alpha \frac{\mathrm{d}^2}{\mathrm{d}x^2} \right)^{-1}.$$

It is easy to see that $\mathcal{R} > 0$ when regarded as an operator from $L^2(0,\pi)$ to $L^2(0,\pi)$. We also define the linear operator $A_0 : \mathcal{D}(A_0) \to H$ by

$$\mathcal{D}(A_0) = \left\{ \varphi \in \mathcal{H}^3(0,\pi) \,\middle|\, \varphi(0) = \varphi(\pi) = 0,\ \frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2}(0) = \frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2}(\pi) = 0 \right\},$$

$$A_0\varphi = \frac{\mathrm{d}^4}{\mathrm{d}x^4}(\mathcal{R}\varphi) \qquad \forall\ \varphi \in \mathcal{D}(A_0).$$

The set $\{\varphi_k \,|\, k \in \mathbb{N}\}$ with

$$\varphi_k(x) = \frac{1}{\sqrt{1+\alpha k^2}} \cdot \sqrt{\frac{2}{\pi}} \sin kx \qquad \forall\ x \in [0,\pi]$$

is an orthonormal basis in H, and we have

$$A_0\varphi_k = \frac{k^4}{1+\alpha k^2}\,\varphi_k, \qquad \mathcal{R}\varphi_k = \frac{1}{1+\alpha k^2}\,\varphi_k$$

so that, in particular, A_0 is self-adjoint, strictly positive and it commutes with $\mathcal{R}$. Defining H_μ and $\|\varphi\|_\mu$ for $\mu \in \mathbb{R}$ by fractional powers of A_0 and duality, as explained after (6.1), we have $H_0 = H = \mathcal{H}_0^1(0,\pi)$ and

$$H_1 = \mathcal{D}(A_0), \quad H_{\frac{1}{2}} = V, \quad H_{-\frac{1}{2}} = L^2(0,\pi), \quad H_{-1} = \mathcal{H}^{-1}(0,\pi), \quad H_{-1.5} = V',$$

the norm $\|\cdot\|_{-\frac{1}{2}}$ being equivalent to the L^2-norm. Here, V' denotes the dual of V with respect to the pivot space $L^2(0,\pi)$. For every $\mu \in \mathbb{R}$, A_0 and $\mathcal{R}$ have extensions (or restrictions) to H_μ and we have $A_0 \in \mathcal{L}(H_\mu, H_{\mu-1})$, $\mathcal{R} \in \mathcal{L}(H_\mu, H_{\mu+1})$.

The spaces U_0 and U_1 from (6.2) are $U_0 = U_1 = \mathbb{C}$, so that $U = \mathbb{C}^2$. Corresponding to the two measurements in (6.13), we define the operators $C_0 \in \mathcal{L}(H_{\frac{1}{2}}, \mathbb{C})$ and $C_1 \in \mathcal{L}(H_1, \mathbb{C})$ by

$$C_0\varphi = \frac{\mathrm{d}\varphi}{\mathrm{d}x}(\xi), \qquad C_1\varphi = \frac{\mathrm{d}^2\varphi}{\mathrm{d}x^2}(\xi).$$

Then $C_0^* \in H_{-\frac{1}{2}}$ and $C_1^* \in H_{-1}$ (the adjoints of C_0 and C_1 with respect to the pivot space H) are

$$C_0^* = -\mathcal{R}\frac{\mathrm{d}}{\mathrm{d}x}\delta_\xi, \qquad C_1^* = \mathcal{R}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\delta_\xi,$$

where δ_ξ is the Dirac mass at the point ξ.

We remark that according to (6.2), the state space of our system is $X = \mathcal{H}^2(0,\pi) \cap \mathcal{H}_0^1(0,\pi) \times \mathcal{H}_0^1(0,\pi)$ with the norm

$$\left\| \begin{bmatrix} q \\ w \end{bmatrix} \right\|^2 = \int_0^\pi \left| \frac{\mathrm{d}^2 q(x)}{\mathrm{d}x^2} \right|^2 \mathrm{d}x + \int_0^\pi |w(x)|^2 \mathrm{d}x + \alpha \int_0^\pi \left| \frac{\mathrm{d}\,w(x)}{\mathrm{d}x} \right|^2 \mathrm{d}x.$$

Assuming that the terms of (6.11) are in $H_{-1.5} = V'$ and applying $\mathcal{R}$ to these terms, we obtain

$$\ddot{q} + A_0 q = \mathcal{R}\mathcal{U} \qquad \text{in } H_{-\frac{1}{2}} = L^2(0,\pi). \tag{6.14}$$

Now the first equation in (6.1) shows that the control terms, represented above by $\mathcal{RU}$, should be

$$\mathcal{RU} = C_0^* u_0 + A_0^{-1} C_1^* \dot{u}_1 = -u_0 \mathcal{R} \frac{\mathrm{d}}{\mathrm{d}x} \delta_\xi - \dot{u}_1 b, \tag{6.15}$$

where $b = -A_0^{-1} C_1^* \in H$ is given by

$$b(x) = \begin{cases} \frac{ax}{\xi} & \text{for } x \le \xi, \\ \frac{a\pi - ax}{\pi - \xi} & \text{for } x > \xi, \end{cases} \qquad \text{where } \frac{1}{a} = \frac{1}{\xi} + \frac{1}{\pi - \xi}.$$

The graph of b consists of two straight lines, with a peak at $x = \xi$. If we apply $\mathcal{R}^{-1}$ to the terms of (6.14), with $\mathcal{RU}$ as in (6.15), we obtain that in $H_{-1.5}$,

$$\ddot{q} - \alpha \frac{\partial^2 \ddot{q}}{\partial x^2} + \frac{\partial^4 q}{\partial x^4} = -u_0 \frac{\mathrm{d}}{\mathrm{d}x} \delta_\xi - \dot{u}_1 [\alpha \delta_\xi + b]. \tag{6.16}$$

This equation must also hold in the sense of distributions on $(0, \pi)$. Together with (6.12) and (6.13), it defines our open-loop system with the "designed" actuators. However, since in general $\frac{\partial^2 q}{\partial x^2}$ will have a discontinuity at ξ, the second part of (6.13) has to be replaced by $y_1(t) = \overline{C}_1 q$, where $\overline{C}_1$ is an extension of C_1 given by

$$\overline{C}_1 \chi = \gamma \chi''(\xi-) + (1 - \gamma) \chi''(\xi+).$$

PROPOSITION 6.4. *For $u \in \mathcal{H}^1_{loc}(0, \infty; \mathbb{C}^2)$, the equations* (6.16) *(in $H_{-1.5}$) and* (6.13) *(with C_1 extended as above) are equivalent to*

$$\dot{z} = Az + Bu \quad (\text{in } X_{-1}), \qquad y = \overline{C} z + Du,$$

with z, u, y as in (6.4) *and with $w = \dot{q} - A_0^{-1} C_1^* u_1 = \dot{q} + b u_1$. Here, A is as in* (6.5), *B is as in* (6.6), *and $\overline{C}$, D are as in* (6.7). *These equations determine a well-posed linear system Σ with input and output space $U = \mathbb{C}^2$ and state space $X = H_{\frac{1}{2}} \times H$. This system is exactly controllable and exactly observable.*

For the proof of this proposition we refer to our paper [50].

PROPOSITION 6.5. *Let Σ be the system from Proposition* 6.4. *Denote*

$$D_1 = \overline{C_1} A_0^{-1} C_0^*, \qquad \kappa_0 = \frac{2}{|D_1|}.$$

Then for every $\kappa \in (0, \kappa_0)$, $K = -\kappa I$ is an admissible feedback operator for Σ and the resulting closed-loop system Σ^κ is exponentially stable.

Proof. We know from Proposition 6.4 that the system Σ is well posed, exactly controllable, and exactly observable. Let $E \in \mathcal{L}(\mathbb{C}^2)$ be the 2×2 matrix from (6.10). Its eigenvalues are $\pm \frac{1}{2} |D_1|$ so that E^+ (the positive part of E) is a matrix of rank one with the eigenvalues $\frac{1}{2}|D_1|$ and 0. Hence, $c = \|E^+\| = \frac{1}{2}|D_1|$, $\kappa_0 = \frac{1}{c} = \frac{2}{|D_1|}$. According to Proposition 6.2, for every $\kappa \in (0, \kappa_0)$, $K = -\kappa I$ is an admissible feedback operator for Σ and Σ^κ is exponentially stable. □

In [50] we derive a simple formula for the number D_1 appearing in the last proposition: $D_1 = \frac{\xi}{\pi} - \gamma$. Note that choosing $\gamma = \frac{\xi}{\pi}$ leads to $\kappa_0 = \infty$.

REFERENCES

[1] K. Ammari, Z. Liu, and M. Tucsnak, *Decay rates for a beam with pointwise force and moment feedback*, Math. Control Signals Systems, 15 (2002), pp. 229–255.

[2] K. Ammari and M. Tucsnak, *Stabilization of second order evolution equations by a class of unbounded feedbacks*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 361–386.

[3] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach*, Prentice-Hall Network Series, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[4] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials—The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[5] T. Bailey and J. E. Hubbard, Jr., *Distributed piezoelectric polymer active vibration control of a cantilever beam*, AIAA J. Guidance Control Dyn., 8 (1985), pp. 605–611.

[6] A. V. Balakrishnan, *Compensator design for stability enhancement with collocated controllers*, IEEE Trans. Automat. Control, 36 (1992), pp. 994–1008.

[7] A. V. Balakrishnan, *Shape control of plates with piezo actuators and collocated position / rate sensors*, Appl. Math. Comput., 63 (1994), pp. 213–234.

[8] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[9] R. J. Benhabib, R. P. Iwens, and R. L. Jackson, *Stability of large space structure control systems using positivity concepts*, AIAA J. Guidance Control Dyn., 4 (1981), pp. 487–494.

[10] G. Chen, *A note on boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106–113.

[11] G. Chen, M. C. Delfour, A. M. Krall, and G. Payre, *Modelling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.

[12] V. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Ouput Properties*, Academic Press, New York, 1975.

[13] W. K. Gawronski, *Dynamics and Control of Structures, A Modal Approach*, Springer-Verlag, New York, 1998.

[14] B-Z. Guo and Z-H. Luo, *Controllability and stability of a second-order hyperbolic system with colocated sensor/actuator*, Systems Control Lett., 46 (2002), pp. 45–65.

[15] A. Haraux, *Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps*, Port. Math., 46 (1989), pp. 245–258.

[16] B. Jacob and J. R. Partington, *The Weiss conjecture on admissibility of observation operators for contraction semigroups*, Integral Equations Operator Theory, 40 (2001), pp. 231–243.

[17] B. Jacob and H. Zwart, *Exact observability of diagonal systems with a finite dimensional output operator*, Systems Control Lett., 43(2001), pp. 101–109.

[18] S. M. Joshi, *Control of Large Flexible Space Structures*, Lecture Notes in Control and Inform. Sci. 131, Springer-Verlag, New York, 1989.

[19] V. Komornik, *Exact Controllability and Stabilization—The Multiplier Method*, John Wiley and Masson, Chichester and Paris, 1994.

[20] I. Lasiecka and R. Triggiani, *$L_2(\Sigma)$-regularity of the boundary to boundary operator B^*L for hyperbolic and Petrowski PDE's*, Abstract and Applied Analysis, 19 (2003), pp. 1061–1139.

[21] I. Lasiecka and R. Triggiani, *The operator B^*L for the wave equation with Dirichlet control*, Abstr. Appl. Anal., 7 (2004), pp. 625–634.

[22] J. E. Lagnese and J.-L. Lions, *Modelling, Analysis and Control of Thin Plates*, Masson and Springer-Verlag, Paris and Berlin, 1988.

[23] J.-L. Lions, *Controlabilité Exacte Perturbations et Stabilisation de Systèmes Distribués*, Volume 1, Masson, Paris, 1988.

[24] K. Liu, *Local distributed control and damping for the conservative systems*, SIAM J. Control Optim., 35 (1997), pp. 1574–1590.

[25] Z-H. Luo, B-Z. Guo, and O. Morgul, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, London, 1999.

[26] L. Miller, *Controllability cost of conservative systems: Resolvent condition and transmutation*, J. Funct. Anal., 218 (2005), pp. 425–444.

[27] R. Rebarber, *Exponential stability of coupled beams with dissipative joints: A frequency domain approach*, SIAM J. Control Optim., 33 (1995), pp. 1–28.

[28] R. Rebarber and S. Townley, *Robustness and continuity of the spectrum for uncertain distributed systems*, Automatica J. IFAC, 31 (1995), pp. 1533–1546.

[29] R. Rebarber and G. Weiss, *Necessary conditions for exact controllability with a finite-dimensional input space*, Systems Control Lett., 40 (2000), pp. 217-227.
[30] D. L. Russell, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Appl., 25 (1969), pp. 663–675.
[31] D. L. Russell and G. Weiss, *A general necessary condition for exact observability*, SIAM J. Control Optim., 32 (1994), pp. 1–23.
[32] D. Salamon, *Infinite dimensional systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[33] M. Slemrod, *Stabilization of boundary control systems*, J. Differential Equations, 22 (1976), pp. 402–415.
[34] M. Slemrod, *Feedback stabilization of a linear control system in Hilbert space with an a priori bounded control*, Math. Control Signals Systems, 2 (1989), pp. 265–285.
[35] O. J. Staffans, *Quadratic optimal control of stable well-posed linear systems*, Trans. Amer. Math. Soc. 349 (1997), pp. 3679–3715.
[36] O. J. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, UK, 2005.
[37] O. J. Staffans and G. Weiss, *Transfer functions of regular linear systems, Part II: The system operator and the Lax-Phillips semigroup*, Trans. Amer. Math. Soc., 354 (2002), pp. 3329–3262.
[38] O. J. Staffans and G. Weiss, *Transfer functions of regular linear systems, Part III: Inversions and duality*, Integral Equations Operator Theory, 49 (2004), pp. 517–558.
[39] V. L. Syrmos, C. T. Abdallah, P. Dorato, and K. Grigoriadis, *Static output feedback - A survey*, Automatica J. IFAC, 33 (1997), pp. 125–137.
[40] R. Triggiani, *Wave equation on a bounded domain with boundary dissipation: An operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438–461.
[41] M. Tucsnak and G. Weiss, *How to get a conservative well-posed linear system out of thin air. Part II: Controllability and stability*, SIAM J. Control Optim., 42 (2003), pp. 907–935.
[42] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[43] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
[44] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, W. Schappacher, eds., Birkhäuser-Verlag, Basel, Switzerland, 1989, pp. 401–416.
[45] G. Weiss, *Transfer functions of regular linear systems, Part I: Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[46] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.
[47] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser-Verlag, Basel, Switzerland, 1991, pp. 367–378.
[48] G. Weiss, *A powerful generalization of the Carleson measure theorem?*, in Open Problems in Mathematical Systems Theory and Control, V. Blondel, E. Sontag, M. Vidyasagar, and J. Willems, eds., Springer-Verlag, London, 1999, pp. 267–272.
[49] G. Weiss, *Optimal control of systems with a unitary semigroup and with colocated control and observation*, Systems Control Lett., 48 (2003), pp. 329–340.
[50] G. Weiss and R. F. Curtain, *Exponential stabilization of a Rayleigh beam using colocated control*, IEEE Trans. Automat. Control, to appear.
[51] G. Weiss and R. Rebarber, *Optimizability and estimatability for infinite-dimensional linear systems*, SIAM J. Control Optim., 39 (2001), pp. 1204-1232.
[52] G. Weiss, O. J. Staffans and M. Tucsnak, *Well-posed linear systems—a survey with emphasis on conservative systems*, Appl. Math. Computer Sci., 11 (2001), pp. 101–127.
[53] G. Weiss and M. Tucsnak, *How to get a conservative well-posed linear system out of thin air. Part I: Well-posedness and energy balance*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 247–274.
[54] E. Zeheb and D. Hertz, *Robust control of the characteristic values of systems with possible parameter variations*, Internat. J. Control, 40 (1984), pp. 81–96.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 298–319

TWO-TIME-SCALE HYBRID FILTERS: NEAR OPTIMALITY*

J. W. WANG†, Q. ZHANG‡, AND G. YIN§

Abstract. This work develops a filtering scheme for hybrid systems. The process dictating the configuration or regimes is a continuous-time Markov chain with a finite state space. Exploiting hierarchical structure of the underlying system, the states of the Markov chain are divided into a number of groups so that it jumps rapidly within each group and slowly among different groups. Focusing on reduction of computational complexity, the filtering scheme includes the following steps: (1) partition the state space of the Markov chain into subspaces, (2) derive a limit system in which the states are averaged out with respect to the invariant distributions of the Markov chain, (3) use the limit system to design quadratic variation test statistics, and (4) use the test statistics to identify which ergodic class the aggregated process belongs to and to construct near-optimal filter. For demonstration, a numerical example is presented.

Key words. hybrid filter, near optimality, Kalman filter, Markov chain, weak convergence, quadratic variation test

AMS subject classifications. 93E11, 93C70, 60J27

DOI. 10.1137/S0363012904443725

1. Introduction.

This paper is concerned with filtering problems of hybrid dynamic systems, in which continuous dynamics and discrete events coexist. Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space and $(x(t), \alpha(t))$ with $t \geq 0$ be a pair of signal processes that is not observable. Here $x(t)$ represents the continuous dynamics part and $\alpha(t)$ describes the discrete events. More specifically, the unknown or hidden process $\alpha(t)$ is a continuous-time Markov chain with a finite state space $\mathcal{M}$. Denote the observation process by $y(t)$ depending on $(x(s), \alpha(s),\ s \leq t)$ together with a Gaussian noise. Throughout the paper, we assume both $x(\cdot)$ and $y(\cdot)$ to be $\mathbb{R}^p$-valued stochastic processes satisfying the system of equations

$$\begin{cases} dx(t) = A(\alpha(t))x(t)dt + \sigma(\alpha(t))dw(t), & x(0) = x_0, \\ dy(t) = H(\alpha(t))x(t)dt + \delta dv(t), & y(0) = 0, \end{cases} \tag{1.1}$$

for $0 \leq t \leq T$, where T is a finite number, x_0 is a given random variable, $w(\cdot)$ and $v(\cdot)$ are independent standard Brownian motions, for $\alpha \in \mathcal{M}$, $A(\alpha)$, $H(\alpha)$, and $\sigma(\alpha)$ are matrices of appropriate dimensions, and $\delta > 0$ is a small parameter representing the intensity of the observation noise.

In the literature, when $\alpha(t) = \alpha_0$ is an unobservable constant, the optimal filtering problem was solved by Hijab [12] and Caines and Chen [3], assuming an a priori distribution of α_0. However, in many practical situations, a priori knowledge of α_0 is not available. To deal with such scenarios, Haussmann and Zhang [11] used two

*Received by the editors May 11, 2004; accepted for publication (in revised form) September 8, 2005; published electronically March 15, 2006. This research was supported in part by the National Science Foundation.

http://www.siam.org/journals/sicon/45-1/44372.html

†Asset Finance Group—New Products, Moodys Investors Service, 99 Church Street, New York, NY 10007 (jianwu.wang@moodys.com).

‡Department of Mathematics, Boyd GSRC, University of Georgia, Athens, GA 30602-7403 (qingz@math.uga.edu).

§Department of Mathematics, Wayne State University, Detroit, MI 48202 (gyin@math.wayne.edu).

statistical hypothesis tests, namely, the quadratic variation test (QVT) and the likelihood ratio test (LRT), to estimate the value of α_0 and to choose among competing filters on successive time intervals. The QVT and LRT were introduced by Fleming and Pardoux [6] to identify the sign of the state variable $x(t)$ in a partially observed system; see also Fleming and Zhang [8, 9] and Fleming et al. [5] for the corresponding discrete-time models and related numerical results along this line. These results are generalized to the case when $\alpha(\cdot)$ is an unknown Markov chain in Zhang [20]. The basic idea of these tests is to examine which state $\alpha(t)$ is more likely to belong to statistically. Then run a number of Kalman filters simultaneously and choose the one that associates with the current estimate of $\alpha(t)$. Therefore, a necessary condition for validating the tests is that $\alpha(t)$ must stay at the current state for some time before jumping to another state. If $\alpha(t)$ jumps too frequently, it is virtually impossible to obtain a meaningful estimate of $x(t)$.

In recent years, hybrid systems of the form (1.1), involving both continuous dynamics and discrete events, have attracted much attention due to their wide range of applicability in financial engineering, wireless communication, and network modeling and optimization. In these applications, the hidden Markov chain represents the random influence due to structural or regime changes in environmental, social, economical, or other factors. Denote the cardinality of the state space of the hidden Markov chain by $|\mathcal{M}|$. For the aforementioned applications, considerations of various factors frequently make $|\mathcal{M}|$ very large resulting in a difficult computational task. The main focus of this paper is to reduce the computational complexity. Within a large-scale system, not all state variables vary at the same speed. We may naturally subdivide the entire system into smaller pieces so that within each subsystem the participating state variables vary rapidly and among different subgroups the jumps take place at a slower pace. Noting the different transition rates of the Markov chain $\alpha(t)$ leads to a two-time-scale formulation involving states having weak and strong interactions. To highlight the contrast of the fast and slow motions, we introduce a small parameter $\varepsilon > 0$ and aim to design near-optimal filters. Our filtering procedure proceeds as follows. First, using weak convergence methods, we obtain certain limit filtering problem as $\varepsilon \to 0$. With the stipulation that the Markov chain varies slowly among different recurrent groups, we then develop quadratic variation test and use it in the process of filtering.

Concerning nonlinear filtering of a hybrid system, much effort has been devoted to finite dimensional approximations. In Blom and Bar-Shalom [2], a discrete-time version of the corresponding filtering problem was considered. They proposed a numerical algorithm to compute the conditional expectation of $(\alpha(t), x(t))$ given observation up to time t. The algorithm seems to perform well numerically. However, there is no theoretical justification for optimality (or near optimality) of these filters; see Li [15] for further discussions. There is substantial literature on related models and problems. For hybrid filter with non-Gaussian noise, see Zhang [22] for a continuous-time model, Zhang [21] for a discrete-time version, and Liu and Zhang [17] for numerical experiments. For a survey of results on nonlinear filtering, we refer to the books by Kallianpur [13] and Liptser and Shiryayev [16]. For recent developments and review of the literature on partially observed systems, see the books by Bensoussan [1], Elliott, Aggoun, and Moore [4], and Kushner [14] and references therein.

The rest of the paper is organized as follows. In the next section, we set up the problem by using a two-time-scale formulation. Preliminary results concerning two-time-scale Markov chains and limit filtering problem are also presented. Section 3 is devoted to an averaged filtering scheme, in which an intermediate filter is developed

for approximation purpose. First, a weak convergence method is used to establish the convergence of the original system to a limit system. Then we construct an averaged filter and show that its output approximates the state process. More specifically, we partition the state space of the Markov chain into subspaces, derive a limit system in which the states are averaged out with respect to the invariant distributions of the Markov chain, design quadratic variation test statistics using the limit system, design associated test statistics for the original process to identify which ergodic class that the aggregated process belongs to, and construct a near-optimal filter. In section 4, we demonstrate the near optimality of the filter by a simple example. Section 5 extends the results to the case, where some transient states are included in the Markov chain. Final remarks are made in section 6. To proceed, a word about the notation is in order. For simplicity, throughout the paper, we use K to denote a generic positive constant, whose values may change for different appearances. Given a matrix A, its transpose is denoted by A' and its trace by $\mathrm{tr}(A)$.

2. Preliminary results. In this section, we present results on two-time-scale Markov chains. In addition, a statistical test for the filtering is also given.

2.1. Two-time-scale Markov chains. To characterize the weak and strong interaction of $\alpha(t)$, we introduce a small parameter $\varepsilon > 0$ and assume $\alpha(\cdot) = \alpha^\varepsilon(\cdot)$, generated by Q^ε, where Q^ε has the form

$$Q^\varepsilon = \frac{1}{\varepsilon}\,\mathrm{diag}(\widetilde{Q}^1, \dots, \widetilde{Q}^l) = \frac{1}{\varepsilon}\begin{pmatrix} \widetilde{Q}^1 & & \\ & \ddots & \\ & & \widetilde{Q}^l \end{pmatrix} + \widehat{Q}, \tag{2.1}$$

where, for each $k = 1, \dots, l$, $\widetilde{Q}^k$ is a generator matrix with dimension $m_k \times m_k$, and $m_1 + \cdots + m_l = m$. Suppose that for $k = 1, \dots, l$, $\widetilde{Q}^k$ are irreducible. The state space of $\alpha^\varepsilon(\cdot)$ is given by

$$\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_l = \{s_{11}, \dots, s_{1m_1}\} \cup \cdots \cup \{s_{l1}, \dots, s_{lm_l}\},$$

where for $k = 1, \dots, l$, $\mathcal{M}_k = \{s_{k1}, \dots, s_{km_k}\}$ are the subspaces corresponding to the recurrent states. Let $\nu^k = (\nu^k_1, \dots, \nu^k_{m_k})$ denote the stationary distribution of $\widetilde{Q}^k$, i.e., ν^k is the only solution to

$$\nu^k \widetilde{Q}^k = 0 \text{ and } \nu^k 1_{m_k} = 1.$$

Let

$$\overline{Q} = \mathrm{diag}(\nu^1, \dots, \nu^l)\widehat{Q}\,\widetilde{1}, \tag{2.2}$$

where

$$\widetilde{1} = \begin{pmatrix} 1_{m_1} & & \\ & \ddots & \\ & & 1_{m_l} \end{pmatrix} \quad \text{and} \quad 1_{m_k} = (1, \dots, 1)' \in \mathbb{R}^{m_k \times 1}.$$

Noting the dominating force in Q^ε, namely, $\mathrm{diag}(\widetilde{Q}^1, \dots, \widetilde{Q}^l)$, the generator Q^ε is almost decomposable. The rationale is that we can take advantage of this structure and use a single state to represent each of the subspaces $\mathcal{M}_k$ (for $k = 1, \dots, l$). The

resulting process will have a much smaller state space. This leads to the following definition of aggregated process.

Let $\overline{\alpha}^\varepsilon(\cdot)$ be defined by $\overline{\alpha}^\varepsilon(t) = k$ if $\alpha^\varepsilon(t) \in \mathcal{M}_k$. Then as shown in Yin and Zhang [18, Thm. 7.4, p. 172] that $\overline{\alpha}^\varepsilon(\cdot)$ converges weakly to $\overline{\alpha}(\cdot)$, a continuous-time Markov chain with generator $\overline{Q}$ and state space $\overline{\mathcal{M}} = \{1, \ldots, l\}$. Moreover, for any deterministic test function $\beta_{kj}(t)$ with $k = 1, \ldots, l$ and $j = 1, \ldots, m_k$, the corresponding occupation measures satisfy

$$E\left(\int_0^T \left(I_{\{\alpha^\varepsilon(t)=s_{kj}\}} - \nu_j^k I_{\{\overline{\alpha}^\varepsilon(t)=k\}}\right)\beta_{kj}(t)dt\right)^2 = O(\varepsilon). \tag{2.3}$$

These estimates may be regarded as functional version of the law of large numbers if one discretizes the above integrals using step size ε.

These results suggest the following two-stage approximation of $\alpha^\varepsilon(\cdot)$ by $\overline{\alpha}(\cdot)$: (i) approximate $\alpha^\varepsilon(\cdot)$ by $\overline{\alpha}^\varepsilon(\cdot)$ under the weak topology in $L^2[0,T]$; (ii) approximate $\overline{\alpha}^\varepsilon(\cdot)$ by its limit $\overline{\alpha}(\cdot)$ in distribution.

2.2. Limit filter. Let $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ be the actual state and observation satisfying (1.1) with $\alpha(\cdot)$ replaced by $\alpha^\varepsilon(\cdot)$, i.e.,

$$\begin{cases} dx^\varepsilon(t) = A(\alpha^\varepsilon(t))x^\varepsilon(t)dt + \sigma(\alpha^\varepsilon(t))dw(t), & x^\varepsilon(0) = x_0, \\ dy^\varepsilon(t) = H(\alpha^\varepsilon(t))x^\varepsilon(t)dt + \delta dv(t), & y^\varepsilon(0) = 0. \end{cases} \tag{2.4}$$

We will show that

$$(x^\varepsilon(\cdot), y^\varepsilon(\cdot), \overline{\alpha}^\varepsilon(\cdot)) \to (\overline{x}(\cdot), \overline{y}(\cdot), \overline{\alpha}(\cdot)) \text{ as } \varepsilon \to 0,$$

in distribution, where $(\overline{x}(\cdot), \overline{y}(\cdot))$ satisfies the limit system of equations

$$\begin{cases} d\overline{x}(t) = \overline{A}(\overline{\alpha}(t))\overline{x}(t)dt + \overline{\sigma}(\overline{\alpha}(t))dw(t), & \overline{x}(0) = x_0, \\ d\overline{y}(t) = \overline{H}(\overline{\alpha}(t))\overline{x}(t)dt + \delta dv(t), & \overline{y}(0) = 0, \end{cases} \tag{2.5}$$

with

$$\begin{cases} \overline{A}(k) = \displaystyle\sum_{j=1}^{m_k} \nu_j^k A(s_{kj}), \\ \overline{H}(k) = \displaystyle\sum_{j=1}^{m_k} \nu_j^k H(s_{kj}), \\ \overline{\Sigma}(k) = \overline{\sigma}(k)\overline{\sigma}'(k) = \displaystyle\sum_{j=1}^{m_k} \nu_j^k \sigma(s_{kj})\sigma'(s_{kj}) \end{cases} \tag{2.6}$$

for $k = 1, \ldots, l$.

In what follows, we briefly summarize related filtering results obtained in Zhang [20] for the limit system (2.5). We make the following assumptions:

- (**A1**) $\overline{H}(k)$ is invertible for $k \in \overline{\mathcal{M}}$.
- (**A2**) $\overline{\sigma}(k) = F(k)\overline{H}'(k)$ for symmetric matrices $F(k) \geq cI > 0$, for some $c > 0$.
- (**A3**) $\left|\mathrm{tr}\big(\overline{H}(k)F(k)\overline{H}'(k)\big)^2 - \mathrm{tr}\big(\overline{H}(k_1)F(k_1)\overline{H}'(k_1)\big)^2\right| \geq c > 0$ for $k \neq k_1$, for some $c > 0$.
- (**A4**) The initial value x_0 is a Gaussian random vector and $E|x_0 - Ex_0|^4 = O(\delta^2)$. Moreover, x_0, $\alpha^\varepsilon(\cdot)$, $w(\cdot)$, and $v(\cdot)$ are independent.

Remark 2.1. The form assumed for $\overline{\sigma}(k)$ in (A2) is not as restrictive as it appears. In fact, in the one-dimensional case, it is equivalent to the condition $\overline{\sigma}(k) \neq 0$. If $F(k) = 1$, then (A3) is equivalent to the condition $|\overline{H}(k)| \neq |\overline{H}(k_1)|$, for $k \neq k_1$. For higher dimensional cases, we refer to the papers by Haussmann and Zhang [10, 11] for related discussions. For notational simplicity, we use the Gaussian initial conditions in (A4). This requirement can be relaxed as in Haussmann and Zhang [10] to non-Gaussian initial cases. We next give an example to further illustrate these conditions in a one-dimensional ($p = 1$) system.

EXAMPLE 2.2. *Let* $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2 = \{s_{11}, s_{12}\} \cup \{s_{21}, s_{22}\}$ *and the generator*

$$Q^\varepsilon = \frac{1}{\varepsilon}\begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix} + \begin{pmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

Then the corresponding stationary distributions $\nu^1 = \nu^2 = (1/2, 1/2)$. *Consider the case when*

$$\sigma(s_{11}) = \sigma(s_{12}) = \sigma(s_{21}) = \sigma(s_{22}) = 1,$$

$$H(s_{11}) = H(s_{12}) = 1,\ H(s_{21}) = H(s_{22}) = 2.$$

Then

$$\overline{\sigma}(1) = \overline{\sigma}(2) = 1,\ \overline{H}(1) = 1,\ \overline{H}(2) = 2,\ \textit{and}\ F(1) = 1,\ F(2) = 1/2.$$

It can be seen that the conditions (A1)–(A3) *are satisfied.*

The parameter δ represents the observation noise level. In practice, the observation noise mainly depends on the sensor measurement characteristics. The development of new technology (such as the use of infrared technology) makes possible fairly small disturbances in observation. Thus it is not only reasonable but practical to consider the models with small observation noise.

If $\overline{\alpha}(t)$ is observable, then the corresponding limit Kalman filter has the form

$$\text{(2.7)}\qquad \begin{cases} d\widehat{x}(t) = \overline{A}(\overline{\alpha}(t))\widehat{x}(t)dt + \dfrac{1}{\delta^2}\overline{R}(t)\overline{H}'(\overline{\alpha}(t))\left(d\overline{y}(t) - \overline{H}(\overline{\alpha}(t))\widehat{x}(t)dt\right), \\ \dot{\overline{R}}(t) = \overline{A}(\overline{\alpha}(t))\overline{R}(t) + \overline{R}(t)\overline{A}'(\overline{\alpha}(t)) \\ \qquad\qquad + \overline{\sigma}(\overline{\alpha}(t))\overline{\sigma}'(\overline{\alpha}(t)) - \dfrac{1}{\delta^2}\overline{R}(t)\overline{H}'(\overline{\alpha}(t))\overline{H}(\overline{\alpha}(t))\overline{R}(t) \end{cases}$$

with $\widehat{x}(0) = Ex_0 \in \mathbb{R}^p$ and $\overline{R}(0) = \mathrm{Cov}(x_0) \in \mathbb{R}^{p\times p}$, where $\mathrm{Cov}(x_0)$ is the covariance matrix $E[(x_0 - Ex_0)(x_0 - Ex_0)']$.

It is important to estimate the value of $\overline{\alpha}(\cdot)$ in the filtering problem under consideration because if an estimate of $\overline{\alpha}(\cdot)$ is given, one can use such estimate to feed to the corresponding Kalman filter. In this paper, we use the QVT to identify the value of the unknown parameter process $\overline{\alpha}(\cdot)$ at a given time t.

For $j = 0, 1, \ldots,$ let

$$\zeta(j) = \frac{1}{\delta}\left(\overline{y}(\delta(j+2)) - 2\overline{y}(\delta(j+1)) + \overline{y}(\delta j)\right).$$

For $n_0, n = 0, 1, \ldots, n > n_0$, we define test statistics

$$\Lambda^{n_0,n} = \frac{1}{\delta(n-n_0)} \sum_{j=n_0}^{n-1} |\zeta(j)|^2. \tag{2.8}$$

Let

$$\mu_k^{n_0,n} = \left(\frac{1}{\delta(n-n_0)} \int_{\delta n_0}^{\delta(n+1)} \rho^{n_0,n}(s) ds \right) \mathrm{tr}(\overline{H}(k) F(k) \overline{H}'(k))^2 + 2p,$$

where

$$\rho^{n_0,n}(s) = \begin{cases} \phi^2(s) I_{[\delta n_0, \delta n]} + \phi^2(s-\delta) I_{[\delta(n_0+1), \delta n]} & \text{if } (n-n_0) \text{ even}, \\ \phi^2(s) I_{[\delta n_0, \delta(n+1)]} + \phi^2(s-\delta) I_{[\delta(n_0+1), \delta(n+1)]} & \text{if } (n-n_0) \text{ odd}, \end{cases}$$

and $\phi(s)$ is a "sawtooth" function on $[0, T]$ such that for any even integer $j = 0, 2, 4, \ldots,$

$$\phi(s) = \begin{cases} \dfrac{s - j\delta}{\delta} & \text{if } j\delta \le s < (j+1)\delta, \\ \dfrac{(j+2)\delta - s}{\delta} & \text{if } (j+1)\delta \le s < (j+2)\delta. \end{cases}$$

(See Haussmann and Zhang [11] for interpretation of these functions.)

Given $\overline{\alpha}(t) = k$, $\delta n_0 \le t < \delta n$, it can be shown as in [11] that for large $(n - n_0)$, $\Lambda^{n_0,n}/\mu_k^{n_0,n}$ is close to 1 by the law of large numbers. Under (A3), it follows that $\mu_{k_1}^{n_0,n} \ne \mu_{k_2}^{n_0,n}$ for $k_1 \ne k_2$. The quadratic variation test is given as follows. Let $\overline{\alpha}^{n_0,n}$ denote a random variable such that $\overline{\alpha}^{n_0,n} = k_0$ if

$$\left| \frac{\Lambda^{n_0,n}}{\mu_{k_0}^{n_0,n}} - 1 \right| = \min \left\{ \left| \frac{\Lambda^{n_0,n}}{\mu_k^{n_0,n}} - 1 \right|, \ k = 1, \ldots, l \right\}. \tag{2.9}$$

Let

$$l_0 = c_0 (\log \delta)^2 + 1 \ \text{ for some } \ c_0 \ge 0. \tag{2.10}$$

Then, for $\delta n \le t < \delta(n+1)$, $n = 0, 1, \ldots,$ we define

$$\widetilde{\alpha}(t) = \begin{cases} 1 & \text{if } t \in [0, \delta), \\ \overline{\alpha}^{0,n} & \text{if } n \le l_0, \\ \overline{\alpha}^{n-l_0,n} & \text{if } n \ge l_0. \end{cases} \tag{2.11}$$

Given $\widetilde{\alpha}(t)$, the corresponding filter is

$$\begin{cases} d\widetilde{x}(t) = \overline{A}(\widetilde{\alpha}(t))\widetilde{x}(t)dt + \dfrac{1}{\delta^2} \widetilde{R}(t) \overline{H}'(\widetilde{\alpha}(t)) \left(d\overline{y}(t) - \overline{H}(\widetilde{\alpha}(t))\widetilde{x}(t)dt \right), \\ \dot{\widetilde{R}}(t) = \overline{A}(\widetilde{\alpha}(t))\widetilde{R}(t) + \widetilde{R}(t)\overline{A}'(\widetilde{\alpha}(t)) \\ \qquad\qquad + \overline{\sigma}(\widetilde{\alpha}(t))\overline{\sigma}'(\widetilde{\alpha}(t)) - \dfrac{1}{\delta^2} \widetilde{R}(t) \overline{H}'(\widetilde{\alpha}(t)) \overline{H}(\widetilde{\alpha}(t)) \widetilde{R}(t) \end{cases} \tag{2.12}$$

with $\widetilde{x}(0) = Ex_0 \in \mathbb{R}^p$ and $\widetilde{R}(0) = \mathrm{Cov}(x_0) \in \mathbb{R}^{p\times p}$. For any given stochastic process $\xi(t)$, $t \geq 0$, we define the norm of $\xi(\cdot)$ as follows:

$$|\xi(\cdot)|_T = \int_0^T \left(E|\xi(t)|^2\right)^{\frac{1}{2}} dt.$$

Using this norm, it can be shown as in Zhang [20] that, under (A1)–(A4), there exist $c_0 > 0$ and $\delta_0 > 0$ such that for $0 < \delta < \delta_0$,

$$|\overline{\alpha}(\cdot) - \widetilde{\alpha}(\cdot)|_T = O(\sqrt{\delta}).$$

3. Near-optimal filter. Note that $\overline{y}(\cdot)$, the weak limit of $y^\varepsilon(\cdot)$, is not available. In order to have a feasible filter, one needs to feed the filter in (2.12) with the actual observation $y^\varepsilon(\cdot)$. Thus, we need to replace $\overline{y}(\cdot)$ by $y^\varepsilon(\cdot)$ in (2.12) and modify the test statistics Λ in (2.8) as follows. For $j = 0, 1, \ldots$, let

$$\zeta^\varepsilon(j) = \frac{1}{\delta}\left(y^\varepsilon(\delta(j+2)) - 2y^\varepsilon(\delta(j+1)) + y^\varepsilon(\delta j)\right).$$

Define

$$\Lambda^{\varepsilon,n_0,n} = \frac{1}{\delta(n-n_0)} \sum_{j=n_0}^{n-1} |\zeta^\varepsilon(j)|^2. \tag{3.1}$$

Let $\alpha^{\varepsilon,n_0,n}$ denote a random variable such that $\alpha^{\varepsilon,n_0,n} = k_0$ if

$$\left|\frac{\Lambda^{\varepsilon,n_0,n}}{\mu_{k_0}^{n_0,n}} - 1\right| = \min\left\{\left|\frac{\Lambda^{\varepsilon,n_0,n}}{\mu_k^{n_0,n}} - 1\right|,\ k = 1, \ldots, l\right\}. \tag{3.2}$$

Then, for $\delta n \leq t < \delta(n+1)$, $n = 0, 1, \ldots$, we define

$$\widetilde{\alpha}^\varepsilon(t) = \begin{cases} 1 & \text{if } t \in [0, \delta), \\ \alpha^{\varepsilon,0,n} & \text{if } n \leq l_0, \\ \alpha^{\varepsilon,n-l_0,n} & \text{if } n \geq l_0, \end{cases} \tag{3.3}$$

where l_0 is defined in (2.10).

Given $\widetilde{\alpha}^\varepsilon(t)$, the averaged filter is

$$\begin{cases} d\widetilde{x}^\varepsilon(t) = \overline{A}(\widetilde{\alpha}^\varepsilon(t))\widetilde{x}^\varepsilon(t)dt + \dfrac{1}{\delta^2}R^\varepsilon(t)\overline{H}'(\widetilde{\alpha}^\varepsilon(t))\left(dy^\varepsilon(t) - \overline{H}(\widetilde{\alpha}^\varepsilon(t))\widetilde{x}^\varepsilon(t)dt\right), \\ \dot{R}^\varepsilon(t) = \overline{A}(\widetilde{\alpha}^\varepsilon(t))R^\varepsilon(t) + R^\varepsilon(t)\overline{A}'(\widetilde{\alpha}^\varepsilon(t)) \\ \qquad\qquad + \overline{\sigma}(\widetilde{\alpha}^\varepsilon(t))\overline{\sigma}'(\widetilde{\alpha}^\varepsilon(t)) - \dfrac{1}{\delta^2}R^\varepsilon(t)\overline{H}'(\widetilde{\alpha}^\varepsilon(t))\overline{H}(\widetilde{\alpha}^\varepsilon(t))R^\varepsilon(t), \end{cases} \tag{3.4}$$

with $\widetilde{x}^\varepsilon(0) = Ex_0 \in \mathbb{R}^p$ and $R^\varepsilon(0) = \mathrm{Cov}(x_0) \in \mathbb{R}^{p\times p}$.

We will show that $\widetilde{x}^\varepsilon(\cdot)$ is an approximation to $x^\varepsilon(\cdot)$. To proceed, we first establish a couple of lemmas. They include the boundedness of the solution of Riccati equations and the moment bounds of the original and averaged processes.

LEMMA 3.1. *The solution of the system of Riccati equations $R^\varepsilon(t)$ in (3.4) is nonnegative definite and satisfies $\sup_{t\leq T}|R^\varepsilon(t)| \leq K$ for all $\omega \in \Omega$.*

Proof. First it is easy to see that $R^\varepsilon(t)$ is symmetric and nonnegative definite. We use the matrix norm $|A| = \sqrt{\mathrm{tr}(AA')}$. It follows that $\mathrm{tr}(AB) \le |A| \cdot |B|$ for any square matrices A and B of same dimension. Under this norm, for any symmetric nonnegative definite matrix R, we have

$$|R| = (\mathrm{tr}(R^2))^{\frac{1}{2}} = (\lambda_1^2 + \cdots + \lambda_n^2)^{\frac{1}{2}} \le (\lambda_1 + \cdots + \lambda_n) = \mathrm{tr}(R),$$

where $\lambda_1 \ge 0, \ldots, \lambda_n \ge 0$ are the eigenvalues of R. Therefore, it suffices to show that $\mathrm{tr}(R^\varepsilon(t))$ is bounded. Let $\phi(t) = \mathrm{tr}(R^\varepsilon(t))$. Take trace on both sides of the Riccati equation in (3.4) to obtain

$$\dot\phi(t) \le \mathrm{tr}(\overline{A}(\alpha^\varepsilon(t))R^\varepsilon(t) + R^\varepsilon(t)\overline{A}'(\alpha^\varepsilon(t))) + \mathrm{tr}(\overline{\sigma}(\alpha^\varepsilon(t))\overline{\sigma}'(\alpha^\varepsilon(t))).$$

Let $\{e_1(t), \ldots, e_p(t)\}$ be a base in $\mathbb{R}^p$ with $e_i(t)$ being eigenvector of $R^\varepsilon(t)$. Then, it follows that

$$\begin{aligned}
&\mathrm{tr}(\overline{A}(\alpha^\varepsilon(t))R^\varepsilon(t) + R^\varepsilon(t)\overline{A}'(\alpha^\varepsilon(t))) \\
&\quad = \sum_{i=1}^p e_i'(t)(\overline{A}(\alpha^\varepsilon(t))R^\varepsilon(t) + R^\varepsilon(t)\overline{A}'(\alpha^\varepsilon(t)))e_i(t) \\
&\quad = \sum_{i=1}^p \lambda_i e_i'(t)\left(\frac{\overline{A}(\alpha^\varepsilon(t)) + \overline{A}'(\alpha^\varepsilon(t))}{2}\right)e_i(t) \\
&\quad \le K_1 \mathrm{tr}(R^\varepsilon(t))
\end{aligned}$$

for some constant K_1. Therefore,

$$\dot\phi(t) \le K_0 + K_1\phi(t).$$

This implies, by Gronwall's inequality, that $\phi(t) \le K$ for some $K < \infty$. □

LEMMA 3.2.

$$\sup_{t\le T}(E|x^\varepsilon(t)|^4 + E|\overline{x}(t)|^4 + E|\widetilde{x}^\varepsilon(t)|^4 + E|\widetilde{x}(t)|^4) < \infty.$$

Proof. We shall verify this for each of the terms above. First, by the classical result on stochastic integration (for example, Fleming and Rishel [7] and Liptser and Shiryayev [16]),

$$\sup_{0\le t\le T} E|x^\varepsilon(t)|^4 \le K(1 + E|x_0|^4) < \infty, \qquad \sup_{0\le t\le T} E|\overline{x}(t)|^4 < \infty.$$

Similar estimates for the observation processes $y^\varepsilon(\cdot)$ and $\overline{y}(\cdot)$ yield that

$$\sup_{0\le t\le T} E|y^\varepsilon(t)|^4 < \infty, \qquad \sup_{0\le t\le T} E|\overline{y}(t)|^4 < \infty.$$

Next, consider $\widetilde{x}^\varepsilon(t)$. We obtain

$$\begin{aligned}
\widetilde{x}^\varepsilon(t) = \ & \widetilde{x}^\varepsilon(0) + \int_0^t \overline{A}(\widetilde{\alpha}^\varepsilon(s))\widetilde{x}^\varepsilon(s)ds + \frac{1}{\delta^2}\int_0^t R^\varepsilon(s)\overline{H}'(\widetilde{\alpha}^\varepsilon(s))H(\alpha^\varepsilon(s))x^\varepsilon(s)ds \\
& + \frac{1}{\delta}\int_0^t R^\varepsilon(s)\overline{H}'(\widetilde{\alpha}^\varepsilon(s))dv(s) - \frac{1}{\delta^2}\int_0^t R^\varepsilon(s)\overline{H}'(\widetilde{\alpha}^\varepsilon(s))\overline{H}(\widetilde{\alpha}^\varepsilon(s))R^\varepsilon(s)ds.
\end{aligned}$$

Taking norms and using the boundedness of

$$\overline{A}(i),\ \overline{H}(i),\ \sup_{0\le t\le T}|R^\varepsilon(t)|,\ \sup_{0\le t\le T}E|x^\varepsilon(t)|^4$$

and

$$E\Big|\int_0^t R^\varepsilon(s)\overline{H}'(\widetilde{\alpha}^\varepsilon(s))dv(s)\Big|^4<\infty$$

yield

$$E|\widetilde{x}^\varepsilon(t)|^4\le K+K\int_0^t E|\widetilde{x}^\varepsilon(s)|^4ds.$$

An application of Gronwall's inequality leads to $E|\widetilde{x}^\varepsilon(t)|^4<\infty$. Furthermore, the bound holds uniformly for $t\in[0,T]$. Similarly, we can establish the upper bound for $\widetilde{x}(t)$. The lemma is concluded. □

We next establish tightness of several pairs of processes. The result is stated in the following lemma.

LEMMA 3.3. *Both $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ and $(y^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ are tight in $D([0,T];\mathbb{R}^p\times\overline{\mathcal{M}})$, where $D([0,T];G)$ denotes the space of functions defined on $[0,T]$ taking values in G that are right continuous have left limits endowed with the Skorohod topology.*

Proof. Let us begin with $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$. For any $\Delta>0$, $0<t$ with $0\le s\le\Delta$, we have

$$\begin{aligned}
&\lim_{\Delta\to0}\limsup_{\varepsilon\to0}E[E[|x^\varepsilon(t+s)-x^\varepsilon(t)|^2|\mathcal{F}_t^\varepsilon]]\\
&\quad\le\lim_{\Delta\to0}\limsup_{\varepsilon\to0}E\left[E\left[\left|\sum_{k=1}^{l}\sum_{j=1}^{m_k}\int_t^{t+s}A(s_{kj})x^\varepsilon(\tau)I_{\{\alpha^\varepsilon(\tau)=s_{kj}\}}d\tau\right|^2\Big|\mathcal{F}_t^\varepsilon\right]\right]\\
&\qquad+K\lim_{\Delta\to0}\limsup_{\varepsilon\to0}E\left[E\left[\left|\sum_{k=1}^{l}\sum_{j=1}^{m_k}\int_t^{t+s}\sigma(s_{kj})I_{\{\alpha^\varepsilon(\tau)=s_{kj}\}}dw(\tau)\right|^2\Big|\mathcal{F}_t^\varepsilon\right]\right]=0.
\end{aligned}$$

In addition, in view of Yin and Zhang [18, Thm. 7.4], $\overline{\alpha}^\varepsilon(\cdot)$ is tight. Then it follows from the tightness criterion (see Kushner [14], Yin and Zhang [18, Appendix], and the references therein), $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ is tight in $D([0,T];\mathbb{R}^p\times\overline{\mathcal{M}})$. Likewise, $(y^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ is tight in $D([0,T];\mathbb{R}^p\times\overline{\mathcal{M}})$. □

We are now in a position to prove the weak convergence of $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ and $(y^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$. The proof will be furnished by using a martingale problem formulation.

Denote the class of real-valued functions that are twice continuously differentiable with compact support by C_0^2. For each $\alpha\in\mathcal{M}$ and $\psi(\cdot,\alpha)\in C_0^2$, define the operator $\mathcal{L}^\varepsilon$ by

(3.5)
$$\mathcal{L}^\varepsilon\psi(x,\alpha)=\psi_x'(x,\alpha)A(\alpha)x+\frac{1}{2}\mathrm{tr}[\psi_{xx}(x,\alpha)\sigma(\alpha)\sigma'(\alpha)]+Q^\varepsilon\psi(x,\cdot)(\alpha),\ \alpha\in\mathcal{M},$$

where $\psi_x(\cdot,\alpha)$ and $\psi_{xx}(\cdot,\alpha)$ denote the gradient and the Hessian of $\psi(\cdot,\alpha)$, respectively. For each $k\in\overline{\mathcal{M}}$, let $f(\cdot,k)\in C_0^2$ and define

(3.6)
$$\overline{f}(x,\alpha)=\sum_{k=1}^{l}f(x,k)I_{\{\alpha\in\mathcal{M}_k\}}.$$

Corresponding to (3.5), we also define

$$\mathcal{L}f(x,\overline{\alpha}) = f'_x(x,\overline{\alpha})\overline{A}(\overline{\alpha})x + \frac{1}{2}\mathrm{tr}[f_{xx}(x,\alpha)\overline{\sigma}(\overline{\alpha})\overline{\sigma}'(\overline{\alpha})] + \overline{Q}f(x,\cdot)(\overline{\alpha}) \tag{3.7}$$

for $\overline{\alpha} \in \overline{\mathcal{M}} = \{1,\dots,l\}$, where $\overline{A}(k)$ and $\overline{\sigma}(k)$ are defined in (2.6), and $\overline{Q}$ is defined in (2.2).

LEMMA 3.4. *Suppose that $(\overline{x}(\cdot),\overline{\alpha}(\cdot))$ is a solution of a martingale problem associated with the operator $\mathcal{L}$. Then it is unique (in the sense of distribution).*

Proof. The proof is a modification of Yin and Zhang [18, Lem. 7.18]. □

THEOREM 3.5. *Assume* (A1)–(A4). *Then as $\varepsilon \to 0$,*

$$(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot)) \Rightarrow (\overline{x}(\cdot),\overline{\alpha}(\cdot)) \quad \textit{and} \quad (y^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot)) \Rightarrow (\overline{y}(\cdot),\overline{\alpha}(\cdot)),$$

where $\overline{x}(\cdot)$ and $\overline{y}(\cdot)$ are solutions of (2.5) *and the symbol $\Rightarrow$ denotes convergence in distribution.*

Proof. Let $f(\cdot,k) \in C_0^2$ and $\overline{f}(\cdot,\alpha)$ as defined in (3.6) for each $k \in \overline{\mathcal{M}}$ and each $\alpha \in \mathcal{M}$. Since the joint pair of processes $(x^\varepsilon(\cdot),\alpha^\varepsilon(\cdot))$ is Markov, for any $t,s \geq 0$ with $t+s \leq T$,

$$\overline{f}(x^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - \overline{f}(x^\varepsilon(t),\alpha^\varepsilon(t)) - \int_0^t \mathcal{L}^\varepsilon \overline{f}(x^\varepsilon(u),\alpha^\varepsilon(u))du$$

is a continuous-time martingale. Thus, for any positive integer κ, any bounded and continuous functions $h_i(\cdot,\alpha)$ for $i \leq \kappa$, and any $t_i \leq t$,

$$E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i),\alpha^\varepsilon(t_i))\Bigg[\overline{f}(x^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - \overline{f}(x^\varepsilon(t),\alpha^\varepsilon(t)) - \int_t^{t+s} \mathcal{L}^\varepsilon \overline{f}(x^\varepsilon(u),\alpha^\varepsilon(u))du\Bigg] = 0,$$

and as a result,

$$\begin{aligned} &E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i),\overline{\alpha}^\varepsilon(t_i))\Bigg[\overline{f}(x^\varepsilon(t+s),\alpha^\varepsilon(t+s)) - \overline{f}(x^\varepsilon(t),\alpha^\varepsilon(t)) - \int_t^{t+s} \mathcal{L}^\varepsilon \overline{f}(x^\varepsilon(u),\alpha^\varepsilon(u))du\Bigg] \\ &= E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i),\overline{\alpha}^\varepsilon(t_i))\Bigg[f(x^\varepsilon(t+s),\overline{\alpha}^\varepsilon(t+s)) - f(x^\varepsilon(t),\overline{\alpha}^\varepsilon(t)) - \int_t^{t+s} \mathcal{L}^\varepsilon f(x^\varepsilon(u),\overline{\alpha}^\varepsilon(u))du\Bigg] = 0. \end{aligned} \tag{3.8}$$

By virtue of Lemma 3.3, $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ is tight. Using Prohorov's theorem, we may extract weakly convergent subsequences. Select such a sequence and still index it by ε for notational simplicity. By the Skorohod representation, $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ converges to $(\overline{x}(\cdot),\overline{\alpha}(\cdot))$ almost surely (a.s). It follows that

$$\begin{aligned} &E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i),\overline{\alpha}^\varepsilon(t_i))[f(x^\varepsilon(t+s),\overline{\alpha}^\varepsilon(t+s)) - f(x^\varepsilon(t),\overline{\alpha}^\varepsilon(t))] \\ &\to E\prod_{i=1}^{\kappa} h_i(\overline{x}(t_i),\overline{\alpha}(t_i))[f(\overline{x}(t+s),\overline{\alpha}(t+s)) - f(\overline{x}(t),\overline{\alpha}(t))]. \end{aligned} \tag{3.9}$$

Detailed calculation shows that

$$
\begin{aligned}
&\int_t^{t+s} \mathcal{L}^\varepsilon f(x^\varepsilon(u), \overline{\alpha}^\varepsilon(u))du \\
&\quad = \sum_{k=1}^{l}\sum_{j=1}^{m_k} \int_t^{t+s} \overline{f}'_x(x^\varepsilon(u), s_{kj})A(s_{kj})x^\varepsilon(u) I_{\{\alpha^\varepsilon(u)=s_{kj}\}}du \\
&\quad + \sum_{k=1}^{l}\sum_{j=1}^{m_k} \int_t^{t+s} \mathrm{tr}[\overline{f}_{xx}(x^\varepsilon(u), s_{kj})\sigma(s_{kj})\sigma'(s_{kj})] I_{\{\alpha^\varepsilon(u)=s_{kj}\}}du \\
&\quad + \sum_{k=1}^{l}\sum_{j=1}^{m_k} \int_t^{t+s} Q^\varepsilon \overline{f}(x^\varepsilon(u), \cdot)(s_{kj}) I_{\{\alpha^\varepsilon(u)=s_{kj}\}}du.
\end{aligned} \tag{3.10}
$$

Let us examine each of the above terms in (3.10). First,

$$
\begin{aligned}
&\int_t^{t+s} \overline{f}'_x(x^\varepsilon(u), s_{kj})A(s_{kj})x^\varepsilon(u) I_{\{\alpha^\varepsilon(u)=s_{kj}\}}du \\
&\quad = \int_t^{t+s} \overline{f}'_x(x^\varepsilon(u), s_{kj})A(s_{kj})x^\varepsilon(u)\nu_j^k I_{\{\alpha^\varepsilon(u)\in\mathcal{M}_k\}}du \\
&\qquad + \int_t^{t+s} \overline{f}'_x(x^\varepsilon(u), s_{kj})A(s_{kj})x^\varepsilon(u)[I_{\{\alpha^\varepsilon(u)=s_{kj}\}} - \nu_j^k I_{\{\alpha^\varepsilon(u)\in\overline{\mathcal{M}}\}}]du.
\end{aligned}
$$

Applying (2.3) to the last term above and recalling that $f(\cdot,\alpha)$ is a C^2-function with compact support,

$$
\begin{aligned}
&E\left|\int_t^{t+s} \overline{f}'_x(x^\varepsilon(u), s_{kj})A(s_{kj})x^\varepsilon(u)[I_{\{\alpha^\varepsilon(u)=s_{kj}\}} - \nu_j^k I_{\{\alpha^\varepsilon(u)\in\overline{\mathcal{M}}\}}]du\right|^2 \\
&\quad = O(\varepsilon) \to 0 \ \text{ as } \varepsilon \to 0.
\end{aligned}
$$

By the weak convergence of $(x^\varepsilon(\cdot), \overline{\alpha}^\varepsilon(\cdot))$, the Skorohod representation, and the continuity of $x^\varepsilon(\cdot)$ and $\overline{x}(\cdot)$, as $\varepsilon \to 0$, $I_{\{\alpha^\varepsilon(u)\in\mathcal{M}_k\}} = I_{\{\overline{\alpha}^\varepsilon(u)=k\}} \to I_{\{\overline{\alpha}(u)=k\}}$, and

$$
\begin{aligned}
&E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i), \overline{\alpha}^\varepsilon(t_i)) \sum_{k=1}^{l}\sum_{j=1}^{m_k} \int_t^{t+s} \overline{f}'_x(x^\varepsilon(u), s_{kj})A(s_{kj})x^\varepsilon(u)\nu_j^k I_{\{\alpha^\varepsilon(u)\in\mathcal{M}_k\}}du \\
&\quad \to E\prod_{i=1}^{\kappa} h_i(\overline{x}(t_i), \overline{\alpha}(t_i)) \sum_{k=1}^{l}\sum_{j=1}^{m_k} \int_t^{t+s} f'_x(\overline{x}(u), s_{kj})A(s_{kj})\overline{x}(u)\nu_j^k I_{\{\overline{\alpha}(u)=k\}}du \\
&\quad = E\prod_{i=1}^{\kappa} h_i(\overline{x}(t_i), \overline{\alpha}(t_i)) \int_t^{t+s} f'_x(\overline{x}(u), \overline{\alpha}(u))du.
\end{aligned} \tag{3.11}
$$

Likewise,

$$
\begin{aligned}
&E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i), \overline{\alpha}^\varepsilon(t_i)) \sum_{k=1}^{l}\sum_{j=1}^{m_k} \mathrm{tr}[\overline{f}_{xx}(x^\varepsilon(u), s_{kj})\sigma(s_{kj})\sigma'(s_{kj})] I_{\{\alpha^\varepsilon(u)=s_{kj}\}}du \\
&\quad \to E\prod_{i=1}^{\kappa} h_i(\overline{x}(t_i), \overline{\alpha}(t_i)) \int_t^{t+s} \mathrm{tr}[f_{xx}(\overline{x}(u), \overline{\alpha}(u))\overline{\Sigma}(\overline{\alpha}(u))]du,
\end{aligned} \tag{3.12}
$$

where $\overline{\Sigma}(k)$ is defined in (2.6).

Next note that the definition of $\overline{f}(x,\alpha)$ in (3.6) implies that $Q^\varepsilon \overline{f}(x,\cdot)(\alpha) = \widehat{Q} f(x,\cdot)(\alpha)$. Using the same averaging argument above with the help of (2.3), we also have

$$\begin{aligned} &E\prod_{i=1}^{\kappa} h_i(x^\varepsilon(t_i),\overline{\alpha}^\varepsilon(t_i))\int_t^{t+s} Q^\varepsilon \overline{f}(x^\varepsilon(u),s_{kj})I_{\{\alpha^\varepsilon(u)=s_{kj}\}}du \\ &\to E\prod_{i=1}^{\kappa} h_i(\overline{x}(t_i),\overline{\alpha}(t_i))\int_t^{t+s} \overline{Q} f(\overline{x}(u),\overline{\alpha}(u))du. \end{aligned} \tag{3.13}$$

Combining (3.9), (3.11), (3.12), and (3.13) leads to the convergence of $(x^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot))$ to $(\overline{x}(\cdot),\overline{\alpha}(\cdot))$ with the desired operator $\mathcal{L}$, which is unique by virtue of Lemma 3.4. Using exactly the same arguments, we also obtain that $(y^\varepsilon(\cdot),\overline{\alpha}^\varepsilon(\cdot)) \Rightarrow (\overline{y}(\cdot),\overline{\alpha}(\cdot))$. The theorem is thus proved. □

COROLLARY 3.6. *Under the conditions of Theorem* 3.5 (*with the use of the Skorohod representation and without changing notation*), *for any* $t\in[0,T]$, $\widetilde{\alpha}^\varepsilon(t)\to\widetilde{\alpha}(t)$, $R^\varepsilon(t)\to R(t)$, *and* $\widetilde{x}^\varepsilon(t)\to\widetilde{x}(t)$ *a.s.*

Proof. In view of the weak convergence $y^\varepsilon(\cdot)\to y(\cdot)$, the Skorohod representation, and the definitions of $\zeta^\varepsilon(j)$ and $\zeta(j)$, $\zeta^\varepsilon(j)\to\zeta(j)$ a.s. as $\varepsilon\to 0$. Recall assumption (A3), which implies that $\mu_{k_1}\neq\mu_{k_2}$ for $k_1\neq k_2$. Owing to the definition of $\Lambda^{\varepsilon,n_0,n}$ in (3.1) and $\widetilde{\alpha}^\varepsilon(\cdot)$ in (3.3), $\Lambda^{\varepsilon,n_0,n}\to\Lambda^{n_0,n}$ a.s. Let

$$G_n=\left\{\left|\frac{\Lambda^{\varepsilon,n_0,n}}{\mu_{k_0}^{n_0,n}}-1\right|=\min\left\{\left|\frac{\Lambda^{\varepsilon,n_0,n}}{\mu_{k}^{n_0,n}}-1\right|,\ k=1,\dots,l\right\}\right\}.$$

It then follows that as $\varepsilon\to 0$,

$$\alpha^{\varepsilon,n_0,n}=\sum_{k_0=1}^{l} k_0 I_{G_n}\to\sum_{k_0=1}^{l} k_0 I_{G_n}=\overline{\alpha}^{n_0,n}\ \text{ a.s.}$$

As a result, $\widetilde{\alpha}^\varepsilon(t)\to\widetilde{\alpha}(t)$ a.s. for each $t\in[0,T]$. The convergence, in fact, is uniform in $[0,T]$.

Since $\overline{A}(k)$, $\overline{\sigma}(k)$, and $\overline{H}(k)$ take values in the finite set $\overline{\mathcal{M}}=\{1,\dots,l\}$,

$$\overline{A}(\widetilde{\alpha}^\varepsilon(t))\to\overline{A}(\widetilde{\alpha}(t)),\ \overline{\sigma}(\widetilde{\alpha}^\varepsilon(t))\to\overline{\sigma}(\widetilde{\alpha}(t)),\ \text{ and }\ \overline{H}(\widetilde{\alpha}^\varepsilon(t))\to\overline{H}(\widetilde{\alpha}(t)).$$

Since $R^\varepsilon(t)$ is uniformly bounded by Lemma 3.1, so is $\dot{R}^\varepsilon(t)$ on $[0,T]$. We may extract a convergent subsequence and still index it by ε for notational simplicity. Writing $R^\varepsilon(t)$ in its integral form leads to

$$\begin{aligned} R^\varepsilon(t) =\ & \mathrm{Cov}(x_0)+\int_0^t \overline{A}(\widetilde{\alpha}^\varepsilon(u))R^\varepsilon(u)du+\int_0^t R^\varepsilon(u)\overline{A}'(\widetilde{\alpha}^\varepsilon(u))du \\ &+\int_0^t \overline{\sigma}(\widetilde{\alpha}^\varepsilon(u))\overline{\sigma}'(\widetilde{\alpha}^\varepsilon(u))du-\frac{1}{\delta^2}\int_0^t R^\varepsilon(u)\overline{H}'(\widetilde{\alpha}^\varepsilon(u))R^\varepsilon(u))du \\ \to\ & \mathrm{Cov}(x_0)+\int_0^t \overline{A}(\widetilde{\alpha}(u))\widetilde{R}(u)du+\int_0^t \widetilde{R}(u)\overline{A}'(\widetilde{\alpha}(u))du \\ &+\int_0^t \overline{\sigma}(\widetilde{\alpha}(u))\overline{\sigma}'(\widetilde{\alpha}(u))du-\frac{1}{\delta^2}\int_0^t \widetilde{R}(u)\overline{H}'(\widetilde{\alpha}(u))\widetilde{R}(u))du \\ =\ & \widetilde{R}(t). \end{aligned}$$

The convergence of $R^\varepsilon(t) \to \widetilde{R}(t)$ is established. Furthermore, the limit does not depend on the chosen subsequence. Likewise, we can show that $\widetilde{x}^\varepsilon(\cdot)$ is tight as in Lemma 3.3. By a similar approach, we obtain $\widetilde{x}^\varepsilon(t) \to \widetilde{x}(t)$ a.s. □

THEOREM 3.7. *For each fixed $\delta > 0$, as $\varepsilon \to 0$,*

$$|x^\varepsilon(\cdot) - \widetilde{x}^\varepsilon(\cdot)|_T \to |\overline{x}(\cdot) - \widetilde{x}(\cdot)|_T = O(\sqrt{\delta}).$$

This, in particular, implies that

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0} |x^\varepsilon(\cdot) - \widetilde{x}^\varepsilon(\cdot)|_T = 0.$$

Proof. First of all, in view of Corollary 3.6 and Lemma 3.2, as $\varepsilon \to 0$,

$$|\widetilde{x}^\varepsilon(\cdot) - \widetilde{x}(\cdot)|_T \to 0.$$

Moreover, in view of Lemma 3.2 and the weak convergence of $x^\varepsilon(\cdot)$ to $\overline{x}(\cdot)$, we have that as $\varepsilon \to 0$,

$$|x^\varepsilon(\cdot) - \overline{x}(\cdot)|_T \to 0.$$

Therefore, it follows that

$$\Big||x^\varepsilon(\cdot) - \widetilde{x}^\varepsilon(\cdot)|_T - |\overline{x}(\cdot) - \widetilde{x}(\cdot)|_T\Big| \le |x^\varepsilon(\cdot) - \overline{x}(\cdot)|_T + |\widetilde{x}^\varepsilon(\cdot) - \widetilde{x}(\cdot)|_T \to 0.$$

Next, we show the error bound of $\widetilde{x}(\cdot) - \overline{x}(\cdot)$ in terms of $\sqrt{\delta}$. It can be shown as in Haussmann and Zhang [10] that

$$|\widehat{x}(\cdot) - \overline{x}(\cdot)|_T = O(\sqrt{\delta}).$$

It suffices to show

$$|\widetilde{x}(\cdot) - \widehat{x}(\cdot)|_T = O(\sqrt{\delta}).$$

Given $\theta = \theta(\cdot) \in D([0,T] : \overline{\mathcal{M}})$ and a, b being the multiples of δ such that $0 \le a < b$, let

$$F^\theta_{a,b} = \Big\{\alpha(\cdot) = \theta(\cdot) \text{ on } [a,b)\Big\}.$$

That is, $F^\theta_{a,b} = \{\alpha(t) = \theta(t) \text{ for all } t \in [a,b)\}$. Let $E^\theta_{a,b}$ denote the conditional expectation given $F^\theta_{a,b}$. We show that for given $0 < \eta < 1$,

$$E^\theta_{a,b}|\widetilde{x}(t) - \overline{x}(t)| = O\left(\delta^{1-\eta} + \exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right)$$

uniformly for $t \in [a + l_0\delta, b]$, here $l_0 = c_0(\log\delta)^2 + 1$.

Consider the differentials of $\overline{H}(k_0)(\widetilde{x}(t) - \widehat{x}(t))$ and $|\overline{H}(k_0)(\widetilde{x}(t) - \widehat{x}(t))|^2$ on the interval when $\overline{\alpha}(t) = k_0$, we obtain

$$\begin{aligned} d(\overline{H}(k_0)(\widetilde{x}(t) - \widehat{x}(t))) \ = \overline{H}(k_0)\Big\{ & (\overline{A}(\widetilde{\alpha}(t))\widetilde{x}(t) - \overline{A}(k_0)\widehat{x}(t))dt \\ & - \frac{1}{\delta^2}\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))(\overline{H}(\widetilde{\alpha}(t))\widetilde{x}(t) - \overline{H}(k_0)\widehat{x}(t))dt \\ & + \frac{1}{\delta^2}\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t)) - \overline{R}(t)\overline{H}'(k_0)\right)\overline{\eta}(t)dt \\ & + \frac{1}{\delta}\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t)) - \overline{R}(t)\overline{H}'(k_0)\right)d\widehat{v}(t)\Big\} \end{aligned}$$

and

$$\begin{aligned} d|\overline{H}(k_0)(\widetilde{x}(t)-\widehat{x}(t))|^2 &= 2(\widetilde{x}(t)-\widehat{x}(t))'\overline{H}'(k_0)d(\overline{H}(k_0)(\widetilde{x}(t)-\widehat{x}(t))) \\ &\quad + \frac{1}{\delta^2}\mathrm{tr}\left(\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))-\overline{R}(t)\overline{H}'(k_0)\right)\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))-\overline{R}(t)\overline{H}'(k_0)\right)'\right) \\ &\quad = A(t)dt + B(t)d\widehat{v}(t), \end{aligned}$$

where $\widehat{v}(t)$ is an innovation process given by $d\widehat{v}(t) = dy(t) - E(\overline{H}(\overline{\alpha}(t))x(t)|\mathcal{Y}_t)dt$, $\overline{\eta}(t) = E(\overline{H}(\overline{\alpha}(t))x(t)|\mathcal{Y}_t) - \overline{H}(\overline{\alpha}(t))\widehat{x}(t)$,

$$\begin{aligned} A(t) = 2(\widetilde{x}(t)-\widehat{x}(t))'\overline{H}'(k_0)\overline{H}(k_0)\Big\{ & (\overline{A}(\widetilde{\alpha}(t))\widetilde{x}(t)-\overline{A}(k_0)\widehat{x}(t)) \\ & - \frac{1}{\delta^2}\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))(H(t,\widetilde{\alpha}(t))\widetilde{x}(t)-\overline{H}(k_0)\widehat{x}(t)) \\ & + \frac{1}{\delta^2}\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))-\overline{R}(t)\overline{H}'(k_0)\right)\overline{\eta}(t)\Big\} \\ & + \frac{1}{\delta^2}\mathrm{tr}\left(\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))-\overline{R}(t)\overline{H}'(k_0)\right)\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))-\overline{R}(t)\overline{H}'(k_0)\right)'\right) \end{aligned}$$

and

$$B(t) = \frac{2}{\delta}(\widetilde{x}(t)-\widehat{x}(t))'\overline{H}'(k_0)\left(\widetilde{R}(t)\overline{H}'(\widetilde{\alpha}(t))-\overline{R}(t)\overline{H}'(k_0)\right).$$

Let $\phi(t) = |\overline{H}(k_0)(\widetilde{x}(t)-\widehat{x}(t))|^2$. Then,

$$\phi(t) = \phi(a+l_0\delta) + \int_{a+l_0\delta}^{t} A(s)ds + \int_{a+l_0\delta}^{t} B(s)d\widehat{v}(s).$$

Write

$$A(t) = A(t)I_{\{\widetilde{\alpha}(\cdot)=k_0 \text{ on } [a+l_0\delta,b]\}} + A(t)I_{\{\widetilde{\alpha}(\cdot)=k_0 \text{ on } [a+l_0\delta,b]\}^c}.$$

As in Lemma 7.1 in the appendix, we can show that $(E^\theta_{a,b}|A(t)|^n)^{1/n} = O(1/\delta)$, for $n = 1, 2, \ldots$ In view of Lemma 7.2, it follows that

$$\begin{aligned} & E^\theta_{a,b}\left|A(t)I_{\{\widetilde{\alpha}(\cdot)=k_0 \text{ on } [a+l_0\delta,b]\}^c}\right| \\ & \quad \le (E^\theta_{a,b}|A(t)|^n)^{\frac{1}{n}}(E^\theta_{a,b}I_{\{\widetilde{\alpha}(\cdot)=k_0 \text{ on } [a+l_0\delta,b]\}^c})^{\frac{n-1}{n}} \\ & \quad \le \frac{K}{\delta}(P(\widetilde{\alpha}(t)\neq k_0 \text{ for some } t\in[a+l_0\delta,b]|F^\theta_{a,b}))^{\frac{n-1}{n}} \\ & \quad \le K\delta^{(2(n-1)/n)-1} \le K\delta^{1-\eta} \text{ for } n \text{ large enough.} \end{aligned} \tag{3.14}$$

Given $\{\widetilde{\alpha}(\cdot) = k_0 \text{ on } [a+l_0\delta, b]\}$, we have

$$\begin{aligned} A(t) = 2(\widetilde{x}(t)-\widehat{x}(t))'\overline{H}'(k_0)\overline{H}(k_0)\Big\{ & (\overline{A}(k_0)\widetilde{x}(t)-\overline{A}(k_0)\widehat{x}(t)) \\ & - \frac{1}{\delta^2}\widetilde{R}(t)\overline{H}'(k_0)(\overline{H}(k_0)(\widetilde{x}(t)-\widehat{x}(t)) \\ & + \frac{1}{\delta^2}\left(\widetilde{R}(t)-\overline{R}(t))\overline{H}'(k_0)\right)\overline{\eta}(t)\Big\} \\ & + \frac{1}{\delta^2}\mathrm{tr}\left((\widetilde{R}(t)-\overline{R}(t))\overline{H}'(k_0)\overline{H}(k_0)(\widetilde{R}(t)-\overline{R}(t))\right). \end{aligned}$$

Recall that $\overline{H}(k_0)$ is invertible and both $\widetilde{R}(t)/\delta$ and $\overline{R}(t)/\delta$ are uniformly bounded. Moreover, in view of Lemma 7.3 (appendix), we have

$$\frac{\widetilde{R}(t)-\overline{R}(t)}{\delta}=O\left(\delta+\exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right).$$

Given $\{\widetilde{\alpha}(\cdot)=k_0 \text{ on } [a+l_0\delta,b]\}$, it follows that

$$A(t)\le K_0\phi(t)-\frac{\kappa_0}{\delta}\phi(t)+K\left(\delta+\frac{|\overline{\eta}(t)|^2}{\delta^2}+\exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right).$$

Hence,

$$\begin{aligned}E^\theta_{a,b}A(t)\ &\le\left(K_0-\frac{\kappa}{\delta}\right)E^\theta_{a,b}\phi(t)\\ &\quad+K\left(\delta^{1-\eta}+\frac{E^\theta_{a,b}|\overline{\eta}(t)|^2}{\delta^2}+\exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right).\end{aligned}\tag{3.15}$$

Next, we claim that

$$E^\theta_{a,b}\int_{a+l_0\delta}^t B(s)d\widehat{v}(s)=0.$$

In fact, if we let $\mathcal{F}^\alpha_{a,b}=\sigma\{\alpha(r):\ a\le r<b\}$ and $\zeta(t)=\int_{a+l_0\delta}^t B(s)d\widehat{v}(s)$, then for $0\le t<b$, $E[\zeta(t)|\mathcal{F}^\alpha_{0,b}]=0$ because $\widehat{v}(\cdot)$ is a Brownian motion given $\alpha(\cdot)$. Note that $E[\zeta(t)|\mathcal{F}^\alpha_{a,b}]=E[E[\zeta(t)|\mathcal{F}^\alpha_{0,b}]|\mathcal{F}^\alpha_{a,b}]$. It follows that $E[\zeta(t)|\mathcal{F}^\alpha_{a,b}]=0$, a.s. Thus,

$$E^\theta_{a,b}\int_{a+l_0\delta}^t B(s)d\widehat{v}(s)=0,\ \text{a.s. with respect to } \widehat{P}(d\theta)=P(\alpha(\cdot)\in d\theta).$$

Using this claim, we have

$$E^\theta_{a,b}\phi(t)=E^\theta_{a,b}\phi(a+l_0\delta)+E^\theta_{a,b}\int_{a+l_0\delta}^t A(s)ds.$$

Thus, in view of (3.15),

$$\begin{aligned}\frac{dE^\theta_{a,b}\phi(t)}{dt}&=E^\theta_{a,b}A(t)\le\left(K_0-\frac{\kappa}{\delta}\right)E^\theta_{a,b}\phi(t)\\ &\quad+K\left(\delta^{1-\eta}+\frac{E^\theta_{a,b}|\overline{\eta}(t)|^2}{\delta^2}+\exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right).\end{aligned}$$

Using Gronwall's inequality, the uniform boundedness of $E^\theta_{a,b}\phi(a+l_0\delta)$, and integration by parts, we obtain

$$E^\theta_{a,b}\phi(t)\le K\left(\delta^{2-\eta}+\exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right)$$

for some $\kappa>0$. Therefore,

$$(E^\theta_{a,b}\phi(t))^{\frac{1}{2}}\le K\left(\delta^{1-\eta/2}+\exp\left(-\frac{\kappa(t-a-l_0\delta)}{\delta}\right)\right).$$

Given a fixed number $0 < \eta < 1$, let

$$\Theta_\eta^\varepsilon = \Big\{\theta(\cdot) \in D([0,T] : \overline{\mathcal{M}}) : \quad \text{the number of jumps of } \theta(\cdot) \le [1/\varepsilon^\eta]\Big\}.$$

Let $\gamma_0 > 0$ be a constant and define

$$\Theta_{\eta,\gamma_0}^\varepsilon = \Big\{\theta = \theta(\cdot) \in \Theta_\eta^\varepsilon : \quad \text{the duration between any two jumps of } \theta(\cdot) \ge \gamma_0\varepsilon\Big\}.$$

If we let $0 = t_0 < t_1 < \cdots < t_n < T$ denote the jump times of $\theta(\cdot)$, then $\theta(\cdot) \in \Theta_{\eta,\gamma_0}^\varepsilon$ implies that $n \le [1/\varepsilon^\eta]$ and $t_{j+1} - t_j \ge \gamma_0\varepsilon$.

Next, we show that

$$\int_0^T E|\widetilde{x}(t) - \widehat{x}(t)| I_{\{\alpha(\cdot)\in\Theta_{\eta,\gamma_0}^\varepsilon\}} dt = O(\delta^{1-\eta}). \tag{3.16}$$

Let $\xi(t) = |\widetilde{x}(t) - \widehat{x}(t)| I_{\{\alpha(\cdot)\in\Theta_{\eta,\gamma_0}^\varepsilon\}}$. Note that

$$E\int_0^T \xi(t)dt \le \sum_{j=0}^{[1/\delta^\eta]} E\int_0^T \xi(t) I_{[\tau_j,\tau_{j+1})}.$$

It suffices to show

$$E\int_0^T \xi(t) I_{[\tau_j,\tau_{j+1})} \le O(\delta^{1-\eta})$$

because

$$E\int_0^T \xi(t)dt \le \sum_{j=0}^{[1/\delta^\eta]} K\delta^{1-\eta} \le O(\delta^{1-2\eta}) = O(\delta^{1-\eta})$$

when $2\eta < \delta$.

For $j = 0$,

$$E\int_0^T \xi(t) I_{[0,\tau_1)} dt = \int_0^\infty E\left[\int_0^{s\wedge T} \xi(t)dt \Big| \tau_1 = s\right] p_1(s)ds,$$

where $p_1(s)$ is the density function of τ_1.

It can be seen that

$$E\left[\int_0^{l_0\delta} \xi(t)dt \Big| \tau_1 = s\right] = O(l_0\delta) = O(\varepsilon(\log\varepsilon)^2).$$

Moreover, we have

$$E\left[\int_{l_0\delta}^{s\wedge T} \xi(t)dt \Big| \tau_1 = s\right] = O\left(\delta^{1-\eta} + \exp\left(-\frac{\kappa(s - l_0\delta)}{\delta}\right)\right).$$

Thus,

$$E\int_0^T \xi(t) I_{[0,\tau_1)} dt = O(\delta^{1-\eta}).$$

For $j \geq 1$, we have

$$E\int_0^T \xi(t)I_{[\tau_j,\tau_{j+1})}dt = \int_0^\infty E\left[\int_r^T \xi(t)dt\Big|\tau_j = r\right]p_j(r)dr$$
$$= \int_0^\infty \int_r^T E\left[\xi(t)dtI_{[r,\tau_{j+1})}\Big|\tau_j = r\right]p_j(r)dr.$$

Similarly as in the case for $j = 0$, we can show

$$\int_r^T E\left[\xi(t)dtI_{[r,\tau_{j+1})}\Big|\tau_j = r\right] = O(\delta^{1-\eta}).$$

It follows that

$$E\int_0^T \xi(t)I_{[\tau_j,\tau_{j+1})}dt = O(\delta^{1-\eta}).$$

Taking $\eta = 1/2$ completes the proof. □

4. A numerical example. In this section we present a simple example to demonstrate the use of the near optimal filter. We consider the following one-dimensional model:

$$\text{(4.1)}\qquad \begin{cases} dx^\varepsilon(t) = A(\alpha^\varepsilon(t))x^\varepsilon(t)dt + \sigma(\alpha^\varepsilon(t))dw(t), & x(0) = x_0, \\ dy^\varepsilon(t) = H(\alpha^\varepsilon(t))x^\varepsilon(t)dt + \delta dv(t), & y(0) = 0, \end{cases}$$

where $\alpha^\varepsilon(t) \in \mathcal{M} = \{1, 2\}$, $t \geq 0$, is a Markov chain generated by

$$Q^\varepsilon = \frac{1}{\varepsilon}Q, \quad \text{with } Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}.$$

Computational experiments and numerical results are reported next.

In this example, it is readily seen that the stationary distribution is

$$\nu = (\nu_1, \nu_2) = \left(\frac{\mu}{\lambda+\mu}, \frac{\lambda}{\lambda+\mu}\right),$$

and

$$\overline{A} = \nu_1 A(1) + \nu_2 A(2),$$
$$\overline{H} = \nu_1 H(1) + \nu_2 H(2),$$
$$\overline{\Sigma} = \overline{\sigma}^2 = \nu_1\sigma^2(1) + \nu_2\sigma^2(2).$$

We use time discretization step size 0.0001. The time horizons in the continuous-time model is $T = 10$. All our results are based on computations with 100 sample paths.

We consider the model with the following specifications:

$$A(1) = -0.01, A(2) = -0.02, H(1) = 1, H(2) = 2, \sigma(1) = \sigma(2) = 1,$$

$$\lambda = \mu = 0.02.$$

TABLE 4.1
Convergence of J^ε to J^0.

ε	0.1	0.05	0.01	0.005	0.001	0.0001
$\lvert J^\varepsilon - J^0\rvert$	6.517	5.295	3.769	3.235	1.745	0.305

TABLE 4.2
Error bound of J^0.

δ	1	0.5	0.1	0.05	0.01	0.005	0.001	0.005
$J^0/\sqrt{\delta}$	8.046	8.071	8.111	8.126	8.170	8.202	8.475	8.450

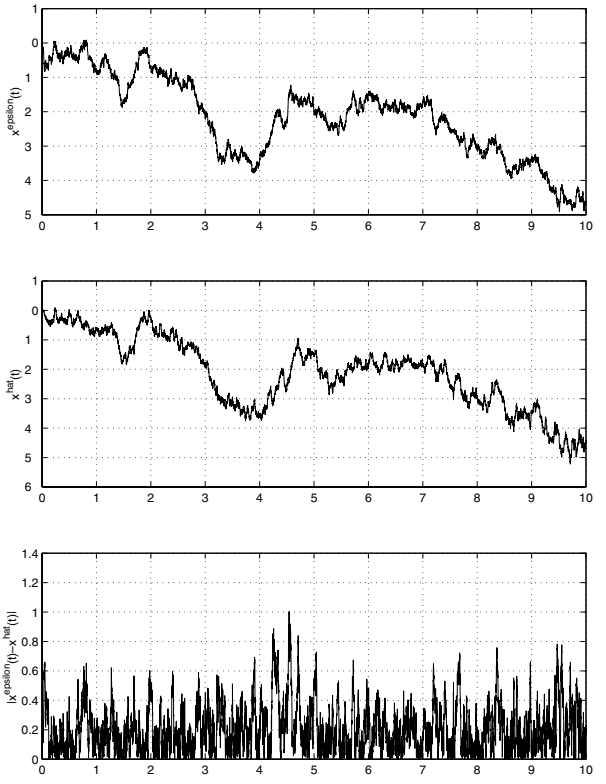


FIG. 4.1. *Sample paths of $x^\varepsilon(\cdot)$, $\widetilde{x}^\varepsilon(\cdot)$ and their difference.*

To demonstrate the convergence of

$$J^\varepsilon = \int_0^T (E|x^\varepsilon(t) - \widetilde{x}^\varepsilon(t)|^2)^{\frac{1}{2}} \quad \text{to} \quad J^0 = \int_0^T (E|\overline{x}(t) - \widetilde{x}(t)|^2)^{\frac{1}{2}},$$

we fix $\delta = 0.1$ and vary ε. In this case, the values of $|J^\varepsilon - J^0|$ are given in Table 4.1. The convergence of $|J^\varepsilon - J^0| \to 0$ can be seen from Table 4.1.

Then to demonstrate the error bound of $J^0 = O(\sqrt{\delta})$, we take $\varepsilon = 0.05$ and vary δ. The results are summarized in Table 4.2. It follows that $|J^0| \le 8.475\sqrt{\delta}$ at these points. Sample paths of $x^\varepsilon(\cdot), -\widetilde{x}^\varepsilon(\cdot)$, and $|x^\varepsilon(\cdot) - \widetilde{x}^\varepsilon(\cdot)|$ are plotted in Figure 4.1.

This example focuses exclusively on approximation of $\widetilde{x}^\varepsilon(\cdot)$ to $x^\varepsilon(\cdot)$ without involving the QVT. There are extensive numerical experiments on QVT in the literature; see, for example, Zhang [20] and references therein.

5. Extensions to Markov chains with transient states. Recall that a finite state Markov chain has at least one recurrent state and either all states are recurrent or there is a class of transient states. In this section, we consider extensions of results obtained in the previous sections to such chains that in addition to recurrent states, there is attached to them a number of transient states. The main idea is still aggregation of states, but we only lump the states in each recurrent class and leave the transient states alone. In fact, these transient states are asymptotically unimportant. Let $\alpha^\varepsilon(\cdot)$ be a finite-state Markov chain with its generator having the following decomposition form:

$$Q^\varepsilon = \frac{1}{\varepsilon}\begin{pmatrix} \widetilde{Q}^1 & & & \\ & \ddots & & \\ & & \widetilde{Q}^l & \\ \widetilde{Q}^1_* & \cdots & \widetilde{Q}^l_* & \widetilde{Q}_* \end{pmatrix} + \widehat{Q}, \tag{5.1}$$

where for each $k = 1, \ldots, l$, $\widetilde{Q}^k$ is a generator with dimension $m_k \times m_k$, $\widetilde{Q}_* \in \mathbb{R}^{m_* \times m_*}$, $\widetilde{Q}^k_* \in \mathbb{R}^{m_* \times m_k}$, and $m_1 + \cdots + m_l + m_* = m$. The state space of $\alpha^\varepsilon(\cdot)$ is given by

$$\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_l \cup \mathcal{M}_* = \{s_{11}, \ldots, s_{1m_1}\} \cup \cdots \cup \{s_{l1}, \ldots, s_{lm_l}\} \cup \{s_{*1}, \ldots, s_{*m_*}\},$$

where $\mathcal{M}_k = \{s_{k1}, \ldots, s_{km_k}\}$, $k = 1, \ldots, l$, are the subspaces corresponding to the recurrent states and $\mathcal{M}_* = \{s_{*1}, \ldots, s_{*m_*}\}$ consists of the transient states.

Suppose that, for $k = 1, \ldots, l$, $\widetilde{Q}^k$ are irreducible and $\widetilde{Q}_*$ is Hurwitz (i.e., all its eigenvalues have negative real parts). Let $a_{m_i} = -\widetilde{Q}_*^{-1}\widetilde{Q}^i_* 1_{m_i}$ for $i = 1, \ldots, l$. The jth component $a_{m_i,j}$ is the probability of $\alpha^\varepsilon(t)$ jumping from s_{*j} to $\mathcal{M}_i$ for small ε. Now define the aggregated process $\overline{\alpha}^\varepsilon(\cdot)$ by

$$\overline{\alpha}^\varepsilon(t) = \begin{cases} i & \text{if } \alpha^\varepsilon(t) \in \mathcal{M}_i, \\ \xi_j & \text{if } \alpha^\varepsilon(t) = s_{*j}, \end{cases}$$

where, for $j = 1, \ldots, m_*$,

$$\xi_j = I_{\{0 \le \xi \le a_{m_1,j}\}} + 2I_{\{a_{m_1,j} < \xi \le a_{m_1,j} + a_{m_2,j}\}} + \cdots + lI_{\{a_{m_1,j} + \cdots + a_{m_{l-1},j} < \xi \le 1\}},$$

and ξ is a random variable with uniform distribution on $[0, 1]$. Let

$$\overline{Q} = \operatorname{diag}(\nu^1, \ldots, \nu^l, 0_{m_* \times m_*})\widehat{Q}\,\widetilde{1}_* = \begin{pmatrix} \overline{Q}_* & \\ & 0_{m_* \times m_*} \end{pmatrix}, \tag{5.2}$$

where

$$\widetilde{1}_* = \begin{pmatrix} 1_{m_1} & & & \\ & \ddots & & \\ & & 1_{m_l} & \\ a_{m_1} & \cdots & a_{m_l} & 0_{m_* \times m_*} \end{pmatrix},$$

and 1_{m_k} is as defined before.

Then it can be shown as in Yin, Zhang, and Badowski [19] that $\overline{\alpha}^\varepsilon(\cdot)$ converges weakly to $\overline{\alpha}(\cdot)$ that is a continuous-time Markov chain with state space $\overline{\mathcal{M}}$ and generator $\overline{Q}_*$ defined in (5.2). Moreover, for any deterministic test functions $\beta_{kj}(t)$, the corresponding occupation measures satisfy

$$\begin{aligned} &E\left(\int_0^T \left(I_{\{\alpha^\varepsilon(t)=s_{kj}\}} - \nu_j^k I_{\{\overline{\alpha}^\varepsilon(t)=k\}}\right)\beta_{kj}(t)dt\right)^2 = O(\varepsilon), \\ &E\left(\int_0^T I_{\{\alpha^\varepsilon(t)=s_{*j}\}}\beta_{*j}(t)dt\right)^2 = O(\varepsilon^2). \end{aligned} \tag{5.3}$$

Recall that $\alpha^\varepsilon(\cdot)$ is a Markov chain generated by Q^ε defined in (5.1) and $\overline{\alpha}(\cdot)$ is a Markov chain generated by $\overline{Q}_*$. Let $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ be the actual state and observation satisfying

$$\begin{cases} dx^\varepsilon(t) = A(\alpha^\varepsilon(t))x^\varepsilon(t)dt + \sigma(\alpha^\varepsilon(t))dw(t), & x^\varepsilon(0) = x_0, \\ dy^\varepsilon(t) = H(\alpha^\varepsilon(t))x^\varepsilon(t)dt + \delta dv(t), & y^\varepsilon(0) = 0. \end{cases}$$

The corresponding limit system is given by

$$\begin{cases} d\overline{x}(t) = \overline{A}(\overline{\alpha}(t))\overline{x}(t)dt + \overline{\sigma}(\overline{\alpha}(t))dw(t), & \overline{x}(0) = x_0, \\ d\overline{y}(t) = \overline{H}(\overline{\alpha}(t))\overline{x}(t)dt + \delta dv(t), & \overline{y}(0) = 0, \end{cases}$$

with $\overline{A}$, $\overline{\sigma}$, and $\overline{H}$ defined in (2.6). Let $\widetilde{x}(\cdot)$ and $\widetilde{x}^\varepsilon(\cdot)$ be defined in (2.12) and (3.4), respectively. Then we can prove the following theorem as in the proof of Theorem 3.7.

THEOREM 5.1. *For each fixed $\delta > 0$, we have*

$$|x^\varepsilon(\cdot) - \widetilde{x}^\varepsilon(\cdot)|_T \to |\overline{x}(\cdot) - \widetilde{x}(\cdot)|_T = O(\sqrt{\delta})$$

as $\varepsilon \to 0$, This, in particular, implies that

$$\lim_{\delta\to 0}\lim_{\varepsilon\to 0} |x^\varepsilon(\cdot) - \widetilde{x}^\varepsilon(\cdot)|_T = 0.$$

6. Conclusions. This paper develops a hybrid filter involving a hidden Markov chain with weak and strong interactions. An averaged filter is constructed based on a quadratic variation test. Such a filter is shown to be near optimal as the observation noise goes to zero. The scheme is easily implementable. Its advantage is significant savings of computational effort, which leads to computable filters for many applications arising in emerging applications of financial engineering and communication networks.

7. Appendix. In this appendix, we present several technical lemmas. The proof of Lemma 7.1 is similar to that of [11, Theorem 4.1]. Providing the error probability of the QVT, Lemma 7.2 can be proved similar to that of [11, Lem. 3.1]. The order of magnitude estimate in Lemma 7.3 may be derived as in [20].

LEMMA 7.1. *Let $x(t)$, $\widehat{x}(t)$, and $\widetilde{x}(t)$, $t \geq 0$, denote the processes defined in* (1.1), (2.7), *and* (2.12), *respectively. Then, for each $n = 1, 2, \ldots$, there exists a constant K such that*

$$E\left(|x(t)|^n + |\widehat{x}(t)|^n + |\widetilde{x}(t)|^n\right) \leq K.$$

LEMMA 7.2. *Assume* (A1)–(A5). *Then for each* $j = 1, 2, \ldots$, *there exist* $k_0 > 0$, $K > 0, \varepsilon_0 > 0$, *such that for* $0 < \varepsilon < \varepsilon_0$, $n - n_0 \geq l_0 := [k_0(\log \varepsilon)^2] + 1$,

$$P(\alpha^{n_0,n} \neq k_0 | \alpha(t) = k_0,\ t \in [\varepsilon n_0, \varepsilon n]) \leq K\varepsilon^2. \tag{7.1}$$

It follows that

$$P(\{\alpha^{n_0,n} \neq k_0\} \cap \{\alpha(t) = k_0,\ t \in [\varepsilon n_0, \varepsilon n]\}) \leq K\varepsilon^2. \tag{7.2}$$

LEMMA 7.3. *Assume* (A1)–(A4). *For each* $\theta \in D([0,T] : \overline{\mathcal{M}})$, *let* $R^\theta(t)$ *denote the solution to the Riccati equation* (3.4) *with* $\widetilde{\alpha}(t)$ *replaced by* $\theta(t)$. *Then the following hold:*

(a) *There exist positive constants* c_1 *and* c_2, *independent of* $0 < \eta < 1$, $\theta \in D([0,T] : \overline{\mathcal{M}})$, *and* $t \in [0,T]$ *such that*

$$c_1 I \leq \frac{R^\theta(t)}{\varepsilon} \leq c_2 I.$$

(b) *There exist* γ_0, K, *and* $\kappa > 0$ *such that, for each* $0 < \eta < 1$ *and* $\theta \in \Theta^\varepsilon_{\eta,\gamma_0}$, *we have*

$$\left| \frac{R^\theta(t)}{\varepsilon} - F(\theta(t)) \right| \leq K \left(\exp\left(-\frac{\kappa(t - t_j)}{\varepsilon} \right) + \varepsilon \right) \tag{7.3}$$

for $t \in [t_j, t_{j+1})$, *where* $\{t_j\}$ *is the set of jump times of* θ *and* $F(k)$ *is given in assumption* (A2).

REFERENCES

[1] A. BENSOUSSAN, *Stochastic Control of Partially Observed Systems*, Cambridge University Press, Cambridge, UK, 1992.

[2] H. A. P. BLOM AND Y. BAR-SHALOM, *The interacting multiple model algorithm for systems with Markovian switching coefficients*, IEEE Trans. Automat. Control, 33 (1988), pp. 780–783.

[3] P. E. CAINES AND H. F. CHEN, *Optimal adaptive LQG control for systems with finite state process parameters*, IEEE Trans. Automat. Control, 30 (1985), pp. 185–189.

[4] R. J. ELLIOTT, L. AGGOUN, AND J. B. MOORE, *Hidden Markov Models: Estimation and Control*, Springer-Verlag, New York, 1995.

[5] W. H. FLEMING, D. JI, P. SALAME, AND Q. ZHANG, *Piecewise monotone filtering in discrete time with small observation noise*, IEEE Trans. Automat. Control, 36 (1991), pp. 1181–1186.

[6] W. H. FLEMING AND E. PARDOUX, *Piecewise monotone filtering with small observation noise*, SIAM J. Control Optim., 27 (1989), pp. 1156–1181.

[7] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[8] W. H. FLEMING AND Q. ZHANG, *Nonlinear filtering with small observation noise: Piecewise monotone observations*, in Stochastic Analysis: Liber Amicorum for Moshe Zakai, E. Merzbach, A. Shwartz, and E. Mayer-Wolf, eds., Academic Press, Boston, 1991, pp. 153–168.

[9] W. H. FLEMING AND Q. ZHANG, *Piecewise filtering with small observation noise: Numerical simulations*, in Applied Stochastic Analysis, Lecture Notes in Control and Information Sci. 177, I. Karatzas and D. L. Ocone, eds., Springer-Verlag, New York, 1992, pp. 108–120.

[10] U. G. HAUSSMANN AND Q. ZHANG, *Optimal control of diffusions with small observation noise*, in Proceedings of the Imperial College Workshop on Applied Stochastic Analysis, M. H. Davis and R. J. Elliott, eds., Stochastics Monog., Vol. 5, 1989, pp. 237–263.

[11] U. G. HAUSSMANN AND Q. ZHANG, *Stochastic adaptive control with small observation noise*, Stoch. Stoch. Rep., 32 (1990), pp. 109–144.

[12] O. Hijab, *The adaptive LQG problem Part* I, IEEE, Trans. Automat. Control, 28 (1983), pp. 171–178.
[13] G. Kallianpur, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.
[14] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.
[15] X. R. Li, *Hybrid Estimation Techniques*, in Control and Dynamic Systems 76, C. T. Leondes, ed., Academic Press, New York, 1996.
[16] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes,* Springer-Verlag, New York, 1977.
[17] R. H. Liu and Q. Zhang, *Nonlinear filtering: A hybrid approximation scheme*, IEEE Trans. Aerospace Electronic Systems, 37 (2001), pp. 470–480.
[18] G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications: A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.
[19] G. Yin, Q. Zhang, and G. Badowski, *Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states*, Ann. Appl. Probab., 10 (2000), pp. 549–572.
[20] Q. Zhang, *Nonlinear filtering and control of a switching diffusion with small observation noise*, SIAM J. Control Optim., 36 (1998), pp. 1738–1768.
[21] Q. Zhang, *Optimal filtering of discrete-time hybrid systems*, J. Optim. Theory Appl., 100 (1999), pp. 123–144.
[22] Q. Zhang, *Hybrid filtering for linear systems with non-Gaussian disturbances*, IEEE Trans. Automat. Control, 45 (2000), pp. 50–61.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 320–342

FINITE-TIME GLOBAL STABILIZATION BY MEANS OF TIME-VARYING DISTRIBUTED DELAY FEEDBACK*

IASSON KARAFYLLIS†

Abstract. The paper contains certain results concerning the finite-time global stabilization for triangular control systems described by retarded functional differential equations by means of time-varying distributed delay feedback. These results enable us to present solutions to feedback stabilization problems for systems with delayed input. The results are obtained by using the backstepping technique.

Key words. nonlinear time delay systems, global feedback stabilization

AMS subject classifications. 93D15, 93C23, 93C10

DOI. 10.1137/040616383

1. Introduction.

It is known that for finite-dimensional continuous-time control systems with locally Lipschitz dynamics (e.g., $\dot{x} = f(x,u)$, where f is locally Lipschitz with respect to $(x,u) \in \Re^n \times \Re^m$), finite-time global stabilization cannot be achieved by means of a locally Lipschitz feedback law. However, it has been shown that finite-time global stabilization is possible by means of continuous (see [2, 3, 4, 6, 8] as well as the reported results in [1]) or discontinuous feedback laws (see [14]).

Recently, the option of using feedback with delays has been considered for the stabilization of continuous-time systems in various problems. The closed-loop system may be considered as a system of time-varying retarded functional differential equations (RFDEs). For example, in [20] analytic driftless control systems of the following form are considered:

$$\dot{x}(t) = f(x(t),\ u(t)) := \sum_{i=1}^{m} f_i(x(t))u_i(t)$$
$$x(t) \in \Re^n,\ u(t) := (u_1(t), \dots, u_m(t))' \in \Re^m.$$

The authors in [20] provide strategies for the construction of control laws of the form $u(t) = k(t, x(t), x(lT))$ for $t \in [lT, (l+1)T)$, where l is a nonnegative integer and $T > 0$ denotes the updating time-period of the control. Notice that this type of feedback is a time-varying feedback with delays of the form $u(t) = k(t, x(t), x([t/T]T))$, where $[t/T]$ denotes the integer part of t/T, which is time-varying even if k is independent of time, i.e., $k(t,x,\xi) = k(x,\xi)$. The same comments apply for the synchronous controller switching strategies proposed in [22]. The possibility of switching control laws using distributed delays was recently exploited in [17]. Observers that make use of past values of the state estimate and guarantee convergence in finite-time were considered in [5, 19]. The ability of output discrete delay feedback to stabilize minimum phase linear systems was studied in [9]. Recently, there has been an increasing interest in the feedback stabilization problems of systems with delayed input (see [15, 18]) as well as the application of the backstepping technique for the stabilization of nonlinear

*Received by the editors October 5, 2004; accepted for publication (in revised form) October 20, 2005; published electronically March 15, 2006.
http://www.siam.org/journals/sicon/45-1/61638.html
†Division of Mathematics, Dept. of Economics, University of Athens, Athens, Greece 10559 (ikarafil@econ.uoa.gr).

time-delay systems (see [10, 16]). An account of the use of delays in linear feedback design is given in [21].

In the present work it is shown that finite-time global stabilization can be achieved by time-varying locally Lipschitz distributed delay feedback. It is known that systems described by time-varying RFDEs admit solutions that converge to the equilibrium point in finite time (e.g., Property 5.1 in Chapter 3 in [7]). Using the backstepping technique (see [13, 24]), the problem of finite-time global stabilization for nonlinear triangular systems is studied and solved. Moreover, the approach proposed in this paper is not limited to triangular finite-dimensional continuous-time control systems but can be directly applied to nonlinear triangular systems described by RFDEs. The case of delayed inputs is also considered. Among other cases, we address the finite-time global stabilization problems for the following cases:

- The case of triangular control systems with no delays,

$$\begin{aligned} &\dot{x}_i(t) = f_i(t, x_1(t), \dots, x_i(t)) + x_{i+1}(t), \quad i = 1, \dots, n-1; \\ &\dot{x}_n(t) = f_n(t, x(t)) + u(t), \\ &x(t) := (x_1(t), \dots, x_n(t)) \in \Re^n,\ u(t) \in \Re,\ t \geq 0. \end{aligned} \tag{1.1}$$

- The case of a chain of delayed integrators with no limitation on the size of the delays,

$$\dot{x}_i(t) = x_{i+1}(t - \tau_i), i = 1, \dots, n-1; \quad \dot{x}_n(t) = u(t - \tau_n), \tag{1.2}$$

  where $x(t) = (x_1(t), \dots, x_n(t)) \in \Re^n, u(t) \in \Re$ and $\tau_i \geq 0$ $i = 1, \dots, n$ are the delays.
- The case of triangular control systems with delayed drift terms:

$$\begin{aligned} &\dot{x}_i(t) = f_i(t, x_1(t - \tau_{i,1}), \dots, x_i(t - \tau_{i,i})) + x_{i+1}(t), \quad i = 1, \dots, n-1, \\ &\dot{x}_n(t) = f_n(t, x_1(t - \tau_{n,1}), \dots, x_n(t - \tau_{n,n})) + u(t), \\ &x(t) := (x_1(t), \dots, x_n(t)) \in \Re^n, u(t) \in \Re, t \geq 0, \end{aligned} \tag{1.3}$$

  where $\min_{i=1,\dots,n-1} \quad \min_{j=1,\dots,i} \tau_{i,j} > 0$.

The construction of the proposed distributed delay feedback control is based on a backstepping method, which is applicable to systems with delays. Roughly speaking, the main idea that lies behind the integrator backstepping lemma is described next (here $(T_r(t)x = \{x(t+\theta); \theta \in [-r, 0]\}$ denotes the *r-history* of the state):

> Suppose that the feedback $y(t) = k(t, T_r(t)x)$ leads the state of the system $\dot{x}(t) = f(t, x(t), y(t))$ to zero in finite time (say $T > 0$) and that the feedback $u(t) = \widetilde{k}(t, T_r(t)x, T_r(t)y)$ guarantees the equality $y(t) = k(t, T_r(t)x)$ in finite time (say $T' > 0$) for the augmented system $\dot{x}(t) = f(t, x(t), y(t)); \dot{y}(t) = u(t)$. Then the feedback $u(t) = \widetilde{k}(t, T_r(t)x, T_r(t)y)$ leads the state of the augmented system to zero in finite time.

The explanation of the conclusion is simple: the first T' time units are used in order to achieve $y(t) = k(t, T_r(t)x)$; the next T time units are used in order to achieve $x(t) = 0$ and finally the last r time units are used in order to achieve $x(t + \theta) = 0; \theta \in [-r, 0]$, which directly implies $y(t) = 0$. The idea above can be applied for the design of controllers using a step-by-step procedure (backstepping method).

The structure of the paper is as follows. In section 2, the reader can find definitions and technical results that will be used in later sections of this work. Section 2 is divided

into two subsections. In the first subsection (section 2.1) we present differentiability notions for functionals. In the second subsection (section 2.2) we give the notion of finite-time stabilizability of control systems as well as some preliminary results for the scalar case. Section 3 contains the statement and the proofs of the main results of the paper. Examples are presented in section 4, where the main results of the paper can be directly applied. The conclusions of the paper are given in section 5.

Notation. Throughout this paper we adopt the following notation:

- For a vector $x \in \Re^n$ we denote by $|x|$ its Euclidean norm. For $x \in C^0([-r,0];\Re^n)$ we define $\|x\|_r := \max_{\theta\in[-r,0]} |x(\theta)|$.
- By $C^j(A)(C^j(A;\Omega))$, where $j \geq 0$ is a nonnegative integer, we denote the class of functions (taking values in Ω) that have continuous derivatives of order j on A.
- We denote by K^+ the class of positive C^∞ functions defined on $\Re^+$. We say that an increasing and continuous function $\rho : \Re^+ \to \Re^+$ with $\rho(0) = 0$ is of class K_∞ if $\lim_{s\to+\infty} \rho(s) = +\infty$.
- Z^+ denotes the set of positive integers and $\Re^+$ the set of nonnegative real numbers. For every real number R, $[R]$ denotes its integer part, i.e., $[R] := \max\{x; x \leq R, x\, integer\}$.
- A continuous mapping $f : I \times C^0([-r,0];\Re^n) \times U \to \Re^k$, where $\Re^+ \subseteq I \subseteq \Re, U \subseteq \Re^m$, is said to be completely locally Lipschitz with respect to $(x,u) \in C^0([-r,0];\Re^n) \times U$ if for every bounded set $S \subset I \times C^0([-r,0];\Re^n) \times U$ there exists $L \geq 0$ such that $|f(t,x,u) - f(t,y,v)| \leq L\|x-y\|_r + L|u-v|$ for all $(t,x,u) \in S, (t,y,v) \in S$. Notice that a mapping $f : I \times C^0([-r,0];\Re^n) \times U \to \Re^k$, where $\Re^+ \subseteq I \subseteq \Re, U \subseteq \Re^m$, which is completely locally Lipschitz with respect to $(x,u) \in C^0([-r,0];\Re^n) \times U$, is also defined on $I \times C^0([-r-\sigma,0];\Re^n) \times U$, for every $\sigma \geq 0$, and is completely locally Lipschitz with respect to $(x,u) \in C^0([-r-\sigma,0];\Re^n) \times U$, for every $\sigma \geq 0$.
- Let $x : [a-r, b) \to \Re^n$ with $b > a \geq 0$ and $r \geq 0$. By $T_r(t)x$ we denote the "history" of x from $t-r$ to t, i.e., $T_r(t)x := \{x(t+\theta); \theta \in [-r,0]\}$, for $t \in [a,b)$.

2. Definitions and technical results. Consider the following control system described by retarded functional differential equations (RFDEs):

$$\dot{x}(t) = f(t, T_r(t)x, u(t-\tau(t))); \quad x(t) \in \Re^n, u(t) \in U, t \geq 0, \tag{2.1}$$

where $0 \in U \subseteq \Re^m$, $f : \Re^+ \times C^0([-r,0];\Re^n) \times U \to \Re^n$ is completely locally Lipschitz with respect to $(x,u) \in C^0([-r,0];\Re^n) \times U$ with $f(t,0,0) = 0$ for all $t \geq 0$ and $\tau : \Re^+ \to \Re^+$ is a bounded continuous function. We denote by $x(t) = x(t,t_0,x_0,u) \in \Re^n$ the solution of (2.1) initiated from $t_0 \geq 0$ with initial condition $T_r(t_0)x = x_0 \in C^0([-r,0];\Re^n)$ and corresponding to $u \in C^0(\Re;U)$. By virtue of Theorem 3.2 in [7], for every $(t_0,x_0,u) \in \Re^+ \times C^0([-r,0];\Re^n) \times C^0(\Re;U)$ there exists $t_{\max}(t_0,x_0,u) > t_0$ (called the maximal existence time) such that the solution $x(t) = x(t,t_0,x_0,u) \in \Re^n$ of (2.1) initiated from $t_0 \geq 0$ with initial condition $T_r(t_0)x = x_0$ and corresponding to $u \in C^0(\Re;U)$, is defined on $[t_0-r, t_{\max})$, is continuous on $[t_0-r, t_{\max})$ and continuously differentiable on $[t_0, t_{\max})$ and cannot be further continued, i.e., if $t_{\max} < +\infty$, then $\limsup_{t\to t^-_{\max}} |x(t)| = +\infty$. When $r = 0$ we identify the space $C^0([-r,0];\Re^n)$ with the finite-dimensional space $\Re^n$ and we obtain the familiar finite-dimensional continuous-time case. Consequently, all of the following definitions and results hold also for finite-dimensional continuous-time systems.

This section is divided into two subsections. In the first subsection (section 2.1) we present the differentiability notions for functionals used in the present paper. In the second subsection (section 2.2) we give the notion of finite-time stabilizability of control systems as well as some preliminary results for the scalar case.

2.1. Differentiability notions. In order to study the properties of control system (2.1), we must clarify the differentiability properties of functionals along the solutions of (2.1). The following definition provides the notions of ultimate differentiability and differentiability of functionals along the solutions of (2.1).

DEFINITION 2.1. *Let a functional $\varphi : I \times C^0([-r,0];\Re^n) \to \Re$, where $\Re^+ \subseteq I \subseteq \Re$, $\varphi(t,x)$ being completely locally Lipschitz with respect to $x \in C^0([-r,0];\Re^n)$, with $\varphi(t,0) = 0$ for all $t \in I$. We say that*

- *φ is ultimately differentiable along the solutions of (2.1) with time constant $T \geq 0$, if there exists a constant $T \geq 0$ and a functional $D\varphi : I \times C^0([-r,0];\Re^n) \times U \to \Re$ (called the derivative of φ along the solutions of (2.1) with $D\varphi(t,x,u)$ being completely locally Lipschitz with respect to $(x,u) \in C^0([-r,0];\Re^n) \times U$ and $D\varphi(t,0,0) = 0$ for all $t \in I$, such that for every $(t_0,x_0,u) \in \Re^+ \times C^0([-r,0];\Re^n) \times C^0(\Re;U)$ for which $t_0 + T < t_{\max}(t_0,x_0,u)$, the mapping $t \to \varphi(t,T_r(t)x)$ is of class C^1 on $[t_0+T, t_{\max})$ and it holds that*

$$\frac{d}{dt}\varphi(t,T_r(t)x) = D\varphi(t,T_r(t)x,u(t-\tau(t))) \quad \forall t \in [t_0+T, t_{\max}).$$

- *φ is differentiable along the solutions of (2.1), if φ is ultimately differentiable along the solutions of (2.1) with time constant $T = 0$.*

If $\varphi(t,x) = q(t,x(0))$, where $q \in C^1(\Re^+ \times \Re^n;\Re)$ is a function with locally Lipschitz derivatives and $q(t,0) = 0$ for all $t \geq 0$, then φ is differentiable along the solutions of (2.1) with $D\varphi(t,x,u) := \frac{\partial q}{\partial t}(t,x(0)) + \frac{\partial q}{\partial x}(t,x(0))f(t,x,u)$ for all $(t,x,u) \in \Re^+ \times C^0([-r,0];\Re^n) \times U$. This remark holds particularly for the case $r = 0$, where the functional $\varphi(t,x)$ is an ordinary function $\varphi : \Re^+ \times \Re^n \to \Re$ with $\varphi(t,0) = 0$ for all $t \geq 0$. However, when delays are involved, the notion of ultimate differentiability of a functional is useful. The following example illustrates this point.

Example 2.2. Consider the following system:

$$\begin{aligned} &\dot{x}_1(t) = x_2(t); \qquad \dot{x}_2(t) = u(t) \\ &(x_1(t),x_2(t)) \in \Re^2, \quad u(t) \in \Re. \end{aligned} \tag{2.2}$$

Let $r > 0$ and consider the functional $\varphi(t,x) := x_1(-r)$ defined on $\Re^+ \times C^0([-r,0];\Re^2)$. It can be shown that φ is ultimately differentiable along the solutions of (2.2) with time constant $T = r > 0$ but φ is *not differentiable* along the solutions of (2.2). Moreover, we have $D\varphi(t,x,u) := x_2(-r)$, since $\dot{x}_1(t-r) = x_2(t-r)$ for all $t \geq t_0 + r$. A *more demanding* notion of differentiability of functionals along the solutions of (2.1) is given next.

DEFINITION 2.3. *Let $0 \leq \mu \leq r$. A functional p defined on $I \times C^0([-(r-\mu),0];\Re^n)$, where $\Re^+ \subseteq I \subseteq \Re$, $p(t,x)$ being completely locally Lipschitz with respect to $x \in C^0([-(r-\mu),0];\Re^n)$, with $p(t,0) = 0$ for all $t \in I$, is called l-differentiable along the solutions of (2.1) with delay μ, if there exist functionals $D^i p$, $i = 1,2,\dots,l$, called the ith derivatives of p, defined on $I \times C^0([-(r-\mu),0];\Re^n)$, each $D^i p(t,x)$ being completely locally Lipschitz with respect to $x \in C^0([-(r-\mu),0];\Re^n)$ with $D^i p(t,0) = 0$ for all $t \in I$, such that for every $(t_0,x_0,u) \in \Re^+ \times C^0([-r,0];\Re^n) \times C^0(\Re;U)$, the*

mapping $t \to p(t, T_{r-\mu}(t-\mu)x)$ *is of class* C^l *on* $[t_0, t_{\max}+\mu)$ *and it holds that*

$$\frac{d^i}{dt^i}p(t, T_{r-\mu}(t-\mu)x) = D^i p(t, T_{r-\mu}(t-\mu)x) \quad \forall t \in [t_0, t_{\max}+\mu), i = 1, 2, \dots, l.$$

Remark 2.4. (i) If $p : I \times C^0([-(r-\mu), 0]; \Re^n) \to \Re$, where $\Re^+ \subseteq I \subseteq \Re$, $0 \leq \mu \leq r$, is l-differentiable along the solutions of (2.1) with delay μ, then its derivatives $D^i p : I \times C^0([-(r-\mu), 0]; \Re^n) \to \Re$, $i = 1, 2, \dots, l-1$, are functionals which are $(l-i)$-differentiable along the solutions of (2.1) with delay μ.

(ii) If $p : I \times C^0([-(r-\mu), 0]; \Re^n) \to \Re$, where $\Re^+ \subseteq I \subseteq \Re$, $0 \leq \mu \leq r$, is l-differentiable along the solutions of (2.1) with delay μ, then for every $\tau \in [0, \mu]$ the functional $k(t, x) := p(t+\tau, x)$ is l-differentiable along the solutions of (2.1) with delay $\mu - \tau$ with derivatives $D^i k(t, x) = D^i p(t+\tau, x)$. Moreover, the functional $k(t, x) := p(t+\tau, T_{r'}(\tau-\mu)x)$, where $r' \geq r+\tau-\mu$ is differentiable along the solutions of (2.1) with derivative $Dk(t, x) = Dp(t+\tau, T_{r'}(\tau-\mu)x)$.

The following definition clarifies the notion of a periodic functional.

Definition 2.5. *Let a functional* $\varphi : I \times C^0([-r, 0]; \Re^n) \to \Re$, *where* $\Re^+ \subseteq I \subseteq \Re$, $\varphi(t, x)$ *being completely locally Lipschitz with respect to* $x \in C^0([-r, 0]; \Re^n)$, *with* $\varphi(t, 0) = 0$ *for all* $t \in I$. *We say that* φ *is* T*-periodic if there exists* $T > 0$ *such that* $\varphi(t+T, x) = \varphi(t, x)$ *for all* $(t, x) \in I \times C^0([-r, 0]; \Re^n)$.

The following lemma provides classes of functionals, which are l-differentiable with delay μ along the solutions of (2.1) and are going to be used extensively in the next section.

Lemma 2.6. *The following statements hold:*

(i) *The functional*

$$k(t, x) := \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) \varphi(t+s, T_R(s+\mu)x) ds \tag{2.3}$$

defined on $\Re \times C^0([-R-\mu, 0]; \Re^n)$, *where* $\mu > 0$, $R \geq 0$, $h \in C^l(\Re; \Re^+)$ *with* $h(s) = 0$ *for all* $s \notin (-2, -1)$ *and* $\varphi : \Re \times C^0([-R, 0]; \Re^n) \to \Re$, $\varphi(t, x)$ *being completely locally Lipschitz with respect to* $x \in C^0([-R, 0]; \Re^n)$, *with* $\varphi(t, 0) = 0$ *for all* $t \in \Re$, *is* l*-differentiable along the solutions of* (2.1) *with delay* μ, *with derivatives for* $i = 1, 2, \dots, l$*:*

$$D^i k(t, x) := \frac{(-1)^i}{\mu^{i+1}} \int_{-2\mu}^{-\mu} \frac{d^i h}{dt^i}\left(\frac{s}{\mu}\right) \varphi(t+s, T_R(s+\mu)x) dw \tag{2.4}$$

Moreover, if φ *is* T*-periodic (or linear), then* k *and its derivatives* $D^i k$ $(i = 1, \dots, l)$ *are* T*-periodic (or linear).*

(ii) *Let* $a \in C^l(\Re; \Re)$ *and* $p : \Re \times C^0([-R-\mu, 0]; \Re^n) \to \Re$ *a* $(l+i)$*-differentiable functional along the solutions of* (2.1) *with delay* μ. *The functional*

$$k(t, x) := D^i p(t, x) + a(t) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) \varphi(t+s, T_R(s+\mu)x) ds \tag{2.5}$$

defined on $\Re \times C^0([-R-\mu, 0]; \Re^n)$, *where* $\mu > 0$, $R \geq 0$, $h \in C^l(\Re; \Re^+)$ *with* $h(s) = 0$ *for all* $s \notin (-2, -1)$, $D^i p$ *is the* i*th derivative of* p *and* $\varphi : \Re \times C^0([-R, 0]; \Re^n) \to \Re$, $\varphi(t, x)$ *being completely locally Lipschitz with respect to* $x \in C^0([-R, 0]; \Re^n)$, *with* $\varphi(t, 0) = 0$ *for all* $t \in \Re$, *is* l*-differentiable along*

the solutions of (2.1) *with delay* μ. *Moreover, if* $a \in C^l(\Re;\Re^+)$, φ, p, *and its derivatives* $D^j p$ $(j = 1,\dots,l+i)$ *are* T*-periodic, then it follows that* k *and its derivatives* $D^j k$ $(j = 1,\dots,l)$ *are* T*-periodic. Finally, if* φ, p, *and its derivatives* $D^j p$ $(j = 1,\dots,l+i)$ *are linear, then it follows that* k *and its derivatives* $D^j k$ $(j = 1,\dots,l)$ *are linear.*

The proof of statement (i) of Lemma 2.6 is an immediate consequence of the following equalities:

$$k(t, T_{R+\mu}(t-\mu)x) = \int_{t-2\mu}^{t-\mu} \mu^{-1} h\left(\frac{w-t}{\mu}\right) \varphi(w, T_R(w)x)dw$$

$$\frac{d^i}{dt^i} k(t, T_{R+\mu}(t-\mu)x) = D^i k(t, T_{R+\mu}(t-\mu)x)$$
$$= \frac{(-1)^i}{\mu^{i+1}} \int_{t-2\mu}^{t-\mu} \frac{d^i h}{dt^i}\left(\frac{w-t}{\mu}\right) \varphi(w, T_R(w)x)dw.$$

The proof of statement (ii) of Lemma 2.6 is an immediate consequence of statement (i) of Lemma 2.6 and is omitted.

2.2. Finite-time stabilizability. Having clarified the notions of differentiability of functionals along the solutions of a control system, we next proceed to the notion of finite-time stabilizability of a control system.

DEFINITION 2.7. *Let* $b := \sup_{t\geq 0} \tau(t)$. *We say that system* (2.1) *is finite-time stabilizable if there exists constant* $T \geq 0$ *and a functional* $k : [-b, +\infty) \times C^0([-R, 0];\Re^n) \to U$, $k(t, x)$ *being completely locally Lipschitz with respect to* $x \in C^0([-R, 0];\Re^n)$ *with* $k(t, 0) = 0$ *for all* $t \geq -b$ *such that*

(P1) *for every* $(t_0, x_0) \in \Re^+ \times C^0([-\widetilde{r}, 0];\Re^n)$, *where* $\widetilde{r} := \max(r, R + b)$, *the solution* $x(t) = x(t, t_0, x_0) \in \Re^n$ *of the closed-loop system* (2.1) *with* $u(t) = k(t, T_R(t)x)$ *initiated from* $t_0 \geq 0$ *with initial condition* $T_{\widetilde{r}}(t_0)x = x_0 \in C^0([-\widetilde{r}, 0];\Re^n)$ *exists for all* $t \geq t_0$ *and satisfies* $x(t) = 0$ *for all* $t \geq t_0 + T$

(P2) $\sup\{|x(t_0 + h, t_0, x_0)|;\ h \in [-\widetilde{r}, s], \|x_0\|_{\widetilde{r}} \leq s, t_0 \in [0, s]\} < +\infty\ \forall s \geq 0$.

Particularly, if system (2.1) *is finite-time stabilizable, then we say that the closed-loop system* (2.1) *with* $u(t) = k(t, T_R(t)x)$ *satisfies the dead-beat property of order* T.

Remark 2.8. Using Lemma 3.3 in [12] it can be shown that if the closed-loop system (2.1) with $u(t) = k(t, T_R(t)x)$ satisfies the dead-beat property of order T, then the equilibrium point $0 \in C^0([-\widetilde{r}, 0];\Re^n)$ is nonuniformly in time globally asymptotically stable for the closed-loop system (2.1) with $u(t) = k(t, T_R(t)x)$. The notion of nonuniform in time robust global asymptotic stability was introduced in [11] for finite-dimensional continuous-time systems and was extended to a wide class of systems in [12], including systems described by RFDEs.

As already remarked in the introduction, it is generally known that for finite-dimensional continuous-time control systems with locally Lipschitz dynamics, finite-time global stabilization cannot be achieved by means of locally Lipschitz feedback. The following lemma studies the scalar case and shows that this is no longer true if time-varying distributed delay feedback is used.

LEMMA 2.9. *Consider the one-dimensional control system:*

$$\dot{x}(t) = u(t-\tau) + v(t); \quad x(t) \in \Re, u(t) \in \Re, v(t) \in \Re, t \geq 0, \tag{2.6}$$

where $\tau \geq 0$ *is a constant. Then for every* $\mu > \tau$, *the solutions of the closed-loop system* (2.6) *with*

$$u(t) = -p_\mu(t + \tau, T_\mu(t + \tau - \mu)x), \tag{2.7}$$

where $p_\mu : \Re \times C^0([-\mu, 0]; \Re) \to \Re$ *is the linear* 3μ*-periodic functional defined by*

$$p_\mu(t, x) := \mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x(s+\mu) ds \tag{2.8}$$

for certain $a \in C^0(\Re; \Re^+)$ *being a periodic function with period* 3 *with* $a(t) = 0$ *for* $t \in [0, 2]$ *and* $\int_2^3 a(t)dt = 1$ *and* $h \in C^0(\Re; \Re^+)$ *with* $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1} h(\frac{s}{\mu})ds = 1$*, initiated from arbitrary* $t_0 \geq 0$ *with arbitrary initial condition* $T_{2\mu}(t_0)x = x_0 \in C^0([-2\mu, 0]; \Re)$ *and corresponding to arbitrary input* $v \in C^0(\Re^+; \Re)$ *satisfy*

(2.9)

$$|x(t)| \leq \exp(3L)\sigma\left(\frac{t - t_0}{\mu}\right) \|x_0\|_{2\mu} + 10\mu \exp(3L) \sup_{\max(t_0, t-6\mu) \leq s \leq t} |v(s)| \quad \forall t \geq t_0,$$

where $\sigma(t) := \left\{ \begin{smallmatrix} 1 & if & t < 6 \\ 0 & if & t \geq 6 \end{smallmatrix} \right.$ *and* $L := \max_{t \in [0,3]} a(t)$.

Proof. The closed-loop system (2.6) with (2.7) is described by the differential equation

$$\dot{x}(t) = -\mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x(t+s) ds + v(t). \tag{2.10}$$

As in the proof of Theorem 1.1, Chapter 6, page 168 in [7] it can be shown that for every initial condition $x_0 \in C^0([-2\mu, 0]; \Re)$ the solution of the closed-loop system (2.6) with (2.7) exists for all $t \geq t_0$ and satisfies

$$|x(t)| \leq \exp(L\mu^{-1}(t - t_0)) \left(\|x_0\|_{2\mu} + (t - t_0) \sup_{t_0 \leq \tau \leq t} |v(\tau)| \right) \quad \forall t \geq t_0. \tag{2.11}$$

Define the time sequence $t_i = 3\mu([\frac{t_0}{3\mu}] + i)$, $i \in Z^+$. We next prove the following claim.

Claim. For the solution of the closed-loop system (2.6) with (2.7), the following inequalities hold for all $i \in Z^+$:

$$\begin{gathered} |x(t)| \leq |x(t_i)| + 5\mu \sup_{t_i \leq \tau \leq t} |v(\tau)| \quad \forall t \in [t_i, t_{i+1}] \quad \text{and} \\ |x(t_{i+1})| \leq 5\mu \sup_{t_i \leq \tau \leq t_{i+1}} |v(\tau)|. \end{gathered} \tag{2.12}$$

Proof of Claim. Let arbitrary $i \in Z^+$. Indeed, since $a \in C^0(\Re; \Re^+)$ is a periodic function with period 3, with $a(t) = 0$ for $t \in [0, 2]$, we have $a(\frac{t}{\mu}) = 0$ for $t \in [t_i, t_i + 2\mu]$. Consequently, it follows that

$$x(t) = x(t_i) + \int_{t_i}^{t} v(\tau) d\tau \quad \forall t \in [t_i, t_i + 2\mu]. \tag{2.13}$$

For $t \in [t_i + 2\mu, t_{i+1}]$ we have

$$x(t) = x(t_i + 2\mu) + \int_{t_i + 2\mu}^{t} v(\tau) d\tau - \mu^{-1} \int_{t_i + 2\mu}^{t} a\left(\frac{\tau}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x(\tau + s) ds \, d\tau.$$

The latter relation in conjunction with (2.13) implies that

$$x(t) = x(t_i)\left[1 - \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) ds \int_{t_i+2\mu}^{t} \mu^{-1} a\left(\frac{\tau}{\mu}\right) d\tau\right] + \int_{t_i}^{t} v(\tau) d\tau$$
$$- \mu^{-1} \int_{t_i+2\mu}^{t} a\left(\frac{\tau}{\mu}\right) \int_{-2\mu}^{-\mu} \left(\mu^{-1} h\left(\frac{s}{\mu}\right) \left(\int_{t_i}^{\tau+s} v(w) dw\right)\right) ds \; d\tau$$
$$\forall t \in [t_i + 2\mu, t_{i+1}]. \tag{2.14}$$

Since $h \in C^0(\Re; \Re^+)$ with $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1} h(\frac{s}{\mu})ds = 1$ and $a \in C^0(\Re; \Re^+)$ is a periodic function with period 3 with $\int_2^3 a(t)dt = 1$, it follows from (2.13), (2.14) that inequalities (2.12) hold. Thus the claim made above holds.

An immediate consequence of the previous claim is the following estimate for the solution of closed-loop system (2.6) with (2.7):

$$|x(t)| \le 10\mu \sup_{t_i \le \tau \le t} |v(\tau)| \quad \forall t \in [t_{i+1}, t_{i+2}]. \tag{2.15}$$

Combining estimates (2.11) (for the interval $[t_0, t_1]$), (2.12) (for the interval $[t_1, t_2]$) and (2.15) (for the interval $[t_2, +\infty)$) we obtain the desired estimate (2.9). The proof is complete. □

Remark 2.10. (a) Although the input of system (2.6) is delayed, when the state of the closed-loop system (2.6) with (2.7) and $v(t) \equiv 0$ hits zero, then the control action is indeed zero. Notice that the control action is given by the following formulae:

$$u(t) = -\mu^{-1} a\left(\frac{t+\tau}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x(t+\tau+s) ds;$$
$$u(t-\tau) = -\mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x(t+s) ds.$$

The reader can verify that under the hypotheses of Lemma 2.9, if $v(t) \equiv 0$ and t_0 is a multiple of 3μ, then we have

$$a\left(\frac{t}{\mu}\right) = 0, u(t-\tau) = 0, x(t) = x(t_0) \quad \forall t \in [t_0, t_0 + 2\mu] \quad \text{and}$$
$$x(t) = x(t_0)\left(1 - \mu^{-1} \int_{t_0+2\mu}^{t} a\left(\frac{\tau}{\mu}\right) d\tau\right),$$
$$u(t-\tau) = -\mu^{-1} a\left(\frac{t}{\mu}\right) x(t_0) \quad \forall t \in [t_0 + 2\mu, t_0 + 3\mu].$$

For $t = t_0 + 3\mu$ both state and input become zero (and thus the control cannot push the system away from zero). The same analysis repeated for the next interval $t \in [t_0 + 3\mu, t_0 + 6\mu]$ (with $x(t_0 + 3\mu) = 0$) shows that $x(t) = 0$, $u(t-\tau) = 0$, for all $t \in [t_0 + 3\mu, t_0 + 6\mu]$. Thus we have $x(t) = 0$, $u(t-\tau) = 0$ for all $t \ge t_0 + 3\mu$.

(b) The functions $a \in C^0(\Re; \Re^+)$ and $h \in C^0(\Re; \Re^+)$ are, in a generalized sense, *time-varying gains* of the linear feedback law $u(t) = -\mu^{-1} a(\frac{t+\tau}{\mu}) \int_{-2\mu}^{-\mu} \mu^{-1} h(\frac{s}{\mu}) x(t + \tau + s) ds$. Clearly, we have $|u(t)| \le \frac{L}{\mu} \|T_{2\mu}(t)x\|_{2\mu}$ for all $t \ge 0$, where $L := \max_{t \in [0,3]} a(t)$.

3. Main results. In the present paper we consider triangular time-varying systems described by RFDEs:

$$\begin{aligned} &\dot{x}_i(t) = f_i(t, T_r(t)x_1, \dots, T_r(t)x_i) + x_{i+1}(t-\tau_i), \quad i = 1, \dots, n-1, \\ &\dot{x}_n(t) = f_n(t, T_r(t)x) + u(t-\tau_n), \\ &x(t) := (x_1(t), \dots, x_n(t)) \in \Re^n, \ u(t) \in \Re, t \geq 0, \end{aligned} \tag{3.1}$$

where $r \geq \tau_i \geq 0, i = 1, \dots, n$, the mappings $f_i : \Re^+ \times C^0([-r, 0]; \Re^i) \to \Re \, i = 1, \dots, n$ are completely locally Lipschitz with respect to $x \in C^0([-r, 0]; \Re^n)$ with $f_i(t, 0) = 0$ for all $t \geq 0$ and satisfy one of the following assumptions:

(A1) *There exist mappings* $\varphi_i : \Re \times C^0([-r+\tau_i, 0]; \Re^i) \to \Re, i = 1, \dots, n$, *which are differentiable along the solutions of* (3.1) *and satisfy the following identities for all* $t \geq 0$ *and* $x \in C^0([-r, 0]; \Re^n)$*:*

$$\varphi_1(t - \tau_1, T_{r-\tau_1}(-\tau_1)x_1) := f_1(t, x_1) \tag{3.2a}$$

$$\begin{aligned} &\varphi_{i+1}(t - \tau_{i+1}, T_{r-\tau_{i+1}}(-\tau_{i+1})x_1, \dots, T_{r-\tau_{i+1}}(-\tau_{i+1})x_i, T_{r-\tau_{i+1}}(-\tau_{i+1})x_{i+1}) \\ &:= D\varphi_i(t, x_1, \dots, x_i, x_{i+1}(-\tau_i)) + f_{i+1}(t, x_1, \dots, x_i, x_{i+1}) \quad i = 1, \dots, n-1. \end{aligned} \tag{3.2b}$$

(A2) *There exist mappings* $\varphi_i : \Re \times C^0([-r+\tau_i, 0]; \Re^i) \to \Re, i = 1, \dots, n$, *which are ultimately differentiable along the solutions of* (3.1) *with time constant* $T > 0$ *and satisfy identities* (3.2). *Moreover, there exists a constant* $R \in (0, r]$ *and a continuous function* $L : \Re^+ \times \Re^+ \to \Re^+$ *such that*

$$\begin{aligned} &\sum_{i=1}^{n-1} x_i(0) f_i(t, x_1, \dots, x_i) - x_n(0) D\varphi_{n-1}(t, x) \\ &\quad \leq L\left(t, \sup_{-r \leq \theta \leq -R} |x(\theta)|\right)(|x(0)|^2 + 1), \\ &\quad \forall (t, x) \in \Re^+ \times C^0([-r, 0]; R^n). \end{aligned} \tag{3.3}$$

Our first main result states that system (3.1) is finite-time stabilizable under assumption (A1). Particularly, we have the following theorem.

THEOREM 3.1. *Consider system* (3.1) *and suppose that assumption* (A1) *holds. Let* $b_{i,m} := \sum_{k=i}^{m} \tau_k$. *Then for every* $\mu > b_{1,n}$ *and* $l \in Z^+$, *there exist functions* $\gamma \in K^+$, $\rho \in K_\infty$, *functionals* $p_i : \Re \times C^0([-r_n + \mu, 0]; \Re^i) \to \Re, i = 1, \dots, n$, *where* $r_n := r + 2n\mu$, *which are* l*-differentiable along the solutions of* (3.1) *with delay* $\mu > 0$ *and a constant* $T > 0$, *such that*

(i) *the closed-loop system* (3.1) *with* $u(t) = k(t, T_{r_n-\mu}(t)x)$ *satisfies the dead-beat property of order* T, *where* $k : \Re \times C^0([-r_n + \mu, 0]; \Re^n) \to \Re$ *is completely locally Lipschitz with respect to* $x \in C^0([-r_n + \mu, 0]; \Re^n)$ *with* $k(\cdot, 0) = 0$ *and is defined by*

$$k(t, x) := -\varphi_n(t, x) - \sum_{i=1}^{n} p_i(t + b_{i,n}, T_{r_n-\mu}(b_{i,n} - \mu)x_1, \dots, T_{r_n-\mu}(b_{i,n} - \mu)x_i), \tag{3.4}$$

(ii) *for every* $(t_0, x_0, v) \in \Re^+ \times C^0([-r_n, 0]; \Re^n) \times C^0(\Re; \Re)$ *the solution of the closed-loop system* (3.1) *with* $u(t) = k(t, T_{r_n-\mu}(t)x) + v(t)$ *satisfies the estimate*

$$|x(t)| \le \gamma(t)\rho\left(\|x_0\|_{r_n} + \sup_{t_0-\tau_n \le s \le t} |v(s)|\right) \quad \forall t \ge t_0. \tag{3.5}$$

Moreover, if the mappings $\varphi_i : \Re \times C^0([-r+\tau_i, 0]; \Re^i) \to \Re$, $i = 1, \dots, n$, *are independent of* t, *then the functionals* p_i $i = 1, \dots, n$ *and* k *as defined by* (3.4) *can be chosen to be* 3μ*-periodic. Finally, if the mappings* $\varphi_i : \Re \times C^0([-r+\tau_i, 0]; \Re^i) \to \Re$, $i = 1, \dots, n$, *are linear, then the functionals* p_i $i = 1, \dots, n$ *and* k *as defined by* (3.4) *can be chosen to be linear.*

The proof of Theorem 3.1 relies on the following technical lemma.

LEMMA 3.2 (adding a delayed integrator). *Consider the system*

$$\dot{x}(t) = f(t, T_r(t)x, y(t-\tau)), \tag{3.6a}$$

$$\begin{aligned} &\dot{y}(t) = g(t, T_r(t)x, T_r(t)y) + u(t-\tau'), \\ &x(t) \in \Re^n, y(t) \in \Re, u(t) \in \Re, t \ge 0, \end{aligned} \tag{3.6b}$$

where $r \ge \tau \ge 0$, $\tau' \ge 0$ *are constants and the mappings* $f : \Re^+ \times C^0([-r, 0]; \Re^n) \times \Re \to \Re^n$, $g : \Re^+ \times C^0([-r, 0]; \Re^n) \times C^0([-r, 0]; \Re) \to \Re$ *are completely locally Lipschitz with respect to* $(x, y) \in C^0([-r, 0]; \Re^n) \times C^0([-r, 0]; \Re)$ *with* $f(t, 0, 0) = 0, g(t, 0, 0) = 0$ *for all* $t \ge 0$. *Suppose there exists a functional* $k : \Re \times C^0([-r+\tau, 0]; \Re^n) \to \Re$, *with* $k(\cdot, 0) = 0$, *which is differentiable along the solutions of* (3.6a) *such that*

(H1) *The closed-loop system* (3.6a) *with* $y(t) = k(t, T_{r-\tau}(t)x)$ *satisfies the dead-beat property of order* $T > 0$.

(H2) *There exist functions* $\gamma \in K^+$ *and* $\rho \in K_\infty$ *such that for every* $(t_0, x_0, z) \in \Re^+ \times C^0([-r, 0]; \Re^n) \times C^0(\Re; \Re)$ *the solution of the closed-loop system* (3.6a) *with* $y(t) = k(t, T_{r-\tau}(t)x) + z(t)$, *initiated from* $t_0 \ge 0$ *with initial condition* $T_r(t_0)x = x_0 \in C^0([-r, 0]; \Re^n)$ *and corresponding to* $z \in C^0(\Re; \Re)$ *satisfies the estimate*

$$|x(t)| \le \gamma(t)\rho\left(\|x_0\|_r + \sup_{t_0-\tau \le s \le t} |z(s)|\right) \quad \forall t \ge t_0. \tag{3.7}$$

(H3) *There exists a functional* $\varphi : \Re \times C^0([-r+\tau', 0]; \Re^{n+1}) \to \Re$, $\varphi(t, x, y)$ *being completely locally Lipschitz with respect to* $(x, y) \in C^0([-r+\tau', 0]; \Re^n) \times C^0([-r-\tau', 0]; \Re)$ *with* $\varphi(t, 0, 0) = 0$ *for all* $t \in \Re$ *and such that the following identity holds for all* $t \ge 0$ *and* $(x, y) \in C^0([-r, 0]; \Re^n) \times C^0([-r, 0]; \Re)$*:*

$$\varphi(t-\tau', T_{r-\tau'}(-\tau')x, T_{r-\tau'}(-\tau')y) := -g(t, x, y) + Dk(t, x, y(-\tau)), \tag{3.8}$$

where $Dk : \Re \times C^0([-r, 0]; \Re^n) \times \Re \to \Re$ *denotes the derivative of* k *along the solutions of* (3.6a).

Then for every $\mu > \tau'$, *there exist functions* $\widetilde{\gamma} \in K^+$, $\widetilde{\rho} \in K_\infty$ *and a functional* $\widetilde{k} : \Re \times C^0([-r-2\mu+\tau', 0]; \Re^{n+1}) \to \Re$, *which is completely locally Lipschitz with respect to* $(x, y) \in C^0([-r-2\mu+\tau', 0]; \Re^{n+1})$ *with* $\widetilde{k}(t, 0, 0) = 0$ *for all* $t \in \Re$ *and is defined by*

$$\begin{aligned} \widetilde{k}(t, x, y) := \varphi(t, x, y) - \mu^{-1} a\left(\frac{t+\tau'}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) (y(\tau'+s) \\ - k(t+\tau'+s, T_{r-\tau}(\tau'+s)x))ds, \end{aligned} \tag{3.9}$$

where $a \in C^0(\Re;\Re^+)$ *is a periodic function with period* 3, *with* $a(t) = 0$ *for* $t \in [0,2]$ *and* $\int_2^3 a(t)dt = 1$, $h \in C^0(\Re;\Re^+)$ *with* $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1}h(\frac{s}{\mu})ds = 1$, *such that the closed-loop system* (3.6) *with* $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y)$ *satisfies the dead-beat property of order* $T + r + 6\mu$ *and for every* $(t_0, x_0, y_0, v) \in \Re^+ \times C^0([-r-2\mu, 0];\Re^{n+1}) \times C^0(\Re;\Re)$ *the solution of the closed-loop system* (3.6) *with* $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y) + v(t)$ *satisfies the estimate*

$$|(x(t), y(t))| \leq \widetilde{\gamma}(t)\widetilde{\rho}\left(\|(x_0, y_0)\|_{r+2\mu} + \sup_{t_0-\tau' \leq s \leq t} |v(s)|\right) \quad \forall t \geq t_0. \tag{3.10}$$

Remark 3.3. Notice that the stabilizing feedback is given by

$$u(t) = \varphi(t, T_r(t)x, T_r(t)y) - \mu^{-1}a\left(\frac{t+\tau'}{\mu}\right)\int_{-2\mu}^{-\mu} \mu^{-1}h\left(\frac{s}{\mu}\right) z(t+\tau'+s)ds,$$

where $z(t) := y(t) - k(t, T_{r-\tau}(t)x)$.

Proof of Lemma 3.2. Let arbitrary $(t_0, x_0, y_0, v) \in \Re^+ \times C^0([-r-2\mu, 0];\Re^{n+1}) \times C^0(\Re;\Re)$ and define

$$z(t) := y(t) - k(t, T_{r-\tau}(t)x). \tag{3.11}$$

By virtue of definitions (3.9), (3.11), and identity (3.8), we guarantee that as long as the solution of the closed-loop system (3.6) with $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y) + v(t)$ and $T_{r+2\mu}(t_0)x = x_0$, $T_{r+2\mu}(t_0)y = y_0$ exists, it coincides with the solution of the following system:

$$\begin{aligned} \dot{x}(t) &= f(t, T_r(t)x, k(t-\tau, T_{r-\tau}(t-\tau)x) + z(t-\tau)) \\ \dot{z}(t) &= -\mu^{-1}a\left(\frac{t}{\mu}\right)\int_{-2\mu}^{-\mu} \mu^{-1}h\left(\frac{s}{\mu}\right) z(t+s)ds + v(t-\tau') \\ x(t) &\in \Re^n,\ z(t) \in \Re,\ v(t) \in \Re,\ t \geq 0 \end{aligned} \tag{3.12}$$

corresponding to the same input $v \in C^0(\Re;\Re)$ with $y(t) = z(t) + k(t, T_{r-\tau}(t)x)$ and $T_{r+2\mu}(t_0)x = x_0$, $z(t_0+\theta) = y_0(\theta) - k(t_0+\theta, T_{r-\tau}(t_0+\theta)x); \theta \in [-\widetilde{\tau}, 0]$, where $\widetilde{\tau} := \max(\tau, 2\mu)$. The solution of (3.12) exists for all $t \geq t_0$, since by virtue of Lemma 2.9, we obtain

$$\begin{aligned} |z(t)| \leq \exp(3L)\sigma\left(\frac{t-t_0}{\mu}\right)\|T_{\widetilde{\tau}}(t_0)z\|_{\widetilde{\tau}} \\ + 10\mu\exp(3L)\sup_{\max(t_0, t-6\mu) \leq s \leq t} |v(s-\tau')| \\ \forall t \geq t_0 - \widetilde{\tau}, \end{aligned} \tag{3.13}$$

where $\sigma(t) := \left\{ \begin{smallmatrix} 1 & if & t < 6 \\ 0 & if & t \geq 6 \end{smallmatrix} \right.$ and $L := \max_{t\in[0,3]} a(t)$ and inequality (3.13) in conjunction with estimate (3.7) implies the following estimate:

$$|x(t)| \leq \gamma(t)\bar{a}\left(\|T_r(t_0)x\|_r + \|T_{\widetilde{\tau}}(t_0)z\|_{\widetilde{\tau}} + \sup_{t_0-\tau' \leq s \leq t} |v(s)|\right) \quad \forall t \geq t_0, \tag{3.14}$$

where $\bar{a}(s) := \rho(\exp(3L)(1+10\mu)s) \in K_\infty$. We next prove the following claim.

CLAIM. There exist functions $\widetilde{\gamma} \in K^+$, $\widetilde{\rho} \in K_\infty$ such that for every $(t_0, x_0, y_0, v) \in \Re^+ \times C^0([-r-2\mu, 0]; \Re^{n+1}) \times C^0(\Re; \Re)$ the solution of (3.12) corresponding to $v \in C^0(\Re; \Re)$ with $y(t) = z(t) + k(t, T_{r-\tau}(t)x)$, $T_{r+2\mu}(t_0)x = x_0$, $z(t_0+\theta) = y_0(\theta) - k(t_0+\theta, T_{r-\tau}(t_0+\theta)x); \theta \in [-\widetilde{\tau}, 0]$ satisfies estimate (3.10).

Proof of Claim. Since $k : [-2\mu, +\infty) \times C^0([-r+\tau, 0]; \Re^n) \to \Re$ is completely locally Lipschitz with respect to $x \in C^0([-r+\tau, 0]; \Re^n)$ with $k(t, 0) = 0$ for all $t \geq -2\mu$, it follows from Lemma 3.2 in [12] that there exist functions $\delta \in K^+$, being nondecreasing and $q \in K_\infty$ such that

$$|k(t,x)| \leq q(\delta(t)\|x\|_{r-\tau}) \quad \forall (t,x) \in [-2\mu, +\infty) \times C^0([-r+\tau, 0]; \Re^n). \tag{3.15}$$

Notice that definition (3.11) in conjunction with (3.15) implies

$$\|T_{\widetilde{\tau}}(t_0)z\|_{\widetilde{\tau}} \leq \|y_0\|_{\widetilde{\tau}} + q(\delta(t)\|x_0\|_{r+2\mu}) \quad \forall t \geq t_0. \tag{3.16}$$

Using Corollary 10 and Remark 11 in [23] we may find a function $\kappa \in K_\infty$ such that

$$q(rs) + \overline{a}(rs) + \exp(3L)(1+10\mu)rs \leq \kappa(r)\kappa(s) \quad \forall r, s \geq 0. \tag{3.17}$$

Without loss of generality we may assume that the function $\gamma \in K^+$ involved in (3.14) is nondecreasing. Thus, using estimate (3.14), we obtain the following estimate for the solution of (3.12):

$$\|T_r(t)x\|_r \leq \gamma(t)\overline{a}\left(\|x_0\|_r + \|T_{\widetilde{\tau}}(t_0)z\|_{\widetilde{\tau}} + \sup_{t_0-\tau' \leq s \leq t} |v(s)|\right) \quad \forall t \geq t_0. \tag{3.18}$$

Combining (3.15), (3.16), (3.17), and (3.18) we obtain for the solution of (3.12),

$$|k(t, T_{r-\tau}(t)x)| \leq p(t)\zeta\left(\|(x_0, y_0)\|_{r+2\mu} + \sup_{t_0-\tau' \leq s \leq t} |v(s)|\right) \quad \forall t \geq t_0, \tag{3.19}$$

where $p(t) := \kappa(\delta(t)\gamma(t)\kappa(1+\kappa(\delta(t))))$ and $\zeta(s) := \kappa(\kappa(2s+\kappa(s)))$. Estimates (3.13), (3.14), (3.19) in conjunction with inequalities (3.16), (3.17) and the trivial inequality $|y(t)| \leq |z(t)| + |k(t, T_{r-\tau}(t)x)|$, show that the solutions of (3.12) satisfy estimate (3.10) for $\widetilde{\gamma}(t) := (1+\gamma(t))\kappa(1+\kappa(\delta(t))) + p(t)$ and $\widetilde{a}(s) := \kappa(2s+\kappa(s)) + \zeta(s)$. Thus the claim is proved.

Since for every $(t_0, x_0, y_0, v) \in \Re^+ \times C^0([-r-2\mu, 0]; \Re^{n+1}) \times C^0(\Re; \Re)$ the solution of the closed-loop system (3.6) with $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y) + v(t)$ and $T_{r+2\mu}(t_0)x = x_0$, $T_{r+2\mu}(t_0)y = y_0$ coincides with the solution of (3.12) corresponding to the same input $v \in C^0(\Re; \Re)$ with $y(t) = z(t) + k(t, T_{r-\tau}(t)x)$ and $T_{r+2\mu}(t_0)x = x_0$, $z(t_0+\theta) = y_0(\theta) - k(t_0+\theta, T_{r-\tau}(t_0+\theta)x); \theta \in [-\widetilde{\tau}, 0]$, where $\widetilde{\tau} := \max(\tau, 2\mu)$, we conclude that for every $(t_0, x_0, y_0, v) \in \Re^+ \times C^0([-r-2\mu, 0]; \Re^{n+1}) \times C^0(\Re; \Re)$ the solution of the closed-loop system (3.6) with $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y) + v(t)$ and $T_{r+2\mu}(t_0)x = x_0$, $T_{r+2\mu}(t_0)y = y_0$ exists for all $t \geq t_0$ and satisfies estimate (3.10). Finally, the fact that the closed-loop system (3.6) with $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y)$ satisfies the dead-beat property of order $T + r + 6\mu$ follows from the observation that $z(t) = 0$ for all $t \geq t_0 + 6\mu$ (recall (3.13) with $v(t) \equiv 0$), which implies $y(t) = k(t, T_{r-\tau}(t)x)$ for all $t \geq t_0 + 6\mu$ (recall definition (3.11)). Thus for all $t \geq t_0 + 6\mu$ the solution of the closed-loop system (3.6) with $u(t) = \widetilde{k}(t, T_{r+2\mu-\tau'}(t)x, T_{r+2\mu-\tau'}(t)y)$ coincides with the solution of (3.6a) with $y(t) = k(t, T_{r-\tau}(t)x)$ initiated from $T_r(t_0+6\mu)x$. It follows that $x(t) = 0$ for all

$t \geq t_0 + T + 6\mu$, which implies $y(t) = k(t, T_{r-\tau}(t)x) = 0$ for all $t \geq t_0 + T + r + 6\mu$. The proof is complete. □

We are now in position to prove Theorem 3.1.

Proof of Theorem 3.1. The proof is made by induction.

Step 1: We show that the statement of Theorem 3.1 holds for $n = 1$. Let $l \in Z^+$ and $\mu > b_{1,n} = \tau_1$. The statement of Theorem 3.1 for $n = 1$ is an immediate consequence of Lemma 2.9, identity (3.2a), (3.4), and the definition

$$p_1(t, x_1) := -\mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x_1(s+\mu) ds, \tag{3.20}$$

where $a \in C^l(\Re; \Re^+)$ is a periodic function with period 3 with $a(t) = 0$ for $t \in [0, 2]$ and $\int_2^3 a(t)dt = 1$ and $h \in C^l(\Re; \Re^+)$ with $h(s) = 0$ for all $s \notin (-2, -1)$, $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1} h(\frac{s}{\mu})ds = 1$. Notice that the fact that $p_1(t, x_1)$, as defined by (3.11), is l-differentiable with delay μ follows directly from Lemma 2.6. Moreover, if $\varphi_1 : \Re \times C^0([-r + \tau_1, 0]; \Re) \to \Re$ is independent of t, then the functionals p_1 and k (defined by (3.4)) are 3μ-periodic. Finally, if $\varphi_1 : \Re \times C^0([-r + \tau_1, 0]; \Re) \to \Re$ is linear, then the functionals p_1 and k (defined by (3.4)) are linear.

Induction step: Suppose that the statement of Theorem 3.1 holds for $n - 1$. We show that the statement of Theorem 3.1 holds for n. By assumption for every $\mu > b_{1,n-1}, l \in Z^+$, there exist functions $\gamma \in K^+, \rho \in K_\infty$, functionals $p_i : \Re \times C^0([-r_{n-1} + \mu, 0]; \Re^i) \to \Re$, $i = 1, \dots, n-1$, where $r_{n-1} := r + 2(n-1)\mu$, which are $(l+1)$-differentiable along the solutions of the following system:

$$\begin{aligned} &\dot{x}_i(t) = f_i(t, T_r(t)x_1, \dots, T_r(t)x_i) + x_{i+1}(t - \tau_i), \quad i = 1, \dots, n-1, \\ &\xi(t) := (x_1(t), \dots, x_{n-1}(t)) \in \Re^{n-1}, x_n(t) \in \Re, t \geq 0 \end{aligned} \tag{3.21}$$

with delay $\mu > 0$ and a constant $T > 0$, such that

(a) the closed-loop system (3.21) with $x_n(t) = k(t, T_{r_{n-1}-\mu}(t)\xi)$ satisfies the deadbeat property of order T, where

$$\begin{aligned} k(t, \xi) := &- \varphi_{n-1}(t, \xi) \\ &- \sum_{i=1}^{n-1} p_i(t + b_{i,n-1}, T_{r_{n-1}-\mu}(b_{i,n-1} - \mu)x_1, \dots, T_{r_{n-1}-\mu}(b_{i,n-1} - \mu)x_i), \end{aligned} \tag{3.22}$$

(ii) for every $(t_0, \xi_0, z) \in \Re^+ \times C^0([-r_{n-1}, 0]; \Re^n) \times C^0(\Re; \Re)$ the solution of the closed-loop system (3.21) with $x_n(t) = k(t, T_{r_{n-1}-\mu}(t)\xi) + z(t)$ satisfies the estimate

$$|\xi(t)| \leq \gamma(t)\rho\left(\|\xi_0\|_{r_{n-1}} + \sup_{t_0 - \tau_{n-1} \leq s \leq t} |z(s)|\right) \quad \forall t \geq t_0. \tag{3.23}$$

Remark 2.4(ii) and hypothesis (A1) show that k as defined by (3.22) is differentiable along the solutions of (3.21) with derivative

$$\begin{aligned} Dk(t, \xi, x_n(-\tau_{n-1})) := &- D\varphi_{n-1}(t, \xi, x_n(-\tau_{n-1})) \\ &- \sum_{i=1}^{n-1} Dp_i(t + b_{i,n-1}, T_{r_{n-1}-\mu}(b_{i,n-1} - \mu)x_1, \dots, \\ &\quad T_{r_{n-1}-\mu}(b_{i,n-1} - \mu)x_i). \end{aligned}$$

Clearly, hypotheses (H1) and (H2) of Lemma 3.2 hold for system (3.1). Next we show that hypothesis (H3) of Lemma 3.2 also holds for system (3.1). Notice that

$$\begin{aligned}&-f_n(t,x)+Dk(t,\xi,x_n(-\tau_{n-1})),\\&=-f_n(t,x)-D\varphi_{n-1}(t,\xi,x_n(-\tau_{n-1}))\\&\quad-\sum_{i=1}^{n-1}Dp_i(t+b_{i,n-1},T_{r_{n-1}-\mu}(b_{i,n-1}-\mu)x_1,\dots,T_{r_{n-1}-\mu}(b_{i,n-1}-\mu)x_i).\end{aligned}$$

Hypothesis (A1) implies that there exists a mapping $\varphi_n : \Re \times C^0([-r+\tau_n,0];\Re^n)\to\Re$, which is differentiable along the solutions of (3.1) and satisfies $\varphi_n(t-\tau_n, T_{r-\tau_n}(-\tau_n)x) := D\varphi_{n-1}(t,\xi,x_n(-\tau_{n-1}))+f_n(t,x)$ for all $t\ge 0$ and $x\in C^0([-r,0];\Re^n)$. Thus we obtain

$$-f_n(t,x)+Dk(t,\xi,x_n(-\tau_{n-1}))=\varphi(t-\tau_n,T_{r_{n-1}-\tau_n}(-\tau_n)x), \tag{3.24}$$

where $\varphi(t,x) := -\varphi_n(t,x)-\sum_{i=1}^{n-1}Dp_i(t+\tau_n+b_{i,n-1},T_{r_{n-1}-\mu}(b_{i,n-1}+\tau_n-\mu)x_1,\dots,T_{r_{n-1}-\mu}(b_{i,n-1}+\tau_n-\mu)x_i)$. Since $b_{i,n-1}+\tau_n=b_{i,n}$ we have

$$\varphi(t,x) := -\varphi_n(t,x)-\sum_{i=1}^{n-1}Dp_i(t+b_{i,n},T_{r_{n-1}-\mu}(b_{i,n}-\mu)x_1,\dots,T_{r_{n-1}-\mu}(b_{i,n}-\mu)x_i). \tag{3.25}$$

Notice that the mapping $\varphi:\Re\times C^0([-r_{n-1}+\tau_n,0];\Re^n)\to\Re$ defined by (3.25) is well defined for all $\mu > b_{1,n}$, with $\varphi(t,x)$ being completely locally Lipschitz with respect to $x\in C^0([-r+\tau_n,0];\Re^n)$ and $\varphi(t,0)=0$ for all $t\in\Re$.

Lemma 3.2 shows that there exist functions $\widetilde{\gamma}\in K^+$, $\widetilde{\rho}\in K_\infty$ and a functional $\widetilde{k}:\Re\times C^0([-r_{n-1}-2\mu,0];\Re^n)\to\Re$, which is completely locally Lipschitz with respect to $x\in C^0([-r_{n-1}-2\mu,0];\Re^n)$ with $\widetilde{k}(t,0)=0$ for all $t\in\Re$ and is defined by

$$\begin{aligned}&\widetilde{k}(t,x) := \varphi(t,x)-p_n(t+\tau_n,T_{r_n-\mu}(\tau_n-\mu)x),\\&p_n(t,x) := \mu^{-1}a\left(\frac{t}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)(x_n(s+\mu)-k(t+s,T_{r_{n-1}}(s+\mu)\xi))ds,\end{aligned} \tag{3.26}$$

where $a\in C^l(\Re;\Re^+)$ is a periodic function with period 3, with $a(t)=0$ for $t\in[0,2]$ and $\int_2^3 a(t)dt=1$, $h\in C^l(\Re;\Re^+)$ with $h(s)=0$ for all $s\notin(-2,-1)$, $\int_{-2}^{-1}h(s)ds=\int_{-2\mu}^{-\mu}\mu^{-1}h(\frac{s}{\mu})ds=1$, such that the closed-loop system (3.1) with $u(t)=\widetilde{k}(t,T_{r_n-\mu}(t)x)$ satisfies the dead-beat property of order $T+r_{n-1}+6\mu$ and for every $(t_0,x_0,v)\in\Re^+\times C^0([-r_n,0];\Re^n)\times C^0(\Re;\Re)$ the solution of the closed-loop system (3.1) with $u(t)=\widetilde{k}(t,T_{r_n-\mu}(t)x)+v(t)$ satisfies estimate (3.5) with $\widetilde{\gamma}\in K^+$ and $\widetilde{\rho}\in K_\infty$ in place of $\gamma\in K^+$ and $\rho\in K_\infty$, respectively. Lemma 2.6 implies that the functional p_n is l-differentiable along the solutions of (3.1) with delay $\mu>0$. Moreover, if φ_n is independent of t and k is 3μ-periodic then the functionals p_n and $\widetilde{k}$ are 3μ-periodic. Finally, if φ_n and k are linear functionals, then the functionals p_n and $\widetilde{k}$ are linear. □

Our second main result states that system (3.1) is finite-time stabilizable under assumption (A2).

THEOREM 3.4. *Consider system* (3.1) *and suppose that assumption* (A2) *holds. Then for every* $\mu \geq R + \sum_{i=1}^{n} \tau_i$, *where* $R > 0$ *is the constant involved in assumption* (A2), *there exists a functional* $k : \Re \times C^0([-\widetilde{r}, 0]; \Re^n) \to \Re$, *where* $\widetilde{r} := 2r + 2n\mu$, *which is completely locally Lipschitz with respect to* $x \in C^0([-\widetilde{r}, 0]; \Re^n)$ *and a constant* $T' > T$ *(where* $T \geq 0$ *is the time constant involved in assumption* (A2)*, such that the closed-loop system* (3.1) *with* $u(t) = k(t, T_{\widetilde{r}}(t)x)$ *satisfies the dead-beat property of order* T'. *Moreover, if the mappings* $\varphi_i : \Re \times C^0([-r + \tau_i, 0]; \Re^i) \to \Re$, $i = 1, \ldots, n$ *are independent of* t, *then the functional* k *can be chosen to be* 3μ*-periodic. Finally, if the mappings* $\varphi_i : \Re \times C^0([-r + \tau_i, 0]; \Re^i) \to \Re$, $i = 1, \ldots, n$ *are linear, then the functional* k *can be chosen to be linear.*

Proof. Consider the system

$$\begin{aligned} &\dot{z}_i(t) = z_{i+1}(t - \tau_i), i = 1, \ldots, n-1;\ \dot{z}_n(t) = v(t - \tau_n) \\ &z(t) = (z_1(t), \ldots, z_n(t)) \in \Re^n, v(t) = \Re, \end{aligned} \tag{3.27}$$

where $\tau_i \geq 0$ are exactly the delays appearing in (3.1). Since assumption (A1) holds for system (3.27) with $\varphi_i \equiv 0$ for $i = 1, \ldots, n$, by virtue of Theorem 3.1 we conclude that for every $\mu \geq \sum_{i=1}^{n} \tau_i + R$ there exists a linear 3μ-periodic time-varying distributed delay feedback $v(t) = \widetilde{k}(t, T_{\widetilde{r-r-R}}(t - R)z)$ and a constant $T' > 0$ such that the closed-loop system (3.27) with $v(t) = \widetilde{k}(t, T_{\widetilde{r-r-R}}(t - R)z)$ satisfies the dead-beat property of order T'.

Consider the mapping $(t, x) \in \Re \times C^0([-\widetilde{r}, 0]; \Re^n) \to P(t, x) = y \in C^0([-(\widetilde{r} - r), 0]; \Re^n)$ defined by

$$\begin{aligned} &y_1(\theta) := x_1(\theta), \quad \theta \in [-(\widetilde{r} - r), 0], \\ &y_i(\theta) := x_i(\theta) + \varphi_{i-1}(t + \theta, T_r(\theta)x_1, \ldots, T_r(\theta)x_{i-1}), \quad i = 2, \ldots, n, \theta \in [-(\widetilde{r} - r), 0]. \end{aligned} \tag{3.28}$$

Notice that since the mappings $\varphi_i : \Re \times C^0([-r + \tau_i, 0]; \Re^i) \to \Re$, $i = 1, \ldots, n$, are all completely locally Lipschitz with respect to $x \in C^0([-r, 0]; \Re^n)$, it follows that for every bounded set $S \subset \Re \times C^0([-\widetilde{r}, 0]; \Re^n)$ there exists $L \geq 0$ such that $\|P(t, x) - P(t, y)\|_{\widetilde{r-r}} \leq L\|x - y\|_{\widetilde{r}}$ for all $(t, x) \in S$, $(t, y) \in S$. Moreover, notice that $P(t, 0) = 0$ for all $t \in \Re$. We next prove that the time-varying distributed delay feedback that induces the dead-beat property for the corresponding closed-loop system is defined by

$$k(t, x) := -\varphi_n(t, T_r(0)x) + \widetilde{k}(t, P(t - R, T_{\widetilde{r}}(-R)x)). \tag{3.29}$$

Notice that if the mappings $\varphi_i : \Re \times C^0([-r + \tau_i, 0]; \Re^i) \to \Re$, $i = 1, \ldots, n$ are independent of t (linear), then the functional k as defined by (3.29) is 3μ-periodic (linear). Let the solution $x(t)$ of the closed-loop system (3.1) with $u(t) = k(t, T_{\widetilde{r+R}}(t)x)$ with arbitrary initial condition $T_{\widetilde{r+R+\tau_n}}(t_0)x = x_0$.

We make the following claim.

CLAIM. *The solution* $x(t)$ *of the closed-loop system* (3.1) *with* $u(t) = k(t, T_{\widetilde{r+R}}(t)x)$ *exists for all* $t \geq t_0$ *and satisfies Property* (P2) *of Definition* 2.7.

Proof of Claim. Since

- $P(t, 0) = 0$ for all $t \in \Re$ and for every bounded set $S \subset \Re \times C^0([-\widetilde{r}, 0]; \Re^n)$ there exists $L \geq 0$ such that $\|P(t, x) - P(t, y)\|_{\widetilde{r-r}} \leq L\|x - y\|_{\widetilde{r}}$ for all $(t, x) \in S$, $(t, y) \in S$.
- $\widetilde{k}(t, z)$ is a linear 3μ-periodic time-varying functional

it follows from Lemma 3.2 in [12] that there exist functions $\delta \in K^+$, being nondecreasing and $q \in K_\infty$ such that

$$|\widetilde{k}(t-\tau_n, P(t-\tau_n-R, T_{\widetilde{r}}(-\tau_n-R)x))| \le q(\delta(t)\|T_{\widetilde{r}}(-\tau_n-R)x\|_{\widetilde{r}}) \quad \forall (t,x) \in \Re^+ \times C^0([-\widetilde{R},0];\Re^n) \tag{3.30}$$

with $\widetilde{R} := \widetilde{r} + \tau_n + R$. In order to prove the claim it suffices to show that if there exists an integer N such that

$$\sup\{|x(t_0+h,t_0,x_0)|;\ h \in [-\widetilde{R}, NR], \|x_0\|_{\widetilde{R}} \le s, t_0 \in [0,s]\} < +\infty \quad \forall s \ge 0, \tag{3.31}$$

then we have

$$\sup\{|x(t_0+h,t_0,x_0)|;\ h \in [-\widetilde{R}, (N+1)R], \|x_0\|_{\widetilde{R}} \le s, t_0 \in [0,s]\} < +\infty \quad \forall s \ge 0, \tag{3.32}$$

where $x(t,t_0,x_0)$ denotes the solution of the closed-loop system (3.1) with $u(t) = k(t, T_{\widetilde{r}}(t)x)$ with initial condition $T_{\widetilde{R}}(t_0)x = x_0$. Indeed, since (3.31) holds for $N = 0$, it will follow by induction that (3.31) holds for all nonnegative integers N.

Suppose that (3.31) holds for some nonnegative integer N. Let arbitrary $s \ge 0$ and define $V(t) := \frac{1}{2}|x(t)|^2$, where $x(t)$ denotes the solution of the closed-loop system (3.1) with $u(t) = k(t, T_{\widetilde{r}}(t)x)$ with arbitrary initial condition $T_{\widetilde{R}}(t_0)x = x_0$ such that $\|x_0\|_{\widetilde{R}} \le s$, $t_0 \in [0,s]$. Using identity (3.2b) for $i = n-1$, inequality (3.3) in conjunction with (3.30) imply that as long as the solution of the closed-loop system (3.1) with $u(t) = k(t, T_{\widetilde{r}}(t)x)$ exists, it holds that

$$\dot{V}(t) \le L\Big(t, \sup_{-r\le\theta\le -R} |x(t+\theta)|\Big)(2V(t)+1) + V(t) + q^2\Big(\delta(t) \sup_{-r\le\theta\le -R} |x(t+\theta)|\Big). \tag{3.33}$$

Let $a := \sup\{|x(t_0+h,t_0,x_0)|;\ h \in [-\widetilde{R}, NR], \|x_0\|_{\widetilde{R}} \le s, t_0 \in [0,s]\} < +\infty$. Clearly, for all $t \in [NR, (N+1)R]$ for which the solution of the closed-loop system (3.1) with $u(t) = k(t, T_{\widetilde{r}}(t)x)$ exists, we have

$$\dot{V}(t) \le \left(1 + q^2(\delta((N+1)R)a) + \max_{\substack{0\le t\le (N+1)R \\ 0\le s\le a}} 2L(t,s)\right)(V(t)+1).$$

The previous differential inequality implies that the solution of the closed-loop system (3.1) with $u(t) = k(t, T_{\widetilde{r}}(t)x)$ exists for all $t \in [NR, (N+1)R]$ and satisfies

$$|x(t)|^2 \le 2V(t) \le \exp(M(N,R,a))(a^2+2) \quad \forall t \in [NR, (N+1)R],$$

where $M(N,R,a) := R(1+q^2(\delta((N+1)R)a) + \max_{\substack{0\le t\le (N+1)R \\ 0\le s\le a}} 2L(t,a))$. The previous property shows that inequality (3.32) holds. Thus the claim is proved.

We continue the proof of Theorem 3.4 by considering the following variables defined by

$$z_1(t) := x_1(t); \quad z_i(t) := x_i(t) + \varphi_{i-1}(t, T_r(t)x_1, \dots, T_r(t)x_{i-1}), \quad i = 2, \dots, n. \tag{3.34}$$

Clearly, by virtue of definitions (3.28) and (3.34) we obtain $T_{r-r}(t-R)z = P(t-R, T_r(t-R)x)$ for all $t \geq t_0$. Moreover, by virtue of (3.34), identities (3.2), and the definition of the notion of ultimate differentiability along the solutions of (3.1) with time constant $T \geq 0$ (Definition 2.1), $z(t)$ as defined by (3.34) for the solution of the closed-loop system (3.1) with $u(t) = k(t, T_{r+R}(t)x)$, coincides for all $t \geq t_0 + T$ with the solution of the closed-loop system (3.27) with $v(t) = \widetilde{k}(t, T_{r-r-R}(t-R)z)$ with initial condition $T_{r-r}(t_0+T)z = P(t_0+T, T_r(t_0+T)x)$. Thus we obtain $z(t) = 0$ for all $t \geq t_0 + T + T'$. It follows from definition (3.34) that $x(t) = 0$ for all $t \geq t_0+T+T'+(n-1)r$. Thus the closed-loop system (3.1) with $u(t) = k(t, T_r(t)x)$ satisfies the dead-beat property of order $\widetilde{T} := T + T' + (n-1)r$. The proof is complete. □

4. Examples. The following examples present systems, which satisfy assumption (A1) and consequently, by virtue of Theorem 3.1, can be finite-time stabilized by means of time-varying distributed delay feedback. Our first example is concerned with finite-dimensional control systems.

Example 4.1. Consider the two-dimensional control system

$$\begin{aligned} &\dot{x}_1(t) = f_1(t, x_1(t)) + x_2(t); \quad \dot{x}_2(t) = f_2(t, x_1(t), x_2(t)) + u(t), \\ &(x_1(t), x_2(t)) \in \Re^2, u(t) \in \Re, t \geq 0, \end{aligned} \tag{4.1}$$

where $f_1 \in C^2(\Re \times \Re; \Re)$ and $f_2 \in C^1(\Re \times \Re^2; \Re)$ with $f_1(t,0) = f_2(t,0,0) = 0$ for all $t \in \Re$. For system (4.1) finite-time global stabilization cannot be achieved by means of a locally Lipschitz feedback law $u(t) = k(t, x_1(t), x_2(t))$. On the other hand Theorem 3.1 guarantees that there exists a time-varying distributed delay feedback law that achieves finite-time stabilization for the closed-loop system. Indeed, system (4.1) satisfies hypothesis (A1) and thus Theorem 3.1 can be applied. Following the proof of Theorem 3.1, it can be shown that for every $\mu > 0$, $l \in Z^+ \cup \{+\infty\}$, there exist functions $\gamma \in K^+$ and $\rho \in K_\infty$ such that the feedback law

$$\begin{aligned} u(t) = v(t) &- f_2(t, x_1(t), x_2(t)) - \frac{\partial f_1}{\partial t}(t, x_1(t)) - \frac{\partial f_1}{\partial x_1}(t, x_1(t)) f_1(t, x_1(t)) \\ &- \frac{\partial f_1}{\partial x_1}(t, x_1(t)) x_2(t) - \mu^{-2} \dot{a}\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) x_1(t+s) ds \\ &+ \mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-2} h'\left(\frac{s}{\mu}\right) x_1(t+s) ds \\ &- \mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) \Big[x_2(t+s) + f_1(t+s, x_1(t+s)) \\ &+ \mu^{-1} a\left(\frac{t+s}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{w}{\mu}\right) x_1(t+s+w) dw \Big] ds, \end{aligned} \tag{4.2}$$

where $a \in C^l(\Re; \Re^+)$ is a periodic function with period 3, with $a(t) = 0$ for $t \in [0, 2]$ and $\int_2^3 a(t)dt = 1$, $h \in C^l(\Re; \Re^+)$ with $h(s) = 0$ for all $s \notin (-2, -1)$, $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1} h(\frac{s}{\mu}) ds = 1$

(i) achieves the dead-beat property of order $T = 14\mu$ for the corresponding closed-loop system with $v(t) \equiv 0$ and

(ii) for every $(t_0, x_0, v) \in \Re^+ \times C^0([-4\mu, 0]; \Re^2) \times C^0(\Re; \Re)$ the solution of the closed-loop system (4.1) with (4.2) satisfies estimate (3.5).

In order to estimate the functions $\gamma \in K^+$ and $\rho \in K_\infty$ we consider the coordinate transformation

$$z_1(t) = x_1(t); z_2(t) = f_1(t, x_1(t)) + \mu^{-1} a\left(\frac{t}{\mu}\right) \int_{t-2\mu}^{t-\mu} \mu^{-1} h\left(\frac{w-t}{\mu}\right) x_1(w)dw + x_2(t). \tag{4.3}$$

The closed-loop system (4.1) with (4.2) is described in z-coordinates by the following set of retarded functional differential equations:

$$\dot{z}_1(t) = -\mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) z_1(t+s)dw + z_2(t);$$
$$\dot{z}_2(t) = -\mu^{-1} a\left(\frac{t}{\mu}\right) \int_{-2\mu}^{-\mu} \mu^{-1} h\left(\frac{s}{\mu}\right) z_2(t+s)dw + v(t).$$

Repeated application of Lemma 2.9 for the above system provides the following estimate:

$$|z(t)| \le 2(1+5\mu)\exp(9L)\sigma\left(\frac{t-t_0}{\mu} - 12\right) \|T_{2\mu}(t_0)z\|_{2\mu} + 100\mu(1+2\mu)\exp(9L) \sup_{t_0 \le s \le t} |v(s)| \quad \forall t \ge t_0, \tag{4.4}$$

where $z := (z_1, z_2)', \sigma(t) := \left\{ \begin{smallmatrix} 1 & if & t < 0 \\ 0 & if & t \ge 0 \end{smallmatrix} \right.$ and $L := \max_{t\in[0,3]} a(t)$. The functions $\gamma \in K^+$ and $\rho \in K_\infty$ can be determined directly from estimate (4.4). For example, consider the case that there exists $b \in K_\infty$ such that $|f_1(t, x_1)| \le b(|x_1|)$ for all $(t, x_1) \in \Re \times \Re$ (this requirement is automatically satisfied if f_1 is independent of time). In this case transformation (4.3) implies the inequalities $|x(t)| \le \beta(\|T_{2\mu}(t)z\|_{2\mu})$ and $|z(t)| \le \beta(\|T_{2\mu}(t)x\|_{2\mu})$, where $\beta(s) := (2 + \mu^{-1}L)s + b(s)$. Using estimate (4.4) and previous inequalities $\|T_{2\mu}(t_0)z\|_{2\mu} \le \beta(\|T_{4\mu}(t_0)x\|_{4\mu})$ (which directly imply the inequality $\|T_{2\mu}(t_0)z\|_{2\mu} \le \beta(\|T_{4\mu}(t_0)x\|_{4\mu})$), we obtain

$$|x(t)| \le \beta\Bigg(2(1+5\mu)\exp(9L)\sigma\left(\frac{t-t_0}{\mu} - 14\right)\beta(\|T_{4\mu}(t_0)x\|_{4\mu}) + 100\mu(1+2\mu)\exp(9L) \sup_{t_0 \le s \le t} |v(s)|\Bigg) \quad \forall t \ge t_0.$$

We conclude that estimate (3.5) holds for the closed-loop system (4.1), (4.2) with $\gamma(t) \equiv 1$ and $\rho(s) := \beta(M(s + \beta(s)))$, where $M := 100(1+5\mu)(1+\mu)\exp(9L)$.

Similarly, we can address the finite-time stabilization problem for the general triangular case (1.1) with no delays, where $f_i \in C^{n-i+1}(\Re \times \Re^i; \Re)$ with $f_i(t, 0, \dots, 0) = 0$ for all $t \in \Re, i = 1, \dots, n$. Proceeding exactly as in the 2-dimensional case, it can be shown that if (1.1) is autonomous then the proposed distributed delay feedback guarantees that estimate (3.5) holds for the corresponding closed-loop system with $\gamma(t) \equiv 1$.

The following example is a triangular autonomous control system with discrete delays.

Example 4.2. Consider the control system:

$$\begin{aligned} &\dot{x}_1(t) = x_1^2(t-\tau_1) + x_2(t-\tau_1); \\ &\dot{x}_2(t) = -2x_1(t)x_1^2(t-\tau_1) - 2x_1(t)x_2(t-\tau_1) + x_2(t-2\tau_2) + u(t-\tau_2), \\ &(x_1(t), x_2(t)) \in \Re^2, u(t) \in \Re. \end{aligned} \tag{4.5}$$

It can be verified that assumption (A1) holds for (4.5) with $\varphi_1(t, x_1) := x_1^2(0)$ and $\varphi_2(t, x_1, x_2) := x_2(-\tau_2)$. Thus Theorem 3.1 guarantees that for every $\mu > \tau_1 + \tau_2$, there exists a 3μ-periodic time-varying distributed delay feedback such that the corresponding closed-loop system satisfies the dead-beat property. The following feedback law:

$$\begin{aligned} u(t) = &-x_2(t-\tau_2) - \mu^{-2}\dot{a}\left(\frac{t+\tau_1+\tau_2}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)x_1(t+\tau_1+\tau_2+s)ds \\ &+ \mu^{-1}a\left(\frac{t+\tau_1+\tau_2}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-2}\dot{h}\left(\frac{s}{\mu}\right)x_1(t+\tau_1+\tau_2+s)ds \\ &- \mu^{-1}a\left(\frac{t+\tau_2}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)\Big[x_2(t+\tau_2+s) + x_1^2(t+\tau_2+s) \\ &+ \mu^{-1}a\left(\frac{t+\tau_1+\tau_2+s}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{w}{\mu}\right)x_1(t+\tau_1+\tau_2+s+w)dw\Big]ds, \end{aligned}$$

where $a \in C^l(\Re; \Re^+)$ is a periodic function with period 3, with $a(t) = 0$ for $t \in [0, 2]$ and $\int_2^3 a(t)dt = 1$, $h \in C^l(\Re; \Re^+)$ with $h(s) = 0$ for all $s \notin (-2, -1)$, $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1}h(\frac{s}{\mu})ds = 1$, achieves the dead-beat property of order $T = 14\mu$ for the corresponding closed-loop system.

Example 4.3. The chain of delayed integrators (1.2), where $\tau_i \geq 0$, has been considered in the literature for the particular case of $\tau_1 = \cdots = \tau_{n-1} = 0$ in [15] with the additional requirement that the stabilizing feedback must be bounded. Here we consider the general case and demand finite-time stabilization of the corresponding closed-loop system. Since assumption (A1) holds with $\varphi_i \equiv 0$ for $i = 1, \ldots, n$, by virtue of Theorem 3.1 we conclude that for every $\mu > \sum_{i=1}^n \tau_i$ there exists a linear 3μ-periodic time-varying distributed delay feedback such that the closed-loop system satisfies the dead-beat property. Notice that there is no limitation on the size of the delays. Specifically, for the case $n = 2$ we obtain the following feedback, which guarantees the dead-beat property of order $T = 14\mu$:

(4.6)

$$\begin{aligned} u(t) = &-\mu^{-2}\dot{a}\left(\frac{t+\tau_1+\tau_2}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)x_1(t+\tau_1+\tau_2+s)ds \\ &+ \mu^{-1}a\left(\frac{t+\tau_1+\tau_2}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-2}h'\left(\frac{s}{\mu}\right)x_1(t+\tau_1+\tau_2+s)ds \\ &- \mu^{-1}a\left(\frac{t+\tau_2}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)\Big[x_2(t+\tau_2+s) \\ &+ \mu^{-1}a\left(\frac{t+\tau_1+\tau_2+s}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{w}{\mu}\right)x_1(t+\tau_1+\tau_2+s+w)dw\Big]ds, \end{aligned}$$

where $a \in C^l(\Re; \Re^+)$ is a periodic function with period 3, with $a(t) = 0$ for $t \in [0, 2]$ and $\int_2^3 a(t)dt = 1$, $h \in C^l(\Re; \Re^+)$ with $h(s) = 0$ for all $s \notin (-2, -1)$, $\int_{-2}^{-1} h(s)ds = \int_{-2\mu}^{-\mu} \mu^{-1}h(\frac{s}{\mu})ds = 1$. Figures 1–3 show the evolution of the state and the input for the C^2 selection:

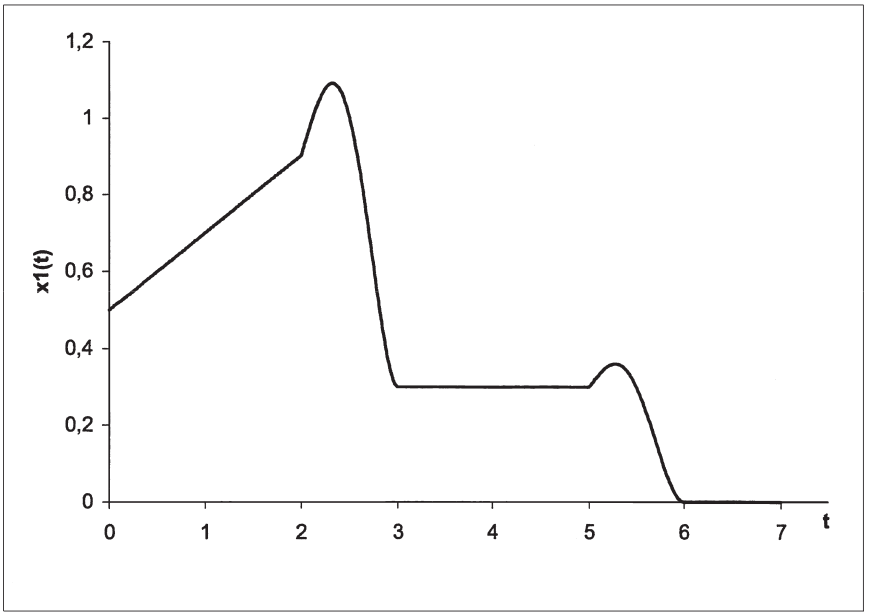


FIG. 1. *Evolution of* $x_1(t)$ *for the case* (1.2) *with* $n = 2$ *and* $\tau_1 = \tau_2 = 0$ *under the feedback law* (4.6), (4.7).

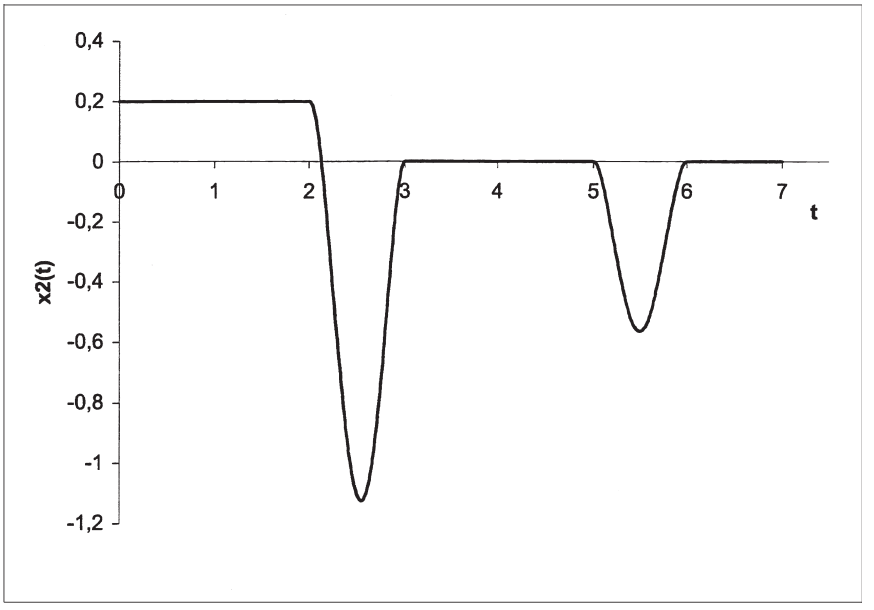


FIG. 2. *Evolution of* $x_2(t)$ *for the case* (1.2) *with* $n = 2$ *and* $\tau_1 = \tau_2 = 0$ *under the feedback law* (4.6), (4.7).

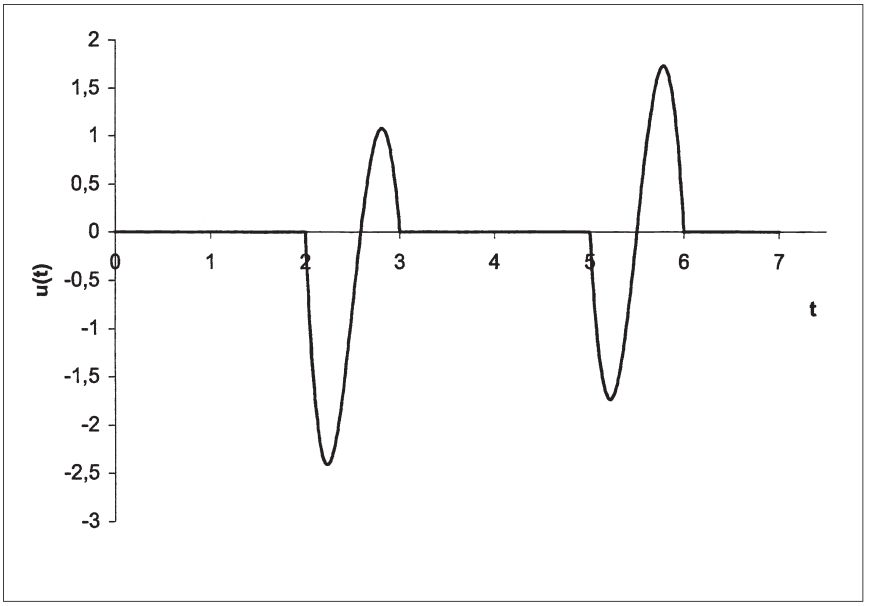


FIG. 3. *Evolution of* $u(t)$ *for the case* (1.2) *with* $n = 2$ *and* $\tau_1 = \tau_2 = 0$ *under the feedback law* (4.6), (4.7).

$$
\begin{aligned}
&a(t) := p\left(t - 3\left[\frac{t}{3}\right]\right) \quad \text{and} \quad h(s) := a(s+4), \text{ where}\\
&p(x) := \begin{cases} 30(x-2)^2(3-x)^2 & \text{if } x \geq 2\\ 0 & \text{if } x < 2 \end{cases}
\end{aligned} \tag{4.7}
$$

for the case (1.2) with $n = 2$ and $\tau_1 = \tau_2 = 0$ with initial condition $x_1(0) = 0.5$ and $x_2(0) = 0.2$.

The case of triangular control systems with delayed drift terms is considered in the following example.

Example 4.4. Consider the planar system

$$
\begin{aligned}
\dot{x}_1(t) &= f_1(t, x_1(t-\tau_{1,1})) + x_2(t);\\
\dot{x}_2(t) &= f_2(t, x_1(t-\tau_{2,1}), x_2(t-\tau_{2,2})) + u(t),\\
&\qquad (x_1(t), x_2(t)) \in \Re^2, u(t) \in \Re, t \geq 0,
\end{aligned} \tag{4.8}
$$

where $\tau_{1,1} > 0, \tau_{2,1} \geq 0, \tau_{2,2} \geq 0$, and $f_1 \in C^2(\Re^+ \times \Re; \Re)$ and $f_2 \in C^1(\Re^+ \times \Re^2; \Re)$ with $f_1(t,0) = f_2(t,0,0) = 0$ for all $t \geq 0$. Clearly, assumption (A2) holds for this system with $R := \tau_{1,1}$, since we have

$$
\begin{aligned}
&x_1(0) f_1(t, x_1(-\tau_{1,1})) - x_2(0)\left(\frac{\partial f_1}{\partial t}(t, x_1(-\tau_{1,1}))\right.\\
&\quad \left. + \frac{\partial f_1}{\partial x_1}(t, x_1(-\tau_{1,1})) f_1(t-\tau_{1,1}, x_1(-2\tau_{1,1})) + \frac{\partial f_1}{\partial x_1}(t, x_1(-\tau_{1,1})) x_2(-\tau_{1,1})\right)\\
&\quad \leq \frac{1}{2}x_1^2(0) + \frac{1}{2}x_2^2(0) + L\left(t, \sup_{-2\tau_{1,1} \leq \theta \leq -\tau_{1,1}} |x(\theta)|\right),
\end{aligned}
$$

where $L(t,s) := \frac{1}{2}\max_{|\xi|\le s}[(f_1(t,\xi_1))^2 + (\frac{\partial f_1}{\partial t}(t,\xi_1) + \frac{\partial f_1}{\partial x_1}(t,\xi_1)f_1(t-\tau_{1,1},\xi_2) + \frac{\partial f_1}{\partial x_1}(t,\xi_1)\xi_3)^2]$. Following the proof of Theorem 3.4, it can be shown that for every $\mu > \tau_{1,1}$ the feedback law

$$\begin{aligned}
u(t) = &-\frac{\partial f_1}{\partial t}(t, x_1(t-\tau_{1,1})) - \frac{\partial f_1}{\partial x_1}(t, x_1(t-\tau_{1,1}))f_1(t-\tau_{1,1}, x_1(t-2\tau_{1,1})) \\
&- \frac{\partial f_1}{\partial x_1}(t, x_1(t-\tau_{1,1}))x_2(t-\tau_{1,1}) - f_2(t, x_1(t-\tau_{2,1}), x_2(t-\tau_{2,2})) \\
&- \mu^{-2}\dot{a}\left(\frac{t}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)x_1(t+s)ds \\
&+ \mu^{-1}a\left(\frac{t}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-2}h'\left(\frac{s}{\mu}\right)x_1(t+s)ds \\
&- \mu^{-1}a\left(\frac{t}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{s}{\mu}\right)\Big[x_2(t+s) + f_1(t+s, x_1(t-\tau_{1,1}+s)) \\
&+ \mu^{-1}a\left(\frac{t+s}{\mu}\right)\int_{-2\mu}^{-\mu}\mu^{-1}h\left(\frac{w}{\mu}\right)x_1(t+s+w)dw\Big]ds
\end{aligned}$$

guarantees the dead-beat property of order $14\mu + 2\tau_{1,1}$. Similarly, we can address the finite-time stabilization problem for the general triangular case with delayed drift terms (1.3), where $\min_{i=1,\dots,n-1}\ \min_{j=1,\dots,i}\tau_{i,j} > 0,\ f_i \in C^{n-i+1}(\Re^+ \times \Re^i; \Re)$ with $f_i(t,0,\dots,0) = 0$ for all $t \ge 0,\ i = 1,\dots,n$.

5. Conclusions. In this paper it is shown that finite-time stabilization by means of time-varying distributed delay feedback is possible for specific classes of systems described by RFDEs. Based on our main results (Theorems 3.1 and 3.4) we have been able to construct locally Lipschitz feedback laws that guarantee the dead-beat property for the corresponding closed-loop systems for important cases such as the general triangular case with no delays, the case of a chain of delayed integrators with no limitation on the size of the delays and the general triangular case with delayed drift terms.

REFERENCES

[1] A. Bacciotti and L. Rosier, *Liapunov Functions and Stability in Control Theory*, Lecture Notes in Control and Information Sciences, 267, Springer-Verlag, London, 2001.

[2] S. P. Bhat and D. S. Bernstein, *Continuous finite-time stabilization of the translational and rotational double integrators*, IEEE Trans. Automat. Control, 43 (1998), pp. 678–682.

[3] S. P. Bhat and D. S. Bernstein, *Finite time stability of continuous autonomous systems*, SIAM J. Control Optim., 38 (2000), pp. 751–766.

[4] J.-M. Coron, *On the stabilization in finite-time of locally controllable systems by means of continuous time-varying feedback law*, SIAM J. Control Optim., 33 (1995), pp. 804–833.

[5] R. Engel and G. Kreisselmeier, *A continuous-time observer which converges in finite time*, IEEE Trans. Automat. Control, 47 (2002), pp. 1202–1204.

[6] V. F. Haimo, *Finite time controllers*, SIAM J. Control Optim., 24 (1986), pp. 760–770.

[7] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[8] Y. Hong, J. Huang, and Y. Xu, *On an output feedback finite-time stabilization problem*, Proceedings of the IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1302–1307.

[9] A. Ilchmann and C. J. Sangwin, *Output feedback stabilization of minimum phase systems by delays*, Systems Control Lett., 52 (2004), pp. 233–245.

[10] M. Jankovic, *Control Lyapunov-Razumikhin functions and robust stabilization of time delay systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1048–1060.

[11] I. Karafyllis and J. Tsinias, *A converse Lyapunov theorem for non-uniform in time global asymptotic stability and its application to feedback stabilization*, SIAM J. Control Optim., 42 (2003), pp. 936–965.
[12] I. Karafyllis, *The non-uniform in time small-gain theorem for a wide class of control systems with outputs*, Eur. J. Control, 10 (2004), pp. 307–323.
[13] M. Krstic, I. Kanellacopoulos, and P. Kokotovic, *Nonlinear and Adaptive Control Design*, Wiley, New York, 1995.
[14] A. Levant, *Higher-order sliding modes, differentiation and output-feedback control*, Internat. J. Control, 76 (2003), pp. 924–941.
[15] F. Mazenc, S. Mondie, and S. I. Niculescu, *Global asymptotic stabilization for chains of integrators with a delay in the input*, IEEE Trans. Automat. Control, 48 (2003), pp. 57–63.
[16] F. Mazenc and P. A. Bliman, *Backstepping design for time-delay nonlinear systems*, Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003.
[17] F. Mazenc and C. Prieur, *Switching using distributed delays*, Proceedings of the 4th IFAC Workshop on Time Delay Systems, Rocquencourt, France, 2003.
[18] F. Mazenc, S. Mondie, and S. I. Niculescu, *Global stabilization of oscillators with bounded delayed input*, Systems Control Lett., 53 (2004), pp. 415–422.
[19] P. H. Menold, R. Findeisen, and F. Allgower, *Finite time convergent observers for nonlinear systems*, Proceedings of the 42nd IEEE Conference on Decision and Control, Hawaii, 2003, pp. 5673–5678.
[20] P. Morin and C. Samson, *Robust point-stabilization of nonlinear affine control systems*, in Stability and Stabilization of Nonlinear Systems, D. Aeyels, F. Lamnabhi-Lagarrigue, and A. van der Schaft, eds., Springer-Verlag, London, 1999, pp. 215–237.
[21] S. I. Niculescu, *Delay Effects on Stability: A Robust Control Approach*, Springer-Verlag, London, 2001.
[22] A. V. Savkin and R. J. Evans, *Hybrid Dynamical Systems*, Birkhauser, Boston, 2002.
[23] E. D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.
[24] J. Tsinias, *Sufficient Lyapunov like conditions for stabilization*, Math. Control Signals and Systems, 2 (1989), pp. 343–357.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 343–367

VELOCITY EXTENSION FOR THE LEVEL-SET METHOD AND MULTIPLE EIGENVALUES IN SHAPE OPTIMIZATION*

FRÉDÉRIC DE GOURNAY†

Abstract. In the context of structural optimization by the level-set method, we propose an extension of the velocity of the underlying Hamilton–Jacobi equation. The gradient method is endowed with a Hilbertian structure based on the H^1 Sobolev space. Numerical results for compliance minimization and mechanism design show a strong improvement of the rate of convergence of the level-set method. Another important application is the optimization of multiple eigenvalues.

Key words. multiple eigenvalues, shape optimization, sensitivity analysis, shape derivative, level-set method, regularization

AMS subject classifications. 74P10, 49Q12, 74H45, 90C90, 74S30

DOI. 10.1137/050624108

1. Introduction.

Optimal design of elastic structures has been widely studied and many different numerical methods are used to solve this problem. Most of the existing methods can be divided into two main classes: topology optimization which optimizes a density of material in each cell and geometric optimization which moves the boundary of the domain.

The most recent topology method, the homogenization method and its variants (power-law method or solid isotropic microstructure with penalization (SIMP) method), may be considered quite classical in view of the number of publications (see, e.g., [1], [2], [6], [8], [7], [12]). The homogenization method seems the most promising because it is independent with respect to initialization and because it gives a strong mathematical result of the existence of solution. Sadly this method bears the difficulty of handling quite badly eigenvalue problems where apparitions of so-called fictitious-material modes prevent stable numerical computation. Another problem is that topology methods give optimal shapes that are composite. Penalization methods allow one to project the composite shape on a classical shape (a black-and-white design).

The problem of fictitious modes is naturally solved with the geometric methods where the shape is clearly identified and the void cannot generate fictitious modes.

The major drawback of geometric methods is their dependency with respect to initialization. Even the most recent level-set method ([4], [5], [34]) is very sensitive to initialization although the topology can change. In order to avoid this problem, a method of topology optimization, the bubble method (or topological gradient method [16], [17], [31], [32]), has been recently coupled with geometric optimization (see [3], [11]). The most recent method of geometric optimization is developed in [23]. But this method has not been numerically implemented yet.

We wish here to correct the so-called void problem that arises when using the level-set algorithm. This problem is generated by the weak material approximation

*Received by the editors February 9, 2005; accepted for publication (in revised form) October 21, 2005; published electronically March 15, 2006.
http://www.siam.org/journals/sicon/45-1/62410.html
†Centre de Mathématiques Appliquées (UMR 7641), Ecole Polytechnique, 91128 Palaiseau, France (degourna@cmapx.polytechnique.fr).

that gives rise to a negative velocity for advecting the shape in the void. The void problem slows down the algorithm when the mesh is refined.

The void problem is presented in section 3, and the chosen solution is presented in section 4. The method mainly consists of applying the Δ^{-1} operator to the velocity which is costless. This method allows one to regularize and to extend and gives a local Hilbertian structure (see the discussion in section 4.2). Numerical results are presented in sections 4.3–4.7; they compare the new method with the previous one and show strong improvements of the level-set algorithm.

The method also has made it possible to deal with the problem of optimizing an eigenvalue when its multiplicity is greater than 1. Theoretical differentiation of the eigenvalue is made in section 5. As can be expected from theory on eigenvalue differentiation, the first eigenvalue is directionally differentiable with respect to shape variation, even when this eigenvalue is multiple. The algorithm that optimizes the first eigenvalue is detailed in section 5.3. It strongly relies on the local Hilbertian structure given by velocity regularization. Indeed, because of this local Hilbertian structure, the differential gives rize to a gradient and therefore to a steepest descent. We show that the choice of the steepest descent is a semidefinite program in low dimension that is easily solved. Some numerical results are presented in section 6.

The method used for optimizing multiple eigenvalues can be extended to other criteria that are not differentiable as the robust compliance criterion in the sense of [13]. The method of velocity regularization presented here is an adaptation of [10], [26], [21]. It is a standard issue in numerical computation.

2. The level-set method in shape optimization.

2.1. Eigenvalue maximization. We set our model problem in linearized elasticity. Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) be a bounded open set occupied by a linear isotropic elastic material with Hooke's law A_0. Recall that, for any symmetric matrix ξ, A_0 is defined by

$$A_0\xi = 2\mu\xi + \lambda\big(\mathrm{Tr}\xi\big)Id,$$

where μ and λ are the Lamé coefficients of the material. The boundary of Ω is made of two disjoint parts

$$\partial\Omega = \Gamma_N \cup \Gamma_D, \tag{1}$$

with Dirichlet boundary conditions on Γ_D and Neumann boundary conditions on Γ_N. The spectral problem is to find an eigenvalue γ and an eigenvector $u \neq 0$ such that

$$\left\{\begin{array}{rcll} -\mathrm{div}\,(\mathrm{A}\,\mathrm{e}(\mathrm{u})) & = & \gamma\rho u & \text{in } \Omega, \\ u & = & 0 & \text{on } \Gamma_D, \\ \big(A\,e(u)\big)n & = & 0 & \text{on } \Gamma_N, \end{array}\right. \tag{2}$$

where ρ is a scalar field on Ω that represents the material density (typically ρ is equal to 1 on Ω).

It is well known that $\mathrm{Sp}(\Omega)$ the set of eigenvalues is a countable set of positive numbers that tends to infinity. The smallest eigenvalue $\gamma_1(\Omega) = \min \mathrm{Sp}(\Omega)$ can then be defined.

A classical way of improving the rigidity of a structure is to maximize the first eigenfrequency. Thus, a natural objective function to be minimized is

$$\mathcal{L}(\Omega) = -\gamma_1(\Omega) + \eta|\Omega|, \tag{3}$$

where η is a given Lagrange multiplier for a volume constraint. We want to minimize $\mathcal{L}$ with respect to Ω with a constraint that $\Omega \subset D$, where D is a given domain of $\mathbb{R}^d$.

2.2. Classical algorithm. The works [24], [29], or [5] give an extensive explanation of the level-set method. Only the main ideas are to be reviewed in this section.

As described in [4], when the dimension of the eigenspace associated to γ_1 is equal to 1, the above functional $\mathcal{L}(\Omega)$ is differentiable with respect to variation of the domain and the geometrical shape optimization method can be applied. It reads, as follows:

1. **Calculation of the gradient.** Let Ω_k be the domain at iteration k. Assuming that $\gamma_1(\Omega_k)$ is simple, for a given $\theta \in W^{1,\infty}(D;D)$ define

$$(Id+\theta)\circ\Omega_k = \{x+\theta(x) \text{ with } x \in \Omega_k\}.$$

Prove that $\mathcal{L}((Id+\theta)\circ\Omega_k)$ is differentiable with respect to θ at the point $\theta = 0$. The value of $\mathcal{L}'$, the differential at the point 0 is given by

$$\mathcal{L}'(\theta) = \int_{\partial\Omega_k} (\theta\cdot n)(-v+\eta) \tag{4}$$

$$\text{with} \quad \begin{cases} v &= Ae(u)\colon e(u) - \gamma_1 \rho u\cdot u & \text{on} \quad \Gamma_N^k, \\ v &= -Ae(u)\colon e(u) & \text{on} \quad \Gamma_D^k, \end{cases} \tag{5}$$

where u is an eigenvector normalized by $\int_{\Omega_k} \rho u\cdot u = 1$ which is, up to a change of sign, by assumption, unique and where n is the outer normal of Ω_k.

2. **Calculation of a descent direction.** Choose θ_k such that $\mathcal{L}'(\theta_k) < 0$, and let $\Omega_{k+1} = (Id + t\theta_k)\circ\Omega_k$, where t is the step of the gradient method.

The level-set method is a geometrical shape optimization method where the domains Ω_k are represented through functions ϕ_k defined on D in the following way: $x \in \Omega_k \Longleftrightarrow \phi_k(x) < 0$. Such a ϕ_k is said to be "a level set function of Ω_k." Of course for a given domain, the associated level-set functions are not unique.

The changes of topology are then handled in a very simple way (see [5] for extensive explanations of numerical advantages). The method relies on the following lemma.

LEMMA 2.1. *If ϕ_k is a level-set function of Ω_k, define ϕ_{k+1} as*

$$\phi_{k+1} - \phi_k + t(\theta_k\cdot n)|\nabla\phi_k| = 0.$$

Then there exists an $O(t^2)$ such that $\phi_{k+1} + O(t^2)$ is a level-set function for the domains $\Omega_{k+1} = (Id + t\theta_k)\circ\Omega_k$.

Thus, the following scheme for choosing ϕ_{k+1} is the most commonly used:

$$\begin{aligned} \phi(0) &= \phi_k, \\ \frac{\partial\phi}{\partial t} + V^*|\nabla\phi| &= 0, \\ \phi_{k+1} &= \phi(T), \end{aligned} \tag{6}$$

where $T > 0$ is the step of the gradient method and V^* is the descent direction chosen according to the calculation (4) of the differential of J. Defining $V^* = v - \eta$ (v being defined in (4)) is the most commonly used choice; we will call this choice the natural extension method. The goal of this paper is to find a different way of defining V^*.

Remark 2.2. In order to avoid multiple definitions of V^*in the natural extension method, it is supposed that the Dirichlet part of the boundary is fixed and that the v used everywhere in the domain is the one defined for Γ_N.

3. The weak material approximation. The stiffness matrix that corresponds to the linear operator of elasticity is not defined on the nodes that does not belong to Ω, i.e., on $x \in D$ such that $\phi(x) > 0$. In other words, this matrix is not invertible. In order to avoid this problem, the so-called weak material (or ersatz material) approximation consists of fixing Hooke's law A and the material density ρ as

$$\begin{array}{rclcrcll} A & = & A_0 & \text{and} & \rho & = & 1 & \text{in} \quad \Omega \\ A & = & \varepsilon A_0 & \text{and} & \rho & = & \varepsilon^\alpha & \text{outside} \quad \Omega \end{array}$$

with a small parameter ε and $\alpha \geq 1$. The fictitious modes are avoided by setting $\alpha = \infty$. In the continuous case, it has been proven (see [28]) that the eigenvector $\tilde{u}$ calculated by using the ersatz material approximation is equal (at first order in ε) to the desired eigenvector u introduced in (2). In $D \setminus \Omega$ $\tilde{u}$ is the lifting of the Dirichlet condition $\tilde{u} = u$ on $\partial\Omega$ (at first order in ε).

Recalling that the Dirichlet part of the boundary of $\partial\Omega$ has been fixed (see Remark 2.2), $v - \eta = Ae(u) : e(u) - \gamma_1 \rho u \cdot u - \eta$ is defined everywhere on D, and the natural extension method consists of defining $V^* = Ae(u) : e(u) - \gamma_1 \rho u \cdot u - \eta$ on every cell.

This nevertheless raises a problem: since u is everywhere of order ε^0, then $Ae(u) : e(u)$ is of order ε outside Ω which means that the velocity extension is almost equal to $-\eta$ outside the domain and makes it very difficult for the shape to increase its volume. Even if the descent step T is increased in order to speed up the method, the parts of the shape where there is a need to decrease the volume will move faster than the parts of the shape where there is a demand on increasing the volume.

In numerical result, it can be seen that between each computation of the eigenvalue, the level-set method cannot move $\partial\Omega$ of more than one cell away from the original boundary when it wants to improve the volume. This leads to a drastically increasing computational time with mesh-refinement.

This remark is true, of course, for every objective function and not only for the minimization of the first eigenvalue.

4. Velocity regularization by the Hilbertian method.

4.1. Definition of the Hilbertian method. We will suppose that the domain Ω has enough regularity so that $v(\cdot, \cdot)$ defined in (4) belongs to $H^{-1/2}(\partial\Omega)$. Even if the optimal domain may possibly be irregular, physical and numerical intelligence tells that throughout optimization $v(\cdot, \cdot)$ has the required regularity but mathematical proof of this fact is still lacking. Let us first define a scalar product.

DEFINITION 4.1. *For $a \in \mathbb{R}^{*+}$, define the following scalar product on $H^1(D)$:*

$$(u, w)_{H^1} = \int_D a \nabla u \cdot \nabla w + uw$$

with the associated norm $\| \cdot \|_{H^1}$.

The velocity V^* in the Hamilton–Jacobi equation (6) is chosen as the unique solution to

$$\int_{\partial\Omega} V^*(-v + \eta) = \min_{\substack{V \in H^1(D) \\ \|V\|_{H^1} = 1}} \int_{\partial\Omega} V(-v + \eta), \tag{7}$$

where v is the differential defined in (4) and Ω is the actual working domain.

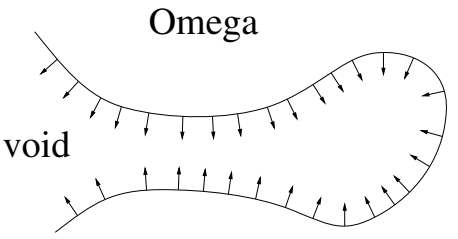


FIG. 1. *Two merging parts of the boundary leads to a nonregular vector field as seen as a distribution on D.*

4.2. Advantages of the method.

Scalar versus vector. We choose to extend the normal velocity which is a scalar field, i.e., if θ is the vector field that advects the domain, its normal velocity is equal to $\theta \cdot n$ on the boundary of the domain. Seen as a distribution over D, the normal velocity is more regular than the vector velocity. A good example is when two parts of the boundary want to merge: $\theta \cdot n$ is positive on the two parts and θ changes orientation. This situation gets worse at the next step of the algorithm (see Figure 1). That is why the normal velocity is extended and not the vector velocity.

How to compute V^*. The problem of finding V^* is not difficult. Let $\bar{V}$ be the unique solution to

$$(\bar{V}, X)_{H^1} = \int_{\partial\Omega} X(-v+\eta) \quad \forall X \in H^1(D),$$

where $(\cdot,\cdot)_{H^1}$ is the scalar product of Definition 4.1. Then $V^* = -\frac{\bar{V}}{\|\bar{V}\|_{H^1}}$ and the inversion of the matrix that corresponds to the scalar product $(\cdot,\cdot)_{H^1}$ has to be done only once in the optimization process.

Extension, regularization, and Hilbertian structure. Three different goals are sought in the Hilbertian extension of the velocity.

First, the formula for the differential of $\mathcal{L}$ gives a velocity that makes sense only on the boundary of the domain, and the Hamilton–Jacobi equation (6) needs a speed defined everywhere on the domain or else the algorithm cannot move the boundary of more than one cell during optimization. This is the extension issue.

Secondly, the velocity is regularized by being diffused with the scalar product defined in Definition 4.1. This is the regularization issue. It is expected to increase the accuracy of the algorithm and is a standard issue in optimization problems (see, e.g., [26], [10], [21]).

Thirdly, the problem is endowed with a Hilbertian structure, and we work with gradient-type methods. This issue will be developed later in section 5.

Hilbertian extension versus other extensions. The most natural way to extend the velocity outside $\partial\Omega$ would have been to extend $v - \eta$ according to the normal of $\partial\Omega$ by a front-propagating-type method such as the fast-marching method described in [25]. This method does not endow the space with a Hilbertian structure. Neither does it regularize the velocity.

V^* is indeed an extension. Because the scalar product diffuses the source term, the velocity is now defined everywhere on D, and the typical problem of null velocity in the void that raises with the natural extension method is now cured.

V^* is indeed a regularization. V^*, the speed used in the Hilbertian extension method, is more regular than $v - \eta$, the speed used in the natural extension method as can be seen in the following formal derivation.

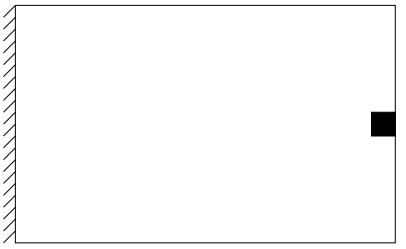


FIG. 2. *Boundary conditions for a two-dimensional (2d) cantilever (the black zone is heavier and not subject to optimization).*

If v belongs to an $H^s(D)$, then $X \mapsto \int_{\partial\Omega} X(-v+\eta)$ is a linear form on $H^{-s+1/2}(D)$, and V^* belongs to an $H^{s+3/2}(D)$ by elliptic regularity on the very smooth domain D. Indeed V^* is obtained by operating an inverse Laplacian on the distribution (on D) $X \mapsto \int_{\partial\Omega} X(-v+\eta)$ and so gains two derivatives.

Consistency of the extension. This algorithm could be seen as a gradient-type algorithm if there were a vector field $\theta^* \in W^{1,\infty}(D;D)$ such that $\theta^* \cdot n = V^*$ on $\partial\Omega$. Two hypotheses would then be needed: a certain regularity on V^* and a certain regularity of the domain Ω itself.

In order to ensure regularity of V^*, a scalar product on H^p, $p \geq 1$, could have been used instead of a scalar product on H^1. In this case, formally, V^* would have $2p - 1/2$ more derivatives in L^2 than v. But numerical computation of V^* may be quite difficult. Indeed, the computation of the matrix of the scalar product of H^p needs finite elements that are chosen according to p.

The extension parameters. The coefficient a characterizes the diffusion of $v-\eta$ in the sense that setting it small compared to 1 will lead to a solution V^* which is pointwise almost equal to $v-\eta$ on $\partial\Omega$ and equal to 0 outside $\partial\Omega$. It must be set small enough so that a big value of $|v-\eta|$ on one part of the boundary does not interfere too much with the values of $v-\eta$ on the other parts of the boundary. But it must be set big enough in order to diffuse the value of $v-\eta$ outside the boundary of Ω.

4.3. Numerical example: Eigenvalues of a cantilever. We study a medium-cantilever problem. The working domain is a 2×1 rectangle, with zero displacement condition on the left side and a small square region at the middle of the right side (see Figure 2) which is 500 times heavier and not subject to optimization. This heavier mass allows one to avoid trivial shapes. Figure 3 is a display of the initialization (left) and the optimal shape at the end of the optimization process (right). The Young modulus of the material is set to 1, the Poisson ratio ν to 0.3, and the Lagrange multiplier to 7×10^{-2}. In the void, the ersatz material has a Young modulus set to 10^{-5} and a density ρ set to 0. The mesh is refined alternatively in each direction by a factor of 2, and the number of transport iterations is adequately increased at each mesh refinement (see the table below).

Explanation of Table 1. Table 1 lists, for different meshes, the number of transport iterations used for each optimization step in the Hamilton–Jacobi equation, the average computed time used per iteration step of the gradient method, and the global computing time. The number of transport iterations is shown for the Hilbertian method only. The number of transport iterations for the natural method is 16 times the number of transport iterations of the Hilbertian method. This explains why the average time per iteration of the gradient algorithm is bigger for the natural method than for the Hilbertian. The last column is the global computing time for obtaining

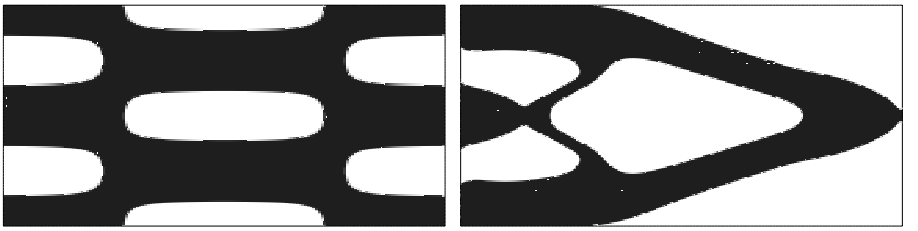


FIG. 3. *Initialization and optimal shape for the first eigenvalue of a cantilever.*

TABLE 1
Data of the numerical test.

Mesh	N^o of transport	Time per iteration Hilbertian / natural	Global time Hilbertian / natural
81×41	16	1.46 / 1.62	30.68 / 50.13
161×81	32	9.35 / 11.37	233.87 / 330.01
321×161	64	72.17 / 94.91	2237.2 / 2657.63
41×41	14	0.74 / 0.84	13.26 / 22.6
81×81	28	4.46 / 6.19	120.62 / 136.1
161×161	56	38.50 / 46.61	1116.57 / 1771.44

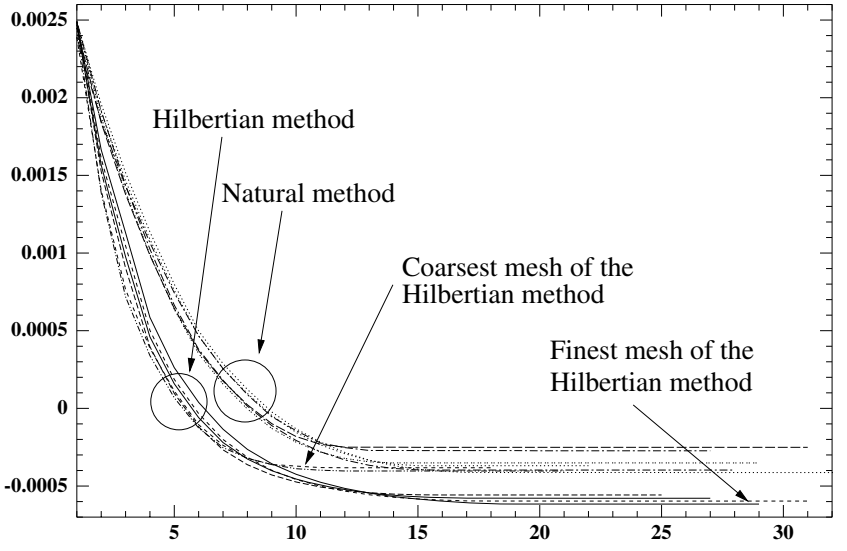


FIG. 4. *Mesh-refinement influence on the velocity for the natural extension (left) and the Hilbertian extension (right).*

the optimal shape. The Hilbertian method takes the same amount of time as the natural method to obtain the optimal shape but only because it is more accurate. As can be seen in Figure 4, the convergence curves of the Hilbertian method are indeed better than the convergence curves of the natural method. Time is given in seconds.

Explanation of Figure 4. There is two sets of curves in Figure 4. The better ones are obtained with the Hilbertian method. It appears that the more the mesh is refined, the better the optimal shape is, even if, when the mesh is refined, the

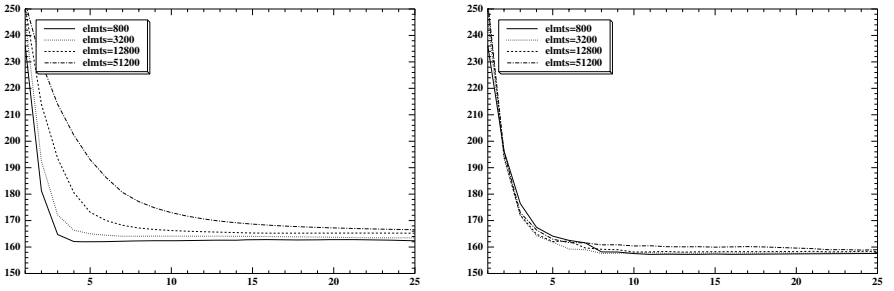


FIG. 5. *Mesh-refinement influence on the natural extension (left) and the Hilbertian extension (right).*

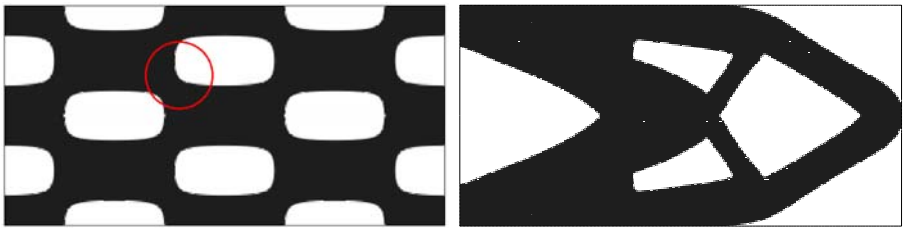


FIG. 6. *Initialization and optimal shape of the cantilever.*

decreasing of the smallest eigenvalue leads to an increase of the criterion. This is easily explained by a better accuracy on the optimal shape itself.

Remark 4.2. In order to prove mesh-consistency of the algorithm and to have comparable curves in Figure 4, the optimal shape has to be the same when the mesh is changed. Thus those examples are chosen so that there is no possibility for the algorithm to create thin structures when the mesh is refined. This explains why there are a very few holes in the initialization. It is well known that the level-set algorithm can produce more complicated structures.

4.4. Numerical example: The compliance of the cantilever. We performed our new velocity extension method on the well-known cantilever problem which is fixed on the left wall and supports a unit vertical point load in the middle of the right wall. The compliance is here optimized. The working domain size is 2×1. The Young modulus is set to 1, and the Poisson ratio to 0.3. The Lagrange multiplier is set to $\eta = 100$.

Optimization is performed for several finer meshes, and the number of transport iterations is multiplied by 2 as each square of the mesh is cut into 4 squares. For the finest mesh (321×161) that corresponds to 51200 elements; the number of transport iterations is equal to 128 for the natural extension and to 16 for the Hilbertian. As a result, the Hilbertian extension is really quicker.

Mesh-refinement influence. The curves in Figure 5 show that the Hilbertian method is less sensitive to mesh-refinement than the natural method. Because there are parts of the boundary that have to increase the volume (one of them is circled in Figure 6) and the natural extension method has problems to improve these parts as was said in section 3, the natural extension method is sensitive to mesh-refinement. In Figure 4 there was no such demand on improving the volume and the mesh-independence of the Hilbertian method was less obvious.

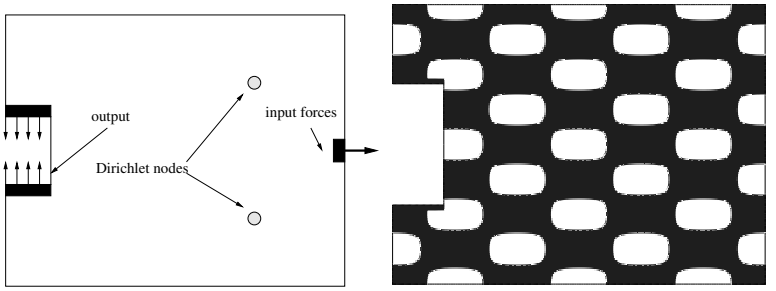


FIG. 7. *The definition of the* 2*d gripping mechanism and its initialization.*

4.5. Numerical example: A two-dimensional (2d) gripping-mechanism. On a 1.2×3.2 rectangle meshed with 241×641 nodes, we give a numerical example for minimizing the least square of a prescribed displacement. This example was given by the Comissariat à l'Energie Atomique for the design of a grip (see [9]). The objective function is

$$\mathcal{J}_{lse}(\Omega) = \int_{\Omega} k(x)|u(x) - u_0(x)|^2, \tag{8}$$

where u is the displacement obtained by a given set of forces, u_0 is a prescribed given displacement, and k is a scalar field that characterizes the zone of interest.

The ponderation k is equal to 10 on the left black box in Figure 7 (left), is equal to 1 on the right black box, and is equal to 0 elsewhere. The prescribed displacement u_0 is equal to $(0, \pm 1)$ on the left black box, and $(0, 0)$ on the right box. Enforcing u to be close to 0 where the force is input allows one to ensure the stiffness and the connectivity of the structure. These black boxes are not subject to optimization.

A force of modulus 1 N is applied in the x-direction on the middle of the left, and a uniform pressure which represents a total force of modulus 5×10^{-2} N is applied between the jaws of the mechanism in order to force the mechanism to hold objects. The prescribed displacement is located on the black box on the left as shown in Figure 7. The Young modulus is set to 1, the Poisson ratio to 0.3, and there is a small Lagrange multiplier of a volume constraint of value 0.05. This Lagrange multiplier helps remove parts of the boundary that are useless.

This is a typical problem where an adjoint is needed. For some reason that is not completely understood, the ratio of the weak material (the factor ε in section 3) cannot be set too low (typically it must be at least 1 percent of the strong material) or the algorithm will not work. One of the explanations that may be given is the tendency of the shape to create hinges. The algorithm then concentrates on the hinges only, ignoring the rest of the shape, and it is believed it is then stuck in a local minimum. If the ratio of the weak material is high, hinges are less efficient and the previous problem is avoided.

Optimization must then be made in two steps. First the shape is optimized with a ratio of the weak material equal to 1 percent. The optimal shape is then reoptimized with a smaller ratio (10^{-5} in this numerical case). In the second optimization procedure, the displacement and the adjoint state are calculated with a more accurate precision which leads to a better precision of the shape-derivative.

In Figure 8 (left), it may seem that the Hilbertian method is less efficient than the natural method. This is explained by the lack of precision in the computation of the

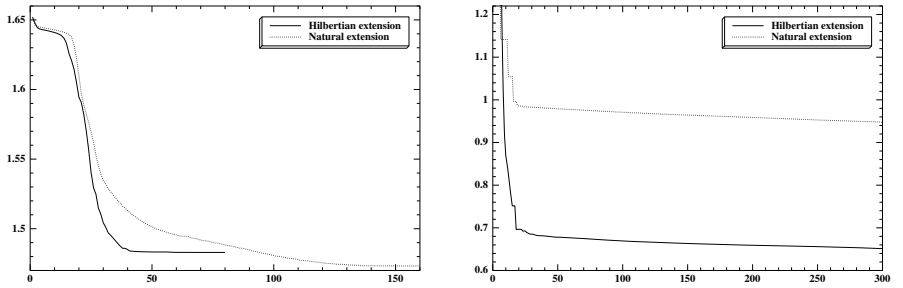


FIG. 8. *Evolution of the objective function for the first step (left) with a ratio for the weak material equal to* 0.01 *and the second step with a ratio of* 10^{-5} *(right).*

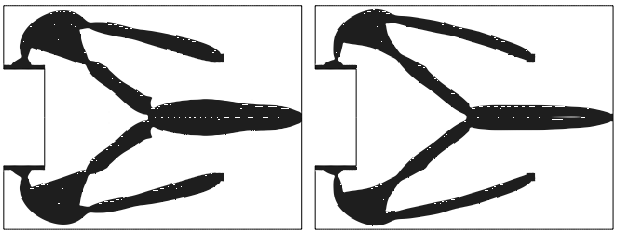


FIG. 9. *Optimal shape for the Hilbertian method (left) and the natural method (right) at the end of the first step of optimization.*

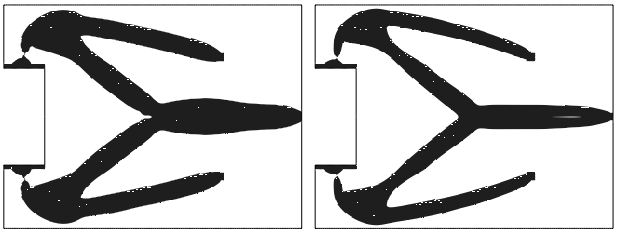


FIG. 10. *Optimal shape for the Hilbertian method (left) and the natural method (right) at the end of the second step of optimization.*

shape-derivative in the first step of the algorithm. But it is also explained by the lack of precision of the computation of the criterion during the first step of the algorithm. Indeed when the two shapes of Figure 9 (obtained with $\varepsilon = 10^{-2}$) are computed using $\varepsilon = 10^{-5}$, the Hilbertian shape is the better one. Figure 10 displays the optimal shape for the two different methods and Figure 11 show how the grip works.

Remark 4.3. The optimal shape for the natural extension method has less volume than the one for the Hilbertian extension (see Figure 10) it is a numerical validation of the problem raised in section 3.

4.6. Numerical example: More 2d mechanisms. We briefly present here some more 2d mechanisms: namely, a negative Poisson modulus cell (Figures 12 to 14) and a force inverter (Figures 15 to 17). These examples are standards of shape optimization problems, and their description can be found in [8]. In order to ensure stiffness of the structures, a small pressure load is applied where the displacement is optimized (on the right black box of Figure 15 for the force inverter and at the

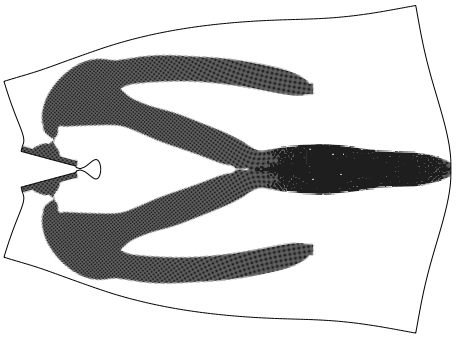


FIG. 11. *Optimal shape displacement.*

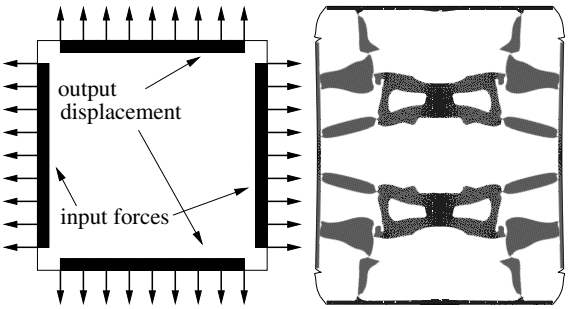


FIG. 12. *The Negative Poisson modulus problem and its deformed solution.*

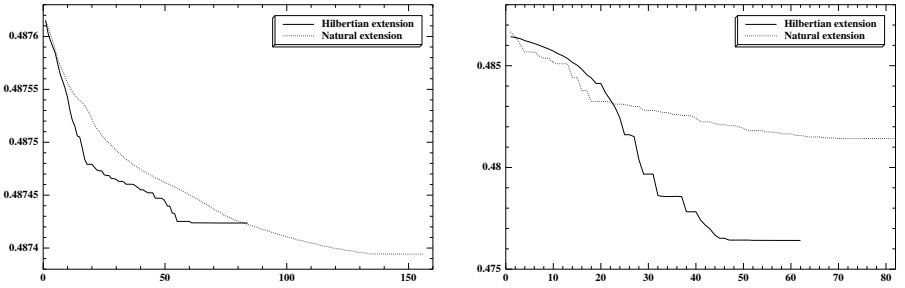


FIG. 13. *Negative Poisson modulus's evolution of the objective function at the first step (left) with a ratio for the weak material equal to* 0.01 *and the second step with a ratio of* 10^{-5} *(right).*

top and bottom of Figure 12 for the negative Poisson modulus mechanism). At the location of the input forces (left black box for the force inverter or the left and right walls for the negative Poisson modulus mechanism), the displacement is enforced to be close to zero in order to ensure connectivity of the shapes. Optimization is made in two steps like the 2d grip of section 4.5, and the behavior of the Hilbertian extension with respect to the natural extension is comparable.

4.7. Numerical example: A three-dimensional (3d) gripping mechanism. The objective function is defined by (8) in section 4.5. The working domain is a $3 \times 2 \times 6$ rectangle. A uniform pressure load is applied on the plane $x = 3$, and the

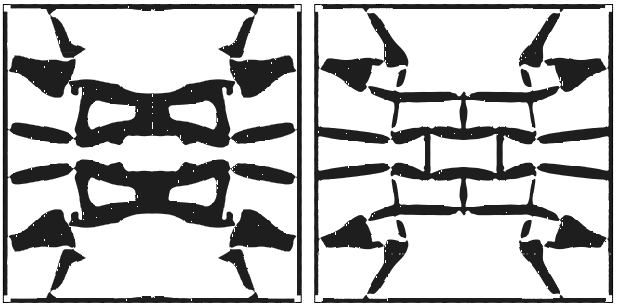


FIG. 14. *Negative Poisson modulus's optimal shape for the Hilbertian method (left) and the natural method (right) at the end of optimization.*

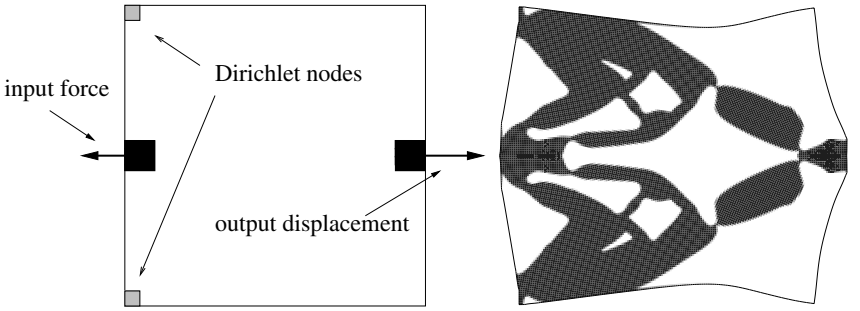


FIG. 15. *The force inverter problem and its deformed solution.*

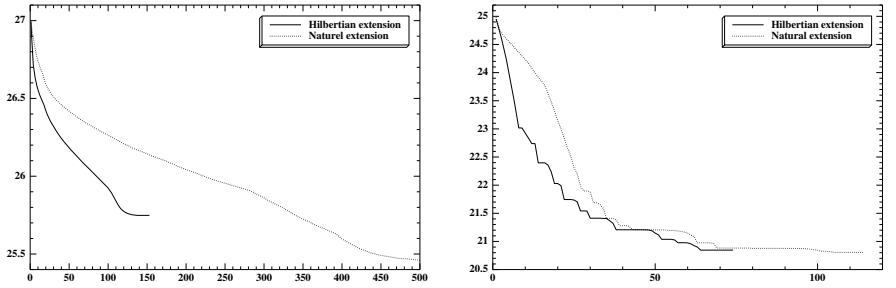


FIG. 16. *Force inverter's evolution of the objective function for the first step (left) with a ratio for the weak material equal to* 0.01 *and the second step with a ratio of* 10^{-5} *(right).*

prescribed displacement is localized on a box at the opposite side (see Figure 18 where a cross-section at $y = 0$ is shown). A uniform pressure load (of order 60 percent of the one on the plane $x = 3$) is also imposed between the jaws of the mechanism so that this mechanism is designed to hold and grip. The Poisson ratio is 0.3, and the Young modulus is 1. The Lagrange multiplier is set to 3, and the mesh used is $31 \times 21 \times 61$. The ratio of the weak material is set to 10^{-5}. There is no need here to perform the 2-step optimization of section 4.5. The reason why things seems to be simpler in 3d is that the changes of topology and the hinges are not of the same nature as in 2d; consequently, it is believed that throughout the process of optimization, the objective

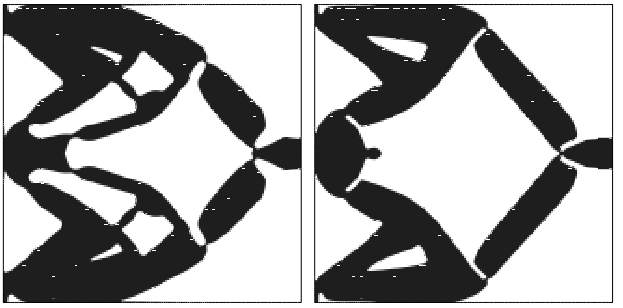


FIG. 17. *Force inverter's optimal shape for the Hilbertian method (left) and the natural method (right) at the end of optimization.*

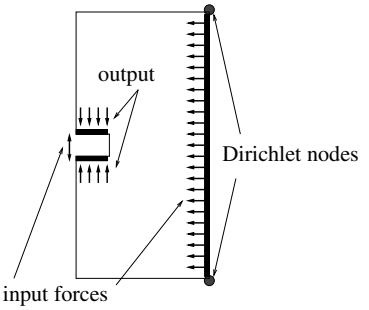


FIG. 18. *The problem of the* 3*d gripping mechanism.*

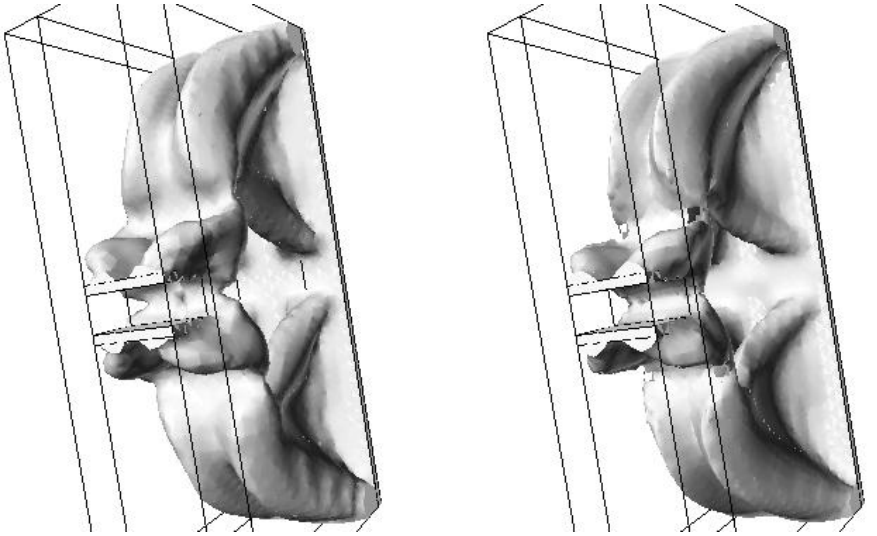


FIG. 19. *Optimal shape for the Hilbertian method (left) and the natural method (right) (isovalue* 0 *of the level-set is shown).*

function is more regular in 3d. Figure 19 displays the resulting optimal shape for the two different methods, as Figure 20 shows the convergence history for both method (left) and how the grip actually works (right).

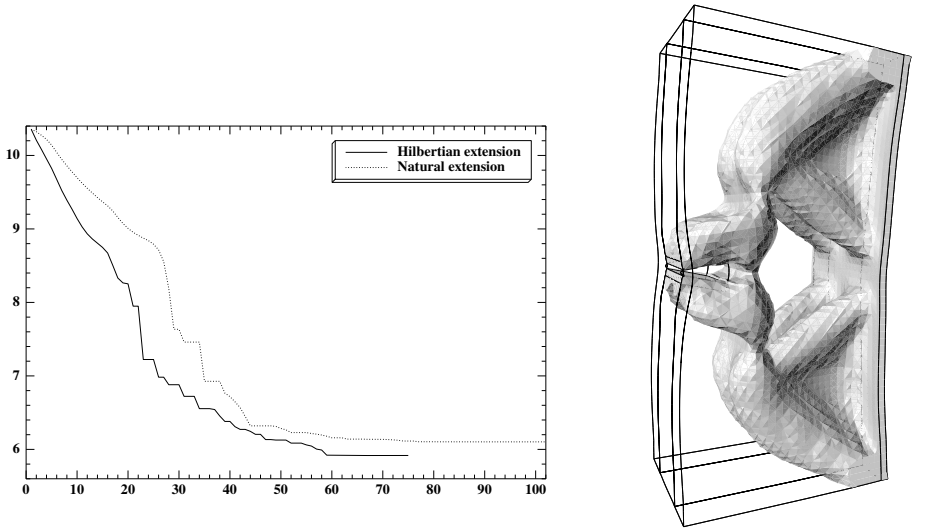


FIG. 20. *Objective function (left) and displacement of the Hilbertian shape (right) (density of material* ≥ 0.3 *is shown).*

5. Optimizing multiple eigenvalues. The development above for optimizing the first eigenvalue stands only when this eigenvalue is of multiplicity equal to 1. When this is not the case, there is no more differentiability of the first eigenvalue with respect to θ, and the above method cannot be applied. Nevertheless it has been proven that $\mathcal{L}$ is directionally differentiable. The velocity extension which endows the space with a Hilbertian structure allows one to find a direction of descent. The algorithm used is now a subgradient-type algorithm. The goal of this section is two-fold:

1. Calculate $\mathcal{L}'(\theta) : W^{1,\infty}(D;D) \to \mathbb{R}$ the nonlinear collection of directional derivative of $\mathcal{L}$. Show that according to Hadamard's structure theorem $\mathcal{L}'(\theta)$ depends only on the value of $\theta \cdot n$ on $\partial\Omega$. Let us denote $j(\theta \cdot n) = \mathcal{L}'(\theta)$.

2. Compute V^* such that $j(V^*) = \min_{\|V\|=1} j(V)$, and advance the domain according to V^*. It is in the computation of V^* that the Hilbertian structure is compulsorily needed.

5.1. A general theorem about eigenvalue differentiation. Differentiating eigenvalues when they are multiple is nowadays quite standard. Two different approaches exist: the one of [19] for the foundations or [27] and [20], or the one of [15] or [18] using the subgradient theory of Clarke [14]. We shall only state their result here.

DEFINITION 5.1. *Define* Ω_0 *as the actual working domain.*

$\diamond$ *Let* $\alpha > 0$, $\beta_1 > 0$ *be constants and* $\mathbb{L}$ *be the space of linear unbounded auto-adjoint operators from* $L^2(\Omega_0) \to L^2(\Omega_0)$ *such that*

$$\forall L \in \mathbb{L}, \quad \forall u \in H^1_D(\Omega_0) \quad \beta_1 \|u\|^2_{H^1_D(\Omega_0)} \geq (Lu,u)_{L^2} \geq \alpha \|u\|^2_{H^1_D(\Omega_0)},$$

i.e., $\mathbb{L}$ *is made of operators "uniformly" coercive with constant* α *and uniformly continuous with constant* β_1

$\diamond$ *Let* $\beta_2 > 0$ *be a constant and* $\mathbb{M}$ *be the space of continuous linear auto-adjoint operators from* $L^2(\Omega_0) \to L^2(\Omega_0)$ *uniformly continuous with constant* β_2*, i.e.,*

$$\forall M \in \mathbb{M}, \quad \forall u \in L^2(\Omega_0) \quad \beta_2 \|u\|^2_{L^2(\Omega_0)} \geq (Mu,u)_{L^2}.$$

◇ *Define the norm in those two spaces as follows: If* $N \in \mathbb{L}$ *or* $N \in \mathbb{M}$*, then*

$$\|N\| = \max_{u \in H^1_D} \frac{(Nu,u)_{L^2}}{\|u\|^2_{H^1_D}},$$

i.e., we endow $\mathbb{M}$ *with the natural norm of* $\mathbb{L}$*.*

◇ *Let*

$$\begin{array}{ccc} U \subset W^{1,\infty}(D;D) & \to & \mathbb{L} \times \mathbb{M} \\ \theta & \mapsto & (L(\theta), M(\theta)) \end{array}$$

be a Fréchet differentiable mapping with respect to the norm just defined. Define $L'(\theta_0)$ *(resp.* $M'(\theta_0)$*) the differential of* $L(\theta)$ *(resp.* $M(\theta)$*) with respect to* θ *at the point* $\theta = 0$ *applied to* θ_0*.*

THEOREM 5.2. *Define* $\gamma_1(\theta)$ *the smallest eigenvalue of the generalized eigenproblem* $L(\theta)u = \gamma M(\theta)u$ *and* E_θ *its eigenspace,* $E_\theta = Ker(L(\theta) - \gamma_1 M(\theta))$*. Then, for all* $\theta_0 \in W^{1,\infty}(D;D)$*,* $\gamma_1(\theta)$ *is directionally differentiable at the point* $\theta = 0$ *in the direction* θ_0 *and the value of the directional derivative is*

$$\gamma_1'(\theta_0) = \min_{\substack{u \in E_0 \\ (M(0)u,u)=1}} \big(L'(\theta_0)u,u\big) - \gamma_1(0)\big(M'(\theta_0)u,u\big).$$

5.2. Calculus of the directional derivative of $\mathcal{L}$. We want to apply the general Theorem 5.2 in the shape sensitivity setting. We want to derivate $\tilde{\mathcal{L}}(\theta)$ defined as below.

DEFINITION 5.3.

◇ *Let* $T = Id + \theta$ *and* $\Omega_\theta = T \circ \Omega_0$*.*

◇ *Let* $\tilde{M}(\theta)$ *and* $\tilde{L}(\theta)$ *be defined as for all* u, v *in* $H^1_D(\Omega_\theta)$*,*

$$(\tilde{M}(\theta)u, v) = \int_{\Omega_\theta} \rho u \cdot v$$
$$(\tilde{L}(\theta)u, v) = \int_{\Omega_\theta} Ae(u) : e(v).$$

◇ *Let* $\tilde{\gamma}_1(\theta)$ *be the smallest eigenvalue associated to the problem*

$$\tilde{L}(\theta)u = \tilde{\gamma}_1(\theta)\tilde{M}(\theta)u.$$

◇ *Let* $\tilde{\mathcal{L}}(\theta) = -\tilde{\gamma}_1(\theta) + \eta|\Omega_\theta|$*.*

The $\tilde{\mathcal{L}}(\theta)$ (resp. $\tilde{\gamma}_1(\theta)$) just defined corresponds to what has been denoted $\mathcal{L}(\Omega_\theta)$ (resp. $\gamma_1(\Omega_\theta)$) in section 2.1.

Theorem 5.2 cannot be applied to $\tilde{\gamma}_1(\theta)$ because the spaces where the operators $\tilde{L}(\theta)$ and $\tilde{M}(\theta)$ are defined changes with θ. That is why we consider the following definition.

DEFINITION 5.4.

◇ *Let* $M(\theta) \in \mathbb{M}$ *and* $L(\theta) \in \mathbb{L}$ *be defined as*

$$(M(\theta)u, v) = (\tilde{M}(\theta)u \circ T^{-1}, v \circ T^{-1})$$
$$(L(\theta)u, v) = (\tilde{L}(\theta)u \circ T^{-1}, v \circ T^{-1}).$$

◇ *Let* $\bar{\gamma}(\theta)$ *be the smallest eigenvalue associated to the problem*

$$L(\theta)u = \bar{\gamma}(\theta)M(\theta)u.$$

We shall work with those operators instead of the classical one. They are defined on a domain independent of θ so that the first eigenvalue can be derivated in the sense of Theorem 5.2. First it must be proven that the introduced eigenvalue $\bar{\gamma}$ is the same as γ_1 the first eigenvalue of the elasticity problem.

LEMMA 5.5.

$\diamond$ *$L(\theta)$ and $M(\theta)$ indeed belongs to $\mathbb{L}$ and $\mathbb{M}$.*

$\diamond$ *$\bar{\gamma}(\theta)$ and $\tilde{\gamma}_1(\theta)$ coincide, where $\tilde{\gamma}_1$ is defined in Definition* 5.3.

Proof of Lemma 5.5.

$\diamond$ Let us first prove that $\tilde{\gamma}_1(\theta) = \bar{\gamma}(\theta)$. Using $u \in H^1(\Omega_\theta) \Longleftrightarrow v = u \circ T^{-1} \in H^1(\Omega_0)$, we have

$$\begin{aligned}\tilde{\gamma}_1(\theta)^{-1} &= \max_{u\in H^1(\Omega_\theta)} \frac{\left(\tilde{M}(\theta)u,u\right)}{\left(\tilde{L}(\theta)u,u\right)} = \max_{u\in H^1(\Omega_\theta)} \frac{\left(M(\theta)u\circ T,u\circ T\right)}{\left(L(\theta)u\circ T,u\circ T\right)}\\ &= \max_{v\in H^1(\Omega_0)} \frac{\left(M(\theta)v,v\right)}{\left(L(\theta)v,v\right)} = \bar{\gamma}(\theta)^{-1}.\end{aligned}$$

$\diamond$ Let us now prove that $L(\theta)$ belongs to $\mathbb{L}$. The fact that $L(\theta)$ is coercive and bounded with respect to the $H^1_D(\Omega_0)$ norm comes from the fact that $\tilde{L}(\theta)$ is coercive and bounded in the $H^1_D(\Omega_\theta)$ norm. We have to show that these constants of coercivity and boundedness are uniform in θ. Let's introduce the tensor A which has the symmetries of the elasticity:

$$A^{ijkl} = A^{jikl} = A^{ijlk} = A^{klij}$$

such that $A^{ijkl}(\partial_j u^i)(\partial_l v^k) = Ae(u)\colon e(v)$. We use the standard tool of shape sensitivity, namely a change of variable:

$$\begin{aligned}(L(\theta)u,v) &= \int_{\Omega_\theta} A^{ijkl}\partial_j(u\circ T^{-1})^i\partial_l(v\circ T^{-1})^k\\ &= \int_{\Omega_0} |\det\nabla T|A^{ijkl}(\partial_s u^i)\left(\partial_j(T^{-1})^s\right)(\partial_m v^k)\left(\partial_l(T^{-1})^m\right) \qquad (9)\\ &= \int_{\Omega_0} l(\theta)^{iskm}(\partial_s u^i)(\partial_m v^k)\end{aligned}$$

with

$$l(\theta)^{iskm} = A^{iskm} + (\partial_l\theta^l)A^{iskm} - A^{ijkm}(\partial_j\theta^s) - A^{iskl}(\partial_l\theta^m) + o(\|\theta\|_{W^{1,\infty}}).$$

Given α and β the coercivity and continuity constant of $L(0)$, there exist $\eta > 0$ such that (s.t.) for every θ with $\|\theta\|_{W^{1,\infty}} < \eta$, we have

$$\forall u \in H^1_D(\Omega_0) \quad 2\beta_0\|u\|^2_{H^1_D(\Omega_0)} \geq (L(\theta)u,u)_{L^2} \geq \frac{\alpha_0}{2}\|u\|^2_{H^1_D(\Omega_0)}.$$

We then verify the hypothesis of Definition 5.1 with $\alpha_1 = \alpha/2$, $\beta_1 = 2\beta$, and $U = \{\theta \text{ s.t. } \|\theta\|_{W^{1,\infty}} < \eta\}$. An analog development stands for M. □

We can now apply Theorem 5.2 to $\bar{\gamma}(\theta) = \tilde{\gamma}_1(\theta)$ and end with the following result.

THEOREM 5.6. *Recalling Definition* 5.3, *$\tilde{\mathcal{L}}(\theta) = -\tilde{\gamma}_1(\theta) + \eta|\Omega_\theta|$ is directionally differentiable with respect to θ, and its directionally derivative at $\theta = 0$ in the direction θ_0 is given by*

$$\mathcal{L}'(\theta_0) = \max_{\substack{u\in E_0\\ \int_{\Omega_0}\rho u\cdot u=1}} \int_{\partial\Omega_0} (\theta_0\cdot n)\left(-v(u,u)+\eta\right),$$

where $v(\cdot,\cdot)$ is a bilinear functional defined by

$$v(u,w) = Ae(u)\colon e(w) - \gamma_1\rho(\Omega_0)u\cdot w \quad \text{on } \Gamma_N,$$
$$v(u,w) = -Ae(u)\colon e(w) \quad \text{on } \Gamma_D,$$

and where E_0 is the first eigenspace associated to $\gamma_1(\Omega_0) = \tilde{\gamma}(0)$ the smallest eigenvalue for $\theta = 0$.

One can of course verify that the formula when the dimension of the eigenspace is greater than 1 is the same as the formula when the dimension is equal to 1.

Proof of Theorem 5.6. We apply Theorem 5.2 to $\tilde{\gamma}_1(\theta) = \bar{\gamma}(\theta)$ in order to calculate $\mathcal{L}'(\theta_0)$. The calculus above for L (and the same for M) shows that

$$\begin{aligned}(L'(\theta)u,v) &= \int_{\Omega_0} \left((\partial_l\theta^l)A^{iskm} - A^{ijkm}(\partial_j\theta^s) - A^{iskl}(\partial_l\theta^m)\right)(\partial_s u^i)(\partial_m v^k)\\ &= \int_{\Omega_0} A^{iskm}(\partial_l\theta^l)(\partial_s u^i)(\partial_m v^k) - A^{iskm}(\partial_s\theta^l)(\partial_l u^i)(\partial_m v^k)\\ &\quad -A^{iskm}(\partial_m\theta^l)(\partial_s u^i)(\partial_l v^k),\\ (M'(\theta)u,v) &= \int_{\Omega_0} (\partial_l\theta^l)\rho u^i v^i.\end{aligned}$$

Applying Theorem 5.2 gives

$$\begin{aligned}\mathcal{L}' = \max_{\substack{u\in E_0\\ \int_{\Omega_0}\rho u\cdot u=1}} &-\int_{\Omega_0} A^{iskm}(\partial_l\theta^l)(\partial_s u^i)(\partial_m v^k) + A^{iskm}(\partial_s\theta^l)(\partial_l u^i)(\partial_m v^k)\\ &+\int_{\Omega_0} A^{iskm}(\partial_m\theta^l)(\partial_s u^i)(\partial_l v^k) + (\partial_l\theta^l)\rho u^i v^i + \eta(\partial_l\theta^l).\end{aligned} \tag{10}$$

We perform an integration by part on θ; the term in $\int_{\Omega_0}$ is equal to

$$\theta^l A^{iskm}\left[\partial_l(\partial_s u^i\partial_m u^k) - \partial_s(\partial_l u^i\partial_m u^k) - \partial_m(\partial_s u^i\partial_l u^k)\right] - \theta^l\gamma_1\rho\partial_l(u^i u^i). \tag{11}$$

Some algebra used in coordination with $\gamma_1\rho u = -\mathrm{divAe(u)}$ allows us to conclude that this term is equal to 0. The remaining term is then equal to

$$\begin{aligned}\mathcal{L}'(\theta) = \max &\int_{\partial\Omega_0} (\theta\cdot n)[-A^{iskm}(\partial_s u^i)(\partial_m u^k) + \gamma_1\rho u\cdot u + \eta]\\ &+\int_{\partial\Omega_0} A^{iskm}\theta^l n^m(\partial_s u^i)(\partial_l u^k) + A^{iskm}\theta^l n^s(\partial_l u^i)(\partial_m u^k).\end{aligned} \tag{12}$$

On the Neumann part of the boundary we use $Ae(u)\cdot n = 0$ and the definition of C to conclude that

$$A^{iskm}(\partial_s u^i)n^m = 0 = A^{iskm}(\partial_m u^k)n^s.$$

On the Dirichlet part of the boundary we use $u = 0$ so that $\nabla u = \frac{\partial u}{\partial n}\otimes n$ to conclude that

$$(\theta\cdot n)(\partial_m u^k) = \theta^l n^m(\partial_l u^k) \text{ and } (\theta\cdot n)(\partial_s u^i) = \theta^l n^s(\partial_l u^i)$$

so that

$$\mathcal{L}'(\theta) = \max\int_{\partial\Omega_0}(\theta\cdot n)[-Ae(u)\colon e(u) + \gamma_1 u\cdot u + \eta] + 2\int_{\Gamma_D}(\theta\cdot n)Ae(u)\colon e(u). \qquad \square$$

Remark 5.7 (*Assumed regularity of* Ω_0). We need regularity on the domain Ω_0 in two occurrences. The first one is when we perform a change of variable ((9) in the demonstration of Lemma 5.5), and the second one is when we perform an integration by part ((11) in the demonstration of Theorem 5.6).

The change of variable is compulsory in order to prove the derivability. It is sufficient to suppose Ω_0 Lipschitz in order to be able to perform it.

The integration by part is used only in order to express the directional derivative (10) as an integral over $\partial\Omega_0$ (see (12)). Performing this integration by part is asking a lot of regularity on Ω_0 because the eigenvector u must be regular enough, i.e., $u \in H^1(\partial\Omega_0)$. But this integration by part is not needed if the only goal is to prove the directional derivability.

5.3. Calculating V^*. The goal of this subsection is to calculate V^* the minimizer to

$$\min_{\|V\|=1} \max_{\substack{u\in E_0 \\ \int_{\Omega_0} u\cdot u=1}} \int_{\partial\Omega_0} V\left(-v(u,u)+\eta\right), \tag{13}$$

where $v(\cdot,\cdot)$ is defined in Theorem 5.6. We will prove in this section that this is a semidefinite programming (SDP) problem in low dimension that is easily solved. The use of SDP for eigenvalue optimization is classical; the goal of this section is to show that solving this SDP problem is a very easy task thanks to the Hilbertian structure endowed by the velocity regularization. A good introduction to SDP is [33] and the references therein. Only some basic facts about SDP problems have been recalled here.

Definition 5.8.

$\diamond$ *Let Y be an unknown vector. Give Y_0 a vector and $E(Y)$ a matrix whose coefficients depend linearly on Y. Let ≥ 0 stand for "symmetric positive." An SDP problem is of the form*

$$\min_{E(Y)\geq 0} Y^T Y_0.$$

$\diamond$ *SDP problems are efficiently solvable by duality methods. In order to ensure that there is no gap of duality, a sufficient condition is to find Y_1 a strictly primal feasible point, i.e., such that $E(Y_1) > 0$ (definite positive).*

We now need to introduce the semidefinite programming problem we will work on.

Definition 5.9. *Recall Definition* 4.1 *of the scalar product $(\cdot,\cdot)_{H^1}$.*

$\diamond$ *Define $(e_i)_{i=1,\dots,d}$ an orthonormal basis of E_0 for the scalar product*

$$(u,w) = \int_{\Omega_0} \rho u\cdot v.$$

$\diamond$ *Define $(a_{ij})_{i,j=1,\dots,d}$ and c as*

$$(a_{ij},X)_{H^1} = \int_{\partial\Omega_0} -Xv(e_i,e_j) \quad \textit{and} \quad (c,X)_{H^1} = \int_{\partial\Omega_0} X\eta.$$

$\diamond$ *Define h_k an orthonormalized basis of $Span(a_{ij},c)_{i,j}$ for the scalar product $(,)_{H^1}$. Let m be the dimension of this space, and let $(a_{ij}^k)_{k=1,\dots,m}$ (resp. $(c^k)_{k=1,\dots,m}$) be the coordinates of a_{ij} (resp. c) on the basis $(h_k)_{k=1,\dots,m}$.*

◇ *For any* $X = (X_1, \dots, X_m)$ *and* $(z, w) \in \mathbb{R}^2$, *let* $Y = [X, w, z]$.
◇ *Let* $A(X)$ *be the* $d \times d$ *matrix* $A(X)_{ij} = a_{ij}^k X_k$, *let* $C(X) = c^k X_k$, *and let*

$$D(Y) = \begin{bmatrix} -A(X) + zId & 0 \\ 0 & -C(X) - z + w \end{bmatrix}, \quad E(Y) = \begin{bmatrix} D(Y) & 0 & 0 \\ 0 & Id & X \\ 0 & X^T & 1 \end{bmatrix}.$$

The coefficients of E *depends linearly on* $Y = [X, w, z]$.
◇ *Let* $Y^* = [X^*, w^*, z^*]$ *be the solution of the following SDP problem:*

$$\min_{E(Y) \geq 0} w \tag{14}$$

THEOREM 5.10. V^*, *the minimizer of the problem* (13), *is given by*

$$V^* = \sum_{k=1}^{m} V_k h_k,$$

where the vector $X^* = [V_1, \dots, V_k]$ *is defined as a solution of the semidefinite problem* (14).
◇ *The problem* $\min_{E(Y) \geq 0} w$ *is strictly feasible; SDP programming can be applied.*
◇ V^* *(or equivalently* X^**) is attained.*

Proof. We transform the problem (13) into (14) by using the fact that $v(\cdot, \cdot)$ is bilinear (see Theorem 5.6 for the definition of v)

$$\begin{aligned} &\min_{\|V\|_{H^1}=1} \max_{\substack{u \in E_0 \\ \int_{\Omega_0} \rho u \cdot u = 1}} \int_{\partial\Omega_0} V[-v(u,u) + \eta] \\ &= \min_{\|V\|_{H^1}=1} \max_{\sum_{i=1}^d \lambda_i^2 = 1} \int_{\partial\Omega_0} V[-\lambda_i \lambda_j v(e_i, e_j) + \eta] \\ &= \min_{\|V\|_{H^1}=1} \max_{\sum_{i=1}^d \lambda_i^2 = 1} \lambda_i \lambda_j (V, a_{ij})_{H^1} + (V, c)_{H^1} \\ &= \min_{\|V\|_{H^1}=1} \max_{\sum_{i=1}^d \lambda_i^2 = 1} (V, h_k)_{H^1} (\lambda_i \lambda_j a_{ij}^k + c^k) \end{aligned}$$

so that V^* is a minimizer of (13) if and only if $X_k^* = (V^*, h_k)_{H^1}$ are minimizers of the following problem:

$$\min_{\sum X_k^2 \leq 1} \max_{\sum_{i=1}^d \lambda_i^2 = 1} \lambda_i \lambda_j a_{ij}^k X_k + c^k X_k. \tag{15}$$

◇ Showing that (15) is equivalent to (14) is a standard issue of SDP: The condition $E([X, w, z]) \geq 0$ is equivalent to $(X, X) \leq 1$ and $D([X, w, z]) \geq 0$. The condition $D([X, w, z]) \geq 0$ is equivalent to $zId \geq A(X)$ and $w \geq z + C(X)$. So

$$E([X, w, z]) \geq 0 \Longleftrightarrow \sum_k X_k^2 \leq 1 \text{ and } w \geq z + X_k c^k \text{ and } z \geq \max_{\sum_{i=1}^d \lambda_i^2 = 1} \lambda_i \lambda_j a_{ij}^k X_k$$

so that minimizing w with the above condition is equivalent to finding X in the problem (15).

◇ Choosing $X = 0$, $z > 0$, and $w > z$ gives a $[X, w, z]$ for which $F([X, w, z]) > 0$. The problem is then strictly feasible, an extended Slater's condition holds, and the dual problem (in terms of semidefinite duality) has the same extremal value.

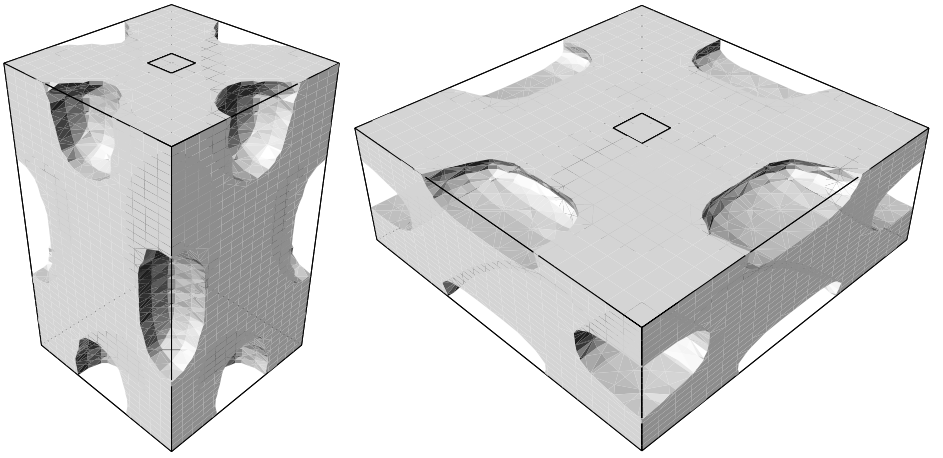


FIG. 21. *Initialization for the narrow domain (left) and for the big domain (right).*

$\diamond$ The condition $\sum_k X_k^2 \leq 1$ ensures that X is bounded and that every minimizing sequence converges up to a subsequence. The maximum is indeed attained. $\square$

Remark 5.11. The SDP problem is not difficult to solve. Recall that d is the dimension of the first eigenspace, then $[X, w, z]$ is of dimension lower or equal to $\frac{d(d+1)}{2} + 3$ and the matrix E is a $\frac{d(d+3)}{2} + 3$ square matrix.

6. Numerical results.

6.1. The 3d eigenvalue of a beam. We naturally set our problem in 3d with symmetries, where we are sure to obtain a multiplicity of the first eigenvalue greater than 1. The first example which will be called the "big domain" problem is a $3 \times 3 \times 1$ rectangle discretized with a $21 \times 21 \times 23$ mesh. A zero displacement boundary condition is imposed on the plane $z = 0$, and four cells on the middle of the plane $z = 1$ are not subject to optimization and are 50 times heavier (see Figure 21). Since the domain is symmetric, the shapes are expected to keep a first eigenvalue of dimension at least 2 along the iterations. The Young's modulus is set to 1 and the Poisson ratio to 0.3. In the void, the density ρ is set to 0 and the parameter ε is equal to 10^{-5}. The second problem is the same as the first except that the rectangle is of dimension $0.6 \times 0.6 \times 1$ (discretized by a $15 \times 15 \times 43$ mesh) with a mass tip that is 200 times heavier. The second problem will be called in this section the "narrow domain" problem.

Discussion about the big domain problem. The Lagrange multiplier being set to 5.3×10^{-8}, Figure 22 is a display of the optimal shape and Figure 23 displays the three smallest eigenvalues. There is one eigenvalue that is always (except on iterations 22 to 28) of multiplicity two and one which is of multiplicity one. The eigenvalue of multiplicity one corresponds to an eigenvector which is localized on the heavy cells and that interferes in the optimization process. The panel on the left-hand side of Figure 23 shows the evolution of the eigenvalues. On each iteration, the value of d the dimension of the subspace of the first eigenvalue is shown. The expected behavior of the algorithm can be verified.

Discussion about the narrow domain problem. Figure 24 shows the optimal shape and Figure 25 shows the convergence history of the narrow domain problem. As can be seen in Figure 25 (left), the global evolution of the algorithm is as follows:

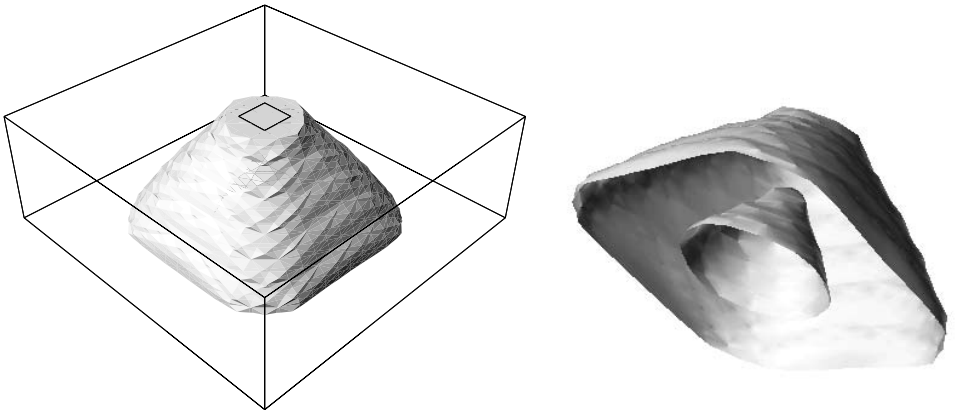


FIG. 22. *The optimal shape for the first eigenvalue in a big domain (left) and its boundary (right).*

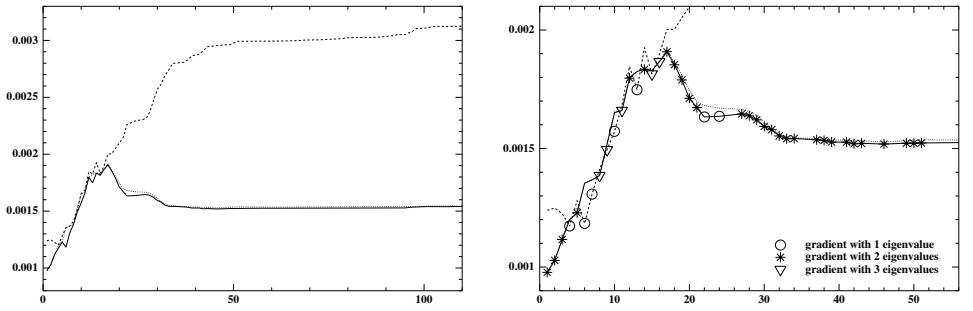


FIG. 23. *Evolution of the three smallest eigenvalues (left) and an interpretation (right) for the big domain.*

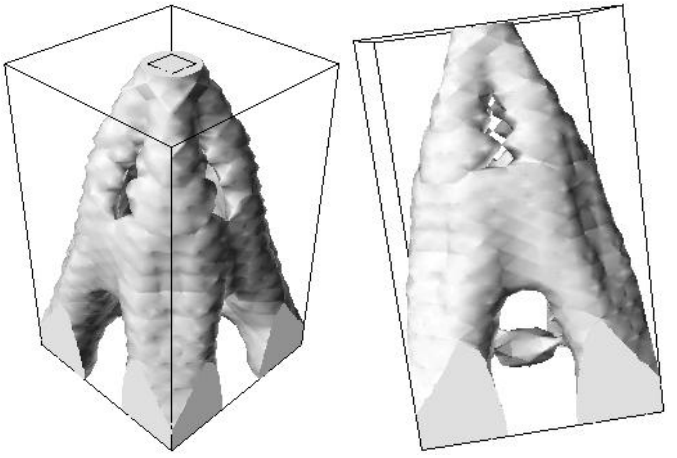


FIG. 24. *The optimal shape for the first eigenvalue in a narrow domain (isovalue* 0.2 *of the density is drawn).*

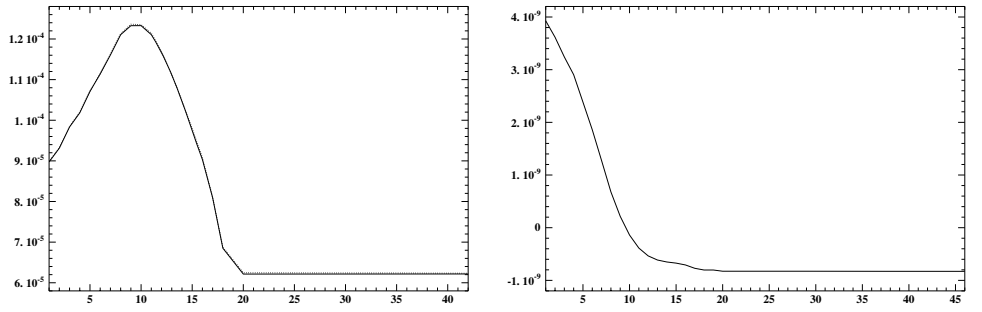


FIG. 25. *Evolution of the two smallest eigenvalues (left) and the objective function (right) for the narrow domain.*

First reinforce the structure so that the first eigenvalue raises, and then optimize the weight of the structure. The Lagrange multiplier is set to 10^{-8} for this example.

Discussion about the symmetries. Neither of the two problems give rise to radially symmetrical shapes. For the narrow-domain problem it can be easily understood by the fact that the shape is constrained into a box. For the big-domain problem one can advance an explanation based on a mesh-effect. But it is known that there exist symmetric problems whose solutions do not respect the symmetries. We still do not know if the optimal shape is or is not radially symmetric for this problem.

6.2. The short cantilever. We run our algorithm on a vibrating cantilever that is the same test case as the one of section 4.3 except that the working domain is of size 1×2 discretized with a regular 80×160 mesh. The other parameters that have changed are the Lagrange multiplier which is set to 0.3 and the heavier mass which has a density 80 times heavier. This test case was introduced in [4] where the authors pointed out the appearance of multiple eigenvalues. It is not the exact same test because when the test of [4] has been run, the improvement of the multiple eigenvalue method is not as obvious as in the test presented here.

The test is run with the standard single eigenvalue optimization (only one eigenvalue is taken into account during optimization) and the multiple eigenvalue optimization processes. Figure 26 (left) is the initializing shape which is the same for the two process and Figure 26 (right) is the optimal shape for the multiple eigenvalue optimization process. Figure 27 shows the evolution of the objectives functions for the two different processes.

At iteration n, the algorithm considers an eigenvalue to be of multiplicity d if and only if the relative differences $\gamma_d/\gamma_1 - 1 \leq \epsilon_0 < \gamma_{d+1}/\gamma_1 - 1$, where ε_0 is a user-defined criterion. If the shape computed at iteration $n+1$ is not better than the shape computed at iteration n and if the eigenvalue at iteration n is multiple, then the parameter ε_0 is decreased. The parameter ε_0 is set to 10 percent at the beginning of the optimization process. This explains the behavior of the algorithm in Figure 28 (right).

It can be seen in Figure 28 (left) that when the first eigenvalue is considered to be always of multiplicity one, the two first eigenvalues have a tendency to merge and the algorithm cannot improve the shape. If we follow the branches of eigenvalues according to the modes, we would see that the smallest eigenvalue does not correspond to the same modes during optimization, i.e., the algorithm optimizes one mode at iteration

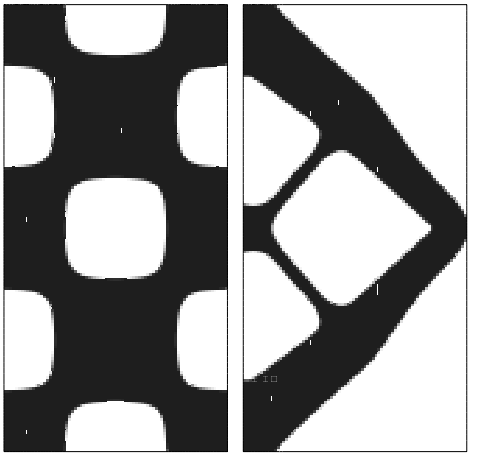


FIG. 26. *Initialization and optimal shape for the first eigenvalue of a short cantilever.*

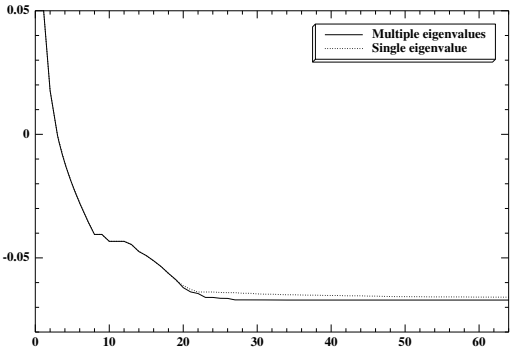


FIG. 27. *Evolution of the objective function for the two different methods.*

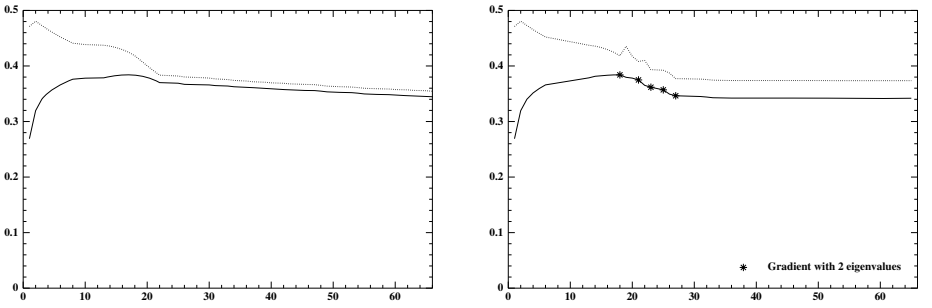


FIG. 28. *Evolution of the two smallest eigenvalues for the single eigenvalue optimization (left) and for the multiple eigenvalue optimization (right).*

n and another one at iteration $n+1$. This is standard of optimization with respect to a maximum of a function when the maximum is multiple and it is well known to slow down the algorithm.

7. Conclusion. The velocity-regularization method presented here improves the speed and accuracy of the level-set method and extends it to new problems. Thanks to the Hilbertian method, three issues can be dealt with: the extension, the regularization, and the endowment of a Hilbertian structure (see the discussion in section 4.2).

The extension issue (i.e., extending everywhere a velocity which is only defined on the boundary) is related only to the speed of the algorithm. The 2d compliant cantilever of section 4.4 which shows the mesh-independence of the Hilbertian method is a good example of the improvements this new method brings to the level-set algorithm.

The regularization issue is about dealing with more regular velocities. It is indeed an improvement as can be seen in the mechanism examples because it allows one to improve the accuracy of the optimal shape. It allows one to also improve the speed of the algorithm by diffusing the peaks of the velocity in the vicinity of the peak. The only test for which velocity regularization is not as efficient in terms of accuracy as the natural extension is the first step of the optimization procedure of the 2d mechanism of section 4.5 (see Figure 8 (left)). This can be explained by the fact that the adjoint is not computed with a precision that is accurate enough when the ratio of the weak material is too high. Sadly for 2d tests which need the computation of an adjoint, it still seems compulsory to perform the 2-steps optimization procedure. It was seen in section 4.7 that this trick is not needed in 3d.

Because the velocity regularization endows the problem with a Hilbertian structure, it allows one to apply the level-set method of optimization for several problems that are not differentiable but whose directional derivative exists. The computation of the descent direction relies on an SDP problem. The transformation of the steepest descent algorithm into an SDP problem can be made because the directional derivative of the problem is a maximum of quadratic functions over a sphere. This is the case for the eigenvalue problem, for the robust compliance problem, and for the buckling load problem which are all based on the generalized eigenvalues problem. The investigation of the two latest problems and application of the Hilbertian method to those issues is in progress.

Acknowledgments. I deeply thank Grégoire Allaire for his advise on this work, for his useful and always pertinent comments and suggestions, and François Jouve for his programming skills and useful critics. Every figure was obtained by using François Jouve's personal software.

REFERENCES

[1] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer-Verlag, New York, 2001.

[2] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve, *Shape optimization by the homogenization method*, Nümer. Math., 76 (1997), pp. 27–68.

[3] G. Allaire, F. de Gournay, F. Jouve, and A.-M. Toader, *Structural optimization using topological and shape sensitivity via a level set method*, Control Cybernet., 34 (2005), pp. 59–80.

[4] G. Allaire and F. Jouve, *A level-set method for vibration and multiple loads in structural optimization*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 3269–3290.

[5] G. Allaire, F. Jouve, and A.-M. Toader, *Structural optimization using sensitivity analysis and a level-set method*, J. Comput. Phys., 194 (2004), pp. 363–393.

[6] M. Bendsoe, *Methods for Optimization of Structural Topology, Shape and Material*, Springer-Verlag, New York, 1995.

[7] M. Bendsoe and N. Kikuchi, *Generating optimal topologies in structural design using a homogenization method*, Comput. Methods Appl. Mech. Eng., 71 (1988), pp. 197–227.

[8] M. Bendsoe and O. Sigmund, *Topology Optimization. Theory, Methods, and Applications*, Springer-Verlag, New York, 2003.
[9] P. Bernardoni, *Outils et méthode de conception de structure mécaniques à déformations et actionnements répartis*, Ph.D. thesis, Université Paris VI, 2004.
[10] M. Burger, *A framework for the construction of level set methods for shape optimization and reconstruction*, Interfaces Free Bound., 5 (2003), pp. 301–329.
[11] M. Burger, B. Hackl, and W. Ring, *Incorporating topological derivatives into level set methods*, J. Comput. Phys., 194 (2004), pp. 344–362.
[12] A. Cherkaev, *Variational Methods for Structural Optimization*, Springer-Verlag, New York, 2000.
[13] A. Cherkaev and E. Cherkaeva, *Principal compliance and robust optimal design*, J. Elasticity, 72 (2003), pp. 71–98.
[14] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[15] S. J. Cox, *The generalized gradient at a multiple eigenvalue*, J. Funct. Anal., 133 (1995), pp. 30–40.
[16] H. Eschenauer and A. Schumacher, *Bubble method for topology and shape optimization of structures*, Struct. Optim., 8 (1994), pp. 42–51.
[17] S. Garreau, P. Guillaume, and M. Masmoudi, *The topological asymptotic for pde systems: the elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.
[18] F. de Gournay, *Optimisation de formes par la méthode des lignes de niveaux*, Ph.D. thesis, Ecole Polytechnique, 2005.
[19] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1976.
[20] G.-W. Litvinov, *Optimization in Elliptic Problems with Applications to Mechanics of Deformable Bodies and Fluid Mechanics*, Oper. Theory Adv. Appl. 119, Birkhäuser, Basel, 2000.
[21] B. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, Clarendon Press, Oxford, 2001.
[22] F. Murat and S. Simon, *Etudes de Problèmes d'Optimal Design*, in Lecture Notes in Comput. Sci. 41, Springer-Verlag, Berlin, 1976, pp. 54–62.
[23] S. A. Nazarov and Y. Sokolovski, *The topological derivative of the dirichlet integral under the formation of a thin bridge*, Siberian Math. J., 45 (2004), pp. 341–355.
[24] S. Osher and F. Santosa, *Level-set methods for optimization problems involving geometry and constraints: Frequencies of a two-density inhomogeneous drum*, J. Comput. Phys., 171 (2001), pp. 272–288.
[25] S. Osher and J.-A. Sethian, *Front propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 78 (1988), pp. 12–49.
[26] B. Protas, T.-R. Bewley, and G. Hagen, *A computationnal framework for the regularization of adjoint analysis in multiscale pde systems*, J. Comput. Phys., 195 (2004), pp. 49–89.
[27] B. Rousselet and D. Chenais, *Continuité et différentiabilité d'élements propres: Application à l'optimisation de structures*, Appl. Math. Optim., 22 (1990), pp. 27–59.
[28] J. Sanchez-Hubert and E. Sanchez-Palencia, *Vibration and Coupling of Continuous Systems. Asymptotic Methods*, Springer-Verlag, Berlin, 1989.
[29] J.-A. Sethian and A. Wiegmann, *Structural boundary design via level-set and immersed interface methods.*, J. Comput. Phys., 163 (2000), pp. 489–528.
[30] J. Simon, *Differentiation with respect to the domain in boundary value problems.*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.
[31] J. Sokolowski and J.-P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, Springer Ser. Comput. Math. 16, Springer-Verlag, New York, 1992.
[32] J. Sokolowski and A. Zochowski, *Topological derivatives of shape functionals for elasticity systems*, Mech. Structures Mach., 29 (2001), pp. 331–349.
[33] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[34] M.-Y. Wang, X. Wang, and D. Guo, *A level-set method for structural topology optimization*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 227–246.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 368–381

ROBUST MATRIX ROOT-CLUSTERING ANALYSIS THROUGH EXTENDED KYP LEMMA*

OLIVIER BACHELIER† AND DRISS MEHDI†

Abstract. This paper is dedicated to robust matrix eigenvalue clustering in a subregion $\mathcal{D}$ of the complex plane. A norm-bounded uncertainty is considered. $\mathcal{D}$ is any intersection and/or union of half planes, discs, and outsides of discs. This work aims at computing the associated complex $\mathcal{D}$-stability radius. This challenge was already tackled in a recent paper but the technique therein may fail or lead to conservative results when the subregions overlap. The present paper proposes to circumvent those limits thanks to an extended version of the Kalman–Yakubovich–Popov (KYP) lemma which is deduced from the so-called generalized S-procedure.

Key words. matrix $\mathcal{D}$-stability, $\partial\mathcal{D}$-regularity, KYP lemma, generalized S-procedure

AMS subject classifications. 93D09, 34D20, 15A39

DOI. 10.1137/S036301290444349X

1. Introduction.

Matrix $\mathcal{D}$-stability appears to be a discerning tool to appreciate the transient performances of a linear model (a matrix is said to be $\mathcal{D}$-stable when its spectrum lies in a region $\mathcal{D}$ of the complex plane). Many contributions propose necessary and sufficient conditions for matrix $\mathcal{D}$-stability such as paper [12], its generalized Lyapunov equations and its Ω-regions, based on the pioneer work of [14], or [9] and its LMI regions (LMI stands for linear matrix inequality: see [8]).

In the presence of uncertainties, the necessity of the above-mentioned conditions cannot be preserved in most cases. Thus, it is important to check robust $\mathcal{D}$-stability through sufficient conditions. The Hurwitz-stability case was tackled in [19, 29, 27, 11]. Ω-regions were considered in [28, 6]. For polytopic uncertainty, LMI conditions given in [9, 20] prove to be very effective. The special case of norm-bounded uncertainty [21] is of interest since the complex $\mathcal{D}$-stability radius can be reached owing to the bounded real lemma when $\mathcal{D}$ is either a disc or a half plane [17, 18, 10] (the complex stability radius corresponds to the largest radius of the ball of complex uncertainties that preserve stability). For the real stability radius, see [13, 22].

Quite few results exist concerning nonconnected regions, especially combinations of several subregions, although it helps to handle the compromise between robustness and satisfactory transient behavior. Since the first contribution [2], other works were presented in [5, 4, 7]. For structured uncertainty, see the work of [25] on the original good ride quality region particularly convenient for aircraft models. One of the most recent contributions dealing with norm-bounded uncertainty is [3] in which the reader can find details and discussions about the above-mentioned papers. Starting from the notion of $\partial\mathcal{D}$-regularity (nonintersection between the matrix spectrum and some curve $\partial\mathcal{D}$), the approach presented in [3] is efficient and often leads to the complex $\mathcal{D}$-stability radius, except when the various subregions of $\mathcal{D}$ overlap. The present paper must be seen as an extension of [3] that circumvents the mentioned limits. It relies on

*Received by the editors May 6, 2004; accepted for publication (in revised form) October 31, 2005; published electronically March 15, 2006.
http://www.siam.org/journals/sicon/45-1/44349.html
†LAII-ESIP, Bâtiment de Mécanique, 40, Avenue du Recteur Pineau, 86022 Poitiers Cedex, France (Olivier.Bachelier@univ-poitiers.fr, Driss.Mehdi@univ-poitiers.fr).

an extended version of the Kalman–Yakubovich–Popov (KYP) lemma deduced from the insightful paper [16] and its generalized S-procedure.

The present paper is organized as follows: section 2 emphasizes the limits of [3] and states the problem under a new and more precise formulation. In section 3, once some background is recalled, an extended version of the KYP lemma is proposed. Section 4 is devoted to numerical examples. Section 5 concludes the paper.

Notation. M' denotes the transpose conjugate of matrix M. Hence, $s' \in \mathbb{C}$ is the conjugate of s. M^H equals matrix $M + M'$. The 2-norm of M (maximal singular value) is denoted by $||M||_2$. $\mathbf{I}$ and 0 are the identity matrix and the null matrix of appropriate dimensions, respectively. In matrix inequalities, > 0 and < 0 mean positive definite and negative definite, respectively; ≥ 0 and ≤ 0 mean positive semidefinite and negative semidefinite, respectively. Symbol $\mathbf{i}$ denotes the imaginary unit. $\lambda(M)$ and $\mathrm{tr}(M)$ are the spectrum and the trace of square matrix M, respectively. Symbol $\otimes$ stands for the matrix Kronecker product. Finally, given a couple of complex matrices (X,Y) and a matrix

$$T = T' = \begin{bmatrix} t_{00} & t_{10} \\ t'_{10} & t_{11} \end{bmatrix} \in \mathbb{C}^{2\times 2},$$

the expressions $\mathcal{F}_T(X,Y)$ and $\mathcal{G}_T(X,Y)$ are defined by

$$\begin{cases} \mathcal{F}_T(X,Y) = t_{00}XX' + (t_{10}YX')^H + t_{11}YY' \\ \mathcal{G}_T(X,Y) = t_{00}X'X + (t_{10}X'Y)^H + t_{11}Y'Y. \end{cases}$$

2. Motivation. In this section, some results proposed in [3] are first recalled. However, the limits inherent to those techniques are highlighted. That is why the problem initially stated in [3] is presented under a new and more complete formulation.

Consider an uncertain matrix that reads:

$$A_c = A + B\Delta C \in \mathbb{C}^{n\times n}. \tag{1}$$

Matrix Δ is constant, unknown, and assumed to belong to the ball $\mathcal{B}(\rho)$ of all matrices $\Delta \in \mathbb{C}^{q\times r}$ satisfying $||\Delta||_2 \leq \rho$. Matrices $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times q}$, and $C \in \mathbb{C}^{r\times n}$ are given. Such a description is referred to as norm-bounded uncertainty [21].

Now consider the next Hermitian matrices:

$$R_k = R'_k = \begin{bmatrix} r_{k_{00}} & r_{k_{10}} \\ r'_{k_{10}} & r_{k_{11}} \end{bmatrix} \in \mathbb{C}^{2\times 2} \quad \forall k \in \{1,\ldots,m\}, \tag{2}$$

from which the following geometric curves are defined:

$$\partial\mathcal{D}_k = \{s \in \mathbb{C} | f_k(s) = \mathcal{F}_{R_k}(1,s) = \mathcal{G}_{R_k}(1,s) = 0\} \quad \forall k \in \{1,\ldots,m\}. \tag{3}$$

Relevant curves are lines or circles which can be nonsymmetric with respect to the real axis. Each curve $\partial\mathcal{D}_k$ enables one to define an associated region:

$$\mathcal{D}_k = \{s \in \mathbb{C} \,|\, f_k(s) < 0\} \quad \forall k \in \{1,\ldots,m\}. \tag{4}$$

Clearly, $\mathcal{D}_k$ denotes either one side or the other side of the boundary $\partial\mathcal{D}_k$. Then it can be a half plane, a disc, or the outside of a disc. It is an open region (i.e., not including $\partial\mathcal{D}_k$) in order to encompass the concept of asymptotic stability for linear time-invariant (LTI) systems.

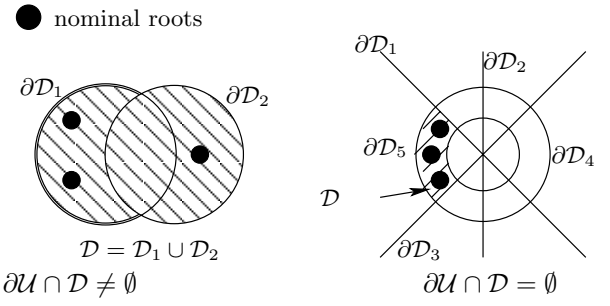


FIG. 1. $\varrho_{\partial\mathcal{U}} = r_{\mathcal{D}}$.

Also define the region $\mathcal{D}$ as a combination, i.e., any union and/or intersection of the various subregions $\mathcal{D}_k$. Such a formulation of $\mathcal{D}$ clearly enables a very large choice of regions.

This contribution aims at computing the complex $\mathcal{D}$-stability radius $r_{\mathcal{D}}$, i.e., the largest value of ρ—the radius of $\mathcal{B}(\rho)$—so that, A being $\mathcal{D}$-stable, A_c defined in (1) remains also $\mathcal{D}$-stable for any $\Delta \in \mathcal{B}(\rho)$. In [3] a technique to derive a lower bound $\rho^\star$ was proposed. This bound may equal the $\mathcal{D}$-stability radius, but it is sometimes conservative. The main result of [3] is now recalled.

THEOREM 1 (see [3]). *Let A_c, $\mathcal{D}$, $\partial\mathcal{D}$ be an uncertain matrix as defined in* (1)*, a clustering region as defined above and its boundary, respectively. Then A_c is robustly $\partial\mathcal{D}$-regular against $\mathcal{B}(\rho)$ if there exist m Hermitian matrices P_k, $k = 1, \ldots, m$, such that*

(5)
$$\mathcal{Q}_k(P_k, \gamma) = \begin{bmatrix} \mathcal{G}_{R_k}(P_k^{1/2}, P_k^{1/2}A) + C'C & r_{k_{10}} P_k B + r_{k_{11}} A' P_k B \\ r'_{k_{10}} B' P_k + r_{k_{11}} B' P_k A & r_{k_{11}} B' P_k B - \gamma \mathbf{I} \end{bmatrix} < 0 \quad \forall k \in \{1, \ldots, m\},$$

where $\gamma = \rho^{-2}$.

In this event, A_c keeps the same number of eigenvalues inside $\mathcal{D}$ for any $\Delta \in \mathcal{B}(\rho)$.

Moreover, if $\partial\mathcal{D}$ can be identified with $\partial\mathcal{U} = \bigcup_{k=1}^{m} \partial\mathcal{D}_k$, then condition (5) *is also necessary.*

To roughly sum up the above theorem and paper [3] in general, one can write that the uncertain matrix A_c defined in (1) is robustly $\mathcal{D}$-stable if the nominal matrix A is $\mathcal{D}$-stable and if A_c is robustly $\partial\mathcal{D}$-regular. The $\partial\mathcal{D}$-regularity of a matrix is simply the nonintersection between its spectrum and the set $\partial\mathcal{D}$. In [3], there are also considerations on the relation between the inertias of matrices P_k and the dispatching of the eigenvalues of A or A_c with respect to the various $\mathcal{D}_k$. But these notions do not need to be developed here. Note that Theorem 1 is deduced from the KYP lemma [26, 23].

The value $\rho^\star$ of ρ, obtained by minimizing γ while satisfying (5), equals $\varrho_{\partial\mathcal{U}}$ (the so-called complex $\partial\mathcal{U}$-regularity radius) which is itself a lower bound of $\varrho_{\partial\mathcal{D}}$ (the complex $\partial\mathcal{D}$-regularity radius). It is shown in [3] that if A is $\mathcal{D}$-stable and if $\partial\mathcal{U} \cap \mathcal{D} = \emptyset$, then $\varrho_{\partial\mathcal{U}}$ actually equals $r_{\mathcal{D}}$, the complex $\mathcal{D}$-stability radius. The $\mathcal{D}$-stability of A is obviously necessary to check the $\mathcal{D}$-stability of A_c but assumption $\partial\mathcal{U} \cap \mathcal{D} = \emptyset$ is not always satisfied; hence the possible conservativeness of Theorem 1. To highlight it, Figure 1 shows both cases: the left side corresponds to $\partial\mathcal{U} \cap \mathcal{D} \neq \emptyset$ so $\varrho_{\partial\mathcal{U}} \leq r_{\mathcal{D}}$, whereas the right side corresponds to $\partial\mathcal{U} \cap \mathcal{D} = \emptyset$ so $\varrho_{\partial\mathcal{U}} = r_{\mathcal{D}}$.

The drawback of Theorem 1 is that no rigorous description of the actual set $\partial\mathcal{D}$ is provided when $\partial\mathcal{U} \cap \mathcal{D} = \emptyset$. For example, on the left side of Figure 1, the parts

of the circles $\partial\mathcal{D}_1$ and $\partial\mathcal{D}_2$ that belong to $\mathcal{D}$ should not be taken into account. This selection of points in $\partial\mathcal{D}_k$ is not made. To remedy this, a new description of $\partial\mathcal{D}$ is now proposed.

Let m curves $\partial\mathcal{D}_k$ be described as in (3). Let the matrices $\Phi_{k,l}$ be defined by

$$\Phi_{k,l} = \begin{bmatrix} \phi_{k,l_{00}} & \phi_{k,l_{10}} \\ \phi'_{k,l_{10}} & \phi_{k,l_{11}} \end{bmatrix} = \Phi'_{k,l} \in \mathbb{C}^{2\times 2}, \quad k = 1,\dots,m,\ l = 1,\dots,p_k. \tag{6}$$

Define m subregions $\bar{\mathcal{E}}_k$ by

$$\bar{\mathcal{E}}_k = \left\{ s \in \mathbb{C} | \mathcal{F}_{\Phi_{k,l}}(1,s) = \mathcal{G}_{\Phi_{k,l}}(1,s) \geq 0 \ \forall l \in \{1,\dots,p_k\} \right\} \ \ \forall k \in \{1,\dots,m\}. \tag{7}$$

Each region $\bar{\mathcal{E}}_k$ is actually the intersection of various closed subregions $\bar{\mathcal{E}}_{k,l}$ characterized by inequalities $\mathcal{G}_{\Phi_{k,l}}(1,s) \geq 0$. Subregions $\bar{\mathcal{E}}_{k,l}$ are half planes, discs, or outsides of discs, not necessarily symmetric. Consider the sets $\partial\Gamma_k$ defined by

$$\partial\Gamma_k = \partial\mathcal{D}_k \cap \bar{\mathcal{E}}_k = \left\{ s \in \mathbb{C} | \mathcal{G}_{R_k}(1,s) = 0, \mathcal{F}_{\Phi_{k,l}}(1,s) \geq 0 \forall l \in \{1,\dots,p_k\} \right\} \tag{8}$$

for all $k \in \{1,\dots,m\}$. Each set $\partial\Gamma_k$ is the intersection of a line or a circle with p_k closed half planes, discs, or outsides of discs. In other words, it is a line or a circle from which only some points are selected through the choice of various $\Phi_{k,l}$. Furthermore, define the set $\partial\mathcal{D}$ as

$$\partial\mathcal{D} = \bigcup_{k=1}^{m} \partial\Gamma_k. \tag{9}$$

If one considers any combination (intersection and/or union) of half planes, discs, and outsides of discs, then its boundary can read formulation (9). For $\mathcal{D}$ as pictured on Figure 1 (left side), the circles $\partial\mathcal{D}_1$ and $\partial\mathcal{D}_2$ can be restricted to the subsets that are not in $\mathcal{D}$, thanks to an appropriate choice of $\Phi_{1,1}$ and $\Phi_{2,1}$ (the first example in section 4 will highlight this assertion). With the description given in (9), the computation of the complex $\mathcal{D}$-stability radius $r_{\mathcal{D}}$ becomes equivalent to that of the complex $\partial\mathcal{D}$-regularity radius $\varrho_{\partial\mathcal{D}}$.

3. Extended KYP lemma and complex $\mathcal{D}$-stability radius. This section is devoted to the computation of $r_{\mathcal{D}}$ through an LMI approach. First, some technical preliminaries are needed. Then, some sort of extended KYP lemma is deduced. Finally, it is shown that this lemma can be directly exploited to solve the stated problem.

3.1. Background and preliminary lemmas. A key point in the present reasoning is that each LMI condition of Theorem 1 enables testing of the robust $\partial\mathcal{D}_k$-regularity of a matrix, whereas it should test the robust $\partial\Gamma_k$-regularity ($\partial\Gamma_k \subset \partial\mathcal{D}_k$, $\forall k$). It is similar to the problem encountered while applying the finite frequency KYP lemma. Such a lemma is proposed in [16] and is proven with the help of the so-called generalized S-procedure. The present paper largely exploits the results provided in [16]. The reader is strongly invited to refer to this paper from which Definition 1 is extracted.

DEFINITION 1 (see [16]). *A subset $\mathcal{S}$ of $n \times n$ Hermitian matrices is said lossless if*

(a) *$\mathcal{S}$ is convex;*

(b) *$S \in \mathcal{S} \Rightarrow \tau S \in \mathcal{S}\ \ \forall \tau > 0$;*

(c) *For each nonzero matrix* $H \in \mathbb{C}^{n\times n}$ *of rank* r *such that*

$$H = H' \geq 0, \quad \mathrm{tr}(SH) \leq 0 \quad \forall S \in \mathcal{S},$$

there exist vectors $\xi_i \in \mathbb{C}^n$, $i = 1, \ldots, r$, *such that*

$$H = \sum_{i=1}^{r} \xi_i \xi_i' \quad \textit{and} \quad \xi_i' S \xi_i \leq 0 \quad \forall S \in \mathcal{S}.$$

The above definition is necessary to apply the so-called generalized S-procedure [16, Theorem 1] implicitly used in forthcoming Theorem 2. This definition and this procedure are not the only borrowings from [16]. Other insights of this discerning paper are exploited in what follows.

LEMMA 1. *Let* W *and* Z *be complex matrices of the same size and* $R \in \mathbb{C}^{2\times 2}$ *be defined by*

$$R = R' = \begin{bmatrix} r_{00} & r_{10} \\ r_{10}' & r_{11} \end{bmatrix}.$$

Then the following statements are equivalent.

(i)

$$\mathcal{F}_R(W, Z) = 0. \tag{10}$$

(ii) *There exists a matrix* E *such that* $Z = WE$ *and*

$$\mathcal{F}_R(\mathbf{I}, E) = 0. \tag{11}$$

Proof. (i)⇒(ii): First assume that $r_{11} = 0$; then from (10)

$$\begin{aligned} r_{00}WW' + r_{10}'WZ' + r_{10}ZW' &= W\left(\frac{1}{2}r_{00}W + r_{10}Z\right)' + \left(\frac{1}{2}r_{00}W + r_{10}Z\right)W' \\ &= WY' + YW' = 0 \end{aligned}$$

with $Y = (\frac{1}{2}r_{00}W + r_{10}Z)$. By virtue of [23, Lemma 3, statement (iii)], this implies the existence of U such that $UU' = \mathbf{I}$ and $Y = W(\mathbf{I} - U)(\mathbf{I} + U)^{-1} = W\bar{U}$. Hence

$$\bar{U} + \bar{U}' = (\mathbf{I} + U')^{-1}\left\{(\mathbf{I} + U')(\mathbf{I} - U) + (\mathbf{I} - U')(\mathbf{I} + U)\right\}(\mathbf{I} + U)^{-1}$$

$$= (\mathbf{I} + U')^{-1}\left\{\mathbf{I} + U' - U - U'U + \mathbf{I} - U' + U - U'U\right\}(\mathbf{I} + U)^{-1} = 0.$$

From $Y = W\bar{U}$, it can be deduced that W and Z comply with

$$W\left(\bar{U} - \frac{1}{2}r_{00}\mathbf{I}\right) = r_{10}Z,$$

which can be written as $Z = WE$ with

$$E = \frac{1}{r_{10}}\left(\bar{U} - \frac{1}{2}r_{00}\mathbf{I}\right) \tag{12}$$

(note that $r_{10} \neq 0$ otherwise W and Z equal 0). Since $\bar{U} + \bar{U}' = 0$,

$$r_{00}\mathbf{I} + r_{10}'E' + r_{10}E = 0 \Leftrightarrow \mathcal{F}_R(\mathbf{I}, E) = 0.$$

Now assume that $r_{11} \neq 0$, then (10) becomes

$$\hat{r}_{00} WW' + \hat{r}'_{10} WZ' + \hat{r}_{10} ZW' + ZZ' = 0$$

with $\hat{r}_{i0} = \dfrac{r_{i0}}{r_{11}}$ for $i = 0, 1$. The previous equation can be rewritten as

$$(Z + \hat{r}'_{10} W)\,(Z + \hat{r}'_{10} W)' = (\hat{r}_{10}\hat{r}'_{10} - \hat{r}_{00})\, WW'$$

from which, by virtue of lemma [23, Lemma 3, statement (i)], one gets

$$Z + \hat{r}'_{10} W = \sqrt{(\hat{r}_{10}\hat{r}'_{10} - \hat{r}_{00})}\, WU$$

for some U such that $UU' = \mathbf{I}$. Finally it leads to $Z = WE$ with

$$E = \left(-\hat{r}'_{10}\mathbf{I} + \sqrt{(\hat{r}_{10}\hat{r}'_{10} - \hat{r}_{00})}\, U \right). \tag{13}$$

It is obvious that E satisfies

$$E + \hat{r}'_{10}\mathbf{I} = \sqrt{(\hat{r}_{10}\hat{r}'_{10} - \hat{r}_{00})}\, U$$

which yields

$$(E + \hat{r}'_{10}\mathbf{I})\,(E + \hat{r}'_{10}\mathbf{I})' = (\hat{r}_{10}\hat{r}'_{10} - \hat{r}_{00})\,\mathbf{I}$$

$$\Leftrightarrow \mathcal{F}_R(\mathbf{I}, E) = 0.$$

To prove (ii)⇒(i), just left- and right-multiply (11) by W and W', respectively, to recover (10). □

An extension of [16, Lemma 2] is now presented.

LEMMA 2. *Let two complex vectors f and $g \neq 0$ be given. Also let $\partial\mathcal{D}$ be defined by* (8) *with $k = m = 1$ (so the dependence of R, P, Φ_l and Q_l on k is omitted). The following statements are equivalent.*

(i) $\exists s \in \partial\mathcal{D} = \partial\Gamma_1 \,|\, f = sg$.

(ii) $$\mathcal{Z} = \begin{bmatrix} g \\ f \end{bmatrix}' \left(R \otimes P - \sum_{l=1}^{p} (\Phi_l \otimes Q_l) \right) \begin{bmatrix} g \\ f \end{bmatrix} \leq 0$$
$$\forall P = P' \text{ and } \forall Q_l = Q'_l > 0,\ l = 1, \ldots, p.$$

Proof.

$$\text{(i)} \Rightarrow \mathcal{Z} = \mathcal{F}_R(1, s) g' P g - \sum_{l=1}^{p} (\mathcal{F}_{\Phi_l}(1, s) g' Q_l g).$$

Since $s \in \partial\Gamma_1$, relations

$$-\sum_{l=1}^{p} (\mathcal{F}_{\Phi_l}(1, s)) \leq 0 \quad \text{and} \quad \mathcal{F}_R(1, s) = 0$$

are satisfied. Besides, since $Q_l > 0\, \forall l \in \{1, \ldots, p\}$, one gets $\mathcal{Z} \leq 0$ and then (ii) is true.

Reciprocally, (ii) $\Rightarrow$

$$\operatorname{tr}\left(\begin{bmatrix} g \\ f \end{bmatrix}\begin{bmatrix} g' & f' \end{bmatrix}\left(R\otimes P-\sum_{l=1}^{p}(\Phi_l\otimes Q_l)\right)\right)\leq 0$$

$$\Leftrightarrow \operatorname{tr}\left(\begin{bmatrix} g \\ f \end{bmatrix}\begin{bmatrix} g' & f' \end{bmatrix}\left[\begin{array}{c|c} r_{00}P-\sum_{l=1}^{p}(\phi_{l_{00}}Q_l) & r_{10}P-\sum_{l=1}^{p}(\phi_{l_{10}}Q_l) \\ \hline r'_{10}P-\sum_{l=1}^{p}(\phi'_{l_{10}}Q_l) & r_{11}P-\sum_{l=1}^{p}(\phi_{l_{11}}Q_l) \end{array}\right]\right)\leq 0$$

$$\Leftrightarrow \operatorname{tr}(\mathcal{F}_R(g,f)P)-\operatorname{tr}\left(\sum_{l=1}^{p}(\mathcal{F}_{\Phi_l}(g,f)Q_l)\right)\leq 0. \tag{14}$$

Clearly, since (14) is true for all $P=P'$ and $Q_l=Q'_l>0$, it holds for any choice $Q_l=w_lI$, $w_l>0$, $l=1,\ldots,p$, and $P=\alpha\mathbf{I}$ with $\alpha>0$ possibly very large. Hence

$$\begin{cases} \mathcal{F}_R(g,f)=0 \\ \displaystyle\sum_{l=1}^{p}(w_l\mathcal{F}_{\Phi_l}(g,f))\geq 0 \quad \forall w_l>0,\ l=1,\ldots,p. \end{cases} \tag{15}$$

From the equality in (15) and Lemma 1, a scalar $s^\bullet\in\partial\mathcal{D}_1$ verfies $f=s^\bullet g$.

Now consider the inequality in (15). Since $f=s^\bullet g$, it comes:

$$\sum_{l=1}^{p}(w_l\mathcal{F}_{\Phi_l}(1,s^\bullet))\geq 0 \quad \forall w_l>0, l=1,\ldots,p.$$

The above inequality holds for any set of $w_l>0$ if and only if $\mathcal{F}_{\Phi_l}(1,s^\bullet)\geq 0$ for all l in $\{1,\ldots,p\}$. Thus, $s^\bullet$ belongs to $\partial\Gamma_1$. □

Now, still inspired from [16], the next lemma is proposed.

LEMMA 3. *Let some matrix R defined in* (2)*, some matrices Φ_l defined in* (6) *(with $k=1$) and a matrix $F\in\mathbb{C}^{2n\times q}$ be given. Also let a subset of Hermitian matrices $\mathcal{S}$ be defined by*

$$\mathcal{S}:=\left\{F'\left(R\otimes P-\sum_{l}^{p}(\Phi_l\otimes Q_l)\right)F\ :\ P=P', Q_l=Q'_l>0, l=1,\ldots,p\right\}. \tag{16}$$

Then $\mathcal{S}$ is lossless.

Proof. Refering to Definition 1, properties (a) and (b) are obvious. For property (c), consider a matrix $H=H'\geq 0$ such that $\operatorname{tr}(HS)\leq 0\,\forall S\in\mathcal{S}$ and $H=GG'$, where $G\in\mathbb{C}^{q\times r}$ is of rank r. One gets:

$$\operatorname{tr}\left(GG'F'\left(R\otimes P-\sum_{l}^{p}(\Phi_l\otimes Q_l)\right)F\right)\leq 0 \quad \forall\, P=P', Q_l=Q'_l>0$$

$$\Leftrightarrow \operatorname{tr}\left(FGG'F'\left(R\otimes P-\sum_{l=1}^{p}(\Phi_l\otimes Q_l)\right)\right)\leq 0 \quad \forall\, P=P', Q_l=Q'_l>0. \tag{17}$$

Define matrices W and Z by

$$\begin{bmatrix} W \\ Z \end{bmatrix} = FG. \tag{18}$$

Thus, inequality (17) can be written

$$\begin{aligned} &\operatorname{tr}\left(\begin{bmatrix} W \\ Z \end{bmatrix}\begin{bmatrix} W \\ Z \end{bmatrix}'\left(R\otimes P - \sum_{l=1}^{p}(\Phi_l\otimes Q_l)\right)\right) \\ &= \operatorname{tr}\left(\begin{bmatrix} WW' & WZ' \\ ZW & ZZ' \end{bmatrix}'\left(R\otimes P - \sum_{l=1}^{p}(\Phi_l\otimes Q_l)\right)\right) \\ &= \operatorname{tr}\left(\begin{bmatrix} WW' & WZ' \\ ZW' & ZZ' \end{bmatrix}(R\otimes P)\right) - \operatorname{tr}\left(\sum_{l=1}^{p}\left(\begin{bmatrix} WW' & WZ' \\ ZW' & ZZ' \end{bmatrix}(\Phi_l\otimes Q_l)\right)\right) \le 0 \end{aligned}$$

$\forall\, P = P', Q_l = Q_l' > 0.$

As in the proof of Lemma 2, it can be deduced that

$$\begin{cases} \mathcal{F}_R(W,Z) = 0 \\ \displaystyle\sum_{l=1}^{p}(w_l\mathcal{F}_{\Phi_l}(W,Z)) \ge 0 \quad \forall w_l > 0,\ \ l = 1,\dots,p. \end{cases} \tag{19}$$

From Lemma 1 and from the equality in (19), there exists some matrix E such that

$$Z = WE \tag{20}$$

and

$$\mathcal{F}_R(\mathbf{I}, E) = 0. \tag{21}$$

Such a matrix E is given by either (12) and (13). Since U satisfies $UU' = \mathbf{I}$, it is easy to check that it has no eigenvalue with geometric multiplicity greater than 1. Then it can be written

$$U = V\Lambda_U V^{-1},$$

where V is a modal matrix and Λ_U is diagonal. Matrix $\bar{U}$ involved in (12) complies with

$$\bar{U} = V(\mathbf{I} - \Lambda_U)(\mathbf{I} + \Lambda_U)^{-1}V^{-1}$$

so it has a similar diagonal form. Expressions (12) and (13) enable one to check that E has also a similar diagonal form with V as a modal matrix. Thus let r vectors v_i, $i = 1,\dots,r$, be the columns of V; one can write the definition of the right eingenstructure of E as follows:

$$Ev_i = \lambda_i v_i,\ \lambda_i \in \mathbb{C} \quad \forall i \in \{1,\dots,r\}.$$

The eigenvectors v_i can be chosen as the orthonormal basis of a subspace of dimension r so that $\sum_{i=1}^{r}(v_i v_i') = \mathbf{I}$. Then, inequality (21) yields

$$v_i'\mathcal{F}_R(I,E)v_i = \mathcal{F}_R(1,\lambda_i)v_i'v_i = 0 \quad \forall i \in \{1,\dots,r\}. \tag{22}$$

The above equality proves that each λ_i belongs to $\partial\mathcal{D}_1$ characterized by R.

Besides, from the inequality given in (19), relation (20) enables one to write:

$$v_i' \left(\sum_{l=1}^{p} (w_l \mathcal{F}_{\Phi_l}(W, WE)) \right) v_i$$

$$= \left(\sum_{l=1}^{p} (w_l \mathcal{F}_{\Phi_l}(1, \lambda_i)) \right) (v_i' W W' v_i) \geq 0 \quad \forall i \in \{1, \dots, r\},\ \forall w_l > 0,\, l = 1, \dots p$$

$$\Rightarrow \sum_{l=1}^{p} (w_l \mathcal{F}_{\Phi_l}(1, \lambda_i)) \geq 0 \quad \forall i \in \{1, \dots, r\},\ \forall w_l > 0,\, l = 1, \dots p$$

$$\Rightarrow \mathcal{F}_{\Phi_l}(1, \lambda_i) \geq 0 \quad \forall l = 1, \dots p. \tag{23}$$

The above inequality proves that each $\lambda_i \in \partial\mathcal{D}_1$ belongs to $\bar{\mathcal{E}}_1$ and therefore to $\partial\Gamma_1$.

The remaining part of the proof is analogous to the one in [16, Lemma 3]. The sum and the Kronecker product in the expressions of $\mathcal{F}_R$ and $\mathcal{F}_{\Phi_l}$ are no real obstacles and bring only slight modifications. ∎

3.2. Extended KYP lemma. Now that the set $\mathcal{S}$ is proved to be lossless, an appropriate application of the generalized S-procedure [16, Theorem 1] enables us to state the next theorem which might be considered as the main result of the paper.

THEOREM 2. *Let $\partial\mathcal{D} = \partial\Gamma_1$ be defined by (8) ($k = m = 1$ so the dependence of R, P, Φ_l and Q_l on k is still omitted). Also let matrices $A \in \mathbb{C}^{n\times n}$, $B \in \mathbb{C}^{n\times q}$ and $\Theta = \Theta' \in \mathbb{C}^{(n+m)\times(n+m)}$ be given. Assume that A is $\partial\mathcal{D}$-regular. Then the following statements are equivalent.*

(i)

$$\begin{bmatrix} (s\mathbf{I} - A)^{-1} B \\ \mathbf{I} \end{bmatrix}' \Theta \begin{bmatrix} (s\mathbf{I} - A)^{-1} B \\ \mathbf{I} \end{bmatrix} < 0 \quad \forall s \in \partial\mathcal{D}. \tag{24}$$

(ii) *There exist $n \times n$ complex Hermitian matrices P and $Q_l = Q_l > 0$, $l = 1, \dots, p$, such that*

$$\begin{bmatrix} \mathbf{I} & 0 \\ A & B \end{bmatrix}' \left(R \otimes P + \sum_{l=1}^{p} (\Phi_l \otimes Q_l) \right) \begin{bmatrix} \mathbf{I} & 0 \\ A & B \end{bmatrix} + \Theta < 0. \tag{25}$$

Proof. The proof is similar to the one of the finite frequency KYP lemma proposed in [16]. However, an outline is proposed for clarity.

Statement (i) is equivalent to $\zeta'\Theta\zeta < 0\ \forall \zeta \in \mathcal{G}$, where

$$\mathcal{G} = \left\{ \zeta = \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{C}^{n+m} \,|\, y \neq 0, sx = Ax + By \text{ for some } s \in \partial\mathcal{D} \right\}.$$

Note that with $F = \begin{bmatrix} \mathbf{I} & 0 \\ A & B \end{bmatrix}$ and by virtue of Lemma 2, the set $\mathcal{G}$ complies with

$$\mathcal{G} = \left\{ \zeta \in \mathbb{C}^{n+m} \,|\, \zeta \neq 0,\ \zeta' S \zeta \leq 0\ \forall S \in \mathcal{S} \right\},$$

where $\mathcal{S}$ is defined in (16). Since it is lossless from Lemma 3, the generalized S-procedure [16] can be applied to prove that there exists $S \in \mathcal{S}$ such that $\Theta < S$ which is equivalent to (ii). ∎

Remark 1. The assumption of $\partial\mathcal{D}$-regularity of A is necessary to ensure that $(sI - A)^{-1}$ exists. However, by virtue of the results of [3] related to the inertia of P, this assumption can be omitted when applying the theorem for the implication (ii) ⇒ (i).

For special applications of this theorem, the following cases can be considered:

(a) $p=1$, $R=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\Phi_1=0$;

(b) $p=1$, $R=\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$, $\Phi_1=0$;

(c) $p=1$, $R=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $\Phi_1=\begin{bmatrix} -1 & 0 \\ 0 & \omega_0^2 \end{bmatrix}$.

Cases (a) and (b) correspond to the continuous-time and discrete-time versions of the well-known KYP lemma [26, 23]. Case (c) corresponds to the so-called finite frequency KYP lemma proposed in [16]. It is not surprising since, as already mentioned, a part of this paper is based on ideas borrowed from [16]. Note that a recent paper [15] generalizes the results in [16]. The reader is strongly advised to refer to this article. In comparison with [16, 15], the present contribution is mostly the involvement of the sum in (25), i.e., the possibility to consider several sets $\bar{\mathcal{E}}_{k,l}$ (i.e., $p>1$).

3.3. Application to the computation of $r_{\mathcal{D}}$. Theorem 2 is now exploited to solve the stated problem.

THEOREM 3. *Let A_c, $\mathcal{D}$, $\partial\mathcal{D}$ be an uncertain matrix as defined in* (1)*, a clustering region and its boundary that reads definition* (9)*, respectively. Assume that A is $\partial\mathcal{D}$-regular. Then A_c is robustly $\partial\mathcal{D}$-regular against $\mathcal{B}(\rho)$ if there exist m Hermitian matrices P_k and positive definite Hermitian matrices $Q_{k,l}$, $k=1,\dots,m$ and $l=1,\dots,p_k$, such that*

(26)
$$\begin{bmatrix} \mathbf{I} & 0 \\ A & B \end{bmatrix}' \left(R_k \otimes P_k + \sum_{l=1}^{p_k} (\Phi_{k,l} \otimes Q_{k,l}) \right) \begin{bmatrix} \mathbf{I} & 0 \\ A & B \end{bmatrix} + \begin{bmatrix} C'C & 0 \\ 0 & -\gamma\mathbf{I} \end{bmatrix} < 0 \ \forall k \in \{1,\dots,m\},$$

where $\gamma=\rho^{-2}$.

Moreover, if A is assumed to be $\mathcal{D}$-stable then A_c remains $\mathcal{D}$-stable over $\mathcal{B}(\rho)$.

Proof. Inequality (26) corresponds to statement (ii) in Theorem 2 with some special choice of Θ and with respect to all sub-boundaries $\partial\Gamma_k$, $k=1,\dots,m$. Equivalent statement (i) holds, i.e.,

$$\Leftrightarrow ||C(s\mathbf{I}-A)^{-1}B||_2 < \sqrt{\gamma} = \rho^{-1} \quad \forall s \in \partial\Gamma_k.$$

At this stage, arguments similar to those in [22, 3] conclude the proof. ∎

Any LMI solver can reach $\rho^\star$, the maximum value of $\rho=(\sqrt{\gamma})^{-1}$ under LMI constraint (26), and then

$$\rho^\star = \varrho_{\partial\mathcal{D}} = \min_{k=1,\dots,m} \varrho_{\partial\Gamma_k}. \tag{27}$$

Furthermore, if A is $\mathcal{D}$-stable, then one gets $\rho^\star = r_{\mathcal{D}}$.

4. Examples.

4.1. First example. This simple example highlights the improvement in comparison with [3]. Consider (1) with

$$A=\begin{bmatrix} -1.1388 & 1.8993 \\ -2.5882 & -2.8612 \end{bmatrix}, \quad B=C=\mathbf{I}.$$

The nominal spectrum is $\lambda(A)=\{-2.5;-1.5\}$. Region $\mathcal{D}$ is the union of two intersecting open discs $\mathcal{D}_1$ and $\mathcal{D}_2$ centered around the eigenvalues of A and both of

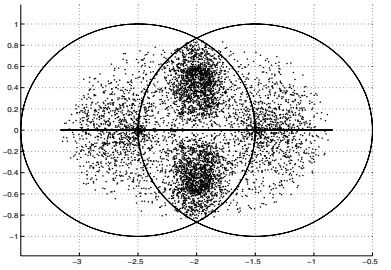


FIG. 2. *Root-locus perturbed with acceptable uncertainties.*

radius 1. Thus, A is $\mathcal{D}$-stable but is both $\partial\mathcal{D}_1$-singular and $\partial\mathcal{D}_2$-singular. That is why LMI condition (5) is found *infeasible*. The reason is that the parts of $\partial\mathcal{D}_1$ and $\partial\mathcal{D}_2$ that belong to $\mathcal{D}$ cannot be excluded before the $\partial\mathcal{D}$-regularity test is achieved. In the present work, one can consider, firstly, $\partial\mathcal{D}_1$ intersecting the outside of $\mathcal{D}_2$ and, secondly, $\partial\mathcal{D}_2$ intersecting the outside of $\mathcal{D}_1$. Then one can apply Theorem 3. It leads to the actual value of the complex $\mathcal{D}$-stability radius:

$$\rho^\star = \varrho_{\partial\mathcal{D}} = r_{\mathcal{D}} = 0.39212.$$

Five hundred instances of uncertainties Δ checking $||\Delta||_2 \leq r_{\mathcal{D}}$ are generated using standard Matlab function `rand`. The associated uncertain root-locus is plotted in Figure 2.

4.2. Second example. Before this example is detailed, it is important to notice that Theorem 2 can be used for various purposes different from pole placement analysis. For instance, assume that A is Hurwitz-stable and that matrix Θ involved in Theorem 2 is

$$\Theta = \begin{bmatrix} C'C & D'C \\ D'C & D'D - \gamma\mathbf{I} \end{bmatrix}.$$

Then, Theorem 2 becomes an extension of the well-known LMI version of the bounded real lemma [1] dedicated to the computation of the $\mathcal{H}_\infty$-norm of a linear model. Indeed, this $\mathcal{H}_\infty$-norm is nothing else than $\sqrt{\gamma}$. In this case, since A is Hurwitz-stable, P is positive definite.

To illustrate the possibilities offered by Theorem 2, a simple SISO model is considered:

$$A = \begin{bmatrix} -1.5000 & -1.0100\times10^4 & -1.50050\times10^4 & -1.0000\times10^6 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 \end{bmatrix}; \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix};$$

$$C = \begin{bmatrix} 0 & 0 & 0 & 10^6 \end{bmatrix}.$$

This realization actually corresponds to the transfer function

$$G(s) = \frac{1}{(1+0.01s+0.01s^2)(1+5\times10^{-5}s+10^{-4}s^2)},$$

which shows that there are two peaks of magnitude in the frequency response, one at $\omega_1 = 10rad/s$ and the other one at $\omega_2 = 100rad/s$. Theorem 2 will be used to analyze

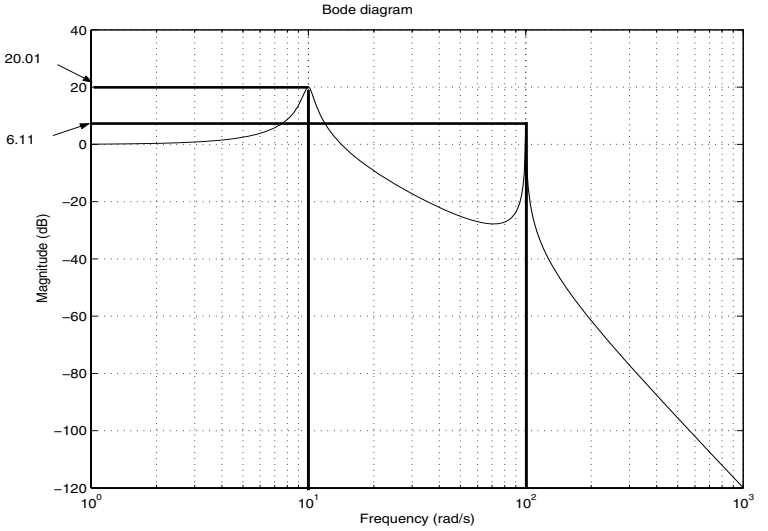


FIG. 3. *Bode diagram of the model.*

the magnitude of $G(s)$ in low frequencies, and around the peaks. The principle consists of considering the imaginary axis as the only curve $\partial\mathcal{D}_1$. To select frequencies, one can make a discerning choice of subregions $\bar{\mathcal{E}}_k$. Several tests are achieved.

(a) The static gain is obtained by considering only one region $\bar{\mathcal{E}}_1$ which is the disc centered around the origin with a radius 0.0001. Then, Theorem 2 corresponds to the finite frequency KYP lemma [16] and here yields

$$\gamma \simeq 1.0000 \Rightarrow T = 20\log(\sqrt{\gamma}) = 3.4375 \times 10^{-6} dB.$$

(b) The magnitude of the first peak is found by considering $\bar{\mathcal{E}}_1$ as the disc centered around $10\mathbf{i}$ with a radius 1. The result is

$$\gamma = 1.0228 \times 10^2 \Rightarrow T = 20\log(\sqrt{\gamma}) = 20.098 dB.$$

The same value is reached when no region $\bar{\mathcal{E}}_k$ is considered since in that case Theorem 2 is reduced to the classical bounded real lemma and the $\mathcal{H}_\infty$-norm is $\sqrt{\gamma} = 10.11$.

(c) The magnitude of the second peak is analyzed with the disc $\bar{\mathcal{E}}_1$ centered around $100\mathbf{i}$ with a radius 10. It leads to

$$\gamma = 4.0881 \Rightarrow T = 20\log(\sqrt{\gamma}) = 6.1077 dB.$$

(d) Both peaks can be analyzed together. One can consider $\bar{\mathcal{E}}_1$ as the intersection of two subregions $\bar{\mathcal{E}}_{1,1}$ and $\bar{\mathcal{E}}_{1,2}$. Here, $\bar{\mathcal{E}}_{1,1}$ is chosen as the disc centered around $55\mathbf{i}$ with a radius 47 while $\bar{\mathcal{E}}_{1,2}$ is chosen as the outside of the disc centered around $55\mathbf{i}$ with a radius 43. Of course the highest magnitude due to the first peak is recovered, i.e., $20.098dB$.

To check the above results, the Bode diagram is given in Figure 3.

5. Conclusions. In this paper, a technique is proposed to compute the complex $\mathcal{D}$-stability radius when $\mathcal{D}$ is a combination (intersection and/or union) of open half planes, discs, and outsides of discs. This point was already addressed in [3] but

an overlap of various subregions could lead to a conservative value. In the present contribution, the inconvenient is circumvented. This is made possible by the use of some extension of the KYP lemma. Such a lemma was deduced by considering insightful results of [16], particularly its generalized S-procedure. This procedure appears to be a catalyst for many future works. The reading of [16] and the more recent paper [15] is highly recommended.

As possible further investigations, the case where A, B and C are affected by polytopic uncertainty could be tackled using the matrix elimination lemma [24, Theorem 2.3.12]. Both the application of this approach in a design context and its possible exploitation for model reduction are a challenge.

REFERENCES

[1] P. Apkarian and P. Gahinet, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.

[2] C. R. Ashokkumar and R. K. Yedavalli, *Eigenstructure perturbation analysis in disjointed domains for linear uncertain systems*, Internat. J. Control, 67 (1997), pp. 887–899.

[3] O. Bachelier, D. Henrion, B. Pradin, and D. Mehdi, *Robust matrix root-clustering of a matrix in intersections or unions of subregions*, SIAM J. Control Optim., 43 (2004), pp. 1078–1093.

[4] O. Bachelier and D. Mehdi, *Robust D_U-stability analysis*, Internat. J. Robust Nonlinear Control, 13 (2003), pp. 533–558.

[5] O. Bachelier and B. Pradin, *Bounds for uncertain matrix root-clustering in a union of subregions*, Internat. J. Robust Nonlinear Control, 9 (1999), pp. 333–359.

[6] W. Bakker, J. S. Luo, and A. Johnson, *Performance robustness bounds for linear systems to guarantee root-clustering in second order subregions of the complex plane*, in Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 3467–3472.

[7] J. Bosche, O. Bachelier, and D. Mehdi, *An approach for robust matrix root-clustering analysis in a union of regions*, IMA J. Math. Control Inform., 22 (2005), pp. 227–239.

[8] S. Boyd, L. El Ghaoui, E. Féron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[9] M. Chilali and P. Gahinet, *H_∞ design with pole placement constraints: An LMI approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 358–367.

[10] M. Chilali, P. Gahinet, and P. Apkarian, *Robust pole placement in LMI regions*, IEEE Trans. Automat. Control, 44 (1999), pp. 2257–2269.

[11] J. D. Gardiner, *Computation of stability bounds for state-space models with structured uncertainty*, IEEE Trans. Automat. Control, 42 (1997), pp. 253–256.

[12] S. Gutman and E. I. Jury, *A general theory for matrix root-clustering in subregions of the complex plane*, IEEE Trans. Automat. Control, 26 (1981), pp. 853–863.

[13] D. Hinrichsen and B. Kelb, *Stability radii and spectral value sets for real matrix perturbations*, in Systems and Networks: Mathematical Theory and Applications, Vol. II, Invited and contributed papers, Akademie-Verlag, Berlin, 1994, pp. 217–220.

[14] J. L. Howland, *Matrix equations and the separation of matrix eigenvalues*, J. Math. Anal. Appl., 33 (1971), pp. 683–691.

[15] T. Iwasaki and S. Hara, *Generalized KYP lemma: Unified frequency domain inequalities with design applications*, IEEE Trans. Automat. Control, 50 (2005), pp. 41–59.

[16] T. Iwasaki, G. Meinsma, and M. Fu, *Generalized S-procedure and finite frequency KYP lemma*, Math. Probl. Engrg., 6 (2000), pp. 305–320.

[17] P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H_∞ control theory*, IEEE Transactions on Automatic Control, 35 (1990), pp. 356–361.

[18] A. Packard and J. Doyle, *Quadratic stability with real and complex perturbations*, IEEE Trans. Automat. Control, 35 (1990), pp. 198–201.

[19] R. V. Patel and M. Toda, *Quantitative measures of robustness for multivariable systems*, in Proceedings of the Joint Automatic Control Conference, TP-8A, San Francisco, CA, 1980.

[20] D. Peaucelle, D. Arzelier, O. Bachelier, and J. Bernussou, *A new robust D-stability condition for real convex polytopic uncertainty*, Systems Control Lett., 40 (2000), pp. 21–30.

[21] I. R. Petersen, *A stabilization algorithm for a class of uncertain linear systems*, Systems Control Lett., 8 (1987), pp. 351–357.

[22] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, and J. C. Doyle, *A formula for computation of the real stability radius*, Automatica J. IFAC, 31 (1995), pp. 879–890.

[23] A. Rantzer, *On the Kalman-Yakubovich-Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.

[24] R. E. Skelton, T. Iwasaki, and K. Grigoriadis, *A Unified Algebraic Approach to Linear Control Design*, Taylor & Francis Series in Systems and Control, Taylor & Francis, London, 1998.

[25] S. Wang, *Robust pole clustering in a good ride quality region of aircraft for matrices with structured uncertainties*, Automatica J. IFAC, 39 (2003), pp. 525–532.

[26] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[27] R. K. Yedavalli, *Flight control application of new stability robustness bounds for linear uncertain systems*, J. Guidance Control Dyn., 16 (1993), pp. 1032–1037.

[28] R. K. Yedavalli, *Robust root clustering for linear uncertain systems using generalized Lyapunov theory*, Automatica J. IFAC, 29 (1993), pp. 237–240.

[29] K. Zhou and P. P. Khargonekar, *Stability robustness bounds for linear state-space models with structured uncertainty*, IEEE Trans. Automat. Control, 32 (1987), pp. 621–623.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 1, pp. 382–387

ERRATA: ELIMINATION OF STRICT CONVERGENCE IN OPTIMIZATION*

DUŠAN BEDNAŘÍK† AND KAREL PASTOR‡

Abstract. We revise our previous paper [*SIAM J. Control Optim.*, 43 (2004), pp. 1063–1077]. We supply an assumption missing in the first part of the assertion of Corollary 4.6, and we prove that the second part of this corollary remains true also under conditions given in the previous paper. Finally, we comment on new contributions and pose the open question in the topic.

Key words. generalized second-order directional derivative, second-order optimality conditions, minimal cusco mapping

AMS subject classifications. 49K35, 26B02, 54C60

DOI. 10.1137/050636309

1. Introduction and preliminaries. In [BP1] we studied the generalized second-order directional derivatives to obtain sufficient and necessary optimality conditions, which we compare with the previous results.

Let X be a real Banach space with the norm $\|\cdot\|$, let X^* be the topological dual space of X, and let $\langle x^*, x\rangle$ be a canonical pair, where $x^* \in X^*$, $x \in X$. S_X denotes the set $\{x \in X : \|x\| = 1\}$.

For a function $f : X \to \mathbb{R}$, we denote the first-order directional derivative of f at $x \in X$ in the direction $v \in X$ by

$$f'(x; v) = \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t}$$

and the lower and upper Clarke directional derivatives of f at $x \in X$ in the direction $v \in X$, respectively, by

$$f_\circ(x; v) = \liminf_{y \to x, t \downarrow 0} \frac{f(y + tv) - f(y)}{t},$$

$$f^\circ(x; v) = \limsup_{y \to x, t \downarrow 0} \frac{f(y + tv) - f(y)}{t}.$$

The Cominetti–Correa [CC] lower and upper generalized second-order directional derivatives of f at x in the direction $(u, v) \in X \times X$ are defined, respectively, by

$$f_\infty(x; u, v) = \liminf_{y \to x,\ s,t \downarrow 0} \frac{f(y + tu + sv) - f(y + tu) - f(y + sv) + f(y)}{st},$$

$$f^\infty(x; u, v) = \limsup_{y \to x,\ s,t \downarrow 0} \frac{f(y + tu + sv) - f(y + tu) - f(y + sv) + f(y)}{st}.$$

*Received by the editors July 18, 2005; accepted for publication (in revised form) September 26, 2005; published electronically March 15, 2006.
http://www.siam.org/journals/sicon/45-1/63630.html

†Department of Mathematics, University of Hradec Králové, Rokitanského 62, 500 03 Hradec Králové, Czech Republic (dbednarik@seznam.cz).

‡Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Tř. Svobody 26, 771 46 Olomouc, Czech Republic (pastor@inf.upol.cz). This author was supported by the Council of Czech Government (MSM 6198959214).

We note that for a set-valued mapping $F : \mathbb{R} \rightsquigarrow \mathbb{R}$, it is a natural to denote

$$\liminf_{t\downarrow 0} F(t) = \liminf_{t\downarrow 0} \inf\{a : a \in F(t)\},$$

etc. Then, we define

$$f'^{L}(x;u,v) = \liminf_{t\downarrow 0} \frac{\langle \partial_c f(x+tu) - \partial_c f(x), v\rangle}{t},$$
$$f'^{U}(x;u,v) = \limsup_{t\downarrow 0} \frac{\langle \partial_c f(x+tu) - \partial_c f(x), v\rangle}{t},$$

where f is Lipschitz near $x \in X$, $(u,v) \in X \times X$ and $\partial_c f(x)$ stands for the Clarke subdifferential [Cl] at x defined by

$$\partial_c f(x) = \{x^* \in X^* : \langle x^*, v\rangle \le f^\circ(x;v) \quad \forall v \in X\}.$$

Finally, for a function which is directonally differentiable, we denote

$$f'^{l}(x;u,v) = \liminf_{t\downarrow 0} \frac{f'(x+tu;v) - f'(x;v)}{t},$$
$$f'^{u}(x;u,v) = \limsup_{t\downarrow 0} \frac{f'(x+tu;v) - f'(x;v)}{t}.$$

Let us recall that a locally Lipschitz function $f : X \mapsto \mathbb{R}$ is said to be regular at $x \in X$ provided that $f'(x;v) = f^\circ(x;v)$ for every $v \in X$. If we say that f is regularly locally Lipschitz function on a subset $A \subset X$, we mean that f is locally Lipschitz on A and regular at every point $x \in A$. Further f is said to be regularly locally Lipschitz near $x \in X$ if f is regularly locally Lipschitz function on some neighborhood of x. If f is assumed to be regularly locally Lipschitz on the whole space X, we simply say that f is regularly locally Lipschitz function.

PROPOSITION 1.1 (see [BP1, Proposition 4.4]). *Let $f : X \mapsto \mathbb{R}$ be a regularly locally Lipschitz function, $x, h \in X$. Then*

$$f'^{L}(x;h,h) = f'^{l}(x;h,h).$$

Because it holds [CC] for f to be directionally differentiable,

$$f_\infty(x;u,v) = \liminf_{y\to x, t\downarrow 0} \frac{f'(y+tu;v) - f'(y;v)}{t},$$

and

$$f^\infty(x;u,v) = \limsup_{y\to x, t\downarrow 0} \frac{f'(y+tu;v) - f'(y;v)}{t},$$

Proposition 1.1 implies the following corollary.

COROLLARY 1.2 (see [BP1, Corollary 4.5]). *Let $f : X \mapsto \mathbb{R}$ be a regularly locally Lipschitz function, $x, h \in X$. Then*

$$f_\infty(x;h,h) \le f'^{L}(x;h,h).$$

Since f can be a regularly locally Lipschitz function, but $(-f)$ cannot, we cannot derive (as we unfortunately did in [BP1, Corollary 4.6]) from Proposition 1.1 by using

lim inf and lim sup calculus that $f'^U(x;h,h) = f'^u(x;h,h)$ and that $f^\infty(x;h,h) \geq f'^U(x;h,h)$. By this way, we can state only the next corollary.

COROLLARY 1.3. *Let $f : X \mapsto \mathbb{R}$ be such function that $(-f)$ is regularly locally Lipschitz, and let $x, h \in X$. Then*

$$f'^U(x;h,h) = f'^u(x;h,h),$$

$$f^\infty(x;h,h) \geq f'^U(x;h,h). \tag{1.1}$$

It follows from definitions that for regularly locally Lipschitz functions we have $f'^u(x;h,h) \leq f'^U(x;h,h)$, but the equality is not true—for example, considering a function $f : \mathbb{R} \mapsto \mathbb{R} : f(x) = |x|$; then $f'^u(0;1,1) = 0 < +\infty = f'^U(0;1,1)$.

Our mistake in [BP1] mentioned above and [BP1, Theorem 6.4] lead to the conclusion that the condition $f'^u(x;h,h) \geq 0$ for every $h \in S_X$ is necessary for x to be a local minimum of a regularly locally Lipschitz function f. As we show by the following example, this conclusion is false.

Example 1.4. Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a function defined as follows:

$$f(x) = \begin{cases} lnx & \text{for } x \geq 1, \\ 0 & \text{for } x < 1. \end{cases}$$

Then f attains a local minimum at 1, and f is regularly locally Lipschitz. On the other hand, because $f'(x;1) = \frac{1}{x}$ for every $x \geq 1$, it holds

$$f'^u(1;1,1) = \limsup_{t \downarrow 0} \frac{\frac{1}{1+t} - 1}{t} = -1.$$

2. Main result. The main aim of this paper is to show that inequality (1.1) from Corollary 1.3 is really true also for regularly locally Lipschitz functions.

For the proof we use some properties of minimal cusco mappings. A set-valued mapping F from a topological space A into nonempty subsets of a Hausdorff locally convex space Y is called cusco if $F(t)$ is compact and convex for each $t \in A$ and F is upper-semicontinuous (i.e., $\{t \in A : F(t) \subset U\}$ is an open subset for each open subset U of Y).

Furthermore, F is said to be *minimal cusco* on A if its graph does not contain the graph of any other cusco on A. The Clarke subdifferential of a regularly locally Lipchitz function [Mo] or maximal monotone operator (see, e.g., [Ph]) are the well-known examples of minimal cusco mappings.

The following lemma is a particular case of a more general result given in [J, Corollary 4.1].

LEMMA 2.1. *Let $D \subset \mathbb{R}$ be an open set, and let $F : D \rightsquigarrow \mathbb{R}$ be minimal cusco. Then the set $\{x \in D : F$ is single-valued at $x\}$ is dense in D.*

LEMMA 2.2 (see [BM, Theorem 3.7]). *For a cusco mapping Φ, from a topological space A into subsets of a separated locally convex topological space X, the following conditions are equivalent:*

(i) *Φ is a minimal cusco on A.*

(ii) *For any densely defined selection σ of Φ, $CSC(\sigma) = \Phi$,*

$$CSC(\sigma)(x) = \bigcap \{\overline{co}\ \sigma(V) : V \text{ is open neighborhood of } x\}.$$

By the analogous way as in the lemma near 2.3.7 in [Cl], we can obtain the following.

LEMMA 2.3. *Let $f : X \mapsto \mathbb{R}$ be a regularly locally Lipschitz function on an open subset containing the line segment $[x, y]$. Then the function $g : [0, 1] \mapsto \mathbb{R}$ defined by $g(t) = f(x + t(y - x))$ is regularly locally Lipschitz on $(0, 1)$, and we have*

$$\partial_c g(t) = \langle \partial_c f(x + t(y - x)), y - x \rangle.$$

THEOREM 2.4. *Let $f : X \mapsto \mathbb{R}$ be a regularly locally Lipschitz function near $x \in X$ and let $h \in S_X$. Then*

$$f'^U(x; h, h) \le f^\infty(x; h, h).$$

Proof. Since

$$f^\circ(x; v) = \max\{\langle x^*, v\rangle; x^* \in \partial_c f(x)\}, \quad f_\circ(x; v) = \min\{\langle x^*, v\rangle, x^* \in \partial_c f(x)\},$$

it holds

$$f'^U(x; h, h) = \limsup_{t \downarrow 0} \frac{f^\circ(x + th; h) - f_\circ(x; h)}{t}.$$

Let $\lambda \in \mathbb{R}$ be such that

$$\lambda < f'^U(x; h, h).$$

Then there exists a sequence $\{t_n\}_{n=1}^{+\infty}$, $t_n > 0$, $\lim_{n\to+\infty} t_n = 0$, with the property

$$\lambda < \frac{f^\circ(x + t_n h; h) - f_\circ(x; h)}{t_n} \quad \forall n \in \mathbb{N}. \tag{2.1}$$

Due to Lemma 2.3, the graph of multivalued mapping $F : \{x + th; t \in \mathbb{R}\} \rightsquigarrow \mathbb{R} : z \rightsquigarrow \langle \partial_c f(z), h\rangle$ is the same as the graph of multivalued mapping $\partial_c g$, where $g : \mathbb{R} \mapsto \mathbb{R} : g(t) = f(x + th)$ is regularly locally Lipschitz function.

Since multivalued mapping $\partial_c g$ is minimal cusco, Lemma 2.1 implies that multivalued mapping F is single-valued on the dense subset of $\{x + th; t \in \mathbb{R}\}$. Thus we can take a densely defined selection σ of F with the property

$$\sigma(z) = f^\circ(z; h) = f_\circ(z; h), \quad z \in D(\sigma), \tag{2.2}$$

where $D(\sigma)$ denotes the domain of σ. Now, for every $n \in \mathbb{N}$, there exists a sequence $\{t_n^k\}_{k=1}^{+\infty}$ such that $F(x + t_n^k h) = \sigma(x + t_n^k h)$, $\lim_{k\to+\infty} t_n^k = t_n$ and $\lim_{k\to+\infty} f^\circ(x + t_n^k h) = f^\circ(x + t_n h)$, because considering a densely defined selection σ_n of F, where

$$\sigma_n(z) = \begin{cases} \sigma(z) & \text{for every } z \in D(\sigma), \\ f^\circ(x + t_n h) & \text{for } z = x + t_n h, \end{cases}$$

it suffices to use Lemma 2.2.

Following inequality (2.1), we can find $k(n) \in \mathbb{N}$ for every $n \in \mathbb{N}$ such that

$$\lambda < \frac{f^\circ(x + t_n^{k(n)} h; h) - f_\circ(x; h)}{t_n^{k(n)}}. \tag{2.3}$$

Using (2.2), it holds

$$f^\circ(x + t_n^{k(n)} h; h) = \lim_{y \to x, s \downarrow 0} \frac{f(y + t_n^{k(n)} h + sh) - f(y + t_n^{k(n)} h)}{s}. \tag{2.4}$$

There exist sequences $\{y_l\}_{l=1}^{+\infty}$ and $\{s_l\}_{l=1}^{+\infty}$, $\lim_{l\to+\infty} y_l = x$, $s_l > 0$, $\lim_{l\to+\infty} s_l = 0$, with the property

$$f_\circ(x;h) = \liminf_{y\to x, s\downarrow 0} \frac{f(y+sh)-f(y)}{s} = \lim_{l\to+\infty} \frac{f(y_l+s_l h)-f(y_l)}{s_l}. \tag{2.5}$$

From formulas (2.3), (2.4), and (2.5), we obtain for every $n \in \mathbb{N}$

$$\lambda < \frac{\lim_{l\to+\infty} \frac{f(y_l+t_n^{k(n)}h+s_l h)-f(y_l+t_n^{k(n)}h)}{s_l} - \lim_{l\to+\infty} \frac{f(y_l+s_l h)-f(y_l)}{s_l}}{t_n^{k(n)}}.$$

Hence,

$$\lambda < \frac{\lim_{l\to+\infty} \frac{f(y_l+t_n^{k(n)}h+s_l h)-f(y_l+t_n^{k(n)}h)-f(y_l+s_l h)+f(y_l)}{s_l}}{t_n^{k(n)}} \quad \forall n \in \mathbb{N}.$$

Thus, for every $n \in \mathbb{N}$, we can find $y_{l(n)} \in X$ and $s_{l(n)} > 0$ such that it holds

$$\lambda < \frac{f(y_{l(n)}+t_n^{k(n)}h+s_{l(n)}h)-f(y_{l(n)}+t_n^{k(n)}h)-f(y_{l(n)}+s_{l(n)}h)+f(y_{l(n)})}{s_{l(n)}t_n^{k(n)}}.$$

Setting $t_n = t_n^{k(n)}$, $y_n = y_{l(n)}$, and $s_n = s_{l(n)}$, for every $n \in \mathbb{N}$, we have

$$\lambda < \frac{f(y_n+t_n h+s_n h)-f(y_n+t_n h)-f(y_n+s_n h)+f(y_n)}{s_n t_n}.$$

Limiting $n \to +\infty$, we obtain

$$\lambda \le \limsup_{n\to+\infty} \frac{f(y_n+t_n h+s_n h)-f(y_n+t_n h)-f(y_n+s_n h)+f(y_n)}{s_n t_n} \le f^\infty(x;h,h).$$

Because λ was arbitrary with the property $\lambda < f'^U(x;h,h)$, it holds

$$f'^U(x;h,h) \le f^\infty(x;h,h). \qquad \square$$

3. Final remarks and open question. In this section, we mention the progress in this topic since the paper [BP1] and set an open question.

Motivated by [GGR], in [BP2] we have generalized the second-order sufficient unconstrained optimality conditions [BZ, Theorem 3.2], [GGR, Theorem 2], [BP1, Theorem 3.2], and [CC, Proposition 5.2] by the following way.

THEOREM 3.1 (see [BP2, Theorem 7]). *Let $f : \mathbb{R}^N \mapsto \mathbb{R}$ be continuous near $x \in \mathbb{R}^N$ and let be ℓ-stable at x. If $f^\ell(x;h) = 0$ for every $h \in S_{\mathbb{R}^N}$, and*

$$\underline{f}'^l_P(x;h) > 0 \quad \forall h \in S_{\mathbb{R}^N},$$

then x is an isolated minimizer of order 2 for f.

In the previous theorem,

$$f^l(x;h) = \liminf_{t\downarrow 0} \frac{f(x+th)-f(x)}{t},$$
$$\underline{f}'^l_P(x;h) = \liminf_{t\downarrow 0} \frac{f(x+th)-f(x)-tf^\ell(x;h)}{t^2/2}.$$

By ℓ-stability of f at x we mean that for some neighborhood U of x and $K > 0$ it holds

$$|f^\ell(y;h) - f^\ell(x;h)| \le K\|y - x\| \quad \forall y \in U, \forall h \in S_X. \tag{3.1}$$

Finally, recall that $x \in X$ is an isolated minimizer of order k ($k \in \mathbb{N}$) for a function $f : X \mapsto \mathbb{R}$ if there are a neighborhood V of x and $A > 0$ satisfying $f(y) \ge f(x)+A\|y-x\|^k$ for every $y \in V$.

Open question. Let $\overline{\ell}$-stability of f at x means that for some neighborhood U of x and $K > 0$ it holds

$$|f^\ell(z; z-y) - f^\ell(y; z-y)| \le K\|z-y\|^2 \quad \forall y, z \in U.$$

It seems that the notions of ℓ-stability and $\overline{\ell}$-stability are quite different (really?) and there arises the question whether we can replace the condition to be ℓ-stable by the condition to be $\overline{\ell}$-stable in Theorem 3.1.

REFERENCES

[BM] J. M. Borwein and W. B. Moors, *Essentially strictly differentiable Lipschitz functions*, J. Funct. Anal., 149 (1997), pp. 305–351.

[BP1] D. Bednařík and K. Pastor, *Elimination of strict convergence in optimization*, SIAM J. Control Optim., 43 (2004), pp. 1063–1077.

[BP2] D. Bednařík and K. Pastor, *Using the Peano derivative in unconstrained optimization*, Nonlinear Programming, submitted.

[BZ] A. Ben-Tal and J. Zowe, *Directional derivatives in nonsmooth optimization*, J. Optim. Theory Appl., 47 (1985), pp. 483–490.

[CC] R. Cominetti and R. Correa, *A generalized second-order derivative in nonsmooth optimization*, SIAM J. Control Optim., 28 (1990), pp. 789–809.

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

[GGR] I. Ginchev, A. Guerraggio, and M. Rocca, *From scalar to vector optimization*, Appl. Math., 51 (2006), pp. 5–36.

[J] L. Jokl, *Minimal convex-valued weak* usco correspondences and the Radon-Nikodym property*, Comment. Math. Univ. Carolin. 28 (1987), pp. 353–375.

[Mo] W. B. Moors, *A characterization of minimal subdifferential mappings of locally Lipschitz functions*, Set-Valued Anal. 3, (1995), pp. 129–141.

[Ph] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, Springer-Verlag, Berlin, 1993.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 389–413

ON THE GEOMETRY OF THE SOLUTIONS OF THE COVER PROBLEM*

F. PUERTA†, X. PUERTA†, AND I. ZABALLA‡

Abstract. For a given system $\Sigma(A, B)$ and a subspace $\mathcal{S}$, the cover problem consists of finding all (A, B)-invariant subspaces containing $\mathcal{S}$. For controllable systems, the set of these subspaces can be suitably stratified. In this paper, necessary and sufficient conditions are given for the cover problem to have a solution on a given strata. Then the geometry of these solutions is studied. In particular, the set of the solutions is provided with a differentiable structure and a parameterization of all solutions is obtained through a coordinate atlas of the corresponding smooth manifold.

Key words. cover problem, controlled invariant subspace, smooth manifold, coordinate chart, Brunovsky indices

AMS subject classifications. 93B29, 15A03

DOI. 10.1137/S0363012904443087

1. Introduction.

Given a time-invariant linear system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

with A, B, $n \times n$, and $n \times m$ matrices, respectively, a subspace $\mathcal{V}$ of $\mathbb{K}^n$ ($\mathbb{K}$ is both the field of real or complex numbers) is said to be controlled invariant with respect to (A, B) or, simply, (A, B)-invariant, if $A\mathcal{V} \subset \mathcal{V} + \operatorname{Im} B$ (see, for example, [3]).

For a given subspace $\mathcal{S}$ of $\mathbb{K}^n$, the cover problem associated to the above system consists of finding all (A, B)-invariant subspaces $\mathcal{V}$ such that $\mathcal{S} \subset \mathcal{V}$. A number of important problems are related to the cover problem. For instance, the *disturbance decoupling problem* (DDP) calls for a state feedback transformation such that an output $y(t) = Cx(t)$ of the closed-loop system does not depend on an input disturbance given by $Qv(t)$. As shown in, for example, [11, Th. 4.2] this is possible if and only if $\operatorname{Im} Q \subset \mathcal{V}^* = \max \mathcal{V}(A, B, \ker C)$, the maximal (A, B)-invariant subspace contained in $\operatorname{Ker} C$. Very nice and efficient algorithms are known in order to compute subspace $\mathcal{V}^*$ (see, for example, [11, 3]). In this paper we are interested in computing not only $\mathcal{V}^*$ but all (A, B)-invariant subspaces containing a given subspace. In fact, important questions concerning this kind of subspace, as, for example, robustness with regard to matrix Q, lead us to the problem of parameterizing the set of solutions of this cover problem. Moreover, some other important problems, like the *observer* and *model-matching* problems, are related to the cover problem (see [1]) as well. Actually, as said by Antoulas in that paper, "The cover problems provide a unifying framework for the solution of a number of important feedforward as well as feedback problems in linear system theory."

*Received by the editors April 20, 2004; accepted for publication (in revised form) July 4, 2005; published electronically March 17, 2006.

http://www.siam.org/journals/sicon/45-2/44308.html

†Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Spain (puerta@ma1.upc.es, coll@ma1.upc.es). This author was partially supported by the MCyT, Proyecto de Investigación MTM2004-06389-C02-02.

‡Departamento de Matemática Aplicada y EIO, Universidad del País Vasco, Apartado 640, 48080 Bilbao, Spain (mepzatej@lg.ehu.es). This author was partially supported by the MCyT, Proyecto de Investigación MTM2004-06389-C02-01, and the UPV/EHU, 9/UPV00100.310-14578/2002.

Antoulas' paper is one of the main contributions to the solution of the cover problem. Antoulas links together the cover problem and the nice partial realization problem providing, through the last one, a parameterization of the solutions. Our approach is based on considering the dual problem. We recall that $\mathcal{V}$ is a conditioned invariant subspace with regard to (C, A) or, simply, a (C, A)-invariant subspace, if $A(\mathcal{V} \cap \operatorname{Ker} C) \subset \mathcal{V}$. By dualilty, $\mathcal{V}$ is (A, B)-invariant if and only if $\mathcal{V}^{\perp}$ is (B^t, A^t)-invariant. Therefore, the map $\mathcal{V} \mapsto \mathcal{V}^{\perp}$ defines a bijection between the solutions of the cover problem of dimension d, with respect to a pair (A, B) and a subspace $\mathcal{S}$, and the set of (B^t, A^t)-invariant subspaces of $\mathcal{S}^{\perp}$ of codimension d. For a give pair (C, A), the problem of parameterizing the set of δ-dimensional (C, A)-invariant subspaces will be called the *dual cover problem*. Of course, the cover problem and the dual one are equivalent. In this note we focus on the last one due to the following interpretation of its solutions. We denote by $Gr_{\delta}(\mathcal{S})$ the Grassmann manifold formed by the δ-dimensional subspaces of $\mathcal{S}$. Then, the set of solutions of the dual cover problem is just the intersection of $Gr_{\delta}(\mathcal{S})$ with the set of δ-dimensional (C, A)-invariant subspaces, which is an stratified manifold as shown in [5]. Our goal here is to gain insight into the parameterization of the solutions of the (dual) cover problem, through the parameterization of the strata of the set of δ-dimensional (C, A)-invariant subspaces given in [6].

A natural requirement in many problems for which the cover problem is a generalization is that of stability. Different approaches can be taken to this issue. For example, for the DDP with stability, the approaches in [11] and [3] are different. In the first one, the solution is connected to the maximal internally stabilizable (A, B)-invariant subspace, while in the second one the minimal (A, B)-invariant subspace self-bounded with respect to $\operatorname{Ker} C$ is considered. Even more general situations can be considered, like, for example, the DDP with pole placement. It is shown in [10] that a necessary and sufficient condition for this problem to be solvable is that (A, B) be controllable and $\operatorname{Im} Q \subset \mathcal{R}^*$, where $\mathcal{R}^*$ is the maximal (A, B)-controllability subspace contained in $\operatorname{Ker} C$. Among other, possibly better known, characterizations, the controllability subspaces can be defined as those (A, B)-invariant subspaces for which the restriction of (A, B) to the subspace is completely controllable (see [6]). The set of controllability susbspaces is also a stratified manifold [6] and for each strata a coordinate atlas that locally parameterizes them has been obtained [8]. This parameterization eventually could be used to study the geometry of the cover problem with pole placement, but it must be said that the differentiable structure of the (A, B)-controllability subspaces is more complicated than that of the (C, A)-invariant ones.

Throughout this note, $\mathcal{M}_{m,n}$ denotes the set of $m \times n$ matrices with elements in $\mathbb{K}$, $\mathcal{M}_n$, the set of the $n \times n$-square ones and $\mathcal{M}^*_{m,n}$, the subset of the full column rank $m \times n$ matrices. If $X \in \mathcal{M}^*_{m,n}$, then $[X]$ denotes the subspace spanned by the columns of X. Finally, if $X \in \mathcal{M}_{m,n}$, $X^* \in \mathcal{M}_{n,m}$ is its conjugate transpose.

The structure of this paper is as follows. In section 2 we recall the parameterization of the set of (C, A)-invariant subspaces given in [7]. In section 3 we describe the solutions of the (dual) cover problem in terms of the solution of a set of algebraic linear equations on the parameters of the mentioned parameterization. From this description we derive explicit necessary and sufficient conditions for the (dual) cover problem to have a solution with given observability indices (Theorem 3.4). In section 4 we prove a simple necessary and sufficient condition in order to ensure that the above conditions are satisfied *generically* (Theorem 4.9). In addition, for the generic case and if the observability indices of the restriction are not required, we give an easy-to-check necessary and sufficient condition for the (dual) cover problem to have

a solution. In section 5 we provide an example of application of the previous results: we propose an algorithm which implements these results to parameterize the output matrices C for which $y = Cx$ can be made insensitive to the input disturbance. In section 6 we show that the set of solutions of the cover problem is a stratified smooth manifold with an orbit space structure similar to the set of (C, A)-invariant subspaces and we compute its dimension (Theorem 6.1). Finally we parameterize this manifold by means of coordinate charts (Theorem 6.9) and we prove that it is connected.

2. Preliminaries. Throughout this note (C, A) is an observable pair of matrices of $\mathcal{M}_{m,n} \times \mathcal{M}_n$ with observability indices $\underline{k} = (k_1, \ldots, k_r)$ which, without loss of generality, we can assume is in the dual Brunovsky canonical form (see [2, Lem. 4.4]).

In this section we recall the results on the differentiable structure and parameterization of the strata of the set of (C, A)-invariant subspaces that we use in the next sections. These strata are defined by fixing specific Brunovsky form for the restriction of (C, A) to the (C, A)-invariant subspaces (see [5] and [7] for more details and proofs). More precisely, we fix a set of indices $\underline{h} = (h_1, \ldots, h_s)$ of a possible restriction of (C, A) (we call $\underline{h}$ *compatible* with $\underline{k}$), which is any finite set of integers with the condition $h_i \leq k_i$ for $i = 1, 2, \ldots, s$. The parameterization of the set of (C, A)-invariant subspaces for which the restriction of (C, A) to them has indices $\underline{h}$ is simpler if it is refered to the dual partitions. So, let the conjugate partitions of $\underline{k}$ and $\underline{h}$ be $\underline{r} = (r_1, \ldots, r_k)$ and $\underline{s} = (s_1, \ldots, s_h)$, respectively, with $r = r_1 \geq \cdots \geq r_k > 0$ and $s = s_1 \geq \cdots \geq s_h > 0$. (Recall that if $\underline{k} = (k_1, k_2, \ldots)$, then $\underline{r} = (r_1, r_2, \ldots)$ is its conjugate partition if and only if $r_j = \#\{i : k_i \geq j\}$, where # stands for cardinality.) From now on we will assume that $s_i := 0$ for $i > h$, $r_i := 0$ for $i > k$, $h_i := 0$ for $i > s$, and $k_i := 0$ for $i > r$. If $(k_1, \ldots, k_r)$ are the observability indices of (C, A), then $(r_1, \ldots, r_k)$ are called the *Brunovsky indices* of (C, A) (see [4]).

Notice that $k_i \geq h_i$ for $i = 1, 2, \ldots, r$ if and only if $r_i \geq s_i$ for $i = 1, 2, \ldots, k$. In other words, $\underline{h}$ and $\underline{k}$ are compatible if and only if $\underline{s}$ and $\underline{r}$ are compatible. In what follows we also assume that $\underline{s}$ and $\underline{r}$ are compatible partitions and we denote by $\mathrm{Inv}\,(\underline{r}, \underline{s})$ the set of (C, A)-conditioned invariant subspaces for which the restriction of (C, A) to each one of them has Brunovsky indices $\underline{s}$.

DEFINITION 2.1. *Let $\mathcal{M}(\underline{r}, \underline{s})$ denote the set of matrices $X \in \mathcal{M}_{n,d}$ which are partitioned into blocks $X = (X_{ij})$, $1 \leq i \leq k$, $1 \leq j \leq h$, such that*

(a) $X_{ij} \in \mathcal{M}_{r_i, s_j}$.

(b) $X_{ij} = 0$ *if* $i < j$.

(c) *If $i \geq j$, X_{ij} can be partitioned into blocks $X_{ij} = (Z_{i\alpha}^{i-j+1})$, $1 \leq \alpha \leq h - j + 1$, where*

$$Z_{i\alpha}^{i-j+1} = \begin{pmatrix} Y_{i\alpha}^{i-j+1} \\ 0 \end{pmatrix}$$

with $Y_{i\alpha}^{i-j+1}$ having size $r_{h-\alpha+i-j+1} \times (s_{h-\alpha+1} - s_{h-\alpha+2})$. That is to say, X_{ij} is a block upper triangular matrix with blocks of sizes $(r_i - r_{i+1}) \times (s_i - s_{i+1})$ (if $r_i = r_{i+1}$ or $s_i = s_{i+1}$, the correspondig block vanishes). We call this block decomposition the standard *block decomposition of X_{ij}.*

(d) *$X_{i+1,j+1}$ is obtained from X_{ij} by removing the last $s_j - s_{j+1}$ columns and the last $r_i - r_{i+1}$ rows (the last column and row blocks) for $1 \leq i \leq k$, $1 \leq j \leq h$.*

(e) $\operatorname{rank} X_{ii} = s_i$, $1 \leq i \leq k$.

We denote by $\mathcal{G}(\underline{s})$ the set $\mathcal{M}(\underline{s}, \underline{s})$, which turns out to be a subgroup of $\mathrm{Gl}(n)$.

Example 2.2. If $k = 4$ and $h = 3$, the elements of $\mathcal{M}(\underline{r}, \underline{s})$ and $\mathcal{G}(\underline{s})$ are, respectively, matrices having the form displayed in the following matrices:

$$\begin{array}{l} r_4 \\ r_3 - r_4 \\ r_2 - r_3 \\ r_1 - r_2 \\ r_4 \\ r_3 - r_4 \\ r_2 - r_3 \\ r_4 \\ r_3 - r_4 \\ r_4 \end{array} \left(\begin{array}{ccc|cc|c} X_{11}^1 & X_{12}^1 & X_{13}^1 & & & \\ X_{21}^1 & X_{22}^1 & X_{23}^1 & & & \\ 0 & X_{32}^1 & X_{33}^1 & \multicolumn{2}{c|}{\mathbf{0}} & \mathbf{0} \\ 0 & 0 & X_{43}^1 & & & \\ \hline X_{11}^2 & X_{12}^2 & X_{13}^2 & X_{11}^1 & X_{12}^1 & \\ 0 & X_{22}^2 & X_{23}^2 & X_{21}^1 & X_{22}^1 & \mathbf{0} \\ 0 & 0 & X_{33}^2 & 0 & X_{32}^1 & \\ \hline 0 & X_{12}^3 & X_{13}^3 & X_{11}^2 & X_{12}^2 & X_{11}^1 \\ 0 & 0 & X_{23}^3 & 0 & X_{22}^2 & X_{21}^1 \\ \hline 0 & 0 & X_{13}^4 & 0 & X_{12}^3 & X_{11}^2 \end{array}\right),$$

$$\begin{array}{l} s_3 \\ s_2 - s_3 \\ s_1 - s_2 \\ s_3 \\ s_2 - s_3 \\ s_3 \end{array} \left(\begin{array}{ccc|cc|c} T_{11}^1 & T_{12}^1 & T_{13}^1 & & & \\ 0 & T_{22}^1 & T_{23}^1 & \multicolumn{2}{c|}{\mathbf{0}} & \mathbf{0} \\ 0 & 0 & T_{33}^1 & & & \\ \hline 0 & T_{12}^2 & T_{13}^2 & T_{11}^1 & T_{12}^1 & \\ 0 & 0 & T_{23}^2 & 0 & T_{22}^1 & \mathbf{0} \\ \hline 0 & 0 & T_{13}^3 & 0 & T_{12}^2 & T_{11}^1 \end{array}\right).$$

Notice that $\operatorname{rank} X_{ii} = s_i$ if and only if $\operatorname{rank} X_{11} = s_1$.

The description of Inv $(\underline{r}, \underline{s})$ as a smooth manifold is given in the next theorem, whose items are proved in [5] and [6].

THEOREM 2.3. *With the notation of Definition* 2.1,

1. $\mathcal{G}(\underline{s})$ *is a subgroup of* $\mathrm{Gl}(\mathbb{K}^d)$.
2. $\mathcal{G}(\underline{s})$ *acts freely on* $\mathcal{M}(\underline{r}, \underline{s})$ *on the right by matrix multiplication.*
3. *If* $X \in \mathcal{M}(\underline{r}, \underline{s})$, $T \in \mathrm{Gl}(\mathbb{K}^d)$, *and* $XT \in \mathcal{M}(\underline{r}, \underline{s})$, *then* $T \in \mathcal{G}(\underline{s})$.
4. *The orbit space* $\mathcal{M}(\underline{r}, \underline{s})/\mathcal{G}(\underline{s})$ *has a differentiable structure such that the natural projection of* $\mathcal{M}(\underline{r}, \underline{s})$ *on it is a submersion.*

There is a natural bijection between Inv $(\underline{r}, \underline{s})$ and $\mathcal{M}(\underline{r}, \underline{s})/\mathcal{G}(\underline{s})$ given by

$$\begin{array}{ccc} \mathcal{M}(\underline{r}, \underline{s})/\mathcal{G}(\underline{s}) & \longrightarrow & \text{Inv } (\underline{r}, \underline{s}), \\ \{XP | X \in \mathcal{M}(\underline{r}, \underline{s}), P \in \mathcal{G}(\underline{s})\} & \rightsquigarrow & [X]. \end{array}$$

(In other words, every (C, A)-conditioned invariant subspace with $\underline{s}$ as Brunovsky indices for the restriction is spanned by the columns of a matrix in $\mathcal{M}(\underline{r}, \underline{s})$ and vice versa.) We consider in Inv $(\underline{r}; \underline{s})$ the differentiable structure defined through this bijection.

COROLLARY 2.4. *If for* $j = 1, \ldots, h$, $m_j = s_j - s_{j+1}$, *then*

$$\dim \text{Inv } (\underline{r}, \underline{s}) = \sum_{i=1}^{h} \sum_{j=1}^{k} m_i (r_{i+j-1} - s_{i+j-1}).$$

The parameterization of Inv $(\underline{r}, \underline{s})$ is based on the construction of a canonical element of each orbit of $\mathcal{M}(\underline{r}, \underline{s})/\mathcal{G}(\underline{s})$. The following definition and proposition deal with such canonical form.

DEFINITION 2.5. *A matrix* $Y \in \mathcal{M}(\underline{r}, \underline{s})$ *is said to be in* reduced *form if there exists a set of pairwise different positive integers* $\underline{n} = \{n_{ij}, 1 \le i \le h, 1 \le j \le m_{h-i+1}\}$ *such that for* $i = 1, 2, \ldots, h$

$$1 \le n_{i1} < n_{i2} < \cdots < n_{im_{h-i+1}} \le r_{h-i+1}$$

and Y can be partitioned into blocks $Y = (Y_{ij})$, $1 \le i \le k$, $1 \le j \le h$, satisfying the following conditions:

1. $Y_{ij} \in \mathcal{M}_{r_i, s_j}$.
2. $Y_{ij} = 0$ *if* $i < j$.
3. *For $i = 1, 2, \dots, h$, Y_{ii} can be partitioned into blocks $Y_{ii} = (L^1_{i\beta})$, $1 \le \beta \le h - i + 1$, in such a way that for $\beta = 1, 2, \dots, h - i + 1$, $L^1_{i\beta} \in \mathcal{M}_{r_i, m_{h-\beta+1}}$ is a matrix whose last $r_i - r_{h-\beta+1}$ rows are zero, the rows n_{ij}, $1 \le i \le \beta - 1$, $1 \le j \le m_{h-i+1}$ are also zero, and the rows $n_{\beta 1}, n_{\beta 2}, \dots, n_{\beta m_{h-\beta+1}}$ are unit vectors $e^\beta_1, e^\beta_2, \dots, e^\beta_{m_{h-\beta+1}}$:*

$$e^\beta_j = (0 \dots 0 \overset{(j}{1} 0 \dots 0) \in \mathbb{K}^{m_{h-\beta+1}}.$$

4. *For $i > j$, Y_{ij} can be partitioned into blocks $Y_{ij} = (L^{i-j+1}_{i\beta})$, $1 \le \beta \le h - j + 1$, in such a way that $L^{i-j+1}_{i\beta} \in \mathcal{M}_{r_i, m_{h-\beta+1}}$ is a matrix whose last $r_i - r_{h-\beta+i-j+1}$ rows are zero and for $\beta \ge i-j+1$, the rows n_{pq}, $1 \le p \le \beta - i + j$, $1 \le q \le m_{h-p+1}$ are also zero. (Note that the entries of Y_{ij} are prescribed by the entries of $Y_{i-j+1,1}$).*
5. *The number of parameters of Y coincide with the dimension of* Inv $(\underline{r}, \underline{s})$.

PROPOSITION 2.6. *For every $X \in \mathcal{M}(\underline{r}, \underline{s})$ there exists a matrix $P \in \mathcal{G}(\underline{s})$ such that XP is a matrix in reduced form. Then, the set of integers $\underline{n}$ is called an* admissible set of indices *for X and XP a* reduced form *of X.*

Example 2.7. Let $k = 4$, $h = 3$, $\underline{r} = (5, 3, 2, 1)$ and $\underline{s} = (3, 3, 1)$. A matrix $X \in \mathcal{M}(\underline{r}, \underline{s})$ has the form

$$\left(\begin{array}{ccc|ccc|c} x_1 & y_1 & y_2 & & & & \\ x_2 & y_3 & y_4 & & & & \\ 0 & y_5 & y_6 & & \mathbf{0} & & \mathbf{0} \\ 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & & & & \\ \hline z_1 & t_1 & t_2 & x_1 & y_1 & y_2 & \\ 0 & t_3 & t_4 & x_2 & y_3 & y_4 & \mathbf{0} \\ 0 & 0 & 0 & 0 & y_5 & y_6 & \\ \hline 0 & u_1 & u_2 & z_1 & t_1 & t_2 & x_1 \\ 0 & 0 & 0 & 0 & t_3 & t_4 & x_2 \\ \hline 0 & 0 & 0 & 0 & u_1 & u_2 & z_1 \end{array}\right).$$

Then, the only possible reduced forms are

$$\left(\begin{array}{ccc|ccc|c} 1 & 0 & 0 & & & & \\ x & 1 & 0 & & \mathbf{0} & & \mathbf{0} \\ 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & & & & \\ \hline y & 0 & 0 & 1 & 0 & 0 & \\ 0 & z & t & x & 1 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 1 & \\ \hline 0 & u & v & y & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & z & t & x \\ \hline 0 & 0 & 0 & 0 & u & v & y \end{array}\right) \text{ and } \left(\begin{array}{ccc|ccc|c} 1 & 0 & 0 & & & & \\ x & 1 & 0 & & \mathbf{0} & & \mathbf{0} \\ 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & & & & \\ 0 & 0 & 0 & & & & \\ \hline y & 0 & 0 & 1 & 0 & 0 & \\ 0 & z & t & x & 1 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 1 & \\ \hline 0 & u & v & y & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & z & t & x \\ \hline 0 & 0 & 0 & 0 & u & v & y \end{array}\right),$$

corresponding to the indices $n_{11} = 1$, $n_{21} = 2$, $n_{22} = 3$ (if $x_1 \neq 0$) and $n_{11} = 2$, $n_{21} = 1$, $n_{22} = 3$ (if $x_2 \neq 0$), respectively.

The admissible set of indices for X depends only on the block X_{11} and completely determines the structure of the reduced form Y. However, they are not uniquely determined by X_{11}; that is to say, several sets may be admissible for the same matrix. In fact, $1 \leq n_{11} < n_{12} < \cdots < n_{1s_h} \leq r_h$ are s_h linearly independent rows of the matrix

$$\begin{pmatrix} X^1_{11} \\ X^1_{21} \\ \vdots \\ X^1_{k-h+1,1} \end{pmatrix},$$

$1 \leq n_{11} < n_{12} < \cdots < n_{1s_h} \leq r_h$, $1 \leq n_{21} < n_{22} < \cdots < n_{2s_{h-1}-s_h} \leq r_{h-1}$ are s_{h-1} linearly independent rows of

$$\begin{pmatrix} X^1_{11} & X^1_{12} \\ X^1_{21} & X^1_{22} \\ \vdots & \vdots \\ X^1_{k-h+1,1} & X^1_{k-h+1,2} \\ 0 & X^1_{k-h+2,2} \end{pmatrix},$$

and so on. Of course, there may be several different ways of choosing these linear independent rows.

PROPOSITION 2.8. (i) *Let $X \in \mathcal{M}(\underline{r}, \underline{s})$ and $Q \in \mathcal{G}(\underline{s})$. If $\underline{n}$ is an admissible set of indices for X, it is also an admissible set of indices for XQ.*

(ii) *Let Y and $\overline{Y}$ be matrices of $\mathcal{M}(\underline{r}, \underline{s})$ in reduced form with the same set of indices $\underline{n}$. If there is a matrix $P \in \mathcal{G}(\underline{s})$ such that $\overline{Y} = YP$, then $Y = \overline{Y}$.*

In other words, once an admissible set of indices (i.e., linearly independent rows in the (1,1)-block) of $X \in \mathcal{M}(\underline{r}, \underline{s})$ has been fixed, the reduced form, Y, of X is unique and the free parameters of Y completely parameterizes the (C, A)-invariant subspace $[X]$. This parameterization can be used to provide the differentiable manifold Inv $(\underline{r}, \underline{s})$ with a coordinate atlas (for details see [7]).

The above proposition shows that all bases of the same (C, A)-invariant subspace have the same admissible set of indices.

DEFINITION 2.9. *Given a (C, A)-invariant subspace $\mathcal{V}$, a multi-index of $\mathcal{V}$ is any admissible set of indices for any matrix $X \in \mathcal{M}(\underline{r}, \underline{s})$ such that $\mathcal{V} = [X]$.*

3. The set of solutions of the (dual) cover problem. Let $\mathcal{S}$ be a subspace of $\mathbb{K}^n$ of codimension $d \leq n$, which is fixed along this note. Let $F \in \mathcal{M}^*_{d,n}$ such that $\mathcal{S} = \text{Ker}F$ or, equivalently, $[F^*] = \mathcal{S}^\perp$. Given $\delta \leq d$ and $X \in \mathcal{M}^*_{n,\delta}$, $[X] \subset \mathcal{S}$ if and only if $FX = 0$. So, the solutions of the (dual) cover problem are given by the set of matrices $X \in \mathcal{M}(\underline{r}, \underline{s})$ such that $FX = 0$ for each partition $\underline{s}$ compatible with $\underline{r}$. The rest of this section is devoted to study the compatibility of these linear algebraic systems.

Let $X \in \mathcal{M}(\underline{r}, \underline{s})$ and put $X^j = (X_{ij})_{1 \leq i \leq k}$, $j = 1, \ldots, h$. For $i = 1, 2, \ldots, h$ let X_i denote the submatrix of X whose jth column block is the ith column block of X^j.

For example, for the matrix in Example 2.2 we would have

$$X_1 = \left(\begin{array}{c|c|c} X_{11}^1 & & \\ X_{21}^1 & \mathbf{0} & \mathbf{0} \\ 0 & & \\ 0 & & \\ \hline X_{11}^2 & X_{11}^1 & \\ 0 & X_{21}^1 & \mathbf{0} \\ 0 & 0 & \\ \hline 0 & X_{11}^2 & X_{11}^1 \\ 0 & 0 & X_{21}^1 \\ \hline 0 & 0 & X_{11}^2 \end{array}\right), \quad X_2 = \left(\begin{array}{c|c} X_{12}^1 & \\ X_{22}^1 & \mathbf{0} \\ X_{32}^1 & \\ 0 & \\ \hline X_{12}^2 & X_{12}^1 \\ X_{22}^2 & X_{22}^1 \\ 0 & X_{32}^1 \\ \hline X_{12}^3 & X_{12}^2 \\ 0 & X_{22}^2 \\ \hline 0 & X_{12}^3 \end{array}\right), \quad X_3 = \left(\begin{array}{c} X_{13}^1 \\ X_{23}^1 \\ X_{33}^1 \\ X_{43}^1 \\ \hline X_{13}^2 \\ X_{23}^2 \\ X_{33}^2 \\ \hline X_{13}^3 \\ X_{23}^3 \\ \hline X_{13}^4 \end{array}\right).$$

Also put

$$R_j^i = \begin{pmatrix} X_{1j}^i \\ X_{2j}^i \\ \vdots \\ X_{k-h-i+j+1,j}^i \end{pmatrix}, \quad j = 1, \ldots, h, \quad i = 1, \ldots, k-h+j,$$

and

$$R_j = \begin{pmatrix} R_j^1 \\ R_j^2 \\ \vdots \\ R_j^{k-h+j} \end{pmatrix}, \quad j = 1, \ldots, h.$$

R_j is of size $(r_{h-j+1}+\cdots+r_k)\times(s_{h-j+1}-s_{h-j+2})$ and will be said to be the *condensed form* of X_j, $j = 1, \ldots, h$. Also the sequence of all these matrices $(R_1, \ldots, R_h)$ will be called the *condensed form* of X. Notice that by knowing the condensed form of X we can easily reconstruct $X \in \mathcal{M}(\underline{r}, \underline{s})$. (The information about $\underline{r}$ and $\underline{s}$ is contained in the condensed form).

In the previous example,

$$R_1^1 = \begin{pmatrix} X_{11}^1 \\ X_{21}^1 \end{pmatrix}, \quad R_1^2 = X_{11}^2, \quad R_1 = \begin{pmatrix} R_1^1 \\ R_1^2 \end{pmatrix} = \left(\begin{array}{c} X_{11}^1 \\ X_{21}^1 \\ \hline X_{11}^2 \end{array}\right),$$

$$R_2^1 = \begin{pmatrix} X_{12}^1 \\ X_{22}^1 \\ X_{32}^1 \end{pmatrix}, \quad R_2^2 = \begin{pmatrix} X_{12}^2 \\ X_{22}^2 \end{pmatrix}, \quad R_2^3 = X_{12}^3, \quad R_2 = \left(\begin{array}{c} X_{12}^1 \\ X_{22}^1 \\ X_{32}^1 \\ \hline X_{12}^2 \\ X_{22}^2 \\ \hline X_{12}^3 \end{array}\right),$$

etc.

The (1,1)-block of X, X_{11} will be also denoted by R:

$$R := \left(\begin{array}{c|c|c|c} R_1^1 & R_2^1 & \cdots & R_h^1 \\ \hline \mathbf{0} & \mathbf{0} & & \end{array}\right) = \left(\begin{array}{c|c|c|c} X_{11}^1 & X_{12}^1 & & X_{1h}^1 \\ X_{21}^1 & X_{22}^1 & & X_{2h}^1 \\ \vdots & \vdots & & \vdots \\ X_{k-h+1,1}^1 & X_{k-h+1,2}^1 & \cdots & X_{k-h+1,h}^1 \\ \hline \mathbf{0} & X_{k-h+2,2}^1 & & X_{k-h+2,h}^1 \\ & \mathbf{0} & & \vdots \\ & & & \vdots \\ & & & X_{k,h}^1 \end{array}\right). \tag{3.1}$$

Now we split the columns of F according to the sizes of the row blocks of X:

$$F = \left(\begin{array}{c|c|c|c|c} F_{11} \quad F_{12} \quad \cdots \quad F_{1k} & F_{21} \quad F_{22} \quad \cdots \quad F_{2,k-1} & \cdots & F_{k-1,1} \quad F_{k-1,2} & F_{k1} \end{array}\right)$$

with F_{ij} of size $d \times (r_{k-j+1} - r_{k-j+2})$, $i = 1, \ldots, k$, $j = 1, \ldots k-i+1$.

Now, $FX = 0$ if and only if $FX_j = 0$ for $j = 1, \ldots, h$. In these homogeneous linear equations the unknown matrix X_j has some repeated entries and other ones must be compulsory zero. However, with the help of the matrices R_j we can avoid these inconveniences. We introduce the following notation. For $\alpha = 1, \ldots, k$ and $\beta = 1, \ldots k-\alpha+1$ let

$$H(\alpha, \beta) = \left(\begin{array}{c} F_{\alpha,1} \; F_{\alpha,2} \; \cdots \; F_{\alpha,\beta} \end{array}\right) \in \mathcal{M}_{d, r_{k-\beta+1}},$$

and for $j = 1, \ldots, h$ let

$$H_j = \begin{pmatrix} H(1, k-j+1) & H(2, k-j) & \cdots & H(k-j+1, 1) \\ H(2, k-j+1) & H(3, k-j) & \cdots & H(k-j+2, 1) \\ \vdots & \vdots & \cdots & \vdots \\ H(j, k-j+1) & H(j+1, k-j) & \cdots & H(k, 1) \end{pmatrix}.$$

Thus H_j is of size $jd \times (r_j + r_{j+1} + \cdots + r_k)$. Notice that this matrix is a kind of truncated block Hankel matrix. In fact

$$H(2, k-j+1) = \left(\begin{array}{c} H(2, k-j) \; F_{2,k-j+1} \end{array}\right),$$
$$H(3, k-j) = \left(\begin{array}{c} H(3, k-j-1) \; F_{3,k-j} \end{array}\right),$$

and so on.

Matrices $H_1, \ldots, H_h$ depend only on the selected basis for $\mathcal{S}^\perp$ and the Brunovsky structure of (C, A). We say that $(H_1, \ldots, H_h)$ is the *block-Hankel structure* associated to F, or *a* block-Hankel structure of $\mathcal{S}$, with regard to (C, A). (Notice that $H_1 = F$).

Now it is easily computed that $FX_j = 0$ for $j = 1, \ldots, h$ if and only if

$$H_{h-j+1} R_j = 0, \quad j = 1, \ldots, h.$$

As said above, the advantage of this expression is that all elements of R_j can be seen as arbitrary unknowns in the open set defined by the condition $\operatorname{rank} R = s_1$.

Following with our example we have that

$$F = \left(\begin{array}{c|c|c|c|c} F_{11} \quad F_{12} \quad F_{13} \quad F_{14} & F_{21} \quad F_{22} \quad F_{23} & F_{31} \quad F_{32} & F_{41} \end{array}\right)$$

and $FX = 0$ if and only if

$$H_3R_1 = \left(\begin{array}{ccc|c} F_{11} & F_{12} & F_{21} & \\ F_{21} & F_{22} & F_{31} & \\ F_{31} & F_{32} & F_{41} & \end{array}\right) \left(\begin{array}{c} X_{11}^1 \\ X_{21}^1 \\ \hline X_{11}^2 \end{array}\right) = 0,$$

$$H_2R_2 = \left(\begin{array}{ccc|cc|c} F_{11} & F_{12} & F_{13} & F_{21} & F_{22} & F_{31} \\ F_{21} & F_{22} & F_{23} & F_{31} & F_{32} & F_{41} \end{array}\right) \left(\begin{array}{c} X_{12}^1 \\ X_{22}^1 \\ X_{32}^1 \\ \hline X_{12}^2 \\ X_{22}^2 \\ \hline X_{12}^3 \end{array}\right) = 0,$$

$$H_1R_3 = \left(\begin{array}{cccc|ccc|ccc|c} F_{11} & F_{12} & F_{13} & F_{14} & F_{21} & F_{22} & F_{23} & F_{31} & F_{32} & F_{41} \end{array}\right) \left(\begin{array}{c} X_{13}^1 \\ X_{23}^1 \\ X_{33}^1 \\ X_{43}^1 \\ \hline X_{13}^2 \\ X_{23}^2 \\ X_{33}^2 \\ \hline X_{13}^3 \\ X_{23}^3 \\ \hline X_{13}^4 \end{array}\right) = 0.$$

As a conclusion we have the following result, whose proof is immediate.

PROPOSITION 3.1. *With the above notation there is a matrix $X \in \mathcal{M}(\underline{r}, \underline{s})$ such that $FX = 0$ if and only if there are matrices $R_1, \ldots, R_h$ such that $H_{h-j+1}R_j = 0$, $j = 1, \ldots, h$, and* $\operatorname{rank} R = s_1$.

Thus the existence of solutions for the cover problem is reduced to the compatibility of a finite number of homogeneous linear equations with the additional restriction $\operatorname{rank} R = s_1$. In what follows we aim to give a more explicit condition with the help of the reduced form introduced in Definition 2.5.

First we pay attention to matrices H_j that form a block-Hankel structure of the subspace $\mathcal{S}$. Let $t_j = \operatorname{rank} H_j$, $j = 1, \ldots, h$. In particular $t_1 = \operatorname{rank} H_1 = \operatorname{rank} F = d$.

DEFINITION 3.2. *An increasing sequence of positive integers $1 \le \ell_1^j < \cdots < \ell_{t_j}^j \le r_j + \cdots + r_k$ is said to be a multi-index for H_j if columns $\ell_1^j, \ldots, \ell_{t_j}^j$ are linearly independent.*

PROPOSITION 3.3. *Let $F, G \in \mathcal{M}_{d,n}^*$ be matrices such that $[F^*] = [G^*] = \mathcal{S}^\perp$. Let $(H_1, \ldots, H_h)$ and $(G_1, \ldots, G_h)$ be two block-Hankel structures associated to F and G (with regard to (C, A)), respectively. Then* $\operatorname{rank} H_j = \operatorname{rank} G_j$ *and $(\ell_1^j, \ldots, \ell_{t_j}^j)$ is a multi-index for H_j if and only if it is a multi-index for G_j, $j = 1, \ldots, h$.*

Proof. In fact $[F^*] = [G^*] = \mathcal{S}^\perp$ if and only if there is an invertible matrix $P \in \mathcal{M}_d^*$ such that $G = PF$. Thus $G_j = \operatorname{Diag}(P, \ldots, P)H_j$ and the proposition follows. □

Then, the following definition makes sense. Let $F \in \mathcal{M}_{d,n}^*$ be such that $[F^*] = \mathcal{S}^\perp$ and $(H_1, \ldots, H_h)$ the block-Hankel structure associated to F. Let $\underline{\ell}^j = (\ell_1^j, \ldots, \ell_{t_j}^j)$ with $1 \le \ell_1^j < \cdots < \ell_{t_j}^j \le r_j + \cdots + r_k$ a multi-index for H_j. Then, we call $\underline{\ell} = (\underline{\ell}^1, \ldots, \underline{\ell}^h)$ a *multi-index of* $\mathcal{S}$ with respect to (C, A).

Similarly we introduce a notation for the multi-indices of a (C, A)-invariant subspace. If $\{(1 \le n_{i1} < n_{i2} < \cdots < n_{im_{h-i+1}} \le r_{h-i+1}) | 1 \le i \le h\}$ (recall that $m_{h-i+1} = s_{h-i+1} - s_{h-i+2}$) is a multi-index of a (C, A)-invariant subspace (Defini-

tion 2.9), then we will write $\underline{n}^i = (n_{i1}, \dots, n_{im_{h-i+1}})$ and $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ denotes a multi-index of that subspace.

In the sequel we will use the same symbol to denote $\underline{\ell}^j$ and $\underline{n}^j$ as tuples and as sets of positive integers. Thus, for example, we will say that that multi-index $\underline{\ell} = (\underline{\ell}^1, \dots, \underline{\ell}^h)$ is either the d-tuple formed by the t_i-tuples $\underline{\ell}^i$, one behind the other, or $\underline{\ell} = \underline{\ell}^1 \cup \underline{\ell}^2 \cup \cdots \cup \underline{\ell}^h$ is the union set of all multi-indices of $\mathcal{S}$. The same applies to $\underline{n}$.

We will also use the following notation: if $\underline{i} = (i_1, \dots, i_r)$ and $\underline{j} = (j_1, \dots . j_s)$ are sequences of positive integers such that $1 \le i_1 < \cdots < i_r \le m$ and $1 \le j_1 < \cdots < j_s \le n$, then $A(\underline{i}, \underline{j})$ will denote the $r \times s$ submatrix of $A \in \mathcal{M}_{m,n}$ formed with rows $i_1, \dots, i_r$ and columns $j_1, \dots, j_s$; and $A[\underline{i}, \underline{j}]$ is the submatrix of A obtained by removing rows $i_1, \dots, i_r$ and columns $j_1, \dots, j_s$. In particular, if $\underline{j} = (1, \dots, n)$, then $A(\underline{i}, \ )$ is the submatrix formed by rows $i_1, \dots, i_r$ and all the columns of A and $A[\underline{i}, \ ]$ is the submatrix obtained by deleting from A rows $i_1, \dots, i_r$ and no columns. Similarly if $\underline{i} = (1, \dots, m)$.

Now we can prove an explicit condition for the cover problem to have a solution

THEOREM 3.4. *Let $\mathcal{S} \in Gr_{n-d}(\mathbb{K}^n)$, (C, A) an observable pair with $r_1 \ge \cdots \ge r_k > 0$ as positive Brunovsky indices and let $s_1 \ge \cdots \ge s_h > 0$ be a partition compatible with $(r_1, \dots, r_k)$. Then, there exists a (C, A)-invariant subspace $\mathcal{V}$ contained in $\mathcal{S}$ such that the restriction of (C, A) to $\mathcal{V}$ has $(s_1, \dots, s_h)$ as positive Brunovsky indices if and only if there are multi-indices $\underline{\ell} = (\underline{\ell}^1, \dots, \underline{\ell}^h)$ and $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ for $\mathcal{S}$ and $\mathcal{V}$, respectively, such that $\underline{\ell}^j \cap \underline{n}^{h-j+1} = \emptyset$ for $j = 1, \dots, h$.*

Proof. Assume first that there is a (C, A)-invariant subspace $\mathcal{V} \in Gr_\delta(\mathbb{K}^n)$ such that $\mathcal{S} \supset \mathcal{V}$ and let $[F^*] = \mathcal{S}^\perp$. For any matrix $X \in \mathcal{M}(\underline{r}, \underline{s})$ (whose columns span $\mathcal{V}$) we have that $FX = 0$. By Proposition 2.6 we can assume that X is in reduced form given by Definition 2.5.

Let $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ be a multi-index of $\mathcal{V}$ and then for X. Since X is in reduced form we have that

$$R_j^1(\underline{n}^j, \ ) = I_{m_{h-j+1}}, \quad j = 1, \dots, h.$$

Thus $\operatorname{rank} R = s_1$ independent of the remaining elements of R_i^1, $i = 1, \dots, h$. We only have to analyze under what conditions equations $H_{h-j+1} R_j = 0$, $j = 1, 2, \dots, h$, are solvable.

Let us consider the equations $H_{h-j+1} R_j = 0$. As $R_j(\underline{n}^j, \ ) = R_j^1(\underline{n}^j, \ ) = I_{m_{h-j+1}}$, if $H_{h-j+1} R_j = 0$, then

$$H_{h-j+1}[\ , \underline{n}^j] R_j[\underline{n}^j, \ ] = -H_{h-j+1}(\ , \underline{n}^j),$$

i.e., the columns of H_{h-j+1} in $\underline{n}^j$ are linearly dependent on the remaining columns of H_{h-j+1}. Since $\operatorname{rank} H_{h-j+1} = t_{h-j+1}$, there must be a multi-index for H_{h-j+1}, $\underline{\ell}^{h-j+1} = (\ell_1^{h-j+1}, \dots, \ell_{t_{h-j+1}}^{h-j+1})$ such that $\underline{\ell}^{h-j+1} \cap \underline{n}^j = \emptyset$, as we wanted to prove.

Conversely, if $\underline{\ell}^{h-j+1} \cap \underline{n}^j = \emptyset$ for $j = 1, \dots, h$, then there are matrices $Z_1, \dots, Z_h$ such that

$$H_{h-j+1}[\ , \underline{n}^j] Z_j = -H_{h-j+1}(\ , \underline{n}^j), \quad j = 1, \dots, h.$$

For $j = 1, \dots, h$ we construct a matrix R_j as follows:

$$\begin{aligned} R_j[\underline{n}^j, \ ] &= Z_j, \\ R_j(\underline{n}^j, \ ) &= I_{m_{h-j+1}}. \end{aligned}$$

Thus $H_{h-j+1}R_j = 0$ and by defining R as in (3.1) we have that $\operatorname{rank} R = s_1$. Now the theorem follows from Proposition 3.1. □

It is worth noticing that in the proof of the above theorem we have not used the whole reduced form of X, just the fact that the rows in $\underline{n}^j$ are canonical vectors. Actually, there are some hidden properties of the admissible multi-indices of H_j behind the block-Hankel structure of these matrices. The following one is relevant.

LEMMA 3.5. *If for some $j = 1, \dots, h$ there is a positive integer α, $1 \le \alpha \le r_j$, such that column α of H_j is a linear combination of the remaining columns of H_j, then for $i = 1, \dots, j-1$ columns α, $r_i + \alpha, \dots, r_i + \cdots + r_{j-1} + \alpha$ of H_i are also linear combinations of the remaining columns of H_i.*

Proof. The proof is straightforward, bearing in mind the block-Hankel structure of matrices H_j. □

A consequence of this lemma is the following necessary condition for the cover problem to have a solution. In fact, this condition is also sufficient generically, as we will show in the next section.

THEOREM 3.6. *Under the assumptions of Theorem* 3.4 *let $(H_1, \dots, H_h)$ be any block-Hankel structure of $\mathcal{S}$ with regard to (C, A). If there exists a (C, A)-invariant subspace $\mathcal{V}$ contained in $\mathcal{S}$, such that the restriction of (C, A) to $\mathcal{V}$ has $(s_1, \dots, s_h)$ as positive Brunovsky indices, then*

$$\operatorname{rank} H_j \le \sum_{i=j}^{k} (r_i - s_i), \quad j = 1, \dots, h. \tag{3.2}$$

Proof. We have seen in the proof of Theorem 3.4 that if $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ is a multi-index for $\mathcal{V}$, then the columns of H_j in $\underline{n}^{h-j+1}$ are a linear combination of the remaining columns of H_j. Thus, if we define that for a positive integer a, $a + \underline{n}^j = \{a + n_{j,1}, a + n_{j,2}, \dots, a + n_{j,m_j}\}$, then, according to Lemma 3.5, the columns of H_j in

$$\begin{aligned}
&\underline{n}^{h-j+1} \cup \\
&\underline{n}^{h-j} \cup (r_j + \underline{n}^{h-j}) \cup \\
&\underline{n}^{h-j-1} \cup (r_j + \underline{n}^{h-j-1}) \cup (r_j + r_{j+1} + \underline{n}^{h-j-1}) \cup \\
&\cdots \\
&\underline{n}^1 \cup (r_j + \underline{n}^1) \cup \cdots (r_j + \cdots + r_{h-1} + \underline{n}^1)
\end{aligned}$$

are linear combination of the remaining columns of H_j. But all these sets of indices are pairwise disjoint (thanks to the relations $n_{i,m_{h-i+1}} \le r_{h-i+1}$) and

$$\begin{aligned}
&\#(\underline{n}^1 \cup \underline{n}^2 \cup \cdots \cup \underline{n}^{h-j+1}) = s_h + m_{h-1} + \cdots + m_j = s_j, \\
&\#(r_j + \underline{n}^1 \cup r_j + \underline{n}^2 \cup \cdots \cup r_j + \underline{n}^{h-j}) = s_h + m_{h-1} + \cdots + m_{j+1} = s_{j+1}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
&\#(r_1 + \cdots + r_{h-1} + \underline{n}^1) = s_h.
\end{aligned}$$

As the number of columns of H_j is $r_j + \cdots + r_k$ we conclude that

$$\operatorname{rank} H_j \le (r_j + \cdots + r_k) - (s_j + \cdots + s_h), \quad j = 1, \dots, k,$$

as we wanted to prove. □

Condition (3.2) is not sufficient in general because not only is the number of linearly independent columns of H_j inportant but also their positions. For example, assume that $\underline{r} = (5,3,2,1)$ and $\underline{s} = (3,3,1)$ as in Example 2.7, and let $\mathcal{S}$ such that $\mathcal{S}^\perp$ is the subspace of $\mathbb{R}^{11}$ generated by the matrix F^*, where $F = \left(\begin{array}{cccccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right)$. Then

$$H_1 = F = \left(\begin{array}{ccccc|ccc|c|c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$H_2 = \left(\begin{array}{ccc|cc|c} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$H_3 = \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

On one hand, $\operatorname{rank} H_1 = 1 \leq r_1 + r_2 + r_3 + r_4 - s_1 - s_2 - s_3 = 4$, $\operatorname{rank} H_2 = 1 \leq r_2 + r_3 + r_4 - s_2 - s_3 = 2$, and $\operatorname{rank} H_3 = 1 \leq r_3 + r_4 - s_3 = 2$. Thus condition (3.2) holds. However, we have seen in Example 2.7 that the only possible multi-indices for all invariant subspaces are $((1),(2,3))$ and $((2),(1,3))$. (Notice that in this case $\underline{n}^3 = \emptyset$.) But $\ell^1 = \ell^2 = \ell^3 = (1)$; and so either $\ell^3 \cap \underline{n}^1 \neq \emptyset$ or $\ell^2 \cap \underline{n}^2 \neq \emptyset$. That is to say, there is not any (C,A)-invariant subspace $\mathcal{V}$ such that the restriction of (C,A) to $\mathcal{V}$ has Brunovsky indices $\underline{s} = (3,3,1)$ and $\mathcal{V} \subset \mathcal{S}$. This does not mean that there is no (C,A)-invariant subspaces of dimension 7 contained in $\mathcal{S}$. In fact, many other partitions $\underline{s}$ are compatible with $\underline{r} = (5,3,2,1)$. For example, $\underline{s} = (4,2,1)$ is compatible with $\underline{r}$ and a possible multi-index may be $\underline{n} = ((2),(3),(4,5))$. Since $\ell^{3-j+1} \cap \underline{n}^j = \emptyset$ for $j = 1,2,3$ we conclude that there is at least one (C,A)-invariant subspace of dimension 7 contained in $\mathcal{S}$. This subspace is generated by a matrix $Y \in \mathcal{M}(\underline{r},\underline{s})$ in the reduced form given in Definition 2.5. In particular, the subspace spanned by

$$Y(y_1,\ldots,y_9) = \left(\begin{array}{cccc|cc|c} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline y_1 & y_2 & y_3 & y_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & y_5 & y_6 & y_7 & y_1 & y_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & y_8 & y_9 & 0 & y_5 & y_1 \end{array}\right)$$

has dimension 7, it is (C,A)-invariant, and it is contained in $\mathcal{S}$. Notice that there are nine free parameters in Y. So, there are many other different choices for matrices Y in reduced form with the same properties; i.e., each one of them spans a (C,A)-invariant subspace of dimension 7 with Brunovsky indices for the restriction $\underline{s} = (4,2,1)$ contained in $\mathcal{S}$. We will see in a later section that the set of all (C,A)-invariant subspace with fixed Brunovsky indices for the restriction contained in a given subspace can be provided with a differentiable structure. As we will prove, the number of free parameters in Y is just the dimension of the corresponding manifold.

4. The generic case. The aim of this section is to show that, generically, that is to say, for $\mathcal{S}$ belonging to an open and dense subset $\mathfrak{S}$ of $\operatorname{Gr}_{n-d}(\mathbb{K}^n)$, condition

(3.2) in Theorem 3.6 is a sufficient condition for the (dual) cover problem to have a solution. We would like to emphasize that genericity is an interesting property from a practical point of view. In fact, it means that if $\mathcal{S} \notin \mathfrak{S}$ we can obtain by a slight modification of $\mathcal{S}$ a subspace $\mathcal{S}' \in \mathfrak{S}$ and that if $\mathcal{S} \in \mathfrak{S}$ the small perturbations of $\mathcal{S}$ remain in $\mathfrak{S}$.

We begin by recalling the definition of genericity. Given a property (P) concerning the elements of a topological space T, (P) is said to be a *generic property* with respect to T if there exists an open and dense subset O of T such that every element of O satisfies this property. Then, we also say that the set O satisfies property (P) *generically*. For example, being invertible is a generic property with respect to the set of square matrices. The following property will play an important role in the achievement of our objective.

(P). We say that a matrix A satisfies property (P) if A has full rank, say r, and all $r \times r$ submatrices of A are invertible.

It is clear that (P) is a generic property with respect to $\mathcal{M}_{m,n}$. However, what we will need is that this property keeps being generic with respect to a smaller set: the set of all block-Hankel structures $(H_1, \ldots, H_h)$ associated to F. The following particular case is essential to this end.

Let $\mathcal{H}_{p,q}$ be the set of $p \times q$ Hankel matrices and ν the number of perpendiculars to the main diagonal. From a topological point of view we identify $\mathcal{H}_{p,q}$ with $\mathbb{K}^\nu$. Let $H(x)$ be the Hankel matrix

$$H(x) = \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_q \\ x_2 & x_3 & \ldots & \ldots & \ldots \\ x_3 & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_p & \ldots & \ldots & \ldots & x_\nu \end{pmatrix} = (x_{i+j-2}),$$

where $x_1, \ldots, x_\nu$ are indeterminates. We will also say that $x_1 \ldots, x_\nu$ are the *parameters* of $H(x)$ and that $H(x)$ is a *parametric* Hankel matrix. A particular Hankel matrix $H(a)$ is obtained by giving to the parameters the values $x_i = a_i$, $i = 1, \ldots, \nu$. We will show that (P) is generic with respect to $\mathcal{H}_{p,q}$. For this we need the following result.

LEMMA 4.1. *Let $H(x)$ be a parametric Hankel matrix with $p \leq q$. Let $I = (i_1, \ldots, i_p)$ be a given sequence of indices, $1 \leq i_1 < \cdots < i_p \leq q$, and let $\Delta_I(x)$ be the determinant formed with the columns in I. Then $\Delta_I(x)$ is not the zero polynomial of $\mathbb{K}[x]$.*

Proof. The proof is by recurrence on the number of rows p. If $p = 1$, the lemma is obvious. Assume now that the lemma is also true for $p - 1 \geq 2$. Then, if p is the number of rows of $H(x)$, we compute $\Delta_I(x)$ by its first column cofactor expansion:

$$\Delta_I(x) = x_{i_1}\Delta_{1,i_1}(x) + \cdots + x_{i_1+p-1}\Delta_{p,i_1}(x).$$

Notice that if we eliminate the first row of a Hankel matrix, the remaining matrix is still a Hankel matrix. Hence, by applying the induction hypothesis we conclude that $\Delta_{1,i_1}(x) \neq 0$. Then taking into account that the indeterminate x_{i_1} does not appear in any of the remaining summands we obtain that $\Delta_I(x) \neq 0$, as we wanted. □

Let $\mathcal{V}_I$ denote the set of matrices $H(a) \in \mathcal{H}_{p,q}$ such that $\Delta_I(a) \neq 0$. The above lemma shows that $\mathcal{V}_I$ is a Zariski open (and hence dense) subset of $\mathcal{H}_{p,q}$. When I runs over all $i_1, \ldots, i_p$ such that $1 \leq i_1 < \cdots < i_p \leq q$, $\bigcap_I \mathcal{V}_I$ is still open and dense.

Since $H \in \mathcal{H}_{p,q}$ satisfies (P) if and only if H^t satisfies the same property, the next lemma follows.

LEMMA 4.2. (P) *is a generic property with respect to* $\mathcal{H}_{p,q}$.

We will have to deal with matrices formed by a finite number of parametric Hankel matrices, one each other with different parameters. For convenience we introduce the following definition.

DEFINITION 4.3. *Let* $\mathcal{H}_{ij}(x_{i,j})$ *be a parametric Hankel matrices with different parameters (that is to say,* $x_{i,j} \neq x_{k,l}$ *if* $(i,j) \neq (k,l)$*). A matrix of the form*

$$\mathcal{H}(x) = \begin{pmatrix} \mathcal{H}_{11}(x_{11}) & \mathcal{H}_{12}(x_{12}) & \cdots & H_{1n}(x_{1n}) \\ \mathcal{H}_{21}(x_{21}) & \mathcal{H}_{22}(x_{22}) & \cdots & H_{2n}(x_{2n}) \\ \vdots & \vdots & \cdots & \vdots \\ \mathcal{H}_{m1}(x_{m1}) & \mathcal{H}_{m2}(x_{m2}) & \cdots & H_{mn}(x_{mn}) \end{pmatrix}$$

will be called an $\mathcal{H}$*-matrix.*

For example,

$$\begin{pmatrix} x_1 & x_2 & x_3 & y_1 & y_2 & v_1 & v_2 & v_3 & v_4 \\ x_2 & x_3 & x_4 & y_2 & y_3 & v_2 & v_3 & v_4 & v_5 \\ \\ z_1 & z_2 & z_3 & u_1 & u_2 & t_1 & t_2 & t_3 & t_4 \\ z_2 & z_3 & z_4 & u_2 & u_3 & t_2 & t_3 & t_4 & t_5 \\ z_3 & z_4 & z_5 & u_3 & u_4 & t_3 & t_4 & t_5 & t_6 \end{pmatrix}$$

is an $\mathcal{H}$-matrix. (For simplicity we have written $x_{11}^i = x_i, x_{12}^j = y_j, \ldots$)

LEMMA 4.4. (P) *is a generic property with respect to the set of* $\mathcal{H}$*-matrices.*

Proof. Let $(x_{ij}^1, \ldots, x_{ij}^{n_{ij}})$ be the parameters of $\mathcal{H}_{ij}(x_{ij})$. Since $x_{ij}^\alpha \neq x_{k\ell}^\alpha$ for $(i,j) \neq (k,\ell)$ we can obtain from $\mathcal{H}(x)$ a Hankel matrix as follows. In the first row block

$$\begin{pmatrix} \mathcal{H}_{11}(x_{11}) & \mathcal{H}_{12}(x_{12}) & \cdots & \mathcal{H}_{1n}(x_{1n}) \end{pmatrix}$$

we modify the parameters of the first column of each one of the blocks $\mathcal{H}_{12}(x_{12})$, $\mathcal{H}_{13}(x_{13}), \ldots$, so that the first row block becomes a Hankel matrix with different parameters. We illustrate this through the example above. Replace y_1 by x_4 and v_1 by y_3:

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 & y_2 & y_3 & v_2 & v_3 & v_4 \\ x_2 & x_3 & x_4 & y_2 & y_3 & v_2 & v_3 & v_4 & v_5 \end{pmatrix}.$$

Next we modify conveniently the parameters of the second row block

$$\begin{pmatrix} \mathcal{H}_{21}(x_{21}) & \mathcal{H}_{22}(x_{22}) & \cdots & \mathcal{H}_{2n}(x_{2n}) \end{pmatrix}$$

so that the resulting matrix formed by the two modified row blocks is a Hankel matrix with different parameters. Again, we illustrate this procedure through the previous example:

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 & y_2 & y_3 & v_2 & v_3 & v_4 \\ x_2 & x_3 & x_4 & y_2 & y_3 & v_2 & v_3 & v_4 & v_5 \\ \\ x_3 & x_4 & y_2 & y_3 & v_2 & v_3 & v_4 & v_5 & t_4 \\ x_4 & y_2 & y_3 & v_2 & v_3 & v_4 & v_5 & t_4 & t_5 \\ y_2 & y_3 & v_2 & v_3 & v_4 & v_5 & t_4 & t_5 & t_6 \end{pmatrix}.$$

Following this way we will obtain from $\mathcal{H}(x)$ a Hankel matrix, $H(x) \in \mathcal{H}_{p,q}$. From the same argument as in the proof of Lemma 4.2, (P) is generic with respect to $\mathcal{H}_{p,q}$, and so it is with respect to the $\mathcal{H}$-matrices. □

Finally, let $\mathcal{O}_j$ be the set of all matrices $F \in \mathcal{M}_{d,n}$ such that if $(H_1 \ \dots \ H_h)$ is the block Hankel structure associated to F, H_j satisfies (P) and set $\mathcal{O} = \cap_j \mathcal{O}_j$.

PROPOSITION 4.5. *$\mathcal{O}_j$ is an open and dense subset of $\mathcal{M}_{d,n}$.*

Proof. Recall that matrices H_j, as defined in section 3, are almost block Hankel matrices. Actually, we can permute the blocks in H_j in order to obtain a matrix with the form

$$H_j P = \begin{pmatrix} H_j^1 & H_j^2 & \cdots & H_j^{k_j} \end{pmatrix}$$

(P a permutation matrix) such that each submatrix H_j^i is a block Hankel matrix for $i = 1, \dots, k_j$.

Following our example in section 3 we would have

$$H_1^1 = \begin{pmatrix} F_{11} & F_{21} & F_{31} & F_{41} \end{pmatrix} \quad H_1^2 = \begin{pmatrix} F_{12} & F_{22} & F_{32} \end{pmatrix} \quad H_1^3 = \begin{pmatrix} F_{13} & F_{23} \end{pmatrix}$$
$$H_1^4 = F_{14}$$
$$H_2^1 = \begin{pmatrix} F_{11} & F_{21} & F_{31} \\ F_{21} & F_{31} & F_{41} \end{pmatrix} \qquad H_2^2 = \begin{pmatrix} F_{12} & F_{22} \\ F_{22} & F_{32} \end{pmatrix} \qquad H_2^3 = \begin{pmatrix} F_{13} \\ F_{23} \end{pmatrix}$$
$$H_3^1 = \begin{pmatrix} F_{11} & F_{21} \\ F_{21} & F_{31} \\ F_{31} & F_{41} \end{pmatrix} \qquad H_3^2 = \begin{pmatrix} F_{12} \\ F_{22} \\ F_{32} \end{pmatrix}.$$

Now, by a suitable permutation of the rows and columns of $H_j P$ we can obtain an $\mathcal{H}$-matrix. The proposition follows from Lemma 4.4, taking into account that the parameters of F not being in H_j are free. □

COROLLARY 4.6. *$\mathcal{O}$ is an open and dense subset of $\mathcal{M}_{d,n}$.*

From this corollary we derive the following result, which is basic to attain our objective.

PROPOSITION 4.7. *Let $\mathfrak{S}$ be the set of subspaces $\mathcal{S} \in Gr_{n-d}(\mathbb{K}^n)$ such that $\mathcal{S}^\perp = [F^*]$ with $F \in \mathcal{O}$. Then $\mathfrak{S}$ is an open and dense subset of $Gr_{n-d}(\mathbb{K}^n)$.*

Proof. Let $F, G \in \mathcal{M}_{d,n}$ be matrices such that $[F^*] = [G^*] = \mathcal{S}^\perp$, and let $(H_1, \dots, H_h)$, $(K_1, \dots, K_h)$ be the block Hankel structures associated to F and G, respectively. From Proposition 3.3 one has that $F \in \mathcal{O}$ if and only if $G \in \mathcal{O}$. Then the proposition follows from the above corollary and the fact that in the description of $\mathrm{Gr}_{n-d}(\mathbb{K}^n)$ as the orbit space $\mathcal{M}^*_{n-d,n}/\operatorname{Gl}(\mathbb{K}^{n-d})$, the natural projection $\mathcal{M}^*_{n-d,n} \longrightarrow \mathcal{M}^*_{n-d,n}/\operatorname{Gl}(\mathbb{K}^{n-d})$ is an open map. □

LEMMA 4.8. *If $\mathcal{S} \in \mathfrak{S}$ and $H_1, \dots, H_h$ is a block Hankel structure of $\mathcal{S}$ satisfying condition* (3.2) *in Theorem* 3.6*, then*

$$\operatorname{rank} H_j = jd.$$

Proof. Since $\mathcal{S} \in \mathfrak{S}$, H_j has full rank. Therefore,

$$\operatorname{rank} H_j = \min\{jd,\, r_j + \cdots + r_k\}$$

and from condition (3.2) we have

$$\min\{jd,\, r_j + \cdots + r_k\} \leq \sum_{i=j}^{k} (r_i - s_i)\,.$$

Hence, $\operatorname{rank} H_j = jd$. □

We are now in position to show the main result of this section.

THEOREM 4.9. *Under the assumptions of Theorem* 3.4 *and with the above notation, for every* $\mathcal{S} \in \mathfrak{S}$ *there is a* (C, A)*-invariant subspace* $\mathcal{V}$ *contained in* $\mathcal{S}$ *such that the restriction of* (C, A) *to* $\mathcal{V}$ *has* $(s_1, \ldots, s_h)$ *as positive Brunovsky indices if and only if*

$$0 \leq \sum_{j=h}^{k} r_j - s_h - hd. \tag{4.1}$$

Proof. (if) Condition (4.1) can be written as

$$hd \leq \sum_{i=h}^{k} (r_i - s_i)$$

and from this relation and bearing in mind that $r_i \geq s_i$, we conclude that for $j = 1, \ldots, h$

$$jd \leq hd \leq \sum_{i=h}^{k} (r_i - s_i) \leq \sum_{i=j}^{k} (r_i - s_i)\,. \tag{4.2}$$

But $jd = \operatorname{rank} H_j = t_j$, because $H_j = \min\{jd, r_1 + \cdots + r_k\}$. Hence,

$$t_j + m_j = jd + s_j - s_{j+1} \leq r_j + \cdots + r_k\,, \qquad j = 1, \ldots, h. \tag{4.3}$$

Since $\mathcal{S} \in \mathfrak{S}$, if $(H_1, \ldots, H_h)$ is a block-Hankel structure of $\mathcal{S}$, for $j = 1, \ldots, h$ any set of t_j columns of H_j is linearly independent. Thus any multi-index is available to be chosen, and, in order to apply Theorem 3.4, we only have to check that $t_j + \#\underline{n}^{h-j+1}$ is less than or equal to the number of columns of H_j. And this is just what relation (4.3) shows.

(only if). From Theorem 3.6 we know that condition (3.2) is satisfied, so that for $j = h$ we have

$$t_h \leq \sum_{j=h}^{k} r_j - s_h$$

and from Lemma 4.8 $\operatorname{rank} H_h = hd$. Hence, condition (4.1) is satisfied. ∎

Remark 4.10. Given $\mathcal{S} \in \operatorname{Gr}_{n-d}(\mathbb{K}^n)$, the existence of a (C, A)-invariant subspace $\mathcal{V}$ contained in $\mathcal{S}$, such that the restriction of (C, A) to $\mathcal{V}$ has $(s_1, \ldots, s_h)$ as positive Brunovsky indices can be stated geometrically as

$$\operatorname{Gr}_\delta(\mathcal{S}) \cap \operatorname{Inv}\,(r, s) \neq \emptyset\,. \tag{4.4}$$

According to the Thom transversality theorem [12, Th. 3.2] we have that for generic submanifolds $\mathcal{T}_1, \mathcal{T}_2$ of $\operatorname{Gr}_\delta(\mathbb{K}^n)$

$$\mathcal{T}_1 \cap \mathcal{T}_2 \neq \emptyset \Rightarrow \dim \mathcal{T}_1 + \dim \mathcal{T}_2 \geq \dim \operatorname{Gr}_\delta(\mathbb{K}^n).$$

So, by analogy, one is tempted to check whether

$$\dim \operatorname{Gr}_\delta(\mathcal{S}) + \dim \operatorname{Inv}\,(r, s) \geq \dim \operatorname{Gr}_\delta(\mathbb{K}^n). \tag{4.5}$$

provided that condition (4.4) holds. We are going to see that this is so in the generic case.

In fact, if $\eta = \dim \mathrm{Inv}\ (r,s)$, inequality (4.5) is equivalent to

$$\delta(n-d-\delta)+\eta \geq \delta(n-\delta)$$

or

$$\delta d \leq \eta. \tag{4.6}$$

We know that $\delta = s_1 + \cdots + s_h$ and

$$\eta = \sum_{i=1}^{h} \left[(s_i - s_{i+1}) \sum_{j=1}^{k} (r_{i+j-1} - s_{i+j-1}) \right]$$

(see Corollary 2.4).

Hence relation (4.6) is equivalent to

$$d\left(\sum_{i=1}^{h} s_i\right) \leq \sum_{i=1}^{h} \left[(s_i - s_{i+1}) \sum_{j=i}^{k} (r_j - s_j) \right].$$

But

$$\sum_{i=1}^{h} s_i = \sum_{i=1}^{h} i(s_i - s_{i+1})$$

so that, finally, (4.6) is equivalent to

$$\sum_{i=1}^{h} \left[(s_i - s_{i+1}) \left(id - \sum_{j=i}^{k} (r_j - s_j) \right) \right] \leq 0. \tag{4.7}$$

Since (4.7) is implied by (4.2) and this is equivalent to (4.1), we conclude that (4.5) follows from (4.1).

Now, by Theorem 4.9, if $S \in \mathfrak{S}$ (i.e., $\mathcal{S}$ is generic), condition (4.1) is necessary and sufficient for $\mathrm{Gr}_\delta(\mathcal{S}) \cap \mathrm{Inv}\ (r,s) \neq \emptyset$. In conclusion, if $S \in \mathfrak{S}$ we have that

$$\mathrm{Gr}_\delta(\mathcal{S}) \cap \mathrm{Inv}\ (r,s) \neq \emptyset \Rightarrow \dim \mathrm{Gr}_\delta(\mathcal{S}) + \dim \mathrm{Inv}\ (r,s) \geq \dim \mathrm{Gr}_\delta(\mathbb{K}^n),$$

i.e., for S generic the submanifolds $\mathrm{Gr}_\delta(\mathcal{S})$ and $\mathrm{Inv}\ (r,s)$ either do not intersect or they intersect transversaly. Following with the analogy with the Thom transversality theorem we can expect that if $\mathrm{Gr}_\delta(\mathcal{S}) \cap \mathrm{Inv}\ (r,s) \neq \emptyset$, then this intersection is a differentiable manifold whose dimension is

$$\dim \mathrm{Gr}_\delta(\mathbb{K}^n) - \dim \mathrm{Gr}_\delta(\mathcal{S}) - \dim \mathrm{Inv}\ (r,s) = \eta - \delta d.$$

We will prove by direct methods in section 6 that $\mathrm{Gr}_\delta(\mathcal{S}) \cap \mathrm{Inv}\ (r,s)$ is a submanifold of $\mathrm{Gr}_\delta(\mathbb{K}^n)$ whose dimension is $\eta - \delta d$, and we will do it not only for generic subspaces but in general.

Theorem 4.9 can be used to provide a solution, in the generic case, to the original (dual) cover problem, that is to say, without prescribing the Brunovsky indices of

the restriction, s. This is stated in the corollary below, whose proof follows from the following observation. Assume that $\underline{s}' = (s'_1, \ldots, s'_{h'})$ is such that $h' < h$. Then,

$$\sum_{j=h}^{k} r_j - s_h - hd < \sum_{j=h'+1}^{k} r_j - h'd < \sum_{j=h'}^{k} r_j - s'_{h'} - h'd.$$

COROLLARY 4.11. *With the notation of the above theorem, for every $\mathcal{S} \in \mathfrak{S}$ there is an (C, A)-invariant subspace $\mathcal{V}$ contained in $\mathcal{S}$ if and only if*

$$0 \leq \sum_{j=h}^{k} r_j - s_h - hd \tag{4.8}$$

for the unique $\underline{s}$ compatible with $\underline{r}$ having minimal h.

We will give an example applying this result. First notice that the unique partition $\underline{s}$ compatible with $\underline{r}$ (i.e., $s_i \leq r_i$ for all i) having minimal h can be explicitly constructed as follows (recall that $s_1 + \cdots + s_h = \delta$):

(i) $s_1 = \min\{r_1, \delta\}$,
(ii) for $j = 2, 3 \ldots$, $s_1 + \cdots + s_j = \min\{r_1 + \cdots + r_j, \delta\}$,

and $h = \min\{j : r_1 + \cdots r_j \geq \delta\}$.

Example 4.12. Let $\underline{r} = (3, 3, 2)$ and $d = 3$. We are going to determine when the cover problem has a generic solution for the possible values of δ. Since $n = 8$, we have $1 \leq \delta \leq 5$. We examine the inequality of formula (4.8) for the partitions $\underline{s}$ compatible with $\underline{r}$ having minimal h, that is to say, for $\delta = 1$, $\underline{s} = (1)$, for $\delta = 2$, $\underline{s} = (2)$, for $\delta = 3$, $\underline{s} = (3)$, for $\delta = 4$, $\underline{s} = (3, 1)$, and for $\delta = 5$, $\underline{s} = (3, 2)$. We conclude that for $\delta = 1, 2, 3$ the cover problem has a generic solution and for $\delta = 4, 5$, it does not.

5. The cover algorithm. In this section we present an application to the disturbance decoupling problem based on the parameterization of the solutions of the dual cover problem given in the previous sections. More precisely, given $\dot{x} = Ax + Bu + Qv$ with Q a full rank $n \times d$ matrix, we aim to parameterize the output matrices C for which Cx can be decoupled from v by state feedback. This is the case if and only if there is an (A, B)-invariant subspace, $\mathcal{V}$, such that $\operatorname{Im} Q \subseteq \mathcal{V} \subseteq \operatorname{Ker} C$. Equivalently, if and only if there is an (A^t, B^t)-invariant subspace, $\mathcal{S}$, such that $(\operatorname{Ker} C)^\perp \subseteq \mathcal{S} \subset (\operatorname{Im} Q)^\perp$. But if $[X]$ is a (B^t, A^t)-invariant subspace contained in $\operatorname{Im} Q^\perp$, then the matrices C such that $[X]^\perp \subset \operatorname{Ker} C$ are those of the form PX^*, where P is any $p \times \delta$ matrix. Thus, the following algorithm provides a paramaterization of the output matrices C for which the output of the system is insensitive to the input disturbance:

Step 1 (data capture)
- compute the controllability indices of (A, B),
- form the corresponding conjugate partition $\underline{r}$,
- form the Young diagram of $\underline{r}$, and
- list the partitions $\underline{s}$ compatible with $\underline{r}$.

Step 2 (robustness with regard to Q)
- for each δ, $1 \leq \delta \leq n - d$ do
  - find $\underline{s}$ compatible with $\underline{r}$ such that $s_1 + \cdots + s_h = \delta$ with h minimal,
  - check the condition $0 \leq \sum_{j=h}^{k} r_j - s_h - hd$,
  - if the condition is not fulfilled, then

- for almost all matrices in any open neighborhood of Q, there is not any (maximal) full rank $(n-\delta)\times n$ matrix C for which the decoupling problem has solution. There is not robustness with respect to Q. End or go to step 3 (you can try anyway).
- else
  - for almost all matrices in an open neighborhood of Q, the decoupling problem has solution with a full rank matrix $(n-\delta)\times n$. The problem will be almost surely solvable, and we have robustness with regard to Q. Go to step 3.
- end if
- end for

Step 3 (data from the particular Q)

- compute S such that $(S^{-1}AS+S^{-1}BF, S^{-1}BT)$ is in the dual Brunovsky canonical form,
- compute $F := (S^{-1}Q)^*$,
- form the block-Hankel sequence associated to F, $H = (H_1, \ldots, H_h)$,
- compute all the possible multi-indices $\underline{l}$ of H.

Step 4 (output matrix parameterization)

- for each δ, $1 \leq \delta \leq n-d$ do
  - for each $\underline{s}$ compatible with $\underline{r}$ such that $s_1+\cdots+s_h = \delta$ do (if we are only interested in a dense set of solutions we can take only $\underline{s}$ having minimal h).
    - for each $\underline{n}$ corresponding to $\underline{s}$ do
      - if $\underline{\ell}^j \cap \underline{n}^{h-j+1} = \emptyset$ for $j = 1, \ldots, h$, then
        - form the parametric matrix $X \in \mathcal{M}(\underline{r}, \underline{s})$ in the standard reduced form with indices $\underline{n}$,
        - solve $FX = 0$,
        - the output matrices C for which the decoupling problem has solution are PX^*S^{-1} where P is any $p \times \delta$ matrix and S the matrix in step 3.
      - end if
    - end for
  - end for
- end for

6. The differentiable structure of the solutions. Along this section we are given a subspace $\mathcal{S} \in Gr_{n-d}(\mathbb{K}^n)$ for which there is at least one $\mathcal{V} \in \text{Inv}\ (\underline{r}, \underline{s})$ such that $\mathcal{V} \subset \mathcal{S}$. A necessary and sufficient condition for such a property to hold was given in Theorem 3.4. Our aim in this section is to show that for a subspace $\mathcal{S}$, the set $\text{Inv}\ (\underline{r}, \underline{s}, \mathcal{S}) = \{\mathcal{V} \in \text{Inv}\ (\underline{r}, \underline{s}) | \mathcal{V} \subset \mathcal{S}\}$ can be provided with a differentiable structure, to compute its dimension and to find an atlas of coordinate charts. This will allow us to parameterize the solutions of the (dual) cover problem.

To begin, let $F \in \mathcal{M}^*_{d,n}$ be a matrix such that $\mathcal{S}^\perp = [F^*]$ and let $X \in \mathcal{M}(\underline{r}, \underline{s})$ be such that $\mathcal{V} = [X]$ with $\mathcal{V} \in \text{Inv}\ (\underline{r}, \underline{s}, \mathcal{S})$. Then $FX = 0$. Let

$$\mathcal{N}(\underline{r}, \underline{s}, \mathcal{S}) := \{X \in \mathcal{M}(\underline{r}, \underline{s}) | FX = 0\},$$

which is an open set of a linear manifold. Notice that if $F, G \in \mathcal{M}^*_{d,n}$ such that $\mathcal{S}^\perp = [F^*] = [G^*]$, then $\{X \in \mathcal{M}(\underline{r}, \underline{s}) | FX = 0\} = \{X \in \mathcal{M}(\underline{r}, \underline{s}) | GX = 0\}$. For notational simplicity we will write $\mathcal{N} := \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ and $\mathcal{M} := \mathcal{M}(\underline{r}, \underline{s})$. Let $\mathcal{G} := \mathcal{G}(\underline{s})$ be the group defined in Definition 2.1. By Theorem 2.3 this group acts freely on $\mathcal{M}$

on the right by matrix multiplication as well as on $\mathcal{N}$. In addition we can identify Inv $(\underline{r}, \underline{s}, \mathcal{S})$ with $\mathcal{N}/\mathcal{G}$ by means of the bijection

$$\begin{array}{ccc} \mathcal{N}/\mathcal{G} & \longrightarrow & \text{Inv } (\underline{r}, \underline{s}, \mathcal{S}), \\ \{XP | X \in \mathcal{N}, P \in \mathcal{G}(\underline{s})\} & \rightsquigarrow & [X]. \end{array}$$

Through this bijection a differentiable structure on Inv $(\underline{r}, \underline{s}, \mathcal{S})$ can be defined.

THEOREM 6.1. (i) *The orbit space $\mathcal{N}/\mathcal{G}$ has a differentiable stucture such that the natural projection $\pi : \mathcal{N} \to \mathcal{N}/\mathcal{G}$ is a submersion.*

(ii) $\dim(\mathcal{N}/\mathcal{G}) = \dim \mathcal{N} - \dim \mathcal{G}$.

Proof. (i) According to [9, Th. 2.9.10] we have to prove

(a) $\Gamma = \{(X, X') \in \mathcal{N} \times \mathcal{N} | X' = XP, P \in \mathcal{G}\}$ is closed in $\mathcal{N} \times \mathcal{N}$, and

(b) the map $\gamma : \mathcal{N} \times \mathcal{G} \to \Gamma$ defined by $\gamma(X, P) = (X, XP)$ is a homeomorphism.

In order to prove (a) notice that if $\Gamma_1 = \{(X, X') \in \mathcal{M} \times \mathcal{M} | X' = XP, P \in \mathcal{G}\}$, then $\Gamma = \Gamma_1 \cap (\mathcal{N} \times \mathcal{N})$. As Γ_1 is closed in $\mathcal{M} \times \mathcal{M}$ (see [5]), Γ is closed in $\mathcal{N} \times \mathcal{N}$.

On the other hand, γ is clearly bijective due to the free action of $\mathcal{G}$ on $\mathcal{N}$, and it continues because it is the restriction of the homeomorphism $\gamma_1 : \mathcal{M} \times \mathcal{G} \to \Gamma_1$ defined by $\gamma_1(X, P) = (X, XP)$ (again see [5]). Its inverse is also to continue for the same reason.

(ii) The dimension formula follows from (i) as in [5]. □

Now we compute the dimension of $\mathcal{N}/\mathcal{G}$. On one hand (see [7]),

$$\dim \mathcal{G} = \sum_{i=1}^{k} \sum_{j=1}^{h} (s_i - s_{i+1}) s_{i+j-1}.$$

On the other hand, $X \in \mathcal{N}$ if and only if $X \in \mathcal{M}$ and $FX = 0$. As shown in the previous section this is equivalent to

$$H_{h-j+1} R_j = 0, \quad j = 1, \ldots, h,$$

where $(R_1, \ldots, R_h)$ is the condensed form of X and $(H_1, \ldots, H_h)$ is a block-Hankel structure of $\mathcal{S}$ with respect to (C, A). This is equivalent, in turns, to

$$H_{h-j+1} R_j(\ , i) = 0, \quad j = 1, \ldots, h, i = 1, \ldots, s_{h-j+1} - s_{h-j+2},$$

where, as in section 3, $R_j(\ , i)$ is the ith column of R_j. The set of solutions of this linear homogeneous system is a vector space of dimension $(r_{h-j+1} + \cdots + r_k) - \operatorname{rank} H_{h-j+1}$; i.e., the solutions depend on $(r_{h-j+1} + \cdots + r_k) - \operatorname{rank} H_{h-j+1}$ parameters. Thus the solutions of $H_{h-j+1} R_j = 0$ depend on $(s_{h-j+1} - s_{h-j+2})\left(\sum_{i=h-j+1}^{k} r_i - \operatorname{rank} H_{h-j+1}\right)$ parameters. Since we have h of these equations,

$$\dim \mathcal{N} = \sum_{j=1}^{h} (s_j - s_{j+1}) \left(\sum_{i=j}^{k} r_i - \operatorname{rank} H_j \right) = \sum_{j=1}^{h} (s_j - s_{j+1}) \left(\sum_{i=1}^{k} r_{i+j-1} - \operatorname{rank} H_j \right).$$

Hence we have proved the following.

PROPOSITION 6.2.

$$\dim \mathcal{N}/\mathcal{G} = \sum_{j=1}^{h}(s_j - s_{j+1})\left(\sum_{i=1}^{k}(r_{i+j-1} - s_{i+j-1}) - \operatorname{rank} H_j\right).$$

Recall that (see [7])

$$\dim \mathcal{M}/\mathcal{G} = \sum_{j=1}^{h}\sum_{i=1}^{k}(s_j - s_{j+1})(r_{i+j-1} - s_{i+j-1}).$$

Hence

$$\dim \mathcal{N}/\mathcal{G} = \dim \mathcal{M}/\mathcal{G} - \sum_{j=1}^{h}(s_j - s_{j+1}) \operatorname{rank} H_j.$$

Notice that, in the generic case, by Lemma 4.8 $\operatorname{rank} H_j = jd$, and so $\sum_{j=1}^{h}(s_j - s_{j+1}) \operatorname{rank} H_j = d\sum_{j=1}^{h} j(s_j - s_{j+1}) = d\sum_{j=1}^{h} s_j = d\delta$. Since $\eta = \dim \operatorname{Inv}(\underline{r}, \underline{s}) = \dim \mathcal{M}/\mathcal{G}$, we recover the result in Remark 4.10.

Since $\mathcal{N}/\mathcal{G} \subset \mathcal{M}/\mathcal{G}$ and the topology of $\mathcal{N}/\mathcal{G}$ is the induced one by that of $\mathcal{M}/\mathcal{G}$ ($\mathcal{N}$ is a submanifold of $\mathcal{M}$ and the projection $\mathcal{M} \to \mathcal{M}/\mathcal{G}$ is an open map), we have the following.

PROPOSITION 6.3. *$\mathcal{N}/\mathcal{G}$ is a regular submanifold of $\mathcal{M}/\mathcal{G}$.*

We are going now to parameterize $\mathcal{N}/\mathcal{G}$ by means of a coordinate atlas.

First we recall an elementary fact from linear algebra. If the homogeneous linear system $Ax = 0$ has a solution, $\{a_{i_1}, \dots, a_{i_r}\}$ is a basis of the column span of A and $\{j_1, \dots, j_{n-r}\} = \{1, \dots, n\} \backslash \{i_1, \dots, i_r\}$, then

$$a_{j_k} = \sum_{t=1}^{r} \alpha_{kt} a_{i_t}, \quad k = 1, \dots, n-r,$$

if and only if

$$x_{i_t} = -\sum_{k=1}^{n-r} \alpha_{tk} x_{j_k}, \quad t = 1, \dots, r.$$

Consider now a matrix $X \in \mathcal{M}(\underline{r}, \underline{s})$ in reduced form (Definition 2.5) such that $FX = 0$, and for $j = 1, \dots, h$ let $(R_1, \dots, R_h)$ be the condensed form of X. Let $\underline{\ell} = (\underline{\ell}^1, \dots, \underline{\ell}^k)$ and $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ be compatible multi-indices of $\mathcal{S}$ and X, respectively; i.e.,

$$\underline{\ell}^j \cap \underline{n}^{h-j+1} = \emptyset, \quad j = 1, \dots, h.$$

If $\underline{\ell}^j = (\underline{\ell}^j_1, \dots, \underline{\ell}^j_{t_j})$, $t_j = \operatorname{rank} H_j$, then $\ell^j_1, \dots, \ell^j_{t_j}$ are linearly independent columns of H_j, $(H_1, \dots, H_h)$ being a block-Hankel structure of $\mathcal{S}$ with respect to (C, A).

Since X is in reduced form, the rows of R_{h-j+1} in $\underline{n}^{h-j+1}$ are canonical vectors; i.e., $R_{h-j+1}(\underline{n}^{h-j+1}, \ ) = I_{m_j}$. And the rows of R_{h-j+1} in the set of indices $\underline{p}^{h-j+1} = \underline{n}^1 \cup \cdots \cup \underline{n}^{h-j} \cup (r_j + \underline{n}^1) \cup \cdots \cup (r_j + \underline{n}^{h-j}) \cup (r_{j+1} + r_j + \underline{n}^1) \cup \cdots \cup (r_{j+1} + r_j + \underline{n}^{h-j-1}) \cup \cdots \cup (r_k + \cdots + r_j + \underline{n}^1)$ are zero rows; i.e., $R_{h-j+1}(\underline{p}^{h-j+1}, \ ) = 0$. (Compare with

the indices in the proof of Theorem 3.6.) But $FX = 0$ is equivalent to $H_j R_{h-j+1} = 0$, $j = 1, \dots, h$. So, if $s^j = \#\underline{p}^{h-j+1} = 2s_{j+1} + s_{j+2} + \cdots + s_h$, $r^j = r_j + \cdots + r_k$ and

$$\{g_1^j, \dots, g_{r^j - t_j - s^j}^j\} = \{1, \dots, r^j\} \backslash (\underline{\ell}^j \cup \underline{p}^{h-j+1}),$$

then columns $H_j(\ , g_i^j)$ are linear combinations of the columns $\ell_1^j, \dots, \ell_{t_j}^j$ of H_j. We can write

$$H_j(\ , g_i^j) = \sum_{q=1}^{t_j} \alpha_{iq} H_j(\ , \ell_q^j), \quad i = 1, \dots, r^j - t_j - s^j, \tag{6.1}$$

and so

$$R_{h-j+1}(\ell_i^j, \ ) = - \sum_{q=1}^{r^j - t_j - s^j} \alpha_{qi} R_{h-j+1}(g_q^j, \ ), \quad i = 1, \dots, t_j. \tag{6.2}$$

This proves the following.

THEOREM 6.4. *For any matrix $X \in \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ there are multi-indices $\underline{\ell} = (\underline{\ell}^1, \dots, \underline{\ell}^h)$ of $\mathcal{S}$ and $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ of X such that*

(i) $\ell^j \cap \underline{n}^{h-j+1} = \emptyset, \quad j = 1, \dots, h.$

(ii) *$Y = XP$ is in reduced form, as give by Definition 2.5, for some $P \in \mathcal{G}(\underline{s})$.*

(iii) *if $(H_1, \dots, H_h)$ is a block-Hankel structure of $\mathcal{S}$, $(R_1, \dots, R_h)$ the condensed form of Y, $\{g_1^j, \dots, g_{r^j - t_j - s^j}^j\} = \{1, \dots, r^j\} \backslash (\underline{\ell}^j \cup \underline{p}^{h-j+1})$ for $j = 1, \dots, h$ and $H_j(\ , g_i^j)$ is given by (6.1), then matrix $R_{h-j+1}(\ell_i^j, \ )$ is given by (6.2).*

DEFINITION 6.5. *With the notation of the above theorem we will say that Y is a reduced form of $X \in \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ with respect to $\underline{n}$ and $\underline{\ell}$.*

Let us compute now the number of free parameters in a reduced form of $X \in \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$. We will show that this is the dimension of $\mathcal{N}/\mathcal{G}$. In fact, the number of free parameters of any reduced form is the same as the number of free parameters in its condensed form. And this is the number of free parameters in the condensed form of the reduced form of $X \in \mathcal{M}(\underline{r}, \underline{s})$ minus the number of elements in $R_{h-j+1}(\ell_i^j, \ )$, $j = 1, \dots, h$, $i = 1, \dots, t_j$. Bearing in mind that the number of free parameters in any reduced form $Y \in \mathcal{M}(\underline{r}, \underline{s})$ is $\dim \mathcal{M}/\mathcal{G}$ (see [7]), we conclude that the number of free parameters in $Y \in \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ is (recall that the number of columns of R_{h-j+1} is $s_j - s_{j+1}$)

$$\dim \mathcal{M}/\mathcal{G} - \sum_{j=1}^{h} t_j (s_j - s_{j+1}) = \dim \mathcal{N}/\mathcal{G}.$$

DEFINITION 6.6. *A multi-index $\underline{n} = (\underline{n}^1, \dots, \underline{n}^h)$ for $X \in \mathcal{M}(\underline{r}, \underline{s})$ will be said an admissible set of indices (or a multi-index) for $X \in \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ if there is a multi-index $\underline{\ell} = (\underline{\ell}^1, \dots, \underline{\ell}^h)$ of $\mathcal{S}$ such that $\underline{\ell}^j \cap \underline{n}^{h-j+1} = \emptyset$, $j = 1, \dots, h$.*

The following lemmas follow immediately from Lemmas 3 and 4 of [7] (see Proposition 2.8).

LEMMA 6.7. *Let $X \in \mathcal{N}$ and $Q \in \mathcal{G}$. If $\underline{n}$ is an admissible set of indices for X it is also an admissible set of indices for XQ.*

LEMMA 6.8. *Let Y and $\tilde{Y}$ be matrices of $\mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ in reduced form with the same multi-index $\underline{n}$. If there is a matrix $P \in \mathcal{G}$ such that $\tilde{Y} = YP$, then P is the identity matrix.*

In order to define the coordinate charts of $\mathcal{N}/\mathcal{G}$ we start as in [7]. Let Λ be the set of multi-indices $\underline{n} = (\underline{n}^1, \ldots, \underline{n}^h)$ for which there is a multi-index, $\underline{\ell} = (\underline{\ell}^1, \ldots, \underline{\ell}^k)$, of $\mathcal{S}$ with respect to (C, A) such that $\underline{\ell}^j \cap \underline{n}^{h-j+1} = \emptyset$, $j = 1, \ldots, h$, i.e., compatible with $\underline{n}$. For $\underline{n} \in \Lambda$ let $\mathcal{U}_{\underline{n}}$ denote the set of matrices $X \in \mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ with $\underline{n}$ as a set of admissible indices. Since all matrices in $\mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$ close enough to X have $\underline{n}$ as a set of admissible indices (because these are just linear independent rows), we have that $\{\mathcal{U}_{\underline{n}} | \underline{n} \in \Lambda\}$ is an open covering of $\mathcal{N}(\underline{r}, \underline{s}, \mathcal{S})$. If $\pi : \mathcal{N} \to \mathcal{N}/\mathcal{G}$ is the natural projection, then π is open and $\{\tilde{\mathcal{U}}_{\underline{n}} = \pi(\mathcal{U}_{\underline{n}}) | \underline{n} \in \Lambda\}$ is an open covering of $\mathcal{N}/\mathcal{G}$.

We must notice that for a multi-index $\underline{n} \in \Lambda$, there may be several multi-indices $\underline{\ell}$ compatible with it. To define the coordinate charts we aim to find a differentiable onto mapping $\phi_{\underline{n}} : \mathcal{U}_{\underline{n}} \to \mathbb{K}^N$, $N = \dim \mathcal{N}/\mathcal{G}$, for each $\underline{n} \in \Lambda$. We proceed as follows. If $\underline{n} \in \Lambda$ admits several compatible multi-indices $\underline{\ell} = (\underline{\ell}^1, \ldots, \underline{\ell}^h)$ we take the one such that for $i = 1, \ldots, h$, $\underline{\ell}^i$ is the smallest in the lexicographical order. In other words, as the elements of $\underline{\ell}^i$ are linearly independent columns of H_i, we are taking its first t_i linearly independent columns compatible with $\underline{n}$. Then for $X \in \mathcal{U}_{\underline{n}}$ we define $\phi_{\underline{n}}(X)$ as the point of $\mathbb{K}^N$ defined by the free parameters of the reduced form of X corresponding to $\underline{n}$ and $\underline{\ell}$, Y, in a certain order. For example, if $(R_1, \ldots, R_h)$ is the condensed form of Y, then the free parameters of Y are in the rows $\{f_1^j, \ldots, f_{r^j - t_j - s'^j}^j\} = \{1, \ldots, r^j\} \backslash (\underline{\ell}^j \cup \underline{p}^{h-j+1} \cup \underline{n}^{h-j+1})$ $(s'^j = s_j + s_{j+1} + \cdots + s_h)$ of R_{h-j+1}, $j = 1, \ldots, h$. Thus we may define

$$\begin{aligned}\phi_{\underline{n}}(X) = (&R_1(f_1^1, \,)\, R_1(f_2^1, \,) \cdots R_1(f_{r^j - t_1 - s'^1}^1, \,) \\ &R_2(f_1^2, \,) \cdots R_2(f_{r^j - t_2 - s'^2}^2, \,) \cdots R_h(f_{r^j - t_h - s'^h}^h, \,)).\end{aligned}$$

In this way $\phi_{\underline{n}}$ is well defined and onto. The differentiability follows from the fact that $Y_{\underline{n}}$ is obtained form X by means of rational operations. This mapping induces the mapping $\theta_{\underline{n}} : \tilde{\mathcal{U}}_{\underline{n}} \to \mathbb{K}^N$.

THEOREM 6.9. *With the above notation $\theta_{\underline{n}}$ is a diffeomorphism and $\{\tilde{\mathcal{U}}_{\underline{n}}; \underline{n} \in \Lambda\}$ is a coordinate atlas of $\mathcal{N}/\mathcal{G}$.*

Proof. The proof is in [7]. □

Example 6.10. Consider a slight modification of the example at the end of the section 3: $\underline{r} = (5, 3, 2, 1)$, $\underline{s} = (3, 3, 1)$, and $\mathcal{S}$ such that $\mathcal{S}^\perp = [F^*]$, where $F = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$. Then

$$H_1 = F = \left(\begin{array}{ccccc|ccc|cc|c} 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$H_2 = \left(\begin{array}{ccc|cc|c} 1 & 0 & 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

$$H_3 = \left(\begin{array}{cc|c} 1 & 0 & 2 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

An admissible multi-index in $\mathcal{M}(\underline{r}, \underline{s})$ is $\underline{n} = ((2), (3), (4, 5))$. But now this multi-index is compatible with two multi-indices $\underline{\ell}$: $\underline{\ell}_1 = ((1), (1, 4), (1, 3))$ and $\underline{\ell}_2 = ((6), (1, 4), (1, 3))$. The one to be chosen to construct the coordinate chart is $\underline{\ell}_1$. The reduced form cor-

responding to $\underline{n}$ and $\underline{\ell}_1$ is

$$Y = \left(\begin{array}{cccc|cc|c} 0 & 0 & -2a_2 & -2a_3 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & a_2 & a_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & a_1 & a_4 & a_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & a_6 & a_7 & a_4 & a_5 & 0 \end{array}\right),$$

and so $\phi_{\underline{n}}(\tilde{\mathcal{U}}_{\underline{n}}) = \{(a_1, a_2, a_3, a_4, a_5, a_6, a_7) \in \mathbb{R}^7\} = \mathbb{R}^7$.

Our last result is about the connection of the manifold $\mathcal{N}/\mathcal{G}$.

THEOREM 6.11. *If $\mathbb{K} = \mathbb{C}$, the manifold $\mathcal{N}/\mathcal{G}$ is connected.*

Proof. It is sufficient to prove that $\mathcal{N}$ is connected.

Assume, on the contrary, that there are open sets $U_1, U_2 \in \mathcal{N}$ such that $U_1 \cup U_2 = \mathcal{N}$ and $U_1 \cap U_2 = \emptyset$. Let $X_1 \in U_1$ and $X_2 \in U_2$ and let $\underline{n}$ be an admissible multi-index for X_1. For a given matrix $X \in \mathcal{N}$, let $M_{\underline{n}}(X)$ denote the submatrix of X formed by the rows in $\underline{n}$ and the first s_1 columns. As $\underline{n}$ is a multi-index admissible for X_1 we have that $\det M_{\underline{n}}(X_1) \neq 0$.

For $z \in \mathbb{C}$ define $p(z) = \det M_{\underline{n}}(X_1 + z(X_2 - X_1))$. Since $p(0) = \det M_{\underline{n}}(X_1) \neq 0$, $p(z)$ is not the zero polynomial, and so $p(z)$ has finitely many zeros. This allows us to define a trajectory $z(t) : [0, 1] \to \mathbb{C}$ such that $z(0) = 0$, $z(1) = 1$ and $p(z(t)) \neq 0$ for all $t \in [0, 1)$. In other words, if $X(t) = X_1 + z(t)(X_2 - X_1)$, then $\det M_{\underline{n}}(X(t)) \neq 0$ for all $t \in [0, 1)$. Notice that for $t \in [0, 1)$, $X(t) \in \mathcal{N} = U_1 \cup U_2$. Furthermore, since $X_1 \in U_1$, $\{X(t) \mid t \in [0, 1)\}$ is connected (because it is the image of a connected set by a continuous map) and $U_1 \cap U_2 = \emptyset$ we conclude that $\{X(t) \mid t \in [0, 1)\} \subset U_1$.

Now as $X_2 \in U_2$ and U_2 is open there is a neighborhood of X_2 contained in U_2. But X_2 is the limit of a sequence of matrices in $\{X(t) \mid t \in [0, 1)\}$. This means that in that neighborhood of X_2 there are matrices of this set. Thus $U_1 \cap U_2 \neq \emptyset$, a contradiction. ∎

7. Conclusions. In this paper the solutions of the cover problem have been studied by taking a geometric approach. The main tool is the parameterization of the set of (C, A)-invariant subspaces according to the Brunovsky indices of the restriction, each strata being a differentiable manifold. The structure of the matrices whose columns form a basis of such (C, A)-invariant subspaces (with respect to the dual Brunovsky basis of $\mathbb{K}^n$ and $\mathbb{K}^{n+m}$) was used to produce a necessary and sufficient condition for the cover problem to have a solution in each strata. This condition was used to provide the set of solutions with a differentiable structure in such a way that this set is a regular submanifold of the previous one. Then, an atlas of coordinate charts for the set of solution was given, obtaining a parameterization of all solutions of the cover problem on each strata. The generic case was also considered. In this case a simple and very easily computable condition for the cover problem to have a solution has been shown to hold.

Acknowledgment. We want to thank the referees for their suggestions and comments, which were very useful to improve the paper.

REFERENCES

[1] A. C. Antoulas, *New results on the algebraic theory of linear systems, The solution of the cover problems*, Linear Algebra Appl., 50 (1983), pp. 1–43.

[2] I. Baragaña and I. Zaballa, *Block similarity invariants of restrictions to* (A, B)*-invariant subspaces*, Linear Algebra Appl., 220 (1995), pp. 31–62.

[3] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1992.

[4] P. Brunovsky, *Classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–188.

[5] J. Ferrer, F. Puerta, and X. Puerta, *Differentiable structure of the set of the controllable* $(A, B)^t$*-invariant subspace*, Linear Algebra Appl., 275-276 (1998), pp. 161–177.

[6] F. Puerta and X. Puerta, *On the geometry of the set of controllability subspaces of a pair* (A, B), Linear Algebra Appl., 351-352 (2002), pp. 585–599.

[7] F. Puerta, X. Puerta, and I. Zaballa, *A coordinate atlas of the manifold of observable conditioned invariant subspaces*, Internat. J. Control, 74 (2001), pp. 1226–1238.

[8] F. Puerta, X. Puerta, and I. Zaballa, *On the parametrization of the controllability subspaces of a controllable pair*, Linear Algebra Appl., 401 (2005), pp. 429–444.

[9] V. S. Varadarajan, *Lie Groups, Lie Algebras and Their Representations*, Springer-Verlag, New York, 1984.

[10] J. C. Willems, *Almost invariant subspaces: An approach to high gain feedback desing—Part* I: *Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235–252.

[11] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1984.

[12] Yung-Chen Lu, *Singularity Theory and an Introduction to Catastrophe Theory*, Springer-Verlag, New York, 1976.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 414–431

A TURNPIKE THEOREM FOR A RISK-SENSITIVE MARKOV DECISION PROCESS WITH STOPPING*

ERIC V. DENARDO† AND URIEL G. ROTHBLUM‡

Abstract. This paper concerns a Markov decision problem with an infinite planning horizon, stationary data, an exponential utility function, and no discounting, but also with the possibility of voluntary or involuntary termination (stopping). This paper establishes conditions under which the expected utility of the income received by repeated use of a well-chosen stationary control equals the limit, as n becomes large, of the optimal expected utility for the n-period problem. Tests for these (turnpike) conditions are provided, as is a linear program whose optimal solution determines such a control.

Key words. Markov control process, risk-sensitive dynamic programming, stopping, turnpike theorem, overtaking optimality

AMS subject classifications. Primary: 90C40; Secondary: 60G40, 90C39, 91B16

DOI. 10.1137/S0363012904442616

1. Introduction.

Howard and Matheson [1972] were the first to analyze a Markov decision problem with a (risk-sensitive) exponential utility function. As a performance criterion, they focused on the "gain rate"—the improvement in certain monetary equivalent (CME) per epoch under steady-state conditions. Their work kindled substantial interest in gain-rate maximization with an exponential utility function: for example, see Jacobson [1972, 1977], Hernández Lerma [1989], Whittle [1990], Fleming and Hernández-Hernández [1997, 1999], Di Masi and Stettner [2000], Cavazos-Cadena and Montes-de-Oca [2003], Cavazos-Cadena and Hernández-Hernández [2004], and the papers they cite. Each of these papers studies gain-rate maximization in a risk-sensitive model with an infinite planning horizon, no discounting, and no possibility of stopping.

Howard and Matheson [1972] rest their analysis on the hypothesis that all states communicate under every control. They demonstrate the equivalence of maximizing the CME and optimizing the spectral radius of the corresponding (positive) transition matrix. This links their work to earlier papers on maximizing the spectral radii of controlled branching systems by Bellman [1957], Mandl [1967], and Mandl and Seneta [1969] and to later work by Sladky [1980], Zijm [1980, 1987], Rothblum and Whittle [1982], Rothblum [1984], and Rothblum and Veinott [forthcoming].

Here, we study a somewhat different model of a Markov decision process with exponential utility with stopping. This model was introduced in Denardo and Rothblum (DR) [1979a]. In it, voluntary or involuntary stopping serves to place an upper bound on the net income that can be earned prior to termination, and the decision maker's goal is to maximize the expected utility of the net income that is earned before stopping occurs. In the context of this model, the widely used gain-rate criterion is unselective: *every* control whose expected net income prior to termination is finite has

*Received by the editors March 30, 2004; accepted for publication (in revised form) October 3, 2005; published electronically March 17, 2006.
http://www.siam.org/journals/sicon/45-2/44261.html
†Center for Systems Science, Yale University, PO 208267, New Haven, CT 06520 (eric.denardo@yale.edu).
‡Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel (rothblum@ie.technion.ac.il).

zero as its gain rate. In DR [1979a] the best of the stationary controls were found, but no comparisons were attempted between stationary controls and more general decision procedures. Here, we establish a *turnpike theorem* by showing that the expected utility of the income received by repeated use of any one of the best stationary controls equals the limit, as n becomes large, of the optimal expected utility for the n-period problem. Turnpike theorems like this one are a prime justification for repeated use of a stationary control when the planning horizon is long and possibly of uncertain length. A corollary (presented in section 4) is that each (stationary) turnpike control overtakes (is overtaking optimal over) all randomized history-remembering decision procedures.

The hypothesis on which this turnpike theorem rests avoids assumptions on the chain structure. In their place are conditions guaranteeing that no control has aggregate payoff that approaches plus infinity but allowing poorly chosen controls to have aggregate payoff that approaches minus infinity. This turnpike theorem sets a proper foundation for the analysis of the multiarmed bandit problem with exponential utility in Denardo, Park, and Rothblum [2004]. This analysis also relates to work by Denardo, Rothblum, and Van der Heyden [2004] on index policies for a search problem with exponential utility and to work by DR [2005] on multiplicative systems.

The model on which this paper is focused is now presented. A system is observed sequentially at epochs labeled $1, 2, \ldots$, while termination has not occurred. At each epoch prior to termination, the system is observed to be in one of N states that are labeled 1 through N. If the system is observed to occupy state i at epoch t, the decision maker must choose some action k in the finite set D_i, with these consequences:

(i) With probability of $P_{iu}^k \geq 0$, the decision maker will earn (possibly negative) reward w_{iu}^k and will observe the system to be in state u at epoch $t+1$.

(ii) With probability of $T_i^k = 1 - \sum_{u=1}^N P_{iu}^k \geq 0$, the decision maker will earn (possibly negative) reward w_i^k, and termination (stopping) will occur.

After termination occurs, the decision maker observes no state, selects no action, and earns no further reward.

The utility $u(x)$ that the decision maker places on x units of income has an exponential form whose shape is determined by a parameter (constant) λ: in the *risk-averse* case, λ is positive and $u(x) = -e^{-\lambda x}$, while in the *risk-seeking* case, λ is negative and $u(x) = e^{-\lambda x}$. To consolidate these cases, we designate $\alpha = \text{sign}(\lambda)$ and write $u(x) = -\alpha e^{-\lambda x}$. The decision maker has no time-preference for money, so the utility $U(w_1, \ldots, w_n)$ of receiving (possibly negative) income $w_1, \ldots, w_n$ in epochs $1, \ldots, n$, respectively, is given by

$$U(w_1, \ldots, w_n) = -\alpha e^{-\lambda(w_1 + \cdots + w_n)} = e^{-\lambda w_1} U(w_2, \ldots, w_n). \tag{1}$$

Equation (1) reflects the multiplicative property of the exponential utility function, and the fact that $e^{-\lambda w_1}$ is nonnegative ensures that the monotonicity property of dynamic programming will be satisfied.

The decision maker's goal is to maximize the expectation of the utility earned until termination. Denote as $f(n)$ the $N \times 1$ vector whose ith element $f(n)_i$ equals the supremum over all randomized history-remembering decision procedures of the expected utility for starting at state i and continuing for n transitions or until termination occurs, whichever comes first. With R_i^k and Q_{iu}^k defined by

$$R_i^k = -T_i^k \alpha e^{-\lambda w_i^k}, \quad Q_{iu}^k = P_{iu}^k e^{-\lambda w_{iu}^k} \qquad \forall\, i,\ \forall\, u, \tag{2}$$

the fact that the Q_{iu}^k are nonnegative and the multiplicative property in (1) lead to the recursion,

$$f(0)_i = -\alpha \qquad \forall\, i, \tag{3}$$

$$f(n)_i = \max_{k \in D_i} \left\{ R_i^k + \sum_u Q_{iu}^k f(n-1)_u \right\} \qquad \forall\, i,\ \forall\, n \geq 1. \tag{4}$$

To relate (4) to a sequential decision problem having a linear utility function, interpret Q_{iu}^k as a *rate* (of transition) and R_i^k as a *reward*. Note from (2) that the rewards (the R_i^k) are nonpositive in the risk-averse case and are nonnegative in the risk-seeking case. Note from (3) that the natural "terminal reward" for this model is $-\alpha$, and not 0 as would be the case for a linear utility function.

This paper is focused on the limiting behavior of $f(n)$ as n becomes large. The $N \times 1$ vector ϕ is specified by

$$\phi_i = \limsup_{n \to \infty} f(n)_i \qquad \forall\, i. \tag{5}$$

Interpret ϕ_i as the least upper bound, as n becomes large, on the expected utility for starting at state i and operating the system for n periods or until termination, whichever comes first. In the risk-averse case, ϕ_i can be as large as zero and as small as $-\infty$. In the risk-seeking case, ϕ_i can be as large as $+\infty$ and as small as 0. This paper establishes conditions under which ϕ is finite and is attained by repeated use of a stationary nonrandomized decision procedure that will be called a "turnpike control."

Throughout, "control" is shorthand for a stationary nonrandomized decision procedure, one whose actions are functions only of the observed state. By contrast, "strategy" is short for a possibly randomized decision procedure whose action can depend on the current state and the prior history of the process. Formally, a *control* δ is an $N \times 1$ vector whose ith element $\delta(i)$ is in D_i for $i = 1, \ldots, N$. To *use* control δ is to select decision $\delta(i)$ at each epoch at which state i is observed. The set D of all such controls is given by $D = D_1 \times \cdots \times D_N = \prod_{i=1}^N D_i$, where $\prod$ denotes the Cartesian product. Each control δ has an $N \times 1$ *reward vector* R^δ and an $N \times N$ *rate* matrix Q^δ whose entries are given by

$$\left[R^\delta\right]_i = R_i^{\delta(i)}, \qquad \left[Q^\delta\right]_{iu} = Q_{iu}^{\delta(i)} \qquad \forall\, i,\ \forall\, u. \tag{6}$$

We designate as $v^\delta(n)$ the $N \times 1$ vector whose ith entry $v^\delta(n)_i$ equals the expected utility for using (stationary) control δ for n epochs given initial state i. Conditioning this expectation on the first transition and using the multiplicative property in (1) leads to the recursion,

$$\begin{aligned} v^\delta(n) &= R^\delta + Q^\delta v^\delta(n-1) \\ &= [1 + Q^\delta + \cdots + (Q^\delta)^{n-1}]R^\delta - (Q^\delta)^n e\alpha, \end{aligned} \tag{7}$$

where e denotes the $N \times 1$ vector of 1's.

A (stationary) control π is now said to be a *turnpike control* if

$$\lim_{n \to \infty} v^\pi(n) = \phi, \tag{8}$$

if each entry in ϕ is finite, and, in the risk-averse case, if no entry in ϕ equals zero. This paper provides conditions under which a turnpike control exists. It provides tests

for those conditions, and it shows how to find the turnpike control (or controls) from the optimal solution to a linear program. Evidently, a (stationary) turnpike control performs well when the actual planning horizon n is long and perhaps of uncertain length.

Before introducing conditions that ensure the existence of a turnpike control, we review some standard nomenclature. If B is an $N \times N$ matrix and S and T are subsets of $\{1, 2, \ldots, N\}$, then B_{ST} denotes the submatrix of B whose rows are indexed by the entries in S and whose columns are indexed by the entries in T. Similar notation applies to subvectors. If J is a subset of $\{1, 2 \ldots, N\}$, then J' denotes the set $\{1, 2, \ldots, N\} \setminus J$. With B as a square matrix, $\rho(B)$ denotes the *spectral radius* of B, namely, the largest absolute value of B's eigenvalues. A square matrix B is called *transient* if $\rho(B) < 1$. A standard result is that a square matrix B is transient if and only if $B^n \to 0$ as $n \to \infty$, and moreover, that if B is transient, then $I - B$ is invertible, with

$$(1 - B)^{-1} = \sum_{n=0}^{\infty} B^n. \tag{9}$$

A nonempty subset J of $\{1, 2, \ldots, N\}$ is said to be *closed* under control δ if

$$0 = T_i^{\delta(i)} + \sum_{u \notin J} P_{iu}^{\delta(i)} \qquad \forall\, i \in J. \tag{10}$$

Evidently, if J is closed under a control, then "escape" from J cannot occur when that control is used repeatedly. For the risk-averse case, the hypothesis that suffices for the existence of a turnpike control includes the following.

Condition I. $\rho[(Q^\pi)_{JJ}] > 1$ for each control π and set J such that J is closed under π.

Two different linear programs test for Condition I; one of them appears in section 2 and the other in section 3.

To explore Condition I, we consider a control π that has a closed set J. Equations (10) and (2) imply $[R^\pi]_J = 0$ and $[Q^\pi]_{JJ'} = 0$, and it follows from (7) that

$$[v^\pi(n)]_J = -[(Q^\pi)_{JJ}]^n e_J \alpha. \tag{11}$$

The risk-averse case has $\alpha = 1$, so Condition I and (11) guarantee that starting in at least one state i in J using control π repeatedly earns expected utility of $-\infty$, in which sense control π is to be avoided, if possible.

In the hypotheses that appear below, RA stands for risk-averse, RS for risk-seeking.

Hypothesis RA. At least one control δ has rate matrix Q^δ that is transient, and Condition I holds.

Hypothesis RS. Each control δ has rate matrix Q^δ that is transient.

The central result in this paper is the following theorem.

THEOREM 1. *Suppose either that* $\lambda > 0$ *and Hypothesis* RA *is satisfied or that* $\lambda < 0$ *and Hypothesis* RS *is satisfied. Then there exists a (stationary) turnpike control.*

In concert, this paper and DR [1979a] demonstrate the existence of a turnpike control, provide tests for Hypotheses RA and RS, and use the optimal solution to a linear program to construct the turnpike control (or controls). In particular, equation

(19) constructs a set D^* of controls from the optimal solution to a linear program, and Theorem 1a (in section 4) shows that D^* is the set of turnpike controls.

That $f(n)$ converges geometrically to ϕ is a theme of discounted dynamic programming. For proofs, see Shapley [1953], Blackwell [1965], Denardo [1967], Veinott [1969, 1971] , and Schweitzer and Federgruen [1978, 1979]. The aforementioned proofs rely, explicitly or implicitly, on monotone contractions. For the risk-averse case, we were unable to obtain a contraction from Hypothesis RA. Our search for a proof for the risk-averse case led us to the linear-algebraic result in Theorem 2, which is presented in section 3. By contrast, for the risk-seeking case, Hypothesis RS allows for the construction of a monotone contraction using an idea of Alan Hoffman that was reported by Veinott [1969] and is reviewed in section 4.

Theorem 1 is reminiscent of work by Blackwell [1967] and Strauch [1966] on positive and negative dynamic programming, respectively. In the risk-averse case, the rewards (the R_i^k) are nonpositive, so the one-stage optimization operation A satisfies $A(0) < 0$, as it does in negative dynamic programming. Similarly, in the risk-seeking case, the rewards are nonnegative, and A satisfies $A(0) > 0$, as it does in positive dynamic programming. There is, however, a key difference. In our model, the terminal reward is $-\alpha e$, not 0; for us, $f(n) = A^n(-\alpha e)$. We require neither $A(-\alpha e) \geq -\alpha e$ nor $A(-\alpha e) \leq -\alpha e$, so the methodologies in Blackwell and Strauch do not seem to be helpful.

Section 2 reviews the material in DR [1979a] about stationary controls that will be needed here. Section 3 develops a test for Condition I. Section 4 uses the results in sections 2 and 3 to prove Theorem 1, along with ancillary results.

2. Stationary controls. When comparing vectors or matrices, $\geq$ refers to elementwise weak inequality, $\gg$ refers to elementwise strict inequality, and $>$ refers to $\geq$ together with $\neq$. The vector x is said to be *semipositive* if $x > 0$. Also, e denotes the vector all of whose elements are 1 (always of appropriate size).

If B is a $K \times K$ nonnegative matrix, the classic *Perron–Frobenius theorem* shows that the spectral radius $\rho(B)$ of B is an eigenvalue of B with semipositive right and left eigenvectors; furthermore, these eigenvectors are strictly positive if B is *irreducible*, that is, if $(\sum_{n=0}^{K} B^n)_{uv} > 0$ for $u, v = 1, 2, \ldots, K$. Condition I can be sharpened to the following equivalent form.

Condition I′. $\rho[(Q^\pi)_{JJ}] > 1$ for each control π and set J such that J is closed under π and $(Q^\pi)_{JJ}$ is irreducible.

To show that Condition I′ suffices, we note that if control π has a closed set J, then it has a closed set $\tilde{J} \subseteq J$ with $(Q^\pi)_{\tilde{J}\tilde{J}}$ irreducible and $\rho[(Q^\pi)_{JJ}] \geq \rho[(Q^\pi)_{\tilde{J}\tilde{J}}]$. Thus, if the latter exceeds 1, so must the former.

Designate as f the $N \times 1$ vector whose ith entry f_i is given by

$$f_i = \max_{\delta \in D} \left\{ \limsup_{n \to \infty} v^\delta(n)_i \right\} \qquad \forall\, i. \tag{12}$$

From $v^\delta(n) \leq f(n)$ and equation (5), it is evident that $f \leq \phi$. Also, for each control δ whose rate matrix Q^δ is transient, the $N \times 1$ vector v^δ is now defined by

$$v^\delta = (1 - Q^\delta)^{-1} R^\delta \qquad \text{if } Q^\delta \text{ is transient.} \tag{13}$$

Letting n approach infinity in (7) and using (9) gives

$$\lim_{n \to \infty} v^\delta(n) = (1 - Q^\delta)^{-1} R^\delta = v^\delta \qquad \text{if } Q^\delta \text{ is transient.} \tag{14}$$

The $N \times 1$ vector v^* is now defined by

$$v_i^* = \max \{[v^\delta]_i : Q^\delta \text{ is transient}\} \qquad \forall i. \tag{15}$$

Interpret v_i^* as the largest infinite-horizon expected utility obtainable by starting in state i and using any (stationary) control δ whose rate matrix Q^δ is transient. A central role is played by the following linear program.

Program 1. Minimize $\sum_i v_i$, subject to the constraints

$$v_i \geq R_i^k + \sum_u Q_{iu}^k v_u \qquad \forall i, k. \tag{16}$$

The use of linear programs to solve Markov decision problems has a long history. We mention the papers that relate most directly to Program 1. D'Epenoux [1960] showed that the correspondent of Program 1 finds the optimal control for a Markov decision process with a linear utility function and discounting. Denardo [1967] generalized this to discounted dynamic programs. Veinott [1968] extended the applicability of Program 1 to identify optimal solutions to Leontief substitution systems. In the context of Markov decision chains with exponential utility, Program 1 is due to DR [1979a], where Proposition 1 is established.

PROPOSITION 1. *Suppose $\lambda > 0$ and at least one control δ has Q^δ transient or that $\lambda < 0$ and every control δ has Q^δ transient. Then Program* 1 *is feasible and bounded, and its unique optimal solution is v^*. Moreover, v^* is a solution to the equation*

$$v_i = \max_k \left\{ R_i^k + \sum_u Q_{iu}^k v_u \right\} \qquad \forall i. \tag{17}$$

Proof. This theorem is an amalgam of Theorems 1 through 3 of DR [1979a]. □

The sets D_i^* and D^* are now defined in terms of v^* by

$$D_i^* = \left\{ k \in D_i : v_i^* = R_i^k + \sum_u Q_{iu}^k v_u^* \right\} \qquad \forall i, \tag{18}$$

$$D^* = \prod_{i=1}^N D_i^*. \tag{19}$$

Proposition 1 shows that v^* satisfies (17), hence that D^* is nonempty. In section 4, Hypotheses RA and RS are shown to imply that D^* is the set of turnpike controls.

For the risk-averse case, the hypothesis of Proposition 1 excludes Condition I, and we pause to indicate what Proposition 1 fails to demonstrate without Condition I. First, it does not conclude that $v^* = f$ because it does not compare v^* with the infinite-horizon utility of any control π whose rate matrix fails to be transient. Second, while it does show that v^* is a solution to (functional equation) (17), it does not guarantee that v^* is the unique solution to this functional equation. Third, while it shows that at least one control π in D^* has Q^π transient and $v^\pi = v^*$, it does not conclude that every control π in D^* has these properties. The following example demonstrates that none of these results can be obtained without strengthening the hypothesis of Proposition 1.

Example 1. There are two states. State 1 has a single decision. State 2 has two decisions, which are a and b. There are only two controls, δ and π, with $\delta(2) = a$ and $\pi(2) = b$. Their rate matrices and reward vectors are

$$(20) \quad Q^\delta = \begin{bmatrix} 1/2 & 0 \\ 2/3 & 0 \end{bmatrix}, R^\delta = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad Q^\pi = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}, \quad R^\pi = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

The matrix Q^δ is transient and (14) indicates that $v^\delta = [-2, -7/3]^T$. (The superscript T denoting transpose.) By contrast, Q^π is not transient (it has 1 as an eigenvalue), and (7) indicates that $v^\pi(n) \to [-2, 0]^T + [0, -1]^T = [-2, -1]^T$ as n becomes large. Evidently, $v^* = [-2, -7/3]^T < f = [-2, -1]^T$. It is easy to check that both v^* and f satisfy functional (17), and moreover, that $D_2^* = \{a, b\}$. For the risk-averse case (in which $\lambda > 0$), it is easy to construct a model whose rates and rewards satisfy (20). □

The risk-averse case. Example 1 violates Condition I because the set $J = \{2\}$ is closed under control π and $\rho[(Q^\pi)_{JJ}] = 1$. For the risk-averse case, Condition I rules out Example 1 and others like it.

DR [1979a] provided a test for Condition I. To review it, we define the sets $S^*(0)$ through $S^*(N)$ from this recursion: $S^*(0) = \{1, 2, \ldots, N\}$ and

$$(21) \qquad S^*(n+1) = \left\{ i \in S^*(n) : 1 = \max_{k \in D_i^*} \sum_{u \in S^*(n)} P_{iu}^k \right\} \qquad \forall\, n \geq 0.$$

Evidently, $S^*(N)$ consists of those states from which it is possible to survive with probability 1 for N transitions using only the decisions in D_1^* through D_N^* that are given in terms of v^* by (18). Consider the following proposition.

PROPOSITION 2. *Suppose $\lambda > 0$ and that at least one control δ has Q^δ transient. Then Condition* I *is satisfied if and only if the set $S^*(N)$ is empty. Furthermore, if Condition* I *is satisfied, then*

(a) $v^* = f \ll 0$.
(b) *f is the unique solution to* (17) *and $v \ll 0$.*
(c) *The set D^* is nonempty, each control π in D^* has Q^π transient, and* $[\pi \in D^*] \Leftrightarrow [f = \lim_{n\to\infty} \{v^\pi(n)\}]$.

Proof. This is an amalgam of Theorems 4 and 5 of DR [1979a]. □

Let us summarize the risk-averse case. Proposition 1 shows that solving Program 1 determined v^*, which makes it easy to compute D_1^* through D_N^* from (18) and then $S^*(0)$ through $S^*(N)$ from (21). If $S^*(N)$ is empty, then Proposition 2 shows that Condition I is satisfied, that $v^* = f \ll 0$, that f is the unique strictly negative solution to (17), and that every control in D^* has $v^\pi = f$. Since f is finite and negative, f is the expected utility of an income stream whose total is bounded, and maximizing expected utility is an appropriate criterion.

The risk-seeking case. For the risk-seeking case, the analogue of Proposition 2 is the following.

PROPOSITION 3. *Suppose $\lambda < 0$ and that every control δ has Q^δ transient. Then*
(a) $v^* = f \geq 0$.
(b) *f is the unique solution to* (17).
(c) *The set D^* is nonempty, and* $[\pi \in D^*] \Leftrightarrow [f = \lim_{n\to\infty} \{v^\pi(n)\}]$.

Proof. This is an amalgam of Theorems 6 and 7 of DR [1979a]. □

The conclusions of Proposition 3 are similar to those for the risk-averse case. One difference is that, for the risk-seeking case, it can occur that some state i has $f_i = 0$. Should $f_i = 0$, it is the expected utility of an income stream whose total approaches $-\infty$ with positive probability, in which case maximizing expected utility would be insensitive to the rate at which costs accrue.

To complete the discussion, we note how Program 1 provides tests for transient matrices. Theorem 3 of DR [1979a] shows that Program 1 is bounded if and only if there exists a control δ for which Q^δ is transient. Theorem 6 of DR [1979a] shows that every control δ has matrix Q^δ that is transient if and only if the variant of Program 1 in which 1 replaces each R_i^k is feasible and bounded.

3. A second test for Condition I. To prove Theorem 1 for the risk-averse case, we will employ a different test for Condition I. Condition I will be shown to hold if and only if a particular set of linear inequalities has a solution. These inequalities will be described in terms of the original decision sets D_1 through D_N, rather than D_1^* through D_N^*.

To prepare, we specify the sets $S(0)$ through $S(N)$ by the recursion: $S(0) = \{1, 2, \ldots, N\}$ and

$$S(n+1) = \left\{ i \in S(n) : 1 = \max_{k \in D_i} \sum_{u \in S(n)} P_{iu}^k \right\}. \tag{22}$$

Evidently, $S(n)$ consists of those states from which it is possible to survive with probability 1 for n transitions. We recall that N is the number of states. The sets C, E_i, and E are now defined in terms of $S(N)$ by

$$C = S(N), \tag{23}$$

$$E_i = \left\{ k \in D_i : 1 = \sum_{u \in C} P_{iu}^k \right\} \qquad \forall\, i \in C, \tag{24}$$

$$E = \prod_{i \in C} E_i. \tag{25}$$

Each element π of E is a *control* for the restricted Markov decision problem whose state space is C and whose decision set for each state j in C is E_j. To *use* a control π in E is to begin at some state i in C and, at each epoch at which state j is observed, select decision $\pi(j)$. The set E contains those controls whose use avoids termination. For each control π in E, the symbol Q^π denotes the $|C| \times |C|$ rate matrix whose rows and columns appear in their natural order, e.g., if state 6 is the lowest-numbered state in C, then the top row of Q^π contains the entries $Q_{6u}^{\pi(6)}$ for $u \in C$, listed in their natural order. We begin with the following lemma.

LEMMA 1. *Suppose $S(N) = C$ is nonempty. Then the following are equivalent:*

(a) *There exists a control π in E and a row vector x such that*

$$x \geq 0 \quad \text{and} \quad x\,Q^\pi < x. \tag{26}$$

(b) *There exists a solution to*

$$\sum_{i \in C} \sum_{k \in E_i} x_i^k Q_{iu}^k - \sum_{k \in E_u} x_u^k + \alpha_u = 0 \qquad \forall\, u \in C\ , \tag{27}$$

$$\sum_{i\in C} \alpha_i = 1 \ , \tag{28}$$

$$x_i^k \geq 0 \ \text{ and } \ \alpha_i \geq 0 \qquad \forall\, i \in C, \quad \forall\, k \in E_i. \tag{29}$$

Proof. First, consider a control π in E and a row vector x that satisfy (26). Rewrite $x\,Q^\pi < x$ as $x\,Q^\pi + \alpha = x$ with $\alpha > 0$, and then normalize the variables x and α, if necessary, so that $\alpha\, e = 1$. Finally, for each i in C, set $x_i^k = 0$ for each $k \neq \pi(i)$ and set $x_i^k = x_i$ for $k = \pi(i)$ to obtain a solution to (27)–(29), thereby proving that (a) implies (b).

To see that (b) implies (a), consider a solution to (27)–(29). Since this system has a solution, it has a solution $(\bar{x},\, \bar{\alpha})$ with minimal support, that is, no other solution has a smaller set of positive variables. For this minimal-support solution, set

$$J = \left\{ u \in C : 0 < \sum_{k\in E_u} \bar{x}_u^k \right\}. \tag{30}$$

All data and decision variables in (27)–(29) are nonnegative. Thus, from (27), we see that each state u in $C \setminus J$ has $\bar{\alpha}_u = 0$ and therefore, from (28), that J is nonempty. Further, by construction, $(\bar{x},\, \bar{\alpha})$ satisfies

$$\sum_{i\in J} \sum_{k\in E_i} \bar{x}_i^k Q_{iu}^k - \sum_{k\in E_u} \bar{x}_u^k + \bar{\alpha}_u = 0 \qquad \forall\, u \in J \ , \tag{31}$$

$$\sum_{i\in J} \bar{\alpha}_i = 1 \ , \tag{32}$$

$$\bar{x}_i^k \geq 0 \ \text{ and } \ \bar{\alpha}_i \geq 0 \qquad \forall\, i \in J, \quad \forall\, k \in E_i. \tag{33}$$

Also, since $(\bar{x},\, \bar{\alpha})$ satisfies (27)–(29), we see that $Q_{iu}^k = 0$ for each triplet (u, i, k) having $u \in C \setminus J$ and $\bar{x}_i^k > 0$. Consequently, each solution of (31)–(33) whose positive variables are a subset of the positive variables in $(\bar{x},\, \bar{\alpha})$ also satisfies (27)–(29). Hence, $(\bar{x},\, \bar{\alpha})$ is also a minimal support solution to system (31)–(33). This system contains $|J|+1$ linear equations. Thus, a standard result is that $(\bar{x},\, \bar{\alpha})$ contains at most $|J|+1$ positive variables.

By construction, for each i in J, at least one of the $\bar{x}_i^k$ is positive. Also, at least one of the $\bar{\alpha}_i$ is positive. We conclude that for each $i \in J$, we have $\bar{x}_i^k > 0$ for exactly one $k \in E_i$, and we have $\bar{\alpha}_i > 0$ for exactly one $i \in J$. These positive variables are now used to specify a control π and a vector x. For each $i \in J$, select the decision $k \in E_i$ having $\bar{x}_i^k > 0$ and set $\pi(i) = k$ and $x_i = \bar{x}_i^k$. For each $i \in C \setminus J$, set $\pi(i) = k$ for any $k \in E_i$ and set $x_i = 0$. These definitions combine with (31)–(33) to give $x\,Q^\pi - x < 0$ and $x > 0$, so that (26) is satisfied. Thus, (b) implies (a), which completes the proof. □

The main result of this section is the next theorem.

THEOREM 2. *Condition* I *holds if and only if there exists a vector w such that*

$$\sum_{u\in C} Q_{iu}^k w_u > w_i \qquad \forall\, i \in C, \ \forall\, k \in E_i, \tag{34}$$

$$w_i > 0 \qquad \forall\, i \in C. \tag{35}$$

Proof. In the case $S(N) = C = \emptyset$ (the empty set), Condition I holds vacuously, and (34)–(35) are vacuous, so the theorem is true. We focus on the remaining case, for which $S(N) = C$ is not empty.

To prove the "if" part, we suppose that w satisfies (34) and (35). Consider any control π for the original decision problem and any nonempty set J of states that is closed under control π. It suffices to prove that the spectral radius of $(Q^\pi)_{JJ}$ exceeds 1. Necessarily, and has $(Q^\pi)_{JJ}$ irreducible J is a subset of C. Further, since $Q^\pi_{ij} = 0$ for each pair (i, j) having $i \in J$ and $j \notin J$, we have that $\pi(i) \in E_i$ for each $i \in J$ and inequalities (35) and (34) give

$$0 \ll w_J \ll (Q^\pi w)_J = (Q^\pi)_{JJ}\, w_J. \tag{36}$$

The Perron–Frobenius theorem guarantees the existence of a row vector $x > 0$ such that $x(Q^\pi)_{JJ} = \mu x$, where μ equals the spectral radius of $(Q^\pi)_{JJ}$. Premultiply (36) by x to obtain $0 < x\, w_J < x(Q^\pi)_{JJ}\, w_J = \mu x\, w_J$, implying $\mu > 1$, which verifies Condition I. This concludes the "if" portion of the proof.

To prove the "only if" part, we assume that Condition I is satisfied. The first step in our proof will be to arrange to use Lemma 1: we assert that there can exist no vector x and control $\hat\pi$ whose restriction π to C is in E and satisfies (26). Suppose there existed such an x and $\hat\pi$. By hypothesis, $\hat\pi_i = \pi_i$ for each $i \in C$. Equations (24)–(25) guarantee that C is closed under control π and thus is closed under $\hat\pi$. Set $J = \{i \in C : x_i > 0\}$. Inequality (26) guarantees that J is nonempty. Applying (26) to $C \setminus J$ gives $0 = x_{C\setminus J} \geq (xQ^\pi)_J = x_J(Q^\pi)_{J,C\setminus J}$, guaranteeing $(Q^\pi)_{J,C\setminus J} = 0$ and, consequently, that J is closed under π. Since $\hat\pi_i = \pi_i$ for each $i \in J$, the set J is closed under $\hat\pi$. By construction, $(Q^\pi)_{JJ} = (Q^{\hat\pi})_{JJ}$. Set $\mu = \rho[(Q^\pi)_{JJ}] = \rho[(Q^{\hat\pi})_{JJ}]$. Condition I guarantees $\mu > 1$. The Perron–Frobenius theorem guarantees the existence of a semipositive right eigenvector v of $(Q^\pi)_{JJ}$ having $(Q^\pi)_{JJ}v = \mu v$. On the other hand, application of (26) to J gives

$$x_J \geq (xQ^\pi)_J = x_J(Q^\pi)_{JJ}, \tag{37}$$

and postmultiplying (37) by v produces $x_J v \geq x_J(Q^\pi)_{JJ}v = \mu x_J v$. Since $x_J \gg 0$ and $v > 0$, we have $x_J v > 0$, and hence that $\mu \leq 1$, which contradicts $\mu > 1$.

Thus, with states restricted to C and controls restricted to E, there can exist no control π and vector x that satisfy (26). Hence Lemma 1 shows that no solution can exist to (27), (28), and (29). Let us consider the linear program whose decision variables are x and α, whose objective is to maximize $\sum_{i\in C} \alpha_i$, and whose constraints are (27) and (29) but not (28). This linear program has 0 as its optimal value. The duality theorem of linear programming shows that its dual also has 0 as its optimal value; as the objective function of the dual program is identically 0, this means that a solution exists to

$$\sum_{u\in C} Q^k_{iu} w_u - w_i \geq 0 \qquad \forall\, i \in C, \quad \forall\, k \in E_i, \tag{38}$$

$$w_i \geq 1 \qquad \forall\, i \in C. \tag{39}$$

With ε as a small positive number, consider the perturbed model in which each datum Q^k_{iu} is replaced by $Q^k_{iu}/(1+\varepsilon)$. This perturbation preserves all eigenvectors and multiplies each eigenvalue by $1/(1+\varepsilon)$. For ε sufficiently close to zero, Condition I is preserved, so the argument in the prior paragraph applied to the model with perturbed data. Hence, there exists a vector w that satisfies (39) and

$$\sum_{u\in C} \frac{Q^k_{iu}}{1+\varepsilon}\, w_u \geq w_i \qquad \forall\, i \in C, \quad \forall\, k \in E_i. \tag{40}$$

Multiply (40) by $(1+\varepsilon)$ to see that w satisfies (34), completing the proof. □

The thrust of (34)–(35) is to provide a lower bound on the growth rate of any state for a system of matrices in product form. Upper bounds are classic, as are least upper bounds. Surprisingly, lower bounds seem to be novel; related results are developed in DR [2005].

4. The turnpike theorem and a corollary. We now state and prove a strengthened form of Theorem 1.

THEOREM 1a. *Suppose either that* $\lambda > 0$ *and Hypothesis* RA *is satisfied or that* $\lambda < 0$ *and Hypothesis* RS *is satisfied. Then the set* D^* *that is defined by* (19) *is nonempty and is the set of turnpike controls.*

The proof of Theorem 1a is parsed into a series of lemmas, for which some nomenclature is necessary. Each control δ has a rate matrix Q^δ that has been defined by (6). Similarly, each control δ has a *transition probability* matrix P^δ that is defined by

$$[P^\delta]_{iu} = P_{iu}^{\delta(i)} \qquad \forall\, i, \quad \forall\, u. \tag{41}$$

Note that P^δ is substochastic, not stochastic; its ith row sums to $1 - T_i^{\delta(i)}$.

Before treating the full class of (randomized, history-remembering) strategies, we focus on "transition-counting" controls, namely, on nonrandomized decision procedures whose decision when state u is observed can depend on the number n of periods that has elapsed since the planning began. Thus, each *transition-counting control* σ is a sequence $\sigma = (\sigma^1, \sigma^2, \ldots)$, where $\sigma^1, \sigma^2, \ldots$ are (stationary) controls and where action $\sigma^t(u)$ is taken if state u is observed at epoch t. With $\sigma = (\sigma^1, \sigma^2, \ldots)$ as any transition-counting control, we now designate $\sigma(t)$ by $\sigma(t) = (\sigma^1, \ldots, \sigma^t)$. For each transition-counting control σ, the matrices $P^{\sigma(t)}$ and $Q^{\sigma(t)}$ are defined by

$$P^{\sigma(t)} = P^{\sigma^1} P^{\sigma^2} \ldots P^{\sigma^t} \qquad \text{for } t = 1, 2, \ldots, \tag{42}$$

$$Q^{\sigma(t)} = Q^{\sigma^1} Q^{\sigma^2} \ldots Q^{\sigma^t} \qquad \text{for } t = 1, 2, \ldots \tag{43}$$

with $P^{\sigma(0)} = Q^{\sigma(0)} = I$. Evidently, $P^{\sigma(t)}$ and $Q^{\sigma(t)}$ are the transition probability matrix and the rate matrix that correspond to using transition-counting control σ for the first t epochs. For $n = 0, 1, \ldots$, designate as $v^{\sigma(n)}$ the $N \times 1$ vector whose ith entry equals the expected utility obtained by using transition-counting control σ for epochs 1 through n, given that state i is observed initially. We have this analogue of (7):

$$v^{\sigma(n)} = R^{\sigma^1} + Q^{\sigma(1)} R^{\sigma^2} + \cdots + Q^{\sigma(n-1)} R^{\sigma^n} - Q^{\sigma(n)} e\alpha. \tag{44}$$

The total (random) reward during the first n periods is bounded, and the exponential utility function guarantees

$$\alpha \left[v^{\sigma(n)} \right]_i < 0 \qquad \forall\, i. \tag{45}$$

Observations about the signs of certain addends in (44) appear in the next lemma.

LEMMA 2. *Consider a state* i *and a transition-counting control* σ *that satisfy* $\sum_u [P^{\sigma(n)}]_{iu} < 1$ *for some positive integer* n*, and let* s *be the smallest integer having* $\sum_u [P^{\sigma(s)}]_{iu} < 1$*. Then*

$$\alpha \sum_u \left[Q^{\sigma(s-1)} \right]_{iu} R_u^{\sigma^s(u)} < 0. \tag{46}$$

Proof. We recall that P_{iu}^k, Q_{iu}^k, and T_i^k are nonnegative numbers. It is immediate from (2) that

$$\left\{P_{iu}^k > 0\right\} \Leftrightarrow \left\{Q_{iu}^k > 0\right\} \quad \text{and} \quad \left\{T_i^k > 0\right\} \Leftrightarrow \left\{\alpha R_i^k < 0\right\}. \tag{47}$$

It follows that

$$\left\{\left[P^{\sigma(t)}\right]_{iu} > 0\right\} \Leftrightarrow \left\{\left[Q^{\sigma(t)}\right]_{iu} > 0\right\}. \tag{48}$$

Aiming to verify (46), note that the selection of s assures that $1 = \sum_v [P^{\sigma(s-1)}]_{iv}$ and that

$$\begin{aligned} 1 > \sum_u \left[P^{\sigma(s)}\right]_{iu} &= \sum_v \left\{\left[P^{\sigma(s-1)}\right]_{iv} \sum_u P_{vu}^{\sigma^s(v)}\right\} = \sum_v \left\{\left[P^{\sigma(s-1)}\right]_{iv} \left[1 - T_v^{\sigma^s(v)}\right]\right\} \\ &= 1 - \sum_v \left[P^{\sigma(s-1)}\right]_{iv} T_v^{\sigma^s(v)}, \end{aligned}$$

which gives $\sum_v [P^{\sigma(s-1)}]_{iv} T_v^{\sigma^s(v)} > 0$, which couples with (48) and (47) to verify (46). □

For the risk-averse case, inequality (45) shows that starting in any state i and using any transition-counting control σ for x periods earns expected utility $[v^{\sigma(x)}]_i$ that is negative. For the risk-averse case, our proof of Theorem 1 requires the *uniform* negative bound in the next lemma.

LEMMA 3. *Suppose $\lambda > 0$, and that Hypothesis* RA *is satisfied. Then there exists a positive number ε such that every transition-counting control σ has*

$$\left[v^{\sigma(x)}\right]_i \le -\varepsilon \qquad \forall\, i,\ \forall\, x \ge 1. \tag{49}$$

Proof. The proof begins by establishing two preliminary claims. □

Claim 1. For each positive integer n, there exists a negative number $A(n)$ such that

$$\left[v^{\sigma(x)}\right]_i \le A(n) \tag{50}$$

for each state i, each transition-counting control σ, and each integer x having $1 \le x \le n + N$.

Proof of Claim 1. Pick any positive integer n. There are finitely many triplets $\{i, x \text{ and } \sigma(n+N)\}$ for which i is a state, x is an integer satisfying $1 \le x \le n + N$, and $\sigma(n+N)$ is a sequence of $n + N$ controls. Inequality (45) implies that each such triplet has $[v^{\sigma(x)}]_i < 0$. Designate as $A(n)$ the largest (least negative) value of $[v^{\sigma(x)}]_i$ over all such triplets. Evidently, (50) is satisfied, which proves Claim 1. □

The next claim applies to states i and transition-counting controls σ whose first n controls $\sigma(n) = (\sigma^1, \ldots, \sigma^n)$ violate

$$\sum_{u \in C} \left[P^{\sigma(t)}\right]_{iu} = 1 \qquad \text{for } t = 0, 1, \ldots, n, \tag{51}$$

where C is (we recall) the set of states from which it is possible to survive with probability of 1 for N transions. Violations of (51) occur if $i \notin C$ or if $i \in C$ and there is a positive probability of "escape" from C within n transitions.

Claim 2. For each positive integer n there exists a negative number $B(n) \geq A(n)$ such that each state i and transition-counting control σ that violate (51) have

$$\left[v^{\sigma(x)}\right]_i \leq B(n) \qquad \forall\, x \geq 1. \tag{52}$$

Proof of Claim 2. Consider any state i and control sequence $\sigma(n)$ that violate (51). Thus, some integer $t \leq n$ satisfies $\sum_{u\in C}[P^{\sigma(t)}]_{iu} < 1$, which assures us that termination occurs with positive probability by epoch $t+N$. Designate as s the first (lowest-numbered) epoch at which termination occurs with positive probability when starting in state i and using control sequence $\sigma(n)$. We have seen that $s \leq N+n$. In addition, Lemma 2 and (44) give

$$0 > \left[Q^{\sigma(s-1)}R^{\sigma^s}\right]_i \geq \left[v^{\sigma(x)}\right]_i \qquad \forall\, x \geq N+n. \tag{53}$$

Over all trios i, $\sigma(n)$ and $s \leq N+n$ for which $[Q^{\sigma(s-1)}R^{\sigma^s}]_i$ is negative, let $C(n)$ be the largest (least negative) value of $[Q^{\sigma(s-1)}R^{\sigma^s}]_i$. Set $B(n) = \max\{A(n), C(n)\}$. Claim 1 and (53) verify (52), which completes the proof of Claim 2.

We are now prepared to prove that (49) holds. If C is empty, Claim 2 with $n=1$ ensures that (52) applies to each state i and transition-counting control σ, so (49) holds with $\varepsilon = -B(1)$.

It remains to establish (49) for the case in which C is nonempty. Applying Theorem 2 to the model with state space C and action set E_i for each $i \in C$ shows that there exists a $|C| \times 1$ vector w and a positive number A such that

$$0 < w_i \leq 1 \qquad \forall\, i \in C, \tag{54}$$

$$w_i(1+A) \leq \sum_{u\in C} Q^k_{iu} w_u \qquad \forall\, i \in C,\ k \in E_i. \tag{55}$$

Having selected w and A that satisfy (54)–(55), we pick any positive integer n large enough that

$$1 \leq w_i(1+A)^n \qquad \forall\, i \in C. \tag{56}$$

With this value of n, we take $A(n)$ and $B(n)$ as negative numbers that satisfy Claims 1 and 2, respectively. We shall establish (49) with $\varepsilon = -B(n)$.

Aiming for a contradiction, consider a triplet (i, s, σ) with $[v^{\sigma(s)}]_i > B(n)$. Among such triplets, we take one that minimizes s. From Claim 1 and $B(n) \geq A(n)$, we have $s > n+N$. From Claim 2, we have that i, n, and $\sigma(n)$ satisfy (51), which guarantees $i \in C$. Equation (51) also ensures that if $u \in C$ and $t \leq n$ and $[P^{\sigma(t-1)}]_{iu} > 0$, then $u \in C$ and $\sigma^t(u) \in E_u$. On the other hand, if $u \in C$ and $t \leq n$ and $[P^{\sigma(t-1)}]_{iu} = 0$, then replacing $\sigma^t(u)$ by any decision in E_u preserves $[v^{\sigma(s)}]_i$. Thus, we can assume that i and $\sigma(n)$ satisfy (51), that $[v^{\sigma(s)}]_i > B(n)$, and, in addition, that

$$\sigma^t(u) \in E_u \qquad \forall\, t \leq n,\ \forall\, u \in C. \tag{57}$$

The set C is closed under every control π having $\pi(u) \in E_u$ for each $u \in C$. Hence, repeated application of (55), coupled with (56) and (54), shows that

$$1 \leq w_i(1+A)^n \leq \sum_{u\in C} \left[Q^{\sigma(n)}\right]_{iu} \qquad \forall\, i \in C. \tag{58}$$

Next, designate transition-counting control γ by $\gamma^x = \sigma^{n+x}$ for $x = 1, 2, \ldots$. From (51), (47), and (48), we have

$$\left[Q^{\sigma(x-1)} R^{\sigma^s}\right]_u = 0 \qquad \text{for } x = 1, \ldots, n,$$

and therefore from $s > n + N$ and (44),

$$\left[v^{\sigma(s)}\right]_i = \sum_{u \in C} \left[Q^{\sigma(n)}\right]_{iu} \left[v^{\gamma(s-n)}\right]_u. \tag{59}$$

The selection of s guarantees

$$\left[v^{\gamma(s-n)}\right]_u \le B(n) \qquad \forall\, u \in C. \tag{60}$$

By hypothesis, $0 > [v^{\sigma(s)}]_i > B(n)$, which combines with (59) and (60) and then (58) to establish the contradiction,

$$\begin{aligned} B(n) < \left[v^{\sigma(s)}\right]_i &= \sum_{u \in C} \left[Q^{\sigma(n)}\right]_{iu} \left[v^{\gamma(s-n)}\right]_u \\ &\le \sum_{u \in C} \left[Q^{\sigma(n)}\right]_{iu} B(n) \le B(n), \end{aligned}$$

which completes the proof. □

The next step in our proof is to study the $N \times 1$ vector ϕ that is defined by (5). Lemma 3 is central to the next lemma.

LEMMA 4. *Suppose $\lambda > 0$ and that Hypothesis* RA *is satisfied. Then,*

$$\phi \ll 0 \qquad \textit{and} \qquad f \le \phi \le \max_{\delta \in D} \left\{R^\delta + Q^\delta \phi\right\}. \tag{61}$$

Proof. Lemma 3 implies that some positive number ε has $f(n) \le -\varepsilon e$, so that $\phi \le -\varepsilon e \ll 0$ is immediate from (5), that $f \le \phi$ is immediate from (12) and (5), and that $\phi \le \max_{\delta \in D}\{R^\delta + Q^\delta \phi\}$ follows directly from the recursion in (3) and from the nonnegativity of the Q^k_{iu}. □

LEMMA 5. *Suppose $\lambda > 0$, and that Hypothesis* RA *is satisfied. Any vector ϕ that satisfies* (61) *has $\phi = f$.*

Proof. From (61), we have $\phi \ll 0$, as well as the existence of a control π such that

$$\phi \le R^\pi + Q^\pi \phi. \tag{62}$$

Set $\mu = \rho(Q^\pi)$, the spectral radius of Q^π. The Perron–Frobenius theorem guarantees the existence of a row vector $x > 0$ such that $x\mu = xQ^\pi$. Set $J = \{j : x_j > 0\}$. Note that J is nonempty, that $x_J \gg 0$, that $x_{J'} = 0$, that $x_J \mu = (xQ^\pi)_J = x_J (Q^\pi)_{JJ}$, and that $0 = x_{J'}\mu = (xQ^\pi)_{J'} = x_J (Q^\pi)_{JJ'}$, which shows that $(Q^\pi)_{JJ'} = 0$. Hence, from (62), we get $\phi_J \le (R^\pi)_J + (Q^\pi)_{JJ}\phi_J$. Premultiply by x_J to get

$$x_J \phi_J (1 - \mu) \le x_J (R^\pi)_J. \tag{63}$$

The hypothesis, $\phi \ll 0$, guarantees $x_J \phi_J < 0$. Since $x_J \gg 0$ and $R^\pi \le 0$, we have $x_J (R^\pi)_J \le 0$. Aiming for a contradiction, suppose $(R^\pi)_J = 0$. In this case, J is closed under control π, so Condition I guarantees $\mu = \rho(Q^\pi) \ge \rho[(Q^\pi)_{JJ}] > 1$, and

(63) establishes the desired contradiction. Thus, $(R^\pi)_J < 0$, ensuring $x_J(R^\pi)_J < 0$, and (63) guarantees $\mu < 1$. Thus, Q^π is transient, and premultiplying (62) by the nonnegative matrix $(I - Q^\pi)^{-1}$ verifies the first inequality in $\phi \leq (I - Q^\pi)^{-1}R^\pi = v^\pi \leq v^*$. Since Proposition 2 and Lemma 4 give $v^* = f \leq \phi$, we have obtained $\phi \leq v^* = f \leq \phi$, so equality holds throughout, completing the proof. □

For the risk-averse case, we are now poised to prove Theorem 1a, which was formulated at the beginning of this section.

Proof of Theorem 1a *under Hypothesis* RA. Lemmas 4 and 5 show that

$$f_i = \phi_i = \limsup_{n\to\infty} \{f(n)_i\} \qquad \forall\, i. \tag{64}$$

Proposition 2 shows that D^* is nonempty and that every control δ in D^* has Q^δ transient and $v^\delta = f$.

Pick any control δ in D^*. From $v^\delta(n) \leq f(n)$, from $v^\delta(n) \to v^\delta = f$, and from (64), we get

$$f_i = (v^\delta)_i = \lim_{n\to\infty} \left\{v^\delta(n)_i\right\} \leq \limsup_{n\to\infty}\{f(n)_i\} = \phi_i = f_i. \tag{65}$$

Thus, equality holds in (65), proving that every control δ in D^* is a turnpike control.

Since $f = \phi$, Proposition 2 implies that no control that is not in D^* can be a turnpike control. This completes the proof. □

The risk-seeking case. For the risk-seeking case, the hypothesis of Theorem 1 is that each control δ has rate matrix Q^δ that is transient. This facilitates an analysis that Veinott [1969] attributed to Alan Hoffman. The idea is so brief and so lovely that we review it. It is to rescale the rate matrices and the rewards in a way that preserves optimizers and converts the model to a discounted Markov decision process. The key lies in equation (66), below.

LEMMA 6. *Suppose* $\lambda < 0$ *and that Hypothesis* RS *is satisfied. Then* $\phi = f$.

Proof. By hypothesis, each control δ has rate matrix Q^δ that is transient. To verify that the equation

$$w_i = \max_k \left\{1 + \sum_j Q^k_{ij} w_j\right\} \qquad \forall\, i \tag{66}$$

has a unique solution, do any one of the following: adapt the control improvement method of Howard [1960], apply Theorem 4 of Denardo [1967], or use a result of Veinott [1969]. It is clear from (66) that $w_i \geq 1$ for each i. The w_i in (66) *rescale* the transition rates and rewards as is indicated by the hats in

$$\hat{Q}^k_{ij} = (1/w_i)Q^k_{ij}w_j, \qquad \hat{R}^k_i = (1/w_i)R^k_i \tag{67}$$

and *rescale* the utility for terminating at state j from $-\alpha$ to $-\hat{\alpha}$, with $\hat{\alpha} = \alpha/w_j$. Thus, the rescaled problem has a state-dependent terminal reward.

It is clear from (44) that the rescaled utility vector $\hat{v}^{\sigma(n)}$ for using transition-counting control σ for n periods satisfies

$$\left[\hat{v}^{\sigma(n)}\right]_i = (1/w_i)\left[v^{\sigma(n)}\right]_i \qquad \forall\, i, \tag{68}$$

so that optimizers for the rescaled problem are identical to those for the original problem. In particular, (68) gives $\hat{f}(n)_i = (1/w_i)f(n)_i$ and $\hat{f}_i = (1/w_i)f_i$ for each i.

Multiply (66) by $(1/w_i)$ and use (67) to obtain $\sum_j \hat{Q}_{ij}^k \leq 1 - 1/w_i$ for each i and k. With $c = \max_i\{1 - 1/w_i\}$, we conclude that

$$\sum_j \hat{Q}_{ij}^k \leq c < 1 \qquad \forall\, i, k. \tag{69}$$

Expression (69) shows that the rescaled problem is a discounted Markov decision process to which the standard theory (see Blackwell [1965] or Denardo [1967]) applies. In particular, $\hat{f}(n)$ converges geometrically to $\hat{f}$, so $f(n)$ converges geometrically to f, implying that $f = \phi$. □

Proof of Theorem 1a *under Hypothesis* RS. The argument of the proof of Theorem 1 under Hypothesis RA applies, with Lemma 6 replacing Lemmas 4 and 5 and with Proposition 3 replacing Proposition 2. □

Overtaking optimality. Now, let σ^* denote a strategy, i.e., a decision procedure in which the decision selected when state i is observed at a particular period can be randomized and can depend on the number n of periods that have elapsed and on the entire history of the process. Designate as $v^{\sigma^*(n)}$ as the $N \times 1$ vector whose ith entry equals the expected utility obtained by starting at state i and using σ^* for periods 1 through n. By a standard argument (cf. section 7 (symmetries) in Denardo [1967]),

$$[v^{\sigma^*(n)}]_i \leq f(n)_i. \tag{70}$$

A (stationary) control π is now said to be *overtaking optimal* if

$$\liminf_{n\to\infty}\{v^\pi(n)_i - v_i^{\sigma^*(n)}\} \geq 0 \qquad \forall\, \sigma,\ \forall\, i. \tag{71}$$

This optimality criterion was introduced by Von Weizsäcker [1965] and Gale [1967].

For the risk-averse case, we now adapt the proof of Theorem 1a to show that D^* is the set of overtaking optimal controls. Evidently, a control π for which Q^π is transient will be overtaking optimal if and only if

$$v^\pi \geq \limsup_{n\to\infty} v^{\sigma^*(n)} \qquad \forall\, \sigma. \tag{72}$$

We designate as g the $N \times 1$ vector whose entries are given by

$$g_i = \sup_{\sigma^*}\left\{\limsup_{n\to\infty}[v^{\sigma^*(n)}]_i\right\} \qquad \forall\, i. \tag{73}$$

In (73), the supremum is taken over all strategies. In the risk-averse case, g_i can be as low as $-\infty$ and as high as 0. Similarly, in the risk-seeking case, g_i lies between 0 and $+\infty$, inclusive. That there is no economic benefit to using a history-remembering control is the thrust of the following corollary.

COROLLARY 1. *Suppose either that* $\lambda > 0$ *and Hypothesis* RA *holds or that* $\lambda < 0$ *and Hypothesis* RS *holds. Then* $f = g$ *and* D^* *is the set of overtaking optimal controls.*

Proof. From (12) and (73), we have $f \leq g$. Next, using (70), we have that every strategy σ^* and state i has

$$\limsup_{n\to\infty}\left[v^{\sigma^*(n)}\right]_i \leq \limsup_{n\to\infty} f(n)_i = \phi_i = f_i, \tag{74}$$

the last from Lemmas 4 through 6. Inequality (74) and (73) give $g \leq f$. Thus, $g = f = \phi$. That D^* is the set of overtaking optimal controls now follows from the proof of Theorem 1. □

To see that there is a difference between turnpike controls and overtaking-optimal controls, consider the undiscounted Markov decision process. DR [1979b] shows that this model has an (average) overtaking-optimal control. But income earned at the end of the planning interval is not attenuated, so this model has no turnpike control, that is, no (stationary) control π having $v^\pi(n) - f(n) \to 0$.

REFERENCES

R. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

D. Blackwell, *Discounted dynamic programming*, Ann. Math. Statist., 36 (1965), pp. 226–235.

D. Blackwell, *Positive dynamic programming*, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, 1967, pp. 415–418.

R. Cavazos-Cadena and R. Montes-de-Oca, *The value iteration algorithm in risk-sensitive average Markov decision chains*, Math. Oper. Res., 28 (2003), pp. 752–776.

R. Cavazos-Cadena and D. Hernández-Hernández, *A characterization of the optimal risk-sensitive average cost in finite controlled Markov chains*, J. Appl. Probab., to appear.

E. V. Denardo, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165–177.

E. V. Denardo, H. Park, and U. G. Rothblum, *Risk-Sensitive and Risk-Neutral Multi-Armed Bandits*, submitted for publication, 2004.

E. V. Denardo and U. G. Rothblum, *Overtaking optimality for Markov decision chains*, Math. Oper. Res., 4 (1979), pp. 144–152.

E. V. Denardo and U. G. Rothblum, *Optimal stopping, exponential utility, and linear programming*, Math. Programming, 16 (1979a), pp. 228–244.

E. V. Denardo and U. G. Rothblum, *Totally expanding multiplicative systems*, Linear Algebra Appl., 406 (2005), pp. 142–158.

E. V. Denardo, U. G. Rothblum, and L. Van der Heyden, *Index policies for stochastic search in a forest with an application to R&D project management*, Math. Oper. Res., 29 (2004), pp. 162–181.

F. D'Epenoux, *Sur un probleme de production de stockage dans l'aleatoire*, Rev. Française Recherche Oper., 14 (1960), pp. 3–16.

G. B. Di Masi and L. Stettner, *Risk sensitive control of discrete-time Markov processes with infinite horizon*, SIAM J. Control Optim., 38 (2000), pp. 61–78.

W. H. Fleming and D. Hernández-Hernández, *Risk sensitive control of the finite state-machine on an infinite horizon* I, SIAM J. Control Optim., 35 (1997), pp. 1790–1810.

W. H. Fleming and D. Hernández-Hernández, *Risk sensitive control of the finite state-machine on an infinite Horizon* II, SIAM J. Control Optim., 37 (1999), pp. 1048–1069.

D. Gale, *On optimal development in a multi-sector economy*, Rev. Econ. Stud., 34 (1967), pp. 1–18.

O. Hernández Lerma, *Adaptive Control Processes*, Springer-Verlag, New York, 1989.

R. A. Howard, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.

R. A. Howard and J. E. Matheson, *Risk sensitive Markov decision processes*, Management Sci., 8 (1972), pp. 356–369.

D. H. Jacobson, *Optimal stochastic linear systems with exponential performance criteria and the relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1972), pp. 124–131.

D. H. Jacobson, *Extensions of Linear-Quadratic Control, Optimizaiton and Matrix Theory*, Academic Press, New York, 1977.

P. Mandl, *An Iterative method for maximizing the characteristic root of positive matrices*, Rev. Roumaine Math. Pures Appl., 12 (1967), pp. 1317–1322.

P. Mandl and E. Seneta, *The theory of nonnegative matrices in a dynamic programming problem*, Austral. J. Statist., 11 (1969), pp. 85–96.

U. G. Rothblum, *Multiplicative Markov decision chains*, Math. Oper. Res., 9 (1984), pp. 6–24.

U. G. Rothblum and A. F. Veinott, Jr., *Markov Population Decision Chains*, to appear.

U. G. Rothblum and P. Whittle, *Growth optimality for branching Markov decision chains*, Math. Oper. Res., 7 (1982), pp. 582–601.

P. Schweitzer and A. Federgruen, *The asymptotic behavior of undiscounted value iteration in Markov decision problems*, Math. Oper. Res., 2 (1978), pp. 360–382.
P. Schweitzer, and A. Federgruen, *Geometric convergence of value iteration in multichain Markov decision problems*, Adv. Appl. Probab., 11 (1979), pp. 188–217.
L. S. Shapley, *Stochastic games*, Proc. Nat. Acad. Sci. USA, 39 (1953), pp. 1905–1100.
K. Sladky, *On the dynamic programming recursion for multiplicative Markov chains with rewards*, Kybernetika, 10 (1980), pp. 350–367.
R. E. Strauch, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871–890.
A. F. Veinott, Jr., *Extreme points of Leontief substitution systems*, Linear Algebra Appl., 1 (1968), pp. 181–194.
A. F. Veinott, Jr., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40 (1969), pp. 1635–1660.
A. F. Veinott, Jr., *Class Notes in Dynamic Programming*, Department of Operations Research, Stanford University, Stanford, CA, 1971.
C. C. Von Weizsäcker, *Existence of optimal programmes of accumulation for an infinite time horizon*, Rev. Econ. Stud., 32 (1965), pp. 85–104.
P. Whittle, *Risk Sensitive Optimal Control*, John Wiley & Sons, New York, 1990.
W. H. M. Zijm, *Determination of the spectral radius of a Markov decision process*, Oper. Res. Verfahren, 37 (1980), pp. 487–501.
W. H. M. Zijm, *Asymptotic expansions for dynamic programming recursions with general nonnegative matrices*, J. Optim. Theory Appl., 54 (1987), pp. 157–191.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 432–456

ON THE EXISTENCE OF A KAZANTZIS–KRAVARIS/LUENBERGER OBSERVER*

VINCENT ANDRIEU† AND LAURENT PRALY‡

Abstract. We state sufficient conditions for the existence, on a given open set, of the extension, to nonlinear systems, of the Luenberger observer as it has been proposed by Kazantzis and Kravaris. We prove it is sufficient to choose the dimension of the system, giving the observer, less than or equal to 2 + twice the dimension of the state to be observed. We show that it is sufficient to know only an approximation of the solution of a PDE, needed for the implementation. We establish a link with high gain observers. Finally we extend our results to systems satisfying an unboundedness observability property.

Key words. nonlinear osbservers, Luenberger observers, high gain observers

AMS subject classifications. 93B07, 93B30, 93C10

DOI. 10.1137/040617066

1. Introduction.

We consider the system

$$\dot{x} = f(x), \qquad y = h(x) \tag{1}$$

with state x in $\mathbb{R}^n$ and output y in $\mathbb{R}^p$ and where the functions f and h are sufficiently smooth. We are concerned with the problem of existence of an observer for x from the measurement y.

In a seminal paper [13], Kazantzis and Kravaris have proposed to extend to the nonlinear case the primary observer introduced by Luenberger in [17] for linear systems. Following this suggestion, the estimate $\widehat{x}$ of x is obtained as the output of the dynamical system

$$\dot{z} = A\,z + B(y), \qquad \widehat{x} = T^*(z), \tag{2}$$

with state z (a complex matrix) in $\mathbb{C}^{m\times p}$ and where A is a Hurwitz complex matrix and B and T^* are sufficiently smooth functions.

In the following we state sufficient conditions on f and h such that we can find (A, B, m) for which there exists T^* guaranteeing the convergence of $\widehat{x}$ to x.

To ease readability, we have divided the paper into two parts. In the first part, we introduce and state our main results, which are proved in the second part. Our first result gives a sufficient condition on f, h, A, and B implying the existence of T^* providing an appropriate observer. This condition involves a partial differential equation whose solution should be injective. In our second result, we propose a set of assumptions guaranteeing the existence of a solution for this equation. Our third and fourth results give two sufficient conditions implying the injectivity property of this solution. Our fifth result shows that an observer can already be obtained if we know only an appropriate approximation of this solution. This latter result allows

*Received by the editors October 15, 2004; accepted for publication (in revised form) August 1, 2005; published electronically, April 7, 2006.
http://www.siam.org/journals/sicon/45-2/61706.html.
†DPRS, ONERA, BP 72, 29 Ave de la Division Leclerc, 92322 Chatillon, France (vincent.andrieu@gmail.com).
‡Centre Automatique et Systèmes, École des Mines de Paris, 35 Rue Saint Honoré, 77305 Fontainebleau, France (Laurent.Praly@ensmp.fr).

us to propose a new insight in the standard high gain observer. Finally we claim that all these statements can be extended to the case where the system satisfies an unboundedness observability property.

Some notations. We assume the functions f and h in (1) are at least locally Lipschitz. So, for each x in $\mathbb{R}^n$, there exists a unique solution $X(x,t)$ to (1), with x as initial condition.

Given an open set $\mathcal{O}$ of $\mathbb{R}^n$, for each x in $\mathcal{O}$, we denote by $(\sigma^-_{\mathcal{O}}(x), \sigma^+_{\mathcal{O}}(x))$ the maximal interval of definition of the solution $X(x,t)$ conditioned to take values in $\mathcal{O}$.

For a set S, we denote by $\mathtt{cl}(S)$ its closure and by $S+\delta$ the open set

$$S+\delta \;=\; \{x \in \mathbb{R}^n : \exists\, \varkappa \in S :\; |x-\varkappa| < \delta\} = \bigcup_{x\in S} \mathcal{B}_\delta(x),$$

where $\mathcal{B}_\delta(x)$ denotes the open ball with center x and radius δ.

By $L_f V$ we denote the Lie derivative of V when it makes sense, i.e.,

$$L_f V(x) \;=\; \lim_{h\to 0} \frac{V(X(x,h)) - V(x)}{h}.$$

Finally, B_{1m} denotes the following vector in $\mathbb{R}^m$:

$$B_{1m} \;=\; (\;1 \quad \dots \quad 1\;)^T. \tag{3}$$

2. Results and comments.

2.1. Existence of a Kazantzis–Kravaris/Luenberger observer. In [13], m, the row dimension of z, is chosen equal to n, the dimension of x, and T^* is the inverse T^{-1} of a function T, solution of the following partial differential equation:

$$\frac{\partial T}{\partial x}(x)\, f(x) \;=\; A\, T(x) \;+\; B(h(x)). \tag{4}$$

The rationale for this equation, as emphasized more in [15] (see also [18]), is that, if T is a diffeomorphism satisfying (4), then the change of coordinates

$$\zeta \;=\; T(x) \tag{5}$$

allows us to rewrite the dynamics (1) equivalently as

$$\dot\zeta \;=\; A\,\zeta \;+\; B(h(T^{-1}(\zeta))), \qquad y \;=\; h(T^{-1}(\zeta)).$$

We then have

$$\dot{\widehat{z-\zeta}} \;=\; A\,(z-\zeta).$$

A being Hurwitz, z in (2) is the state of an asymptotically convergent observer of $\zeta = T(x)$. Then, if the function $T^* = T^{-1}$ is uniformly continuous, $\widehat{x} = T^*(z)$ is an asymptotically convergent observer of $x = T^*(\zeta) = T^*(T(x))$.

This way of finding the function T^* has motivated active research on the problem of the existence of an analytic and invertible solution to (4) (see [13, 15], for instance). But it turns out that having a (weak) solution to (4) which is only continuous and uniformly injective is already sufficient. By sufficiency we mean, here, that an observer is appropriate if we have convergence to zero of the observation error associated to any solution which remains in a given open set $\mathcal{O}$. For the latter, we need the following.

DEFINITION 1 (completeness within $\mathcal{O}$). *The system* (1) *is forward (resp., backward) complete within* $\mathcal{O}$ *if we have the implication, for each* x *in* $\mathcal{O}$,

$$\sigma^+_{\mathcal{O}}(x) \ < \ +\infty \qquad \Longrightarrow \qquad \sigma^+_{\mathcal{O}}(x) \ < \ \sigma^+_{\mathbb{R}^n}(x). \tag{6}$$

In other words, completeness within $\mathcal{O}$ says that any solution $X(x,t)$ which exits $\mathcal{O}$ in finite time must cross the boundary of $\mathcal{O}$ (at a finite distance). An usual case where this property holds is when f has an at most linear growth within $\mathcal{O}$.

THEOREM 1 (sufficient condition of existence of an observer). *Assume the system* (1) *is forward complete within* $\mathcal{O}$ *and there exist an integer* m, *a Hurwitz complex* $m \times m$ *matrix* A, *and functions* $T : \mathtt{cl}(\mathcal{O}) \to \mathbb{C}^{m\times p}$, *continuous,* $B : \mathbb{R}^p \to \mathbb{C}^{m\times p}$, *continuous, and* ρ, *of class* $\mathcal{K}_\infty$, *satisfying*

$$L_f T(x) = AT(x) + B(h(x)) \qquad \forall x \in \mathcal{O}, \tag{7}$$

$$|x_1 - x_2| \le \rho(|T(x_1) - T(x_2)|) \qquad \forall (x_1, x_2) \in \mathtt{cl}(\mathcal{O})^2. \tag{8}$$

Under these conditions, there exists a continuous function $T^* : \mathbb{C}^{m\times p} \to \mathtt{cl}(\mathcal{O})$ *such that for each* x *in* $\mathcal{O}$ *and* z *in* $\mathbb{C}^{m\times p}$ *the (unique) solution* $(X(x,t), Z(x,z,t))$ *of*

$$\dot{x} \ = \ f(x), \qquad \dot{z} \ = \ Az + B(h(x)) \tag{9}$$

is right maximally defined on $[0, \sigma^+_{\mathbb{R}^n}(x))$. *Moreover, we have the implication*

$$\sigma^+_{\mathcal{O}}(x) \ = \ \sigma^+_{\mathbb{R}^n}(x) \qquad \Longrightarrow \qquad \lim_{t\to\sigma^+_{\mathbb{R}^n}(x)} |T^*(Z(x,z,t)) - X(x,t)| \ = \ 0. \tag{10}$$

Remark 1. 1. With the forward completeness within $\mathcal{O}$ (6), the condition on the left in (10) implies that the solution $X(x,t)$ never exits $\mathcal{O}$ and so

$$\sigma^+_{\mathcal{O}}(x) \ = \ \sigma^+_{\mathbb{R}^n}(x) \ = \ +\infty. \tag{11}$$

2. Theorem 1 extends readily to the case where (a) y is a scalar, (b) the state x can be decomposed in $x = (\xi_1, \xi_2)$ and satisfies

$$\dot{\xi}_1 \ = \ f_1(\xi_1, u) \ + \ h(\xi_2), \qquad \dot{\xi}_2 \ = \ f_2(\xi_2), \qquad y \ = \ \xi_1,$$

and (c) the function B can be chosen linear. In this case the observer is implemented as the reduced order observer:

$$\dot{\widehat{z - By}} \ = \ Az \ + \ Bf_1(y,u), \qquad \widehat{\xi}_2 \ = \ T^*(z).$$

Assuming we have a continuous function T satisfying (7), to implement the observer, we have to find a function T^* satisfying

$$|T^*(z) - x| \ \le \ \rho^*(|z - T(x)|) \qquad \forall (x,z) \in \mathcal{O} \times \mathbb{C}^{m\times p}$$

for some function ρ^* of class $\mathcal{K}_\infty$. As shown by Kreisselmeier and Engel in [14], such a function T^* exists if T is continuous and uniformly injective as prescribed by (8).

In conclusion, a Kazantzis–Kravaris/Luenberger observer exists mainly if we can find a continuous function T solving (7) and uniformly injective in the sense of (8).

2.2. Existence of T solving (7). To exhibit conditions guaranteeing the existence of a function T solution of (7), we abandon the interpretation above of a change of coordinates (see (5)) and come back to the original idea in [17] (see also [13] and [4]) of dynamic extension. Namely, we consider the augmented system (9). Because of its triangular structure and the fact that A is Hurwitz, we may expect this system to have, at least maybe only locally, an exponentially attractive invariant manifold in the augmented (x, z) space which could even be described as the graph of a function as

$$\left\{(x,z) \in \mathbb{R}^n \times \mathbb{C}^{m\times p} \,:\, z = T(x)\right\}.$$

In this case, the function T would satisfy the following identity, for all t in the domain of definition of the solution $(X(x,t), Z((x,z),t))$ of (9) issued from (x,z) (compare with [18, Definition 5]),

$$T(X(x,t)) \;=\; Z((x,T(x)),t),$$

or equivalently,

$$T(X(x,t)) \;=\; \exp(At)\,T(x) \;+\; \int_0^t \exp(As)B(h(X(x,s)))ds. \tag{12}$$

From this identity, (7) is obtained by derivation with respect to t. But since we need (7) to hold only on $\mathcal{O}$, from (12), it is sufficient that T satisfies

$$T(x) \;=\; \exp(-At)\,T(\breve{X}(x,t)) \;-\; \int_0^t \exp(-As)B(h(\breve{X}(x,s)))ds,$$

where $\breve{X}(x,s)$ is a solution of the modified system

$$\dot{x} \;=\; \breve{f}(x) \;=\; \chi(x)\,f(x), \tag{13}$$

where $\chi : \mathbb{R}^n \to \mathbb{R}$ is an arbitrary locally Lipschitz function satisfying

$$\chi(x) \;=\; 1 \quad \text{if} \quad x \in \mathcal{O}, \qquad \chi(x) \;=\; 0 \quad \text{if} \quad x \notin \mathcal{O} + \delta_u, \tag{14}$$

for some positive real number δ_u. So, as standard in the literature on invariant manifolds, by letting t go to $-\infty$, we get the following candidate expression for T:

$$T(x) \;=\; \int_{-\infty}^0 \exp(-As)B(h(\breve{X}(x,s)))ds. \tag{15}$$

The above nonrigorous reasoning can be made correct as follows

THEOREM 2 (existence of T). *Assume the existence of a strictly positive real number δ_u such that the system* (1) *is backward complete within $\mathcal{O} + \delta_u$. Then, for each Hurwitz complex $m\times m$ matrix A, we can find a C^1 function $B : \mathbb{R}^p \to \mathbb{C}^{m\times p}$ such that the function $T : \mathtt{cl}(\mathcal{O}) \to \mathbb{C}^{m\times p}$, given by* (15)*, is continuous and satisfies* (7).

Remark 2. All that is needed here about the function B is that it guarantees that the function $t \mapsto |\exp(-At)B(h(\breve{X}(x,t)))|$ is exponentially decaying with t going to $-\infty$. So in particular (see Remark 6) when $\mathtt{cl}(\mathcal{O})$ is bounded, B can be chosen simply as a linear function.

Approaching the problem from another perspective, Kreisselmeier and Engel have introduced in [14] this same expression (15) (but with X instead of $\breve{X}$ and B the identity function). Another link between [13] and [14] has been established in [16].

2.3. T injective. Assuming now we have at our disposal the continuous function T, we need to make sure that it is injective, if not uniformly injective as specified by (8). Here is where observability enters the game. Following [17], in [13, 15], when $m = n$, observability of the first order approximation at an equilibrium together with an appropriate choice of A and B is shown to imply injectivity of the solution T of (4) in a neighborhood of this equilibrium. In [14], uniform injectivity of T is obtained under the following two assumptions:

1. The past output path $t \mapsto h(X(x,t))$ is uniformly injective in x with the set of past output paths equipped with an exponentially weighted L^2-norm.

2. The system (1) has finite complexity, i.e., there exists a (finite) number M of piecewise continuous function ϕ_i in $L^2(\mathbb{R}_-;\mathbb{R}^p)$ and a strictly positive real number δ such that we have, for each pair (x_1,x_2) in $\mathcal{O}^2$,

$$\sum_{i=1}^{M}\left[\int_{-\infty}^{0}\exp(-\ell s)\phi_i(s)^T[h(X(x_1,s))-h(X(x_2,s))]ds\right]^2 \geq \delta\int_{-\infty}^{0}\exp(-2\ell s)|h(X(x_1,s))-h(X(x_2,s))|^2ds.$$

Our next result states that with the only assumption that the past output path $t \mapsto h(X(x,t))$ is injective in x, it is sufficient to choose $m = n+1$ generic complex eigenvalues for A to get T injective. The specific injectivity condition we need is as follows.

DEFINITION 2 (backward $\mathcal{O}$-distinguishability). *There exist two strictly positive real numbers $\delta_\Upsilon < \delta_d$ such that for each pair of distinct points x_1 and x_2 in $\mathcal{O}+\delta_\Upsilon$, there exists a time t, in $(\max\{\sigma^-_{\mathcal{O}+\delta_d}(x_1),\sigma^-_{\mathcal{O}+\delta_d}(x_2)\},0]$, such that we have*

$$h(X(x_1,t)) \neq h(X(x_2,t)).$$

This distinguishability assumption says that the present state x can be distinguished from other states in $\mathcal{O}+\delta_\Upsilon$ by looking at the past output path restricted to the negative time interval where the solution $X(x,t)$ is in $\mathcal{O}+\delta_d$.

THEOREM 3 (injectivity). *Assume the system (1) is backward complete within $\mathcal{O}+\delta_u$ and backward $\mathcal{O}$-distinguishable with the corresponding δ_d in $(0,\delta_u)$. Assume also the existence of an injective C^1 function $b:\mathbb{R}^p\to\mathbb{C}^p$, a continuous function $M:\mathcal{O}+\delta_\Upsilon\to\mathbb{R}^+$, and a negative real number ℓ such that, for each x in $\mathcal{O}+\delta_\Upsilon$, the two functions $t\mapsto\exp(-\ell t)b(h(\breve{X}(x,t)))$ and $t\mapsto\exp(-\ell t)\frac{\partial b\circ h\circ\breve{X}}{\partial x}(x,t)$ satisfy, for each t in $(\breve{\sigma}^-_{\mathbb{R}^n}(x),0]$,*

$$|\exp(-\ell t)b(h(\breve{X}(x,t)))| + \left|\exp(-\ell t)\frac{\partial b\circ h\circ\breve{X}}{\partial x}(x,t)\right| \leq M(x), \tag{16}$$

where again $\breve{X}$ is a solution of (13), but this time with the function χ satisfying

$$\chi(x) = 1 \quad \text{if} \quad x\in\mathcal{O}+\delta_d, \qquad \chi(x) = 0 \quad \text{if} \quad x\notin\mathcal{O}+\delta_u. \tag{17}$$

Under these conditions, there exists a subset S of $\mathbb{C}^{n+1}$ of zero Lebesgue measure such that the function $T:\mathtt{cl}(\mathcal{O})\to\mathbb{C}^{(n+1)\times p}$ defined, with the notation (3), by

$$T(x) = \int_{-\infty}^{0}\exp(-As)B_{1m}\,b(h(\breve{X}(x,s)))ds \tag{18}$$

is injective provided A is a diagonal matrix with $n+1$ complex eigenvalues λ_i arbitrarily chosen in $\mathbb{C}^{n+1}\setminus S$ and with real part strictly smaller than ℓ.

Remark 3. 1. Condition (16) holds, for instance, if f, h, and b have bounded derivative on $\mathtt{cl}(\mathcal{O} + \delta_\Upsilon)$ (see [16]).

2. Theorem 3 gives injectivity, not uniform injectivity. As already mentioned, if $\mathcal{O}$ is bounded, continuity and the former imply the latter.

Following Theorem 3, for any generic choice of $n+1$ complex eigenvalues for the matrix A, the function T given by (18) (or, equivalently, (15)) is injective. This says that the (real) row dimension of z is $m = 2n+2$. It is a well-known fact in observer theory that it is generically sufficient to extract $m = 2n+1$ pieces of information from the output path (with h generically chosen) to observe a state of dimension n (see, for instance, [1, 23, 10, 7, 22]). It can be understood from the adage that the relation $T(x_1) = T(x_2)$ between the two states x_1 and x_2 in $\mathbb{R}^n$, i.e., for $2n$ unknowns, has generically the unique trivial solution $x_1 = x_2$ if we have strictly more than $2n$ equations, i.e., $T(x)$ has strictly more than $2n$ components.

2.4. Injectivity in the case of complete observability. Another setup where injectivity can be obtained is when we have complete observability. Namely, we can find a row dimension m and a function $b : y \in \mathbb{R}^p \mapsto b(y) = (b_1(y), \dots, b_p(y)) \in \mathbb{R}^p$ so that the following function $H : \mathbb{R}^n \to \mathbb{R}^{m\times p}$ is injective when restricted to $\mathtt{cl}(\mathcal{O})$:

$$H(x) = \begin{pmatrix} b_1(h(x)) & \dots & b_p(h(x)) \\ L_f b_1(h(x)) & \dots & L_f b_p(h(x)) \\ \vdots & \vdots & \vdots \\ L_f^{m-1} b_1(h(x)) & \dots & L_f^{m-1} b_p(h(x)) \end{pmatrix}. \tag{19}$$

Here $L_f^i h$ denotes the ith iterate Lie derivative, i.e., $L_f^{i+1} h = L_f(L_f^i h)$. Of course, for this to make sense, the functions b, f, and h must be sufficiently smooth. This setup has been popularized and studied in detail by Gauthier and his coworkers (see [11] and the references therein; see also [18]). In particular, it is established in [10] that when $p = 1$, for any generic pair (f, h) in (1), it is sufficient to pick $m = 2n+1$.

With a Taylor expansion of the output path at $t = 0$, we see that the injectivity of H implies that the function which associates the initial condition x to the output path, restricted to a very small time interval, is injective. This property is nicely exploited by observers with very fast dynamics as high gain observers (see [11]). Specifically, we have the next theorem.

THEOREM 4 (injectivity in the case of complete observability). *Assume the existence of a sufficiently smooth function $b : \mathbb{R}^p \to \mathbb{C}^p$ such that for the function H defined in* (19)*, there exist a positive real number L and a class $\mathcal{K}_\infty$ function ρ such that we have*

$$|L_f^m b(h(x_1) - L_f^m b(h(x_2))| \le L|H(x_1) - H(x_2)| \quad \forall (x_1, x_2) \in \mathtt{cl}(\mathcal{O})^2, \tag{20}$$

$$|x_1 - x_2| \le \rho(|H(x_1) - H(x_2)|) \quad \forall (x_1, x_2) \in \mathtt{cl}(\mathcal{O})^2. \tag{21}$$

Then, for any diagonal Hurwitz complex $m \times m$ matrix A, with m the row dimension of H, there exists a real number k^* such that for any k strictly larger than k^*, there exists a function $T : \mathtt{cl}(\mathcal{O}) \to \mathbb{C}^{m\times p}$ which is continuous, uniformly injective, and satisfies (see (3)*)*

$$L_f T(x) \;=\; kAT(x) + B_{1m} b(h(x)) \qquad \forall x \in \mathcal{O}. \tag{22}$$

2.5. Approximation. Fortunately for the implementation of the observer, knowing a function T satisfying (7) only approximately is sufficient. But, in this case, we have to modify the observer dynamics.

THEOREM 5 (approximation). *Assume the system* (1) *is forward complete within* $\mathcal{O}$. *Assume also the existence of an integer* m, *a Hurwitz complex* $m \times m$ *matrix* A, *and functions* $T_a : \mathtt{cl}(\mathcal{O}) \to \mathbb{C}^{m\times p}$, *continuous,* $B : \mathbb{R}^p \to \mathbb{C}^{m\times p}$, *continuous, and* ρ *of class* $\mathcal{K}_\infty$, *such that*

$$|x_1 - x_2| \;\le\; \rho(|T_a(x_1) - T_a(x_2)|) \qquad \forall (x_1, x_2) \in \mathtt{cl}(\mathcal{O})^2, \tag{23}$$

the function $L_f T_a$ *is well defined on* $\mathcal{O}$, *and the function* $\mathfrak{E} : \mathtt{cl}(\mathcal{O}) \to \mathbb{C}^{m\times p}$ *defined as*

$$\mathfrak{E}(x) = L_f T_a(x) - [AT_a(x) + B(h(x))] \qquad \forall x \in \mathcal{O} \tag{24}$$

satisfies

$$|\mathfrak{E}(x_1) - \mathfrak{E}(x_2)| \;\le\; N\,|T_a(x_1) - T_a(x_2)| \qquad \forall (x_1, x_2) \in \mathtt{cl}(\mathcal{O})^2, \tag{25}$$

where N *is a positive real number satisfying*

$$2N\,\lambda_{\max}(P) \;<\; 1 \tag{26}$$

with $\lambda_{\max}(P)$ *the largest eigenvalue of the Hermitian matrix* P *solution of*

$$\overline{A}^{\top} P + PA = -I. \tag{27}$$

Under these conditions, there exists a function $T_a^* : \mathbb{C}^{m\times p} \to \mathtt{cl}(\mathcal{O})$ *and a locally Lipschitz function* $\mathfrak{F} : \mathbb{C}^{m\times p} \to \mathbb{C}^{m\times p}$ *such that for each* x *in* $\mathcal{O}$ *and* z *in* $\mathbb{C}^{m\times p}$ *each solution* $(X(x,t), Z(x,z,t))$ *of*

$$\dot{x} \;=\; f(x), \qquad \dot{z} \;=\; Az + \mathfrak{F}(z) + B(h(x)) \tag{28}$$

is right maximally defined on $[0, \sigma^+_{\mathbb{R}^n}(x))$. *Moreover, we have the implication*

$$\sigma^+_{\mathcal{O}}(x) \;=\; \sigma^+_{\mathbb{R}^n}(x) \qquad \Longrightarrow \qquad \lim_{t\to\sigma^+_{\mathbb{R}^n}(x)} |T_a^*(Z(x,z,t)) - X(x,t)| \;=\; 0. \tag{29}$$

Remark 4. 1. In (24), $\mathfrak{E}$ represents the error in (7) given by the approximation T_a of T. This error should not be too large in an incremental sense as specified by (25) and (26). This indicates that one way to approximate T is to look for T_a in a set of functions minimizing the L^∞ norm on $\mathtt{cl}(\mathcal{O})$ of the gradient of the associated error $\mathfrak{E}$. In particular, in the case where $\mathcal{O}$ is bounded, it follows from the Weierstrass approximation theorem that we can always choose a Hurwitz complex matrix A and a linear function B so that the constraint (26) can be satisfied by restricting ourself to choose the function T_a as a polynomial in x.

2. The function $\mathfrak{F}$ in the observer (28) can be chosen as any Lipschitz extension of $\mathfrak{E} \circ T_a^*$ outside $T_a(\mathtt{cl}(\mathcal{O}))$. This is very similar to what is done in [18] where a constructive procedure for this extension is proposed. Fortunately, this Lipschitz extension is not needed in the case where the function $\mathfrak{E}$ satisfies

$$|\mathfrak{E}(x_1) - \mathfrak{E}(x_2)| \le \frac{N}{4}\,\rho^{-1}(|x_1 - x_2|) \quad \forall (x_1, x_2) \in \mathtt{cl}(\mathcal{O})^2,$$

where ρ is the function satisfying (23). In this case we take simply (see (67))

$$\mathfrak{F}(z) = \mathfrak{E}(T_a^*(z)) \quad \forall z \in \mathbb{C}^{m\times p}.$$

The combination of Theorems 4 and 5 gives us a new insight in the classical high gain observer of order m as studied in [10] or [18], for instance. Specifically, we have the following.

COROLLARY 1 (high gain observer). *Assume the system* (1) *is forward complete within* $\mathcal{O}$ *and there exist a sufficiently smooth function* $b : \mathbb{R}^p \to \mathbb{C}^p$, *a class* $\mathcal{K}_\infty$ *function* ρ, *and a positive real number* L *such that* (20) *and* (21) *hold with* H *and* m *given by* (19). *Under these conditions, for any diagonal Hurwitz complex* $m \times m$ *matrix* A, *there exists a real number* k^* *such that for any real number* k *strictly larger than* k^*, *there exist a function* $T_a^* : \mathbb{C}^{m\times p} \to \mathtt{cl}(\mathcal{O})$, *left inverse on* $T_a(\mathtt{cl}(\mathcal{O}))$ *of the function* $T_a : \mathtt{cl}(\mathcal{O}) \to \mathbb{C}^{m\times p}$ *defined as*

$$T_a(x) \;=\; -\sum_{i=1}^{m} (kA)^{-i} B_{1m} L_f^{i-1} b(h(x)) \tag{30}$$

and a function $\mathfrak{F} : \mathbb{R}^{m\times p} \to \mathbb{R}^{m\times p}$ *such that for each* x *in* $\mathcal{O}$ *and* z *in* $\mathbb{R}^{m\times p}$ *each solution* $(X(x,t), Z(x,z,t))$ *of*

$$\dot{x} \;=\; f(x), \qquad \dot{z} \;=\; kAz \;+\; \mathfrak{F}(z) \;+\; B_{1m}\, b(h(x)) \tag{31}$$

is right maximally defined on $[0, \sigma^+_{\mathbb{R}^n}(x))$. *Moreover, we have the implication*

$$\sigma^+_{\mathcal{O}}(x) \;=\; \sigma^+_{\mathbb{R}^n}(x) \qquad \Longrightarrow \qquad \lim_{t\to\sigma^+_{\mathbb{R}^n}(x)} |T_a^*(Z(x,z,t)) - X(x,t)| \;=\; 0. \tag{32}$$

Remark 5. When $\mathcal{O}$ is bounded and H is injective, uniform injectivity (21) and forward completeness within $\mathcal{O}$ hold necessarily. Thus, in this case, we recover [18, Lemma 1].

2.6. Extension to boundedness observability. Completeness is a severe restriction. Instead, it is proved in [4] that a necessary condition for the existence of an observer providing the convergence to zero of the observation error within the domain of definition of the solutions is the forward unboundedness observability property.

DEFINITION 3 (unboundedness observable within $\mathcal{O}$). *The system* (1) *is forward (resp., backward) unboundedness observable within* $\mathcal{O}$ *if there exists a proper and* C^1 *function* $V_{\mathfrak{f}} : \mathbb{R}^n \to \mathbb{R}_+$ *(resp.,* $V_{\mathfrak{b}} : \mathbb{R}^n \to \mathbb{R}_+$*) and a continuous function* $\gamma_{\mathfrak{f}} : \mathbb{R}^p \to \mathbb{R}_+$ *(resp.,* $\gamma_{\mathfrak{b}} : \mathbb{R}^p \to \mathbb{R}_+$*) such that*

$$L_f V_{\mathfrak{f}}(x) \le V_{\mathfrak{f}}(x) \;+\; \gamma_{\mathfrak{f}}(h(x)) \qquad \forall x \in \mathcal{O}, \tag{33}$$

$$(\text{ resp. } L_f V_{\mathfrak{b}}(x) \ge -V_{\mathfrak{b}}(x) \;-\; \gamma_{\mathfrak{b}}(h(x)) \qquad \forall x \in \mathcal{O}).$$

Fortunately, all our previous results still hold if completeness is replaced by unboundedness observability but provided[1]

1. the observer is modified in

$$\dot{z} \;=\; \gamma(y)\,[Az + B(y)]\,, \qquad \widehat{x} \;=\; T^*(z),$$

[1]The interested reader will find in [3] the precise statements of the corresponding results.

where γ is a C^1 function satisfying

$$\gamma(h(x)) \geq 1 + \gamma_{\mathfrak{f}}(h(x)) \quad \text{(resp., and } \gamma(h(x)) \geq 1 + \gamma_{\mathfrak{b}}(h(x))) \quad \forall x \in \mathtt{cl}(\mathcal{O}). \tag{34}$$

As suggested in [4], the introduction of γ takes care of possible finite escape time. This has nothing in common with the objective of error dynamics linearization as considered in [19].

2. In most occurrences, e.g., (7), (13), (19), (22), (24), ..., f is replaced by f_γ defined as

$$f_\gamma(x) \;=\; \frac{f(x)}{\gamma(h(x))}.$$

3. Proofs.

3.1. Proof of Theorem 1. Because of the triangular structure of the system (9), for each x in $\mathcal{O}$ and z in $\mathbb{C}^{m\times p}$, the component $Z(x,z,t)$ of the corresponding solution of this system is defined as long as $h(X(x,t))$ is defined. So this solution $(X(x,t), Z(x,z,t))$ is right maximally defined on the same interval $[0, \sigma^+_{\mathbb{R}^n}(x))$ as $X(x,t)$, solution of (1).

Let us now restrict our attention to points x in $\mathcal{O}$ satisfying the condition on the left in (10). In this case, with the forward completeness within $\mathcal{O}$, we have (11). On the other hand, from (7) and (9), we obtain, for each x in $\mathcal{O}$, z in $\mathbb{C}^{m\times p}$ and t in $[0, \sigma^+_{\mathbb{R}^n}(x))$,

$$T(X(x,t)) - Z(x,z,t) \;=\; \exp(At)\,(T(x) - z). \tag{35}$$

As A is a Hurwitz matrix, this and (11) yield

$$\lim_{t\to\sigma^+_{\mathbb{R}^n}(x)} |Z(x,z,t) - T(X(x,t))| \;=\; 0.$$

From this, the implication (10) follows readily if there exist a continuous function $T^* : \mathbb{C}^{m\times p} \to \mathtt{cl}(\mathcal{O})$ and a class $\mathcal{K}_\infty$ function $\rho^* : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying

$$|T^*(z) - x| \;\leq\; \rho^*(z - T(x)) \qquad \forall z \in \mathbb{C}^{m\times p}\,,\ \forall x \in \mathtt{cl}(\mathcal{O}). \tag{36}$$

To find such functions, we first remark, as in [14], that (8) and completeness of $\mathbb{C}^{m\times p}$ and $\mathbb{R}^n$ imply that $T(\mathtt{cl}(\mathcal{O}))$ is a closed subset of $\mathbb{C}^{m\times p}$. It follows that for each z in $\mathbb{C}^{m\times p}$, the infimum, in x in $\mathtt{cl}(\mathcal{O})$, of $|T(x) - z|$ is achieved by at least one point, denoted $T^*_p(z)$ (in $\mathtt{cl}(\mathcal{O})$). This defines a function $T^*_p : \mathbb{C}^{m\times p} \to \mathtt{cl}(\mathcal{O})$ satisfying

$$T(T^*_p(z)) \;=\; z \qquad \forall z \in T(\mathtt{cl}(\mathcal{O})), \tag{37}$$

$$|T(T^*_p(z)) - z| \;\leq\; |T(x) - z| \qquad \forall z \in \mathbb{C}^{m\times p}\,,\ \forall x \in \mathtt{cl}(\mathcal{O}). \tag{38}$$

With (8), (37) implies that the restriction T^*_p to $T(\mathtt{cl}(\mathcal{O}))$ is continuous. Also, with the triangle inequality, (38) gives, for each z in $\mathbb{C}^{m\times p}$ and each x in $\mathtt{cl}(\mathcal{O})$,

$$|x - T^*_p(z)| \;\leq\; \rho(|T(x) - z| + |z - T(T^*_p(z))|) \;\leq\; \rho(2|T(x) - z|). \tag{39}$$

Now we build the function T^* by smoothing out T^*_p. For each z in $\mathbb{C}^{m\times p}$, let

$$\epsilon(z) \;=\; \frac{1}{2} \inf_{x\in\mathtt{cl}(\mathcal{O})} |T(x) - z|.$$

$\big(\mathcal{B}_{\epsilon(z)}(z)\big)_{z\in\mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O}))}$ is a covering of the open set $\mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O}))$ by open subsets. From the Lindelöf theorem (see [6, Lemma 4.1], for instance), there exists a sequence $\{z_i\}_{i\in\mathbb{N}}$ such that $\big\{\mathcal{B}_{\epsilon(z_i)}(z_i)\big\}_{i\in\mathbb{N}}$ is a countable and locally finite covering by open subsets of $\mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O}))$.

For each x in $\mathtt{cl}(\mathcal{O})$, each z_i in $\{z_i\}_{i\in\mathbb{N}}$ and each z in $\mathcal{B}_{\epsilon(z_i)}(z_i)$, we have

$$|z_i - z| \;<\; \epsilon(z_i) \;\le\; \frac{1}{2}\,|T(x) - z_i| \;\le\; \frac{1}{2}\,[|T(x) - z| + |z - z_i|] \;\le\; |T(x) - z|.$$

With (39), this gives

$$|x - T_p^*(z_i)| \;\le\; \rho(2|T(x) - z_i|) \;\le\; \rho(2(|T(x) - z| + |z - z_i|)) \;\le\; \rho(4|T(x) - z|).$$

From [6, Theorem IV.4.4], we know that there exists a countable set of C^∞ functions $\{\phi_i\}_{i\in\mathbb{N}} : \mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O})) \to [0,1]$ satisfying, for each z in $\mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O}))$,

$$\sum_i \phi_i(z) \;=\; 1, \qquad\qquad \phi_i(z) \;=\; 0 \quad \forall z \notin \mathcal{B}_{\epsilon(z_i)}(z_i).$$

We define the function $T^* : \mathbb{C}^{m\times p} \to T(\mathtt{cl}(\mathcal{O}))$ as

$$\begin{aligned} T^*(z) &= \sum_i \phi_i(z) T_p^*(z_i) \quad \text{if } z \in \mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O})),\\ &= T_p^*(z) \qquad\qquad\quad \text{if } z \in T(\mathtt{cl}(\mathcal{O})). \end{aligned}$$

It is continuous when restricted to the open set $\mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O}))$ and to the closed set $T(\mathtt{cl}(\mathcal{O}))$. Also, for each z in $\mathbb{C}^{m\times p}\setminus T(\mathtt{cl}(\mathcal{O}))$ and each x in $\mathtt{cl}(\mathcal{O})$, we get

$$\begin{aligned} |T^*(z) - x| \;&=\; \left|\sum_i \phi_i(z) T_p^*(z_i) - x\right| \le \sum_i \phi_i(z)|T_p^*(z_i) - x|,\\ &\le \sum_i \phi_i(z)\rho(4|z - T(x)|) \;\le\; \rho(4|z - T(x)|). \end{aligned} \tag{40}$$

And, for each z in $T(\mathtt{cl}(\mathcal{O}))$ and each x in $\mathtt{cl}(\mathcal{O})$, we get readily from (8) and (37)

$$|T^*(z) - x| \;=\; |T_p^*(z) - x| \;\le\; \rho(|T(T_p^*(z)) - T(x)|) \;=\; \rho(|z - T(x)|). \tag{41}$$

With (40) and (41), (36) is established. This proves also that T^* is continuous on whole $\mathbb{C}^{m\times p}$.

3.1.1. Proof of Theorem 2. With [12, Corollary I.4.7], we know there exists a locally Lipschitz function $\chi : \mathbb{R}^n \to \mathbb{R}$ satisfying (14). It follows that the function $\breve{f}$ in (13) is locally Lipschitz. Thus, for each x in $\mathbb{R}^n$ there exists a unique solution $\breve{X}(x,t)$ of (13), with initial condition x, maximally defined on $(\breve{\sigma}^-_{\mathbb{R}^n}(x), \breve{\sigma}^+_{\mathbb{R}^n}(x))$. Moreover, backward completeness within $\mathcal{O} + \delta_u$ of (1) implies backward completeness of (13), i.e., $\breve{\sigma}^-_{\mathbb{R}^n}(x) = -\infty$. Following [2], this implies the existence of a proper and C^1 function $V_{\mathfrak{b}} : \mathbb{R}^n \to \mathbb{R}_+$ and a continuous function $\gamma_{\mathfrak{b}} : \mathbb{R}^p \to \mathbb{R}_+$ satisfying

$$L_{\breve{f}} V_{\mathfrak{b}}(x) \;\ge\; -V_{\mathfrak{b}}(x) \;-\; 1 \qquad \forall x \in \mathbb{R}^n. \tag{42}$$

Let α be a strictly positive real number so that $A + \alpha I$ is a Hurwitz matrix. We define the function $W_{\mathfrak{b}} : \mathbb{R}^n \to \mathbb{R}$ as

$$W_{\mathfrak{b}}(x) \;=\; (V_{\mathfrak{b}}(x) \;+\; 1)^\alpha.$$

With the help of Gronwall's lemma, (42) yields

$$W_{\mathfrak{b}}(\breve{X}(x,t)) \leq W_{\mathfrak{b}}(x)\exp(-\alpha t) \qquad \forall x \in \mathbb{R}^n,\ \forall t \in (-\infty, 0]. \tag{43}$$

Since $W_{\mathfrak{b}}$ is a proper function and h is continuous, we can find a C^1 function $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ of class $\mathcal{K}_\infty$ and a real number β_0 such that for each component h_i of h, we have

$$|h_i(x)| \ \leq\ \beta(W_{\mathfrak{b}}(x)) + \beta_0 \qquad \forall x \in \mathbb{R}^n.$$

Let $\breve{\beta} : \mathbb{R}_+ \to \mathbb{R}_+$ be the function defined as

$$\breve{\beta}(w) \ =\ \sqrt{w} \ +\ \beta(w) \ +\ \beta_0.$$

This function is strictly increasing, C^1 on $(0, +\infty)$ and its derivative $\breve{\beta}'$ satisfies

$$\lim_{x \to 0} \breve{\beta}'(x) \ =\ +\infty.$$

It admits an inverse $\breve{\beta}^{-1} : \mathbb{R}_+ \to \mathbb{R}_+$ which satisfies

$$\breve{\beta}^{-1}(|h_i(x)|) \ \leq\ W_{\mathfrak{b}}(x) \qquad \forall x \in \mathbb{R}^n. \tag{44}$$

Moreover the function $\eta \mapsto \frac{\eta \breve{\beta}^{-1}(|\eta|)}{|\eta|}$ is C^1 on $\mathbb{R}\backslash\{0\}$ and can be extended by continuity on $\mathbb{R}$ as a C^1 injective function. So, with p arbitrary vectors b_j in $\mathbb{R}^m$, we define the function $B : \mathbb{R}^p \to \mathbb{R}^{m \times p}$ as

$$B(h) = \left(\begin{array}{ccc} \frac{h_1 \breve{\beta}^{-1}(|h_1|)}{|h_1|} b_1 & \dots & \frac{h_p \breve{\beta}^{-1}(|h_p|)}{|h_p|} b_p \end{array} \right).$$

Since $A + \alpha I$ is a Hurwitz matrix, (43), (44), and the backward completeness imply

1. The existence of strictly positive real numbers c_0, c_1, and ε such that we have

$$|\exp(-As)B(h(\breve{X}(x,s)))| \leq c_0\, |\exp(-As)|\, W_{\mathfrak{b}}(\breve{X}(x,s)), \tag{45}$$

$$\leq c_1\, W_{\mathfrak{b}}(x)\, \exp(\varepsilon s) \qquad \forall (s,x) \in \mathbb{R}_- \times \mathbb{R}^n. \tag{46}$$

2. For each fixed s in $\mathbb{R}_-$, the function $x \mapsto \exp(-As)B(h(\breve{X}(x,s)))$ is continuous.

So the Lebesgue dominated convergence Theorem (see [8, Théorème (3.149)], for instance) implies that the following expression defines properly a continuous function $T : \mathbb{R}^n \to \mathbb{C}^{m \times p}$:

$$T(x) \ =\ \int_{-\infty}^{0} \exp(-As)\, B(h(\breve{X}(x,s)))ds. \tag{47}$$

Then, for each x in $\mathbb{R}^n$ and for each t in $(-\infty, \breve{\sigma}^{+}_{\mathbb{R}^n}(x))$, we get

$$\begin{aligned} T(\breve{X}(x,t)) - T(x) &= \int_{-\infty}^{0} \exp(-As)\, B(h(\breve{X}(\breve{X}(x,t),s)))ds \ -\ T(x), \\ &= \exp(At) \int_{-\infty}^{t} \exp(-Au)\, B(h(\breve{X}(x,u)))du \ -\ T(x), \\ &= (\exp(At) - I)\, T(x) \ +\ \exp(At) \int_{0}^{t} \exp(-Au) B(h(\breve{X}(x,u)))du. \end{aligned}$$

Thus, we obtain, for all x in $\mathbb{R}^n$,

$$\chi(x)\, L_f T(x) = L_{\breve{f}} T(x) = \lim_{t\to 0} \frac{T(\breve{X}(x,t)) - T(x)}{t} = AT(x) + B(h(x)). \tag{48}$$

With (14), this implies (7) is satisfied. □

Remark 6. 1. For the case where A is diagonalizable, with eigenvalue λ_i, and where the vectors b_j are chosen so that the p pairs (A, b_j) are controllable, our expression for T gives for its ith component in the diagonalizing coordinates:

$$T_i(x) \;=\; \int_{-\infty}^{0} \exp(-\lambda_i s) B_i(h(\breve{X}(x,s)))ds \tag{49}$$

with

$$B_i(h) \;=\; \left(\begin{array}{ccc} \frac{h_1 \breve{\beta}^{-1}(|h_1|)}{|h_1|} b_{i1} & \dots & \frac{h_p \breve{\beta}^{-1}(|h_p|)}{|h_p|} b_{ip} \end{array} \right),$$

where each b_{ij} is nonzero. Note that each of the m rows of the function B is an injective function from $\mathbb{R}^p$ to $\mathbb{R}^p$.

2. If $\mathcal{O}$ is bounded, the function $s \in \mathbb{R}_- \mapsto h(\breve{X}(x,s)) \in \mathbb{R}^p$ is a bounded function, uniformly in x in $\mathtt{cl}(\mathcal{O})$. It follows that the inequality (46) holds by choosing the function $\breve{\beta}^{-1}$ simply as the identity function. This says that, in this case, the function B is linear.

3.2. Proof of Theorem 3. We first remark that backward $\mathcal{O}$-distinguishability property of the original system (1) implies the same property for the modified system (13). Then we need the following lemma.

LEMMA 1 (due to Coron). *Let Ω and Υ be open subsets of $\mathbb{C}$ and $\mathbb{R}^{2n}$, respectively. Let $g : \Upsilon \times \Omega \to \mathbb{C}^p$ be a function which is holomorphic in λ for each x in Υ and C^1 in x for each λ in Ω. If for each pair (x,λ) in $\Upsilon \times \Omega$ for which $g(x,\lambda)$ is zero we can find, for at least one of the p components g_j of g, an integer k satisfying*

$$\frac{\partial^i g_j}{\partial \lambda^i}(x,\lambda) \;=\; 0 \qquad \forall i \in \{0,\dots,k-1\}, \qquad\qquad \frac{\partial^k g_j}{\partial \lambda^k}(x,\lambda) \;\neq\; 0, \tag{50}$$

then the following set has zero Lebesgue measure in $\mathbb{C}^{n+1}$:

$$S \;=\; \bigcup_{x\in\Upsilon} \left\{ (\lambda_1,\dots,\lambda_{n+1}) \in \Omega^{n+1} :\; g(x,\lambda_1) = \cdots = g(x,\lambda_{n+1}) = 0 \right\}. \tag{51}$$

This result has been established by Coron in [7, Lemma 3.2] in a stronger form except for the very minor point that, here, g is not C^∞ in both x and λ. To make sure that this difference has no bad consequence and for the sake of completeness, we give an ad hoc proof in the appendix.

To complete the proof of Theorem 3 all we have to do is to generate an appropriate function g satisfying all the required assumptions of this Lemma 1.

Let Ω and Υ be the following open subsets of $\mathbb{C}$ and $\mathbb{R}^{2n}$, respectively:

$$\Omega \;=\; \{\lambda \in \mathbb{C} :\, \mathtt{Re}(\lambda) < \ell\}, \qquad \Upsilon \;=\; \left\{ x = (x_1, x_2) \in (\mathcal{O} + \delta_\Upsilon)^2 :\, x_1 \neq x_2 \right\}.$$

By following the same arguments as in the proof of Theorem 2, the backward completeness allows us to conclude

$$\breve{\sigma}^{-}_{\mathbb{R}^n}(x) = -\infty \qquad \forall x \in \mathcal{O} + \delta_\Upsilon.$$

Then, with (16), we get, for each (x,λ,t) in $(\mathcal{O}+\delta_\Upsilon)\times\Omega\times(-\infty,0]$,

$$\begin{aligned}|\exp(-\lambda t)\, b(h(\breve{X}(x,t)))| &\leq \exp([\ell-\mathtt{Re}(\lambda)]t)|\exp(-\ell t) b(h(\breve{X}(x,t)))|,\\ &\leq \exp([\ell-\mathtt{Re}(\lambda)]t)\, M(x).\end{aligned}$$

So the Lebesgue dominated convergence theorem implies that for each fixed λ in Ω, the expression

$$T_\lambda(x) \;=\; \int_{-\infty}^{0} \exp(-\lambda s)\, b(h(\breve{X}(x,s)))ds$$

defines properly a continuous function $T_\lambda : \mathcal{O}+\delta_\Upsilon \to \mathbb{C}^p$. With similar arguments (see [8, Théorème (3.150)], for instance), with (16), we can establish that this function is actually C^1.

Now, let $\mathcal{D}T : (\mathcal{O}+\delta_\Upsilon)^2\times\Omega\to\mathbb{C}^p$ be the function defined as

$$\begin{aligned}\mathcal{D}T(x,\lambda) &= T_\lambda(x_1)-T_\lambda(x_2),\\ &= \int_{-\infty}^{0}\exp(-\lambda s)[b(h(\breve{X}(x_1,s)))-b(h(\breve{X}(x_2,s)))]ds\end{aligned} \tag{52}$$

with $x=(x_1,x_2)$. It is C^1 in x in $(\mathcal{O}+\delta_\Upsilon)^2$ for each λ in Ω. Also, as proved in [21, chapter 19, p. 367] with the help of the Morera and Fubini theorems, it is holomorphic in λ in Ω for each x in $(\mathcal{O}+\delta_\Upsilon)^2$. Moreover, since we have for each a in $(-\infty,\ell)$

$$\int_{-\infty}^{0}\exp(-2as)|b(h(\breve{X}(x_1,s)))-b(h(\breve{X}(x_2,s)))|^2\,ds \;\leq\; \frac{M(x_1)^2+M(x_2)^2}{2(\ell-a)} \;<\; +\infty,$$

we can apply the Plancherel theorem to obtain, for each a in $(-\infty,\ell)$ and each x in $(\mathcal{O}+\delta_\Upsilon)^2$,

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\infty}^{+\infty}|\mathcal{D}T(x,a+is)|^2\,ds\\ &\qquad=\int_{-\infty}^{0}\exp(-2as)|b(h(\breve{X}(x_1,s)))-b(h(\breve{X}(x_2,s)))|^2\,ds.\end{aligned} \tag{53}$$

Now, for x in Υ, with the distinguishability property, continuity with respect to time and injectivity of b imply the existence of an open time interval (t_0,t_1) such that

$$|b(h(\breve{X}(x_1,s)))-b(h(\breve{X}(x_2,s)))| \;>\; 0 \qquad \forall s\in(t_0,t_1).$$

It follows with (53) that we have

$$\int_{-\infty}^{+\infty}|\mathcal{D}T(x,a+is)|^2\,ds \;>\; 0.$$

This says that for each x in Υ, the function $\lambda\mapsto\mathcal{D}T(x,\lambda)$ is not identically equal zero on Ω. Since it is holomorphic, this implies that for each (x,λ) in $\Upsilon\times\Omega$, we can find, for at least one of the p components $\mathcal{D}T_j$ of $\mathcal{D}T$, an integer k satisfying

$$\frac{\partial^i\mathcal{D}T_j}{\partial\lambda^i}(x,\lambda) \;=\; 0 \qquad \forall i\in\{0,\dots,k-1\}, \qquad\qquad \frac{\partial^k\mathcal{D}T_j}{\partial\lambda^k}(x,\lambda) \;\neq\; 0.$$

Thus we can invoke Lemma 1 with $\mathcal{D}$ as function g. With (52), it allows us to conclude that the following set S has zero Lebesgue measure in $\mathbb{C}^{n+1}$:

$$S \;=\; \bigcup_{(x_1,x_2)\in\Upsilon}\left\{(\lambda_1,\dots,\lambda_{n+1})\in\Omega^{n+1}:\quad T_{\lambda_i}(x_1)=T_{\lambda_i}(x_2)\quad \forall i\in\{1,\dots,n+1\}\right\}.$$

3.3. Proof of Theorem 4. Our first step consists in proposing a function T solution of (22). The definition (19) of H and the inequality (20), give, for each pair (x_1, x_2) in $\mathtt{cl}(\mathcal{O})^2$,

$$\begin{aligned}|L_f H(x_1) - L_f H(x_2)| &\leq |H(x_1) - H(x_2)| + |L_f^m b(h(x_1)) - L_f^m b(h(x_2))|,\\ &\leq (1+L)|H(x_1) - H(x_2)|.\end{aligned}$$

Also (21) implies that for each Y in $H(\mathtt{cl}(\mathcal{O}))$, there exists a unique x in $\mathtt{cl}(\mathcal{O})$ solution of $Y = H(x)$. Hence we can define a Lipschitz function $F : H(\mathtt{cl}(\mathcal{O})) \to \mathbb{R}^{m\times p}$ satisfying

$$F(H(x)) \;=\; L_f H(x) \qquad \forall x \in \mathtt{cl}(\mathcal{O}). \tag{54}$$

Furthermore, as in the proof of Theorem 1, continuity and uniform injectivity of the function H on $\mathtt{cl}(\mathcal{O})$ as given by (21) imply that $H(\mathtt{cl}(\mathcal{O}))$ is closed. Then it follows from Kirszbraun's Lipschitz extension theorem (see [9, Theorem 2.10.43], for instance) that F can be extended as a function $\breve{F} : \mathbb{R}^{m\times p} \to \mathbb{R}^{m\times p}$ satisfying

$$|\breve{F}(Y_1) - \breve{F}(Y_2)| \leq (1+L)\,|Y_1 - Y_2| \qquad \forall (Y_1, Y_2) \in \mathbb{R}^{m\times p} \times \mathbb{R}^{m\times p}, \tag{55}$$

$$\breve{F}(Y) = F(Y) \qquad \forall Y \in H(\mathtt{cl}(\mathcal{O})). \tag{56}$$

Let $\mathfrak{Y}(Y,t)$ denote a solution of the following system on $\mathbb{R}^{m\times p}$:

$$\dot{Y} = \breve{F}(Y).$$

With (55), such a solution is unique for each Y in $\mathbb{R}^{m\times p}$ defined on $(-\infty, +\infty)$ and satisfies, for some fixed matrix Y_0 in $\mathbb{R}^{m\times p}$ and for each pair (Y_1, Y_2) in $\mathbb{R}^{m\times p} \times \mathbb{R}^{m\times p}$,

$$\begin{aligned}|\mathfrak{Y}(Y_1,t) - Y_0| &\leq |Y_1 - Y_0| \;+\; \int_t^0 |\breve{F}(\mathfrak{Y}(Y_1,s)) - \breve{F}(Y_0)|ds \;-\; |\breve{F}(Y_0)|\,t,\\ &\leq |Y - Y_0| \;+\; (1+L)\int_t^0 |\mathfrak{Y}(Y,s) - Y_0|ds \;-\; |\breve{F}(Y_0)|\,t,\\ |\mathfrak{Y}(Y_1,t) - \mathfrak{Y}(Y_2,t)| &\leq (1+L)\int_t^0 |\mathfrak{Y}(Y_1,s) - \mathfrak{Y}(Y_2,t)|ds.\end{aligned}$$

With Gronwall's inequality, this gives for all $t \leq 0$,

$$|\mathfrak{Y}(Y,t) - Y_0| \leq \exp(-(1+L)t)\left[|Y - Y_0| + \frac{|\breve{F}(Y_0)|}{1+L}\right] \;-\; \frac{\breve{F}(Y_0)}{1+L}, \tag{57}$$

$$|\mathfrak{Y}(Y_1,t) - \mathfrak{Y}(Y_2,t)| \leq \exp(-(1+L)t)\,|Y_1 - Y_2|. \tag{58}$$

So, in particular, we have, for each $t \leq 0$ and Y in $\mathbb{R}^{m\times p}$,

$$|\breve{F}(\mathfrak{Y}(Y,t))| \;\leq\; (1+L)\left(\exp(-(1+L)t)\left[|Y - Y_0| + \frac{|\breve{F}(Y_0)|}{1+L}\right]\right). \tag{59}$$

Hence, given any diagonal Hurwitz $m \times m$ matrix A, with eigenvalue λ_i, for each real number $k \geq \frac{1+L}{-\max_i\{\mathtt{Re}(\lambda_i)\}}$, we can properly define a continuous function $\mathfrak{R} : \mathbb{R}^{m\times p} \to \mathbb{C}^m$ as

$$\mathfrak{R}(Y) \;=\; \int_{-\infty}^0 \exp(-skA)\,B_{1m}\,\breve{F}(\mathfrak{Y}(Y,s))_m ds \tag{60}$$

with the notation (3), and where $\breve{F}(Y)_m$ denotes the mth row of $\breve{F}(Y)$. As for (48), we can prove that we have

$$L_{\breve{F}}\mathfrak{R}(Y) \;=\; kA\,\mathfrak{R}(Y) \;+\; B_{1m}\,\breve{F}(Y)_m \qquad \forall Y \in \mathbb{R}^{m\times m}.$$

But with (19), (54), and (56), this yields

$$L_F\mathfrak{R}(H(x)) \;=\; kA\,\mathfrak{R}(H(x)) \;+\; B_{1m}\,L_f^m b(h(x)) \qquad \forall x \in \mathtt{cl}(\mathcal{O}). \tag{61}$$

Let now $T : \mathtt{cl}(\mathcal{O}) \to \mathbb{R}^n$ be the continuous function defined as

$$T(x) \;=\; (kA)^{-m}\mathfrak{R}(H(x)) - K^{-1}SH(x) \tag{62}$$

with the notations:

$$S \;=\; \begin{pmatrix} \lambda_1^{-1} & \dots & \lambda_1^{-m} \\ \vdots & \vdots & \vdots \\ \lambda_m^{-1} & \dots & \lambda_m^{-m} \end{pmatrix}, \qquad K \;=\; \mathrm{diag}\,(k,\dots,k^m). \tag{63}$$

We want to show that T is a solution of (22). We have

$$K^{-1}SH(x) = \begin{pmatrix} \sum_{i=1}^m (k\lambda_1)^{-i} L_f^{i-1} b(h(x)) \\ \vdots \\ \sum_{i=1}^m (k\lambda_m)^{-i} L_f^{i-1} b(h(x)) \end{pmatrix}.$$

Thus, for each x in $\mathbb{R}^n$, we get

$$K^{-1}SL_fH(x) \;=\; kAK^{-1}SH(x) \;-\; B_{1m}b(h(x)) \;+\; (kA)^{-m}B_{1m}L_f^m b(h(x)). \tag{64}$$

In view of (62), it remains to compute the Lie derivatives of $(kA)^{-m}\mathfrak{R}(H(x))$. From (14), (19), (54), and (56), we get the identity

$$\mathfrak{Y}(H(x),t) \;=\; H(\breve{X}(x,t)) \qquad \forall t \in (\breve{\sigma}^{\,-}_{\mathcal{O}}(x), \breve{\sigma}^{\,+}_{\mathcal{O}}(x)) \qquad \forall x \in \mathcal{O}.$$

This gives readily for all t in $(\breve{\sigma}^{\,-}_{\mathcal{O}}(x), \breve{\sigma}^{\,+}_{\mathcal{O}}(x))$ and x in $\mathcal{O}$,

$$\mathfrak{R}(\mathfrak{Y}(H(x),t)) \;-\; \mathfrak{R}(H(x)) \;=\; \mathfrak{R}(H(\breve{X}(x,t)) \;-\; \mathfrak{R}(H(x))$$

and therefore

$$L_F\mathfrak{R}(H(x)) \;=\; L_f\mathfrak{R}(H(x)) \qquad \forall x \in \mathcal{O}.$$

By continuity this identity extends to $\mathtt{cl}(\mathcal{O})$. So, with (61), we get

$$L_f\mathfrak{R}(H(x)) \;=\; kA\,\mathfrak{R}(H(x)) \;+\; B_{1m}\,L_f^m(b(h(x))) \qquad \forall x \in \mathtt{cl}(\mathcal{O}).$$

Consequently, with (62) and (64), we finally obtain, for each x in $\mathtt{cl}(\mathcal{O})$,

$$\begin{aligned} L_fT(x) &= (kA)^{-m}L_f\mathfrak{R}(H(x)) - K^{-1}SL_fH(x), \\ &= kA\left[(kA)^{-m}\,\mathfrak{R}(H(x)) - K^{-1}SH(x)\right] + B_{1m}b(h(x)), \\ &= kAT(x) + B_{1m}b(h(x)). \end{aligned}$$

This proves that the function T defined by (62) is a solution of (22).

Our second step in this proof is to show that by picking k large enough, the function T given by (62) is uniformly injective. To simplify the following notation, to a function f, we associate the function Δf as follows:

$$\Delta f(x_1, x_2) \;=\; f(x_1) \;-\; f(x_2).$$

So, for instance, for each pair (x_1, x_2) in $\mathcal{O}^2$, we have

$$T(x_1) - T(x_2) \;=\; (kA)^{-m} \Delta(\mathfrak{R} \circ H)(x_1, x_2)) \;+\; K^{-1} S \Delta H(x_1, x_2).$$

With (55), (60), and (58), we get, for each (Y_1, Y_2) in $\mathbb{R}^{m\times p} \times \mathbb{R}^{m\times p}$,

$$\begin{aligned}
|\Delta\mathfrak{R}(Y_1, Y_2)| &\leq \int_{-\infty}^{0} \left|\exp(-skA)\, B_{1m} \,\left[\breve{F}(\mathfrak{Y}(Y_1, s))_m - \breve{F}(\mathfrak{Y}(Y_2, s))_m\right]\right| ds \\
&\leq (1+L) \int_{-\infty}^{0} |\exp(-skA)|\, |B_{1m}|\, [\mathfrak{Y}(Y_1, s) - \mathfrak{Y}(Y_2, s)|\, ds \\
&\leq (1+L)\, |B_{1m}|\, |Y_1 - Y_2| \int_{-\infty}^{0} \exp\left(-s\left[1 + L + k \max_i\{\mathtt{Re}(\lambda_i)\}\right]\right) ds \\
&\leq \frac{(1+L)|B_{1m}|}{-\left[1 + L + k \max_i\{\mathtt{Re}(\lambda_i)\}\right]}\, |Y_1 - Y_2|.
\end{aligned}$$

This yields, for each pair (x_1, x_2) in $\mathtt{cl}(\mathcal{O})^2$,

$$\begin{aligned}
|T(x_1) - T(x_2)| &\geq |K^{-1} S \Delta H(x_1, x_2)| \;-\; |(kA)^{-m} \Delta(\mathfrak{R} \circ H)(x_1, x_2)|, \\
&\geq \frac{|\Delta H(x_1, x_2)|}{|K|\, |S^{-1}|} \;-\; \frac{|(kA)^{-m}|\, (1+L)|B_{1m}|}{-\left[1 + L + k \max_i\{\mathtt{Re}(\lambda_i)\}\right]}\, |\Delta H(x_1, x_2)|, \\
&\geq k^{-m} \left(\frac{1}{|S^{-1}|} - \frac{|A|^{-m}(1+L)|B_{1m}|}{-\left[1 + L + k \max_i\{\mathtt{Re}(\lambda_i)\}\right]}\right) |H(x_1 - H(x_2)|.
\end{aligned}$$

So, with (21), the function T is uniformly injectivity on $\mathtt{cl}(\mathcal{O})$ for all k large enough.

3.4. Proof of Theorem 5. Following the same arguments as in the proof of Theorem 1, continuity and uniform injectivity (23) of the function T_a on $\mathtt{cl}(\mathcal{O})$ imply that $T_a(\mathtt{cl}(\mathcal{O}))$ is closed and we can construct a continuous function $T_a^* : \mathbb{C}^{m\times p} \to \mathtt{cl}(\mathcal{O})$ satisfying

$$|T_a^*(z) - x| \;\leq\; \rho(4|z - T_a(x)|) \qquad\qquad \forall (x, z) \in \mathtt{cl}(\mathcal{O}) \times \mathbb{C}^{m\times p}. \tag{65}$$

This implies

$$T_a^*(T_a(x)) \;=\; x \qquad \forall x \in \mathtt{cl}(\mathcal{O}). \tag{66}$$

Now, let us assume for the time being there exists a function $\mathfrak{F} : \mathbb{C}^{m\times p} \to \mathbb{C}^{m\times p}$ to be used in (28) and satisfying

$$|\mathfrak{E}(x) - \mathfrak{F}(z)| \;\leq\; N\, |T_a(x) - z| \qquad \forall (x, z) \in \mathtt{cl}(\mathcal{O}) \times \mathbb{C}^{m\times p}. \tag{67}$$

As a direct consequence, we get the inequality

$$|\mathfrak{F}(z)| \;\leq\; N\, |z| \;+\; M \qquad z \in \mathbb{C}^{m\times p}$$

for some real number M $(= |\mathfrak{E}(x_0)| + N|T_a(x_0)|$, with some arbitrarily fixed x_0 in $\mathtt{cl}(\mathcal{O}))$. It follows that the z dynamics in the system (28) satisfy

$$|\dot{z}| \;\leq\; (|A| + N)\,|z| \;+\; (M \;+\; |B(h(x))|).$$

Hence, for each x in $\mathcal{O}$ and z in $\mathbb{C}^{m\times p}$, the component $Z(x,z,t)$ of a solution $(X(x,t), Z(x,z,t))$ of (28) is defined as long as $h(X(x,t))$ is defined. So this solution is right maximally defined on the same interval $[0, \sigma^+_{\mathbb{R}^n}(x))$ as $X(x,t)$ solution of (1).

With (65), (29) holds if we have

$$\lim_{t\to\sigma^+_{\mathbb{R}^n}(x)} T_a(X(x,t)) - Z(x,z,t) \;=\; 0 \qquad\qquad \forall (x,z)\in\mathcal{O}\times\mathbb{C}^{m\times p}. \tag{68}$$

To establish this limit, we associate to each pair (x,z) in $\mathcal{O}\times\mathbb{C}^{m\times p}$ the matrix e in $\mathbb{C}^{m\times p}$:

$$e \;=\; T_a(x) - z.$$

With (24) and (28), we get

$$\dot{e} \;=\; Ae + \mathfrak{E}(x) - \mathfrak{F}(z).$$

Let $U : \mathbb{C}^{m\times p} \to \mathbb{R}_+$ be the positive definite and proper function defined as

$$U(e) \;=\; \sum_{i=1}^{p} \overline{e}_i^{\top} P e_i,$$

where e_i denotes the ith column of e, $\overline{e_i}$ denotes its complex conjugate and P is given by (27). Using (67) and completing the squares, we get

$$\begin{aligned}\dot{\overline{U(e)}} \;=\; \sum_{i=1}^{p} \left[-|e_i|^2 \;+\; 2\overline{e_i}^{\top} P\left(\mathfrak{E}(x) - \mathfrak{F}((z))_i\right]\right. &\leq -[1 - 2N\lambda_{\max}(P)]\,|e|^2,\\ &\leq -\frac{1 - 2N\lambda_{\max}(P)}{\lambda_{\min}(P)}\,U(e).\end{aligned}$$

So, with (26), we have established the existence of a strictly positive real number ε such that we have

$$\dot{\overline{U(e)}} \;\leq\; -\varepsilon\, U(e) \qquad \forall (x,z)\in\mathcal{O}\times\mathbb{C}^{m\times p}. \tag{69}$$

This implies for all t in $[0, \sigma^+_{\mathcal{O}}(x))$ and (x,z) in $\mathcal{O}\times\mathbb{C}^{m\times p}$,

$$\exp(-\varepsilon t)\; U(e) \;\geq\; U(E(x,z,t)) \qquad (\;=\; U\left(T_a(X(x,t)) - Z(x,z,t)\right)). \tag{70}$$

With forward completeness within $\mathcal{O}$ and the condition in the left of (29) (see (11)), this implies (68) holds.

It remains to establish the existence of a function $\mathfrak{F} : \mathbb{C}^{m\times p} \to \mathbb{C}^{m\times p}$ satisfying (67). With (66), we see that (25) becomes

$$|\mathfrak{E}(T_a^*(z_1)) - \mathfrak{E}(T_a^*(z_2))| \;\leq\; N\,|z_1 - z_2| \qquad \forall (z_1,z_2)\in T_a(\mathtt{cl}(\mathcal{O}))^2.$$

Thus, $\mathfrak{E}\circ T_a^*$ is a Lipschitz function on the closed subset $T_a(\mathtt{cl}(\mathcal{O}))$. From Kirszbraun's Lipschitz extension theorem, $\mathfrak{E}\circ T_a^*$ can be extended as a function $\mathfrak{F} : \mathbb{C}^{m\times p} \to \mathbb{C}^{m\times p}$ satisfying

$$|\mathfrak{F}(z_1) \;-\; \mathfrak{F}(z_2)| \leq N\,|z_1 \;-\; z_2| \quad \forall (z_1,z_2)\in(\mathbb{C}^{m\times p})^2,$$

$$\mathfrak{F}(z) = \mathfrak{E}(T^*(z)) \quad \forall z \in T_a(\mathtt{cl}(\mathcal{O})).$$

So, in particular, we get (67).

3.5. Proof of Corollary 1. Let λ_i be the eigenvalues of a given diagonal Hurwitz complex $m \times m$ matrix A. With the notation (63), the function $T_a : \mathtt{cl}(\mathcal{O}) \to \mathbb{R}^{m\times p}$ defined in (30) can be rewritten as

$$T_a(x) \;=\; -K^{-1}SH(x). \tag{71}$$

In the following we show that we can find a real number $k^* \geq 1$ such that if k is strictly larger than k^*, then the triple (kA, T_a, B_{1m}) satisfies all the assumptions of Theorem 5:

1. The forward completeness within $\mathcal{O}$ is satisfied by assumption.

2. (23) is satisfied since, using (21) and the definition of T_a in (71), we get, for each pair (x_1, x_2) in $\mathtt{cl}(\mathcal{O})^2$,

$$\begin{aligned} |x_1 - x_2| &\leq \rho(|S^{-1}K|\,|K^{-1}SH(x_1) - K^{-1}SH(x_2)|), \\ &\leq \rho(|S^{-1}K|\,|T_a(x_1) - T_a(x_2)|). \end{aligned}$$

3. Let the function $\mathfrak{E} : \mathtt{cl}(\mathcal{O}) \to \mathbb{R}^{m\times p}$ be defined as

$$\mathfrak{E}(x) \;=\; -(kA)^{-m}B_{1m}L_f^m b(h(x)).$$

We have to show that this function satisfies (24) and (25). Using (64), we get, for each x in $\mathcal{O}$,

$$\mathfrak{E}(x) + B_{1m}b(h(x)) = -K^{-1}SL_fH(x) + kAK^{-1}SH(x) = L_fT_a(x) - kAT_a(x).$$

So, (24) does hold. Also, with (20) and $k^{-m}|K| \leq 1$, which holds for $k \geq 1$, we get, for each (x_1, x_2) in $\mathtt{cl}(\mathcal{O})^2$,

$$\begin{aligned} |\mathfrak{E}(x_1) \;-\; \mathfrak{E}(x_2)| &= |(kA)^{-m}(B_{1m}(L_f^m(h(x_1)) \;-\; L_f^m(h(x_2))))|, \\ &\leq |(kA)^{-m}|\,|B_{1m}|\,L\,|H(x_1) - H(x_2)|, \\ &\leq \frac{1}{\min_i |\lambda_i|^m}\,|B_{1m}|\,L\,|S^{-1}|\,|T_a(x_1) - T_a(x_2)|. \end{aligned}$$

Hence, (25) is satisfied with $N = \frac{1}{\min_i |\lambda_i|^m}\,|B_{1m}|\,L\,|S^{-1}|$, which does not depend on k.

4. It remains to show that by choosing k sufficiently large, the constraint (26) is satisfied. As kA is a diagonal complex matrix, the inequality (26) is simply

$$\frac{1}{\min_i |\lambda_i|^m}\,|B_{1m}|\,L\,|S^{-1}|\,\frac{1}{k\,(-\max_i \mathtt{Re}(\lambda_i))} \;<\; 1.$$

Clearly this inequality holds for all k large enough.

3.6. Technical comments on section 2.6. Due to space limitations, we give here only some hints on how the results established for the case of completeness can be extended to the case of boundedness observability.

The introduction of γ in the observer has mainly two consequences:

1. For the error convergence, t, in $\exp(At)$ in (35) or $\exp(-\varepsilon t)$ in (70), is replaced by the integral $\int_0^t \gamma(h(X(x,s)))ds$. If $\sigma^+_{\mathbb{R}^n}(x) = +\infty$, then,$\gamma$ being larger than 1, this integral goes to $+\infty$ as t goes to $\sigma^+_{\mathbb{R}^n}(x)$. If, instead, $\sigma^+_{\mathbb{R}^n}(x)$ is finite, then $V_{\mathfrak{f}}(X(x,t))$ goes to $+\infty$ as t goes to $\sigma^+_{\mathbb{R}^n}(x)$. From (33) this is possible only if the above integral tends again to $+\infty$.

2. The function T given in (18) is defined in terms of the solutions $\breve{X}(x,t)$ of the modified system (13) with f_γ instead of f. So we must show that this latter system shares the backward $\mathcal{O}$-distinguishability property of the original system (1). This can be done by associating, to each x in $\mathcal{O}+\delta_d$, the function $\tau_x : (\sigma^-_{\mathcal{O}+\delta_d}(x), \sigma^+_{\mathcal{O}+\delta_d}(x)) \to \mathbb{R}$ defined as

$$\tau_x(t) \;=\; \int_0^t \gamma(h(X(x,s)))ds.$$

It admits an inverse τ^{-1} which is such that we have

$$X(x, \tau_x^{-1}(t)) \;=\; \breve{X}(x,t) \qquad \forall x \in \mathcal{O}+\delta_d \;,\; \forall t \in \tau_x(\sigma^-_{\mathcal{O}+\delta_d}(x), \sigma^+_{\mathcal{O}+\delta_d}(x)).$$

Then it is possible to prove that, if for some pair (x_1, x_2) in $\mathcal{O}^2$ we have

$$h(\breve{X}(x_1,t)) \;=\; h(\breve{X}(x_2,t)) \qquad \forall t \;\in\; (\breve{\sigma}\,{}^-_{\mathcal{O}+\delta_d}(x_1), 0] \cap (\breve{\sigma}\,{}^-_{\mathcal{O}+\delta_d}(x_2), 0],$$

then we have also

$$\tau_{x_1}^{-1}(t) \;=\; \tau_{x_2}^{-1}(t) \qquad \forall t \;\in\; (\breve{\sigma}\,{}^-_{\mathcal{O}+\delta_d}(x_1), 0] \cap (\breve{\sigma}\,{}^-_{\mathcal{O}+\delta_d}(x_2), 0].$$

4. Conclusion. We have stated sufficient conditions under which the extension to nonlinear systems of the Luenberger observer, as it has been proposed by Kazantzis and Kravaris in [13], can be used as long as the state to be observed remains in a given open set. In doing so, we have exploited the fact, already mentioned in [4, 16], that the observer proposed by Kreisselmeier and Engel in [14] is a possible way of implementing the Kazantzis–Kravaris/Luenberger observer.

We have established that a sufficient (row) dimension of the dynamic system giving the observer is 2 + twice the dimension of the state to be observed. This is in agreement with many other results known on the generic number of pieces of information to be extracted from the output paths to be able to reconstruct the state.

We have also shown that it is sufficient to know only an approximation of a solution of a partial differential equation which we need to solve to implement the observer. In this way, we have been able to make a connection with high gains observers.

Finally, to get less restrictive sufficient conditions, we have found it useful to modify the observer in a way which induces a time rescaling as already suggested in [4].

At this stage, our results are mainly theoretical. They are concerned with existence. Several problems of prime importance for practice remain to be addressed, like type and speed of convergence. In these regards, the contribution of Rapaport and Maloum in [18] is an important starting point.

Even for the purpose of showing the existence, we have to note that the conditions we have given can be strongly relaxed if an estimation of the norm of the state is available. This idea has been exploited in [4], where a truly global observer has been proposed under the assumption of global complete observability and unboundedness observability.

Appendix. Proof of Coron's Lemma 1. The idea of the proof is to show that the set

$$S \;=\; \bigcup_{x\in\Upsilon} \big\{\Lambda = (\lambda_1, \ldots, \lambda_{n+1}) \in \Omega^{n+1} : \quad g(x,\lambda_\ell) = 0 \quad \forall \ell \in \{1, \ldots, n+1\}\big\}$$

defined in (51) is contained in a countable union of sets which have zero Lebesgue measure.

Given $(\underline{x}, \underline{\Lambda}, \epsilon)$ in $\Upsilon \times \Omega^{n+1} \times \mathbb{R}_{+*}$, we denote by $S_{\epsilon,\underline{x},\underline{\Lambda}}$ the set

$$S_{\epsilon,\underline{x},\underline{\Lambda}} \;=\; \bigcup_{x \in \mathcal{B}_\epsilon(\underline{x})} \{\Lambda \in \mathcal{B}_\epsilon(\underline{\Lambda}) : \quad g(x,\lambda_\ell) = 0 \quad \forall \ell \in \{1,\ldots,n+1\}\}\,. \tag{72}$$

Assume for the time being that for each pair $(\underline{x}, \underline{\Lambda})$ in $\Upsilon \times \Omega^{n+1}$, we can find a strictly positive real number ϵ and a countable family of C^1 functions $\sigma_i : \mathcal{B}_\epsilon(\underline{x}) \to \Omega^{n+1}$, such that we have

$$S_{\epsilon,\underline{x},\underline{\Lambda}} \subset \bigcup_{i \in \mathbb{N}} \sigma_i(\mathcal{B}_\epsilon(\underline{x})). \tag{73}$$

The family $(\mathcal{B}_\epsilon(\underline{x}) \times \mathcal{B}_\epsilon(\underline{\Lambda}))_{(\underline{x},\underline{\Lambda}) \in \Upsilon \times \Omega^{n+1}}$ is a covering of $\Upsilon \times \Omega^{n+1}$ by open subsets. From the Lindelöf theorem (see [6, Lemma 4.1], for instance), there exists a countable family $\left\{(\underline{x}_j, \underline{\Lambda}_j)\right\}_{j \in \mathbb{N}}$ such that we have

$$\Upsilon \times \Omega^{n+1} \;\subset\; \bigcup_{j \in \mathbb{N}} \mathcal{B}_{\epsilon_j}(\underline{x}_j) \times \mathcal{B}_{\epsilon_j}(\underline{\Lambda}_j),$$

where ϵ_j denotes the ϵ associated to the pair $(\underline{x}_j, \underline{\Lambda}_j)$. With (73), it follows that we have

$$S \;\subset\; \bigcup_{j \in \mathbb{N}} \bigcup_{i \in \mathbb{N}} \sigma_{i,j}(\mathcal{B}_{\epsilon_j}(\underline{x}_j)),$$

where $\sigma_{i,j}$ denotes the ith function σ associated with the pair $(\underline{x}_j, \underline{\Lambda}_j)$. The set $\sigma_{i,j}(\mathcal{B}_{\epsilon_j}(\underline{x}_j))$ is the image, contained in $\mathbb{C}^{n+1}$, a real manifold of dimension $2(n+1)$, by a C^1 function of $\mathcal{B}_{\epsilon_j}(\underline{x}_j)$, a real manifold of dimension $2n$. From a variation on Sard's theorem (see [20, Theorem 3, paragraph 3], for instance), this image $\sigma_{i,j}(\mathcal{B}_{\epsilon_j}(\underline{x}_j))$ has zero Lebesgue measure in $\mathbb{C}^{n+1}$. So S, being a countable union of such zero Lebesgue measure subsets, has zero Lebesgue measure.

So all we have to do to establish Lemma 1 is to prove the existence of ε and the functions σ_i satisfying (73) for each pair $(\underline{x}, \underline{\Lambda})$in $\Upsilon \times \Omega^{n+1}$. For ε, we consider two cases:

1. Consider a pair $(\underline{x}, \underline{\Lambda})$ such that $g_j(\underline{x}, \underline{\lambda}_\ell)$ is nonzero for some component λ_ℓ of $\underline{\Lambda}$ and g_j of g. By continuity of g_j, we can find a strictly positive real number ϵ such that $g(x, \lambda_\ell)$ is also nonzero for all x in $\mathcal{B}_\epsilon(\underline{x})$ and Λ in $\mathcal{B}_\epsilon(\underline{\lambda})$. In this case, the set $S_{\epsilon,\underline{x},\underline{\Lambda}}$ is empty.

2. Consider a pair $(\underline{x}, \underline{\Lambda})$ such that $g(\underline{x}, \underline{\lambda}_\ell)$ is zero for each of the $n+1$ components of $\underline{\lambda}_\ell$ of $\underline{\Lambda}$. From the assumption (50), for each ℓ, we can find a component g_{j_ℓ} of g and an integer k_ℓ satisfying

$$\frac{\partial^i g_{j_\ell}}{\partial \lambda^i}(\underline{x}, \underline{\lambda}_\ell) \;=\; 0 \qquad \forall i \in \{0,\ldots,k_\ell - 1\}, \qquad\qquad \frac{\partial^{k_\ell} g_{j_\ell}}{\partial \lambda^{k_\ell}}(\underline{x}, \underline{\lambda}_\ell) \;\neq\; 0.$$

In this case, following the Weierstrass preparation theorem (see [12, Theorem IV.1.1],[2] for instance), for each ℓ in $\{1,\ldots,n+1\}$, we know the existence of a strictly positive

[2]In [12, Theorem IV.1.1], this theorem is stated with the assumption that g_j is holomorphic in both x and λ. However, as far as x is concerned, it can be seen in the proof of this theorem that we need only the implicit function theorem to apply. So continuous differentiability in x for each λ is enough.

real number ϵ_ℓ, a function $q_\ell : \mathcal{B}_{\epsilon_\ell}(\underline{x}) \times \mathcal{B}_{\epsilon_\ell}(\underline{\lambda}_\ell) \to \mathbb{C}$, and k_ℓ C^1 functions $a_j^\ell : \mathbb{R}^{2n} \to \mathbb{C}$ satisfying for all (x, λ) in $\mathcal{B}_{\epsilon_\ell}(\underline{x}) \times B_{\epsilon_\ell}(\underline{\lambda}_\ell)$

$$q_\ell(x,\lambda)\, g_{j_\ell}(x,\lambda) \;=\; (\lambda - \underline{\lambda}_\ell)^{k_\ell} \;+\; \sum_{j=0}^{k_\ell - 1} a_j^\ell(x)(\lambda - \underline{\lambda}_\ell)^j. \tag{74}$$

We choose the real number ϵ, to be associated to $(\underline{x}, \underline{\Lambda})$ in the definition of $S_{\epsilon,\underline{x},\underline{\Lambda}}$, as

$$\epsilon \;=\; \inf_{\ell \in \{1,\ldots,n+1\}} \epsilon_\ell.$$

In the following $P_\ell : \mathcal{B}_\epsilon(\underline{x}) \times \mathbb{C} \to \mathbb{C}$ and $a^\ell(x) : \mathcal{B}_\epsilon(\underline{x}) \to \mathbb{C}^{k_\ell}$ denote

$$P_\ell(x,\lambda) \;=\; (\lambda - \underline{\lambda}_\ell)^{k_\ell} \;+\; \sum_{j=0}^{k_\ell - 1} a_j^\ell(x)(\lambda - \underline{\lambda}_\ell)^j, \qquad a^\ell(x) \;=\; (a_0^\ell(x), \ldots, a_{k_\ell - 1}^\ell(x)).$$

With this definition of ε, we have the following implication for Λ in $\mathcal{B}_\epsilon(\underline{\Lambda})$ and x in $\mathcal{B}_\epsilon(\underline{x})$:

$$g(x,\lambda_\ell) = 0 \quad \forall \ell \in \{1,\ldots,n+1\} \quad \Rightarrow \quad (\lambda_\ell, a^\ell(x)) \in M^\ell \quad \forall \ell \in \{1,\ldots,n+1\}, \tag{75}$$

where M^ℓ is the set

$$M^\ell \;=\; \left\{ (\lambda, (b_0, \ldots, b_{k_\ell - 1})) \in \mathbb{C} \times \mathbb{C}^{k_\ell} \;:\; (\lambda - \underline{\lambda}_\ell)^{k_\ell} + \sum_{j=0}^{k_\ell - 1} b_j (\lambda - \underline{\lambda}_\ell)^j = 0 \right\}. \tag{76}$$

Our interest in this set follows from the following lemma, which we prove later on.

LEMMA 2. *Let M be the set defined as*

$$M \;=\; \left\{ (\lambda, b_0, \ldots, b_{k-1}) \in \mathbb{C} \times \mathbb{C}^k \;:\; \lambda^k + \sum_{j=0}^{k-1} b_j \lambda^j = 0 \right\}.$$

There exists a countable family $\{M_m\}_{m \in \mathbb{N}}$ of regular submanifolds of $\mathbb{C}^k$ and a countable family of C^1 functions $\rho_m : M_m \to \mathbb{C}$ such that we have the inclusion

$$M \;\subset\; \bigcup_{m \in \mathbb{N}} \bigcup_{b \in M_m} \{(\rho_m(b), b)\}. \tag{77}$$

So, for each ℓ in $\{1,\ldots,n+1\}$ we have a countable family $\{M_{m_\ell}^\ell\}_{m_\ell \in \mathbb{N}}$ of regular submanifolds of $\mathbb{C}^{k_\ell}$ and a countable family of C^1 functions $\rho_{m_\ell}^\ell : M_{m_\ell}^\ell \to \mathbb{C}$ such that for each x in $\mathcal{B}_\epsilon(\underline{x})$, if $P_\ell(x, \lambda_\ell)$ is zero, then there exists an integer m_ℓ such that we have

$$a^\ell(x) \;\in\; M_{m_\ell}^\ell, \qquad \lambda_\ell \;=\; \rho_{m_\ell}^\ell(a^\ell(x)). \tag{78}$$

Hence, with (75), if

$$g(x,\lambda_\ell) \;=\; 0 \qquad \forall \ell \in \{1,\ldots,n+1\},$$

then there exists an $(n+1)$-tuple $\mu = (m_1, \ldots, m_{n+1})$ of integers satisfying

$$a^\ell(x) \;\in\; M_{m_\ell}^\ell, \quad \lambda_\ell \;=\; \rho_{m_\ell}^\ell(a^\ell(x)) \qquad \forall \ell \in \{1,\ldots,n+1\}.$$

So, by letting

$$S^{\mu}_{\epsilon,\underline{x},\underline{\Lambda}} = \bigcup_{\left\{x\in\mathcal{B}_\epsilon(\underline{x}):\ a^\ell(x)\in M^\ell_{m_\ell}\ \forall \ell\in\{1,\ldots,n+1\}\right\}} \left\{(\rho^1_{m_1}(a^1(x)),\ldots,\rho^{n+1}_{m_{n+1}}(a^{n+1}(x))\right\} \tag{79}$$

we have established

$$S_{\epsilon,\underline{x},\underline{\Lambda}} \subset \bigcup_{\mu\in\mathbb{N}^{n+1}} S^{\mu}_{\epsilon,\underline{x},\underline{\Lambda}}. \tag{80}$$

Comparing (73) with (80) and using the definition (79), we see that a candidate for the function σ_i is

$$\sigma_i(x) = \left(\rho^\ell_{m_\ell}\left(R_{M^\ell_{m_\ell}}(a^\ell(x))\right)\right)_{\ell\in\{1,\ldots,n+1\}},$$

where i happens to be the $(n+1)$-tuple μ and $R_{M^\ell_{m_\ell}} : \mathbb{C}^{k_\ell} \to M^\ell_{m_\ell}$ is a "restriction" to $M^\ell_{m_\ell}$ since we have to consider only those $a^\ell(x)$ which are in $M^\ell_{m_\ell}$. Finding such functions $R_{M^\ell_{m_\ell}}$ such that σ_i is C^1 may not be possible. But, following [7, Lemma 3.3], we know the existence, for each ℓ, of a countable family of C^1 functions $R^\ell_\nu : \mathbb{C}^{k_\ell} \to M^\ell_{m_\ell}$ such that we get

$$S^{\mu}_{\epsilon,\underline{x},\underline{\Lambda}} \subset \bigcup_{\nu\in\mathbb{N}} \left\{\left(\rho^\ell_{m_\ell}\left(R^\ell_\nu(a^\ell(\mathcal{B}_\epsilon(\underline{x})))\right)\right)_{\ell\in\{1,\ldots,n+1\}}\right\}.$$

In other words, the family of functions σ_i is actually given by the family

$$\sigma_{\mu,\nu} = \left(\rho^\ell_{m_\ell}\circ R^\ell_\nu \circ a_\ell\right)_{\ell\in\{1,\ldots,n+1\}},$$

i.e., we have

$$S_{\epsilon,\underline{x},\underline{\Lambda}} \subset \bigcup_{\mu\in\mathbb{N}^{n+1}} \bigcup_{\nu\in\mathbb{N}} \sigma_{\mu,\nu}(\mathcal{B}_\epsilon(\underline{x})). \qquad \square$$

Proof of Lemma 2: The inclusion (77) says that we are looking for a covering of the set M with some special structure. A covering easily found, but not having this special structure, is obtained by choosing a first complex number λ, denoted λ_1, as well as $k-1$ other complex numbers λ_j. Then the b_j are given by the identity

$$\prod_{j=1}^{k}(\lambda-\lambda_j) = \lambda^k + \sum_{j=0}^{k-1} b_j\lambda^j, \quad \lambda\in\mathbb{C}. \tag{81}$$

In other words, if we denote by $\eta : \mathbb{C}^k \to \mathbb{C}^k$ the function which gives the b_j from the λ_j, we have

$$\bigcup_{(\lambda_1,\ldots,\lambda_k)\in\mathbb{C}^k} \{(\lambda_1,\eta(\lambda_1,\ldots,\lambda_k))\} \subseteq M.$$

Specifically, given the elementary symmetric functions s_i, sum of all the products of i distinct λ_j,

$$s_i = \lambda_1\ldots\lambda_{i-1}\lambda_i + \lambda_1\ldots\lambda_{i-1}\lambda_{i+1} + \cdots + \lambda_{k-i+1}\lambda_{k-i+2}\ldots\lambda_k,$$

the b_j are obtained as

$$b_j \;=\; (-1)^{k-j} s_{k-j} \qquad \forall j \in \{1, \dots, k-1\}.$$

Also, the elementary symmetric functions are related to the sum of similar powers σ_p,

$$\sigma_p \;=\; \sum_{j=1}^{k} \lambda_j^p,$$

via the Newton equations:

$$\sigma_i \;-\; \sigma_{i-1} s_1 \;+\; \sigma_{i-2} s_2 \;+\; \cdots \;+\; (-1)^{i-1} \sigma_1 s_{k-1} \;+\; (-1)^i i s_i \;=\; 0.$$

The corresponding functions $(\sigma_p) \mapsto (s_i)$ and $(s_i) \mapsto (b_j)$ are C^∞ diffeomorphisms.

To obtain the result stated in the lemma, we need to invert the function $\eta : (\lambda_\ell) \mapsto (b_j)$. This function is known to be a homeomorphism if the λ_j are defined up to permutations (see [5, Proposition 1.5.5], for instance). But unfortunately we cannot go beyond continuity of the inverse because of possible multiple roots. To round this problem, we choose the multiplicity c_ℓ of the root λ_ℓ so that the sum of the c_ℓ is k. So, except if they are all 1, some of them must be 0. Maybe after reordering, we can assume that each c_1 to c_q is nonzero and satisfies

$$c_1 \;+\; \cdots \;+\; c_q \;=\; k.$$

Then we choose q different complex numbers ϖ_ℓ and we let

$$\lambda_1 = \cdots = \lambda_{c_1} = \varpi_1,\ \lambda_{c_0+1} = \cdots = \lambda_{c_1+c_2} = \varpi_2,\ \dots \\ \dots\ ,\ \lambda_{c_1+\cdots+c_{q-1}+1} = \ldots = \lambda_{c_1+\cdots+c_q} = \varpi_q.$$

This yields

$$\sigma_p \;=\; \sum_{\ell=1}^{q} c_\ell \varpi_\ell^p. \tag{82}$$

We stress at this point that to any k-tuple of λ_ℓ in $\mathbb{C}^k$, we can associate, maybe after a permutation θ of its components, such q-tuples of $c = (c_r)$ and $\varpi = (\varpi_r)$, with $\varpi_i \neq \varpi_j$ if $i \neq j$. It follows that the function η can be decomposed as follows:

$$(\lambda_\ell) \underbrace{\;\mapsto\; \theta(\lambda_\ell) \;\mapsto\; (c, \varpi) \;\mapsto\; (\sigma_p) \;\mapsto\; (s_i) \;\mapsto\;}_{\eta} (b_j).$$

This way, given a permutation θ and a root multiplicity vector c, with no zero component, we have defined a function $\gamma : \mathbb{C}^q \setminus \{\varpi_i = \varpi_j\} \to \mathbb{C}^k$ which maps the ϖ_r into the b_j:

$$\gamma : \varpi \;\mapsto\; (\sigma_p) \;\mapsto\; (s_i) \;\mapsto\; (b_j).$$

This function has rank q. Indeed, we know that the last two functions above are diffeomorphisms and, for the first one, we get from (82)

$$\frac{\partial \sigma_p}{\partial \varpi_r} \;=\; p\, c_r\, \varpi_r^{p-1}.$$

Since the p and c_r are not zero, we see that the matrix $(\frac{\partial \sigma_p}{\partial \varpi_r})$ has full rank since it has a Vandermonde structure and the ϖ_r are different. Consequently, the jacobian matrix $(\frac{\partial b_j}{\partial \varpi_r})$ of the function γ has full rank $q \leq k$. It follows from [6, Theorem III, 4.12, Theorem III, 5.5] that for each q-tuple ϖ in $\mathbb{C}^q \setminus \{\varpi_i = \varpi_j\}$, there exists a strictly positive real number $\epsilon(\varpi)$ such that

- $\mathcal{B}_{\epsilon(\varpi)}(\varpi)$ is a subset of $\mathbb{C}^q \setminus \{\varpi_i = \varpi_j\}$,
- $\gamma\left(\mathcal{B}_{\epsilon(\varpi))}(\varpi)\right)$ is a regular submanifold of (the real manifold) $\mathbb{C}^k$,
- the restriction of $\gamma : \mathcal{B}_{\epsilon(\varpi)}(\varpi) \to \gamma\left(\mathcal{B}_{\epsilon(\varpi)}(\varpi)\right)$ is a diffeomorphism.

We denote by γ^{-1} the "inverse" function.

The family $\left\{\mathcal{B}_{\epsilon(\varpi)}(\varpi)\right\}_{\varpi \in \mathbb{C}^q \setminus \{\varpi_i = \varpi_j\}}$ is a covering by open subsets of $\mathbb{C}^q \setminus \{\varpi_i = \varpi_j\}$. So there exists a countable family $(\varpi^i)_{i \in \mathbb{N}}$ such that the family $\left\{\mathcal{B}_{\epsilon(\varpi^i)}(\varpi^i)\right\}_{i \in \mathbb{N}}$ is covering by open subsets of $\mathbb{C}^q \setminus \{\varpi_i = \varpi_j\}$. Moreover, since to each k-tuple (b_j) in $\mathbb{C}^k$ we can associate a pair (c, ϖ) with $\varpi_i \neq \varpi_j$ if $i \neq j$, any such k-tuple (b_j) is in at least one set $\gamma\left(\mathcal{B}_{\epsilon(\varpi^i)}(\varpi^i)\right)$. So, since the number of permutations θ in $\mathbb{C}^k$ and multiplicity vectors c is finite, with varying i and q, we get a countable family $\{M_m\}_{m \in \mathbb{N}}$ of regular submanifold of $\mathbb{C}^k$ defined as

$$M_m \ := \ \gamma\left(\mathcal{B}_{\epsilon(\varpi^i)}(\varpi^i)\right)$$

and a countable family of C^1 functions ρ_m defined as

$$\begin{array}{ccccccc} & & \gamma^{-1} & & & (c,\theta) & \\ \rho_m \ : \ (b_0, \ldots, b_{k-1}) \in M_m & \mapsto & \varpi \in \mathcal{B}_{\epsilon(\varpi^i)}(\varpi^i) & \mapsto & (\lambda_1, \ldots, \lambda_k) \in \mathbb{C}^k \ . & & \end{array}$$

Each $b = (b_0, \ldots, b_{k-1})$ in $\mathbb{C}^k$ is in least one M_m and we have

$$\eta(\rho_m(b)) = b.$$

Our result follows then from

$$\mathbb{C}^k \ = \ \bigcup_{m \in \mathbb{N}} \bigcup_{b \in M_m} \{\rho_m(b)\}.$$

Acknowledgments. This paper was initiated by many discussions with Alessandro Astolfi about the gap between the Kazantzis–Kravaris/Luenberger observer, introduced as a local object in [13], and the global observer in [4]. We are also very grateful to Jean-Michel Coron, who was our guide all along our way in the quite involved technicalities needed in the proof of most of our results.

REFERENCES

[1] D. Aeyels, *Generic observability of differentiable systems*, SIAM J. Control Optim., 19 (1981), pp. 595–603.

[2] D. Angeli and E. Sontag, *Forward completeness, unboundedness observability, and their Lyapunov characterizations*, Systems Control Lett., 38 (1999), pp. 209–217

[3] V. Andrieu and L. Praly, *Remarks on the existence of a Kazantzis-Kravaris/Luenberger observer*, in Proceedings of the 43rd IEEE Conference on Decision and Control, 2004.

[4] A. Astolfi and L. Praly, *Global complete observability and output-to-state stability imply the existence of a globally convergent observer*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003.

[5] R. Benedetti and J.-J. Risler, *Real Algebraic and Semi-Algebraic Sets*, Hermann, Paris, 1990.

[6] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
[7] J.-M. Coron, *On the stabilization of controllable and observable systems by output feedback law*, Math. Control Signals Systems, 7 (1994), pp. 187–216.
[8] P. Deheuvels, *L'Intégrale*, Presses Universitaires de France, Paris 1980.
[9] H. Federer, *Geometric Measure Theory*, Classics in Math., Vol. XVI, Springer-Verlag, New York, 1996.
[10] J. P. Gauthier, H. Hammouri, and I. Kupka, *Observers for nonlinear systems*, in Proceedings of the 30th IEEE Conference on Decision and Control, Vol. 2, 1991, pp. 1483–1489.
[11] J. P. Gauthier and I. Kupka, *Deterministic Observation Theory and Applications*, Cambridge University Press, Cambridge, UK, 2001.
[12] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Graduate Texts in Math. 14, Springer-Verlag, New York, 1973.
[13] N. Kazantzis and C. Kravaris, *Nonlinear observer design using Lyapunov's auxiliary theorem*, Systems Control Lett., 34 (1998), pp. 241–247.
[14] G. Kreisselmeier and R. Engel, *Nonlinear observers for autonomous Lipschitz continuous systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 451–464.
[15] A. Krener and M. Xiao, *Observers for linearly unobservable nonlinear systems*, Systems Control Lett., 46 (2002), pp. 281–288.
[16] A. Krener and M. Xiao, *Nonlinear observer design for smooth systems*, Chaos in Automatic Control, W. Perruquetti and J.-P. Barbot, eds., Control Engineering, Vol. 18, Taylor & Francis Group, NY, 2005.
[17] D. Luenberger, *Observing the state of a linear system*, IEEE Trans. Military Electronics, MIL-8, (1964), pp. 74–80.
[18] A. Rapaport and A. Maloum, *Design of exponential observers for nonlinear systems by embedding*, Internat. J. Robust Nonlinear Control, 14 (2004), pp. 273–288.
[19] W. Respondek, A. Pogromsky, and H. Nijmeijer, *Time scaling for observer design with linearizable error dynamics*, Automatica, 40 (2004), pp. 277–285.
[20] G. de Rham, *Variétés Différentiables*, Hermann, Paris, 1960.
[21] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[22] E. D. Sontag, *For differential equations with r parameters, $2r+1$ experiments are enough for identification*, J. Nonlinear Sci., 12 (2002), pp. 553–583.
[23] F. Takens, *Detecting Strange Attractors in Turbulence*, in Proceedings of the Symposium on Dynamical Systems and Turbulence, University of Warwick, 1979–1980, D. A. Rand and L. S. Young, eds., Springer, Berlin, 1981.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 457–482

DUBINS' PROBLEM ON SURFACES II: NONPOSITIVE CURVATURE*

MARIO SIGALOTTI† AND YACINE CHITOUR‡

Abstract. Let M be a complete, connected, two-dimensional Riemannian manifold with nonpositive Gaussian curvature K. We say that M satisfies the *unrestricted complete controllability* (UCC) property for the Dubins problem if the following holds: given any (p_1, v_1) and (p_2, v_2) in TM, there exists, for every $\varepsilon > 0$, a curve γ in M, with geodesic curvature smaller than ε, such that γ connects p_1 to p_2 and, for $i = 1, 2$, $\dot\gamma$ is equal to v_i at p_i. Property UCC is equivalent to the complete controllability of a family of control systems of Dubins' type, parameterized by ε. It is well known that the Poincaré half-plane does not verify property UCC. In this paper, we provide a complete characterization of the two-dimensional nonpositively curved manifolds M, with either uniformly negative or bounded curvature, that satisfy property UCC. More precisely, if $\sup_M K < 0$ or $\inf_M K > -\infty$, we show that UCC holds if and only if (i) M is of the first kind or (ii) the curvature satisfies a suitable integral decay condition at infinity.

Key words. controllability, geometric control, geodesic curvature, nonpositive curvature

AMS subject classifications. 14H55, 53C21, 93B05, 93B27

DOI. 10.1137/040619739

1. Introduction.

Let (M, m) be a connected, oriented, complete Riemannian manifold and let $N = UM$ be its unit tangent bundle. Points of N are pairs (p, v), where $p \in M$, $v \in T_pM$, and $m(v, v) = 1$. Given $\varepsilon > 0$, Dubins' problem consists of finding, for every $(p_1, v_1), (p_2, v_2) \in N$, a curve $\gamma : [0, T] \to M$, parameterized by arc-length and whose geodesic curvature is bounded by ε, such that $\gamma(0) = p_1$, $\dot\gamma(0) = v_1$, $\gamma(T) = p_2$, $\dot\gamma(T) = v_2$, and which minimizes T. Although Dubins' problem makes sense in any dimension, the present paper deals only with the two-dimensional case, which can be seen as the time-optimal control problem for the following control system:

$$(D_\varepsilon): \qquad \dot q = f(q) + ug(q),\ q \in N,\ u \in [-\varepsilon, \varepsilon],$$

where f is the geodesic spray on N (i.e., the infinitesimal generator of the geodesic flow on M), g is the smooth vector field generating the fiberwise rotation with angular velocity equal to one, and the admissible controls are measurable functions $u : [0, T] \to [-\varepsilon, \varepsilon]$.

In the robotics literature (cf. [20]), the dynamics defined by (D_ε) when M is the Euclidean plane represent the motion of a unicycle (or a rolling penny) and the projections of the trajectories of (D_ε) on the plane are planar curves parameterized by arc-length with curvature bounded by ε. In 1957, Dubins [9] determined the global structure of the corresponding time-optimal trajectories: he showed that they are concatenations of at most three pieces made of circles of radius $1/\varepsilon$ or straight lines. Further restrictions on the length of the arcs of an optimal concatenation have been obtained by Sussmann and Tang [30].

*Received by the editors November 22, 2004; accepted for publication (in revised form) August 25, 2005; published electronically April 7, 2006. This work was partially supported through a European Community Marie Curie Fellowship, contract ALTOT.

http://www.siam.org/journals/sicon/45-2/61973.html

†Institut Élie Cartan, UMR 7502 UHP/CNRS/INRIA, POB 239, 54506 Vandœuvre-lés-Nancy, France (mario.sigalotti@inria.fr).

‡Laboratoire des Signaux et Systèmes, Université Paris Sud-CNRS-Supélec, Gif-sur-Yvette, France (chitour@lss.supelec.fr).

Dubins'-like problems have been proposed by considering more general manifolds M. For instance, the case where M is a two-dimensional manifold of constant Gaussian curvature was investigated in [5, 6, 12, 17], and the case where $M = \mathbf{R}^n, S^n$, $n \geq 3$, was studied in [18, 19, 29]. The controllability issue is not difficult to solve, because Dubins' problem can be reformulated as a control system (left-)invariant under the action of a Lie group. Let us mention another line of generalization, which consists of considering the distributional version of (D_ε), whose best-known example is the so-called Reeds–Shepp car [22]. In dimension two, the distributional dynamics can be represented by the two-input control system $(DD_\varepsilon) : \dot{q} = uf(q) + vg(q)$ with $|u|, |v| \leq \varepsilon$. The controllability of (DD_ε) simply results from the fact that the distribution (f, g) is strong bracket generating, i.e., for every $q \in N$, the triple $(f(q), g(q), [f, g](q))$ spans T_qN.

A motivation for the present work is the remark that when generalizing Dubins' problem from the Euclidean plane to a general Riemannian surface M, the complete controllability of (D_ε) stops to be a trivial issue. Indeed, when M has no particular symmetry, it is not clear how to design a trajectory starting from and ending at the fiber of a given point $p \in M$, with "small" geodesic curvature. We are naturally faced with an intrinsic property of M, independent of global rescaling of its metric m, that is, whether (D_ε) is completely controllable for every $\varepsilon > 0$. We refer to such property as the *unrestricted complete controllability* (UCC) of Dubins' problem on M. The present paper primarily focuses on the controllability of Dubins' problem, and our main scope is to find geometric or topological conditions on M characterizing UCC.

The Lie algebraic structure of (D_ε) reveals quickly that the Gaussian curvature K plays a crucial role in the characterization of controllable Dubins' problems. Indeed, as it follows from the structural equations satisfied by Cartan's connections (cf. Spivak [28, Volume II]),

$$[f, [f, g]](q) = -K(\pi(q))g(q) \tag{1}$$

for every $q \in N$, where $\pi : N \to M$ denotes the bundle projection. Notice that the bracket relation (1) identifies K uniquely. A generalized version of it was proposed by Agrachev and Sachkov ([1]; see also [26]) in order to introduce a notion of curvature for two-dimensional optimal control systems.

The role of K in the study of UCC of Dubins' problem is further suggested by the following standard example of noncontrollability: if M is the Poincaré half-plane H^2, then (D_ε) is completely controllable if and only if $\varepsilon > 1$ (cf. [6, 17]). Roughly speaking, the negativeness of K is an obstacle for the controlled turning action (modeled by g) to overcome the spreading of the geodesics. When $|u| \leq \varepsilon \leq 1$, the effect of ug is not strong enough and complete controllability fails to hold.

In a previous paper [7], we proposed some geometric and topological conditions on M, which ensure that the UCC property holds true. In particular, it was proved that any compact M satisfies the UCC property. In the unbounded case, the same conclusion was obtained when M is *asymptotically flat*, i.e., K tends to zero at infinity. The control strategy goes as follows. It is first shown that the controllability of (D_ε) is equivalent to the fact that every $q = (p, v) \in N$ can be steered to $q^- = (p, -v)$ by an admissible trajectory of (D_ε). The latter property, called *weak controllability* of (D_ε), is proved by tracking a teardrop loop (see Figure 1) in a covering domain over a piece of M at infinity. A suitable covering manifold can be globally described by a single appropriate geodesic chart.

If one keeps in mind the noncontrollability of Dubins' problem on H^2, it is reasonable to consider the situation where $K \geq 0$. Indeed, in that case, no local spreading

effect due to the drift term has to be compensated. Moreover, if M is an open manifold with nonnegative curvature, then the theorems of Cohn-Vossen [8] and Huber [11] imply that

$$\int_M K dA < \infty, \tag{2}$$

where dA is the surface element in M (see [7] for details). The integral decay of K to zero at infinity can be interpreted as a sort of asymptotic flatness condition and it suggests that (D_ε) should be completely controllable for every $\varepsilon > 0$. In [7], we were able to confirm that intuition under the additional assumption that K is bounded over M, i.e., $\sup_M K$ is finite, and also to extend that result to the case where $K \geq 0$ outside a compact set. The strategy of the proof still consists in tracking a teardrop loop in a covering domain D but the construction of D becomes much more delicate than that of the asymptotic flat case.

Recently, Pansu pointed out how the existence of a closed geodesic on a surface M with constant negative curvature implies that for every point $p \in M$ and every $\varepsilon > 0$, there exists a noncontractible admissible trajectory of (D_ε) which steers some point $q \in U_p M$ to the point q^-. Such a feature is particularly interesting, taking into account that UCC does not hold for the Poincaré half-plane H^2. An immediate question is whether UCC can be recovered for a quotient manifold $M = H^2/\Gamma$, provided that we add enough topology to H^2 (i.e., provided that the group Γ is large enough), or if, on the contrary, Dubins' problem on any quotient of H^2 is not UCC. It turns out that the answer can be easily stated: $M = H^2/\Gamma$ has the UCC property if and only if it is of the first kind, i.e., its limit set (see Definition 3.1) is the entire boundary at infinity of H^2.

In the present work, we investigate the more general case of Riemannian surfaces with nonpositive curvature. It is well known that a surface M of such kind can be identified with the quotient space X/Γ, where X is a Hadamard surface (i.e., a simply connected, complete Riemannian manifold of nonpositive curvature) and Γ is a group of orientation-preserving isometries which acts freely and discontinuously on X. (For all the definitions, see section 3.) We provide conditions on M, sufficient for UCC to hold. The first one generalizes the hyperbolic case. It says that (i) $M = X/\Gamma$ is of the first kind, i.e., its limit set $\mathcal{L}(\Gamma)$ is equal to the ideal boundary $X(\infty)$ of X. The second condition is reminiscent of the nonnegative curvature case, namely, it states that (ii) for every $r > 0$ and every sector S in X, $\sup_{p \in S} \int_{B_X(p,r)} K dA = 0$, where $B_X(p, r)$ denotes the ball of center p and radius r. Then, we present necessary conditions on M for property UCC to hold. A quite surprising result is the following. Under the assumption that either K is bounded or $\sup_M K < 0$, if M verifies property UCC, then either condition (i) or condition (ii) must hold true. In that way, we exactly characterize the surfaces M, with nonpositive curvature K either bounded or such that $\sup_M K < 0$, verifying property UCC. In addition, some controllability and noncontrollability results in the case where K is nonpositive outside a compact subset of M are given. We also provide sufficient conditions ensuring that M has the UCC property, with no sign assumption on K, namely, (a) M has finite area, (b) the geodesic flow on M is topologically transitive. We conclude with some remarks on the structure of time-optimal trajectories for (D_ε). If $\sup_M K \leq -\varepsilon$, we show that they follow Dubins' pattern, namely, that they are concatenation of a bang, a singular, and a bang arc (where some arc can possibly have zero length), with the (possible) singular arc being a geodesic of the surface.

The paper is organized as follows. In section 2, we formulate the control problem at hand and describe its basic features. In section 3, general facts on Riemannian surfaces of nonpositive curvature are recalled. The main results of the paper are stated in section 4 and their proofs are provided in section 5. Finally, section 6 contains some remarks on time-optimal trajectories.

2. Formulation of the problem. Let (M, m) be a two-dimensional Riemannian manifold, or, as we will equivalently call it, a *Riemannian surface*. Notice that M is not required to be embedded nor immersed in $\mathbf{R}^3$. Assume that M is oriented and that its metric m is complete. Denote by N the unit tangent bundle $UM = \{q \in TM |\ m(q,q) = 1\}$ and by $\pi : N \to M$ the canonical bundle projection of N onto M. Let K be the Gaussian curvature on M. We will use the symbol K also for the trivial extension (constant on fibers) of K on N. The distance on M induced by m is denoted by $d_M(\cdot,\cdot)$ (or $d(\cdot,\cdot)$ when no confusion is possible), and, for every $p \in M$ and $r > 0$, $B_M(p, r)$ stands for the ball of center p and radius r.

Let f be the *geodesic spray* on TM, whose restriction to N (still denoted by f) is a well-defined vector field on N. Recall that f is characterized by the following property: $p(\cdot)$ is a geodesic on M if and only if $(p(\cdot), \dot{p}(\cdot))$ is an integral curve of f. In particular, f satisfies the relation

$$\pi_\star(f(q)) = q \tag{3}$$

for every $q \in N$. The *exponential map on* M is defined by

$$\exp(t, q) = \pi(e^{tf}(q)), \tag{4}$$

where $e^{tf} : N \to N$ denotes the flow of the vector field f at time t.

Let g be the smooth vector field on N, whose corresponding flow at time t is the fiberwise rotation of angle t. For every $q \in N$, we set $q^- = e^{\pi g}(q)$ and $\mathcal{R}q = e^{-\frac{\pi}{2}g}(q)$.

For $\varepsilon > 0$, let $(\mathcal{D}_\varepsilon^M)$ be the control system

$$(\mathcal{D}_\varepsilon^M): \qquad \dot{q} = f(q) + ug(q), \quad q \in N, \quad u \in [-\varepsilon, \varepsilon].$$

An *admissible control* is a measurable function $u(\cdot)$, defined on some interval of $\mathbf{R}$, with values in $[-\varepsilon, \varepsilon]$. The solutions of $(\mathcal{D}_\varepsilon^M)$ corresponding to admissible controls are called *admissible trajectories*. (Sometimes, to prevent any confusion, we will speak also of *ε-admissible controls* and *ε-admissible trajectories*).

For every $q \in N$ and $T > 0$, the *attainable set from q up to time T* is the set $A_q^T = A_q^T(M, \varepsilon)$ consisting of the endpoints of all admissible trajectories for $(\mathcal{D}_\varepsilon^M)$, starting from q, of length smaller than or equal to T. We also write

$$A_q = A_q(M, \varepsilon) = \cup_{T>0} A_q^T(M, \varepsilon).$$

The control system $(\mathcal{D}_\varepsilon^M)$ is called *completely controllable* if $A_q = N$ for every $q \in N$.

Definition 2.1. *We say that the Dubins problem on M has the* unrestricted complete controllability (UCC) property *(or, equivalently, that it is* UCC*) if, for every $\varepsilon > 0$, $(\mathcal{D}_\varepsilon^M)$ is completely controllable.*

We stress that the UCC notion corresponds to the property of existence of solutions to the original, control-free Dubins problem on M: UCC holds if and only if, for every $\varepsilon > 0$ and every (p_1, v_1), (p_2, v_2) in TM, there exists a curve $\gamma : [0, T] \to M$ with geodesic curvature smaller than ε such that $\gamma(0) = p_1$, $\gamma(T) = p_2$ and $\dot{\gamma}(0) = v_1$,

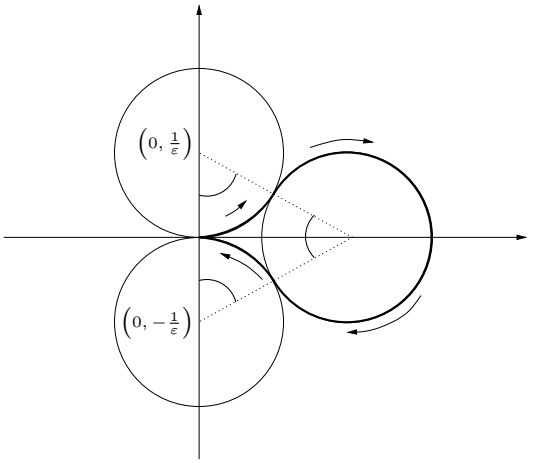


FIG. 1. *The teardrop trajectory of size* $1/\varepsilon$.

$\dot\gamma(T) = v_2$. Up to a reparameterization by arc-length, $(\gamma, \dot\gamma)$ is an ε-admissible trajectory in N, whose corresponding control $u(t)$ is equal, at almost every $t \in [0, T]$, to the geodesic curvature of γ at the point $\gamma(t)$.

It is easy to check that the distribution $\{f, g\}$ is *bracket generating*, i.e., that, for every $q \in N$, $T_qM = \text{span}\{X(q) |\ X \in \text{Lie}(f, g)\}$, where $\text{Lie}(f, g)$ denotes the Lie algebra of vector fields on N generated by f and g. Indeed, for every $q \in N$, we have that $\pi_*([f, g](q)) = \pi_*(f(\mathcal{R}q))$. Therefore, f, g, and $[f, g]$ span T_qM at every $q \in N$, proving not only that $\{f, g\}$ is bracket generating but also that it is a *contact distribution* on N. As a consequence, for every $0 < t < T$ and every $q \in N$,

$$e^{tf}(q) \in \text{Int}(A_q^T), \tag{5}$$

where $\text{Int}(A_q^T)$ denotes the interior of A_q^T. This follows, for instance, from the description of small-time attainable sets for single-input nondegenerate three-dimensional control systems given by Lobry in [15].

Since f, g, and $[f, g]$ are linearly independent at every point, they can be used to introduce a metric on N, requiring $(f(q), g(q), [f, g](q))$ to be an orthonormal basis of T_qN, for every $q \in N$. Such metric endows N with a complete Riemannian structure (see, for instance, [23]). In accordance with the notations introduced above, we will denote by $d_N(\cdot, \cdot)$ the induced distance on N and, for every $q \in N$ and $\rho > 0$, by $B_N(q, \rho)$ the ball of center q and radius ρ.

Remark 2.2. In [7] we noticed that, since (D_ε^M) has the property of being bracket generating, Dubins' problem on M is UCC if and only if for every $\varepsilon > 0$ and every $q \in N$, there exists $\hat q \in A_q(M, \varepsilon)$ such that $\hat q^- \in A_{\hat q}(M, \varepsilon)$. In order to prove this last property, we will often mimic on M the behavior of a teardrop trajectory on the Euclidean plane. We call teardrop trajectory of size $1/\varepsilon$ (see Figure 1) the bang-bang trajectory of $(D_\varepsilon^{\mathbf{R}^2})$ whose control u, defined by

$$u(t) = \begin{cases} \varepsilon & \text{if} \quad 0 \le t \le \frac{\pi}{3\varepsilon}, \\ -\varepsilon & \text{if} \quad \frac{\pi}{3\varepsilon} < t \le \frac{2\pi}{\varepsilon}, \\ \varepsilon & \text{if} \quad \frac{2\pi}{\varepsilon} < t \le \frac{7\pi}{3\varepsilon}, \end{cases}$$

steers $(1, 0) \in U_{(0,0)}\mathbf{R}^2$ to $(-1, 0) \in U_{(0,0)}\mathbf{R}^2$.

Remark 2.3. If M_1, M_2 are two Riemannian manifolds and $P : M_1 \to M_2$ is a local isometry at every point of M_1, then every admissible trajectory for $(D_\varepsilon^{M_1})$ is transformed by $P_* : UM_1 \to UM_2$ in an admissible trajectory for $(D_\varepsilon^{M_2})$. In particular, if P is onto and the Dubins problem on M_1 is UCC, then the same is true for the Dubins problem on M_2.

3. General facts on Riemannian surfaces of nonpositive curvature. Of special interest for the present study are Riemannian surfaces of nonpositive curvature. We collect in this section some definitions and known results about them, which will be used in what follows. When no explicit source is given, see [4].

A simply connected, complete Riemannian manifold of nonpositive sectional curvature is called a *Hadamard manifold*. If M is a complete, connected, oriented Riemannian manifold of nonpositive sectional curvature, then M can be described as X/Γ, where X is a Hadamard manifold and Γ is a group of orientation-preserving isometries which acts freely and discontinuously on X. We will denote by $\Pi : X \to M$ the canonical projection of X onto M. When the sectional curvature is constant on M, then M is said a *hyperbolic manifold* and X a *hyperbolic space*.

3.1. Hadamard surfaces. From now on, X will denote a Hadamard surface, that is, a Hadamard two-dimensional manifold. If $M = X/\Gamma$ is a complete Riemannian surface, then we will denote by the same letters f, g, K, m, and π the corresponding objects on X and M. This is motivated by the fact that $\Pi : X \to M$, being a local isometry, identifies such objects, at least at a local level.

It is well known that X is diffemorphic to $\mathbf{R}^2$. Any geodesic segment in X is the unique length-minimizing trajectory between its extremes. In particular, all complete geodesics are simple.

A *ray* is a half-geodesic on X. A *sector* is a region of X bounded by two distinct rays starting at the same point, which is called the *vertex of* S.

The *ideal boundary* of X, denoted by $X(\infty)$, is defined as the quotient of the set of all rays parameterized by arc-length by the equivalence relation

$$c_1 \sim c_2 \iff \limsup_{t\to\infty} d_X(c_1(t), c_2(t)) < \infty.$$

The equivalence class of a parameterized ray c is denoted by $c(+\infty)$ and it is called the *endpoint of* c. If $c : \mathbf{R} \to X$ is a parameterized complete geodesic, then $c(-\infty)$ denotes the equivalence class of $[0, \infty) \ni t \mapsto c(-t)$.

For a given point $p \in X$ there is a one-to-one correspondence ψ_p between U_pX and $X(\infty)$, which assigns to $v \in U_pX$ the equivalence class of $[0, \infty) \ni t \mapsto \exp(t, v)$. The correspondence defines a topology on $X(\infty)$, which is called the *sphere topology*. The sphere topology extends to the so-called cone topology on $\overline{X} = X \cup X(\infty)$. The cone topology is generated by the open sets of X and the sets

$$\psi_p(U) \cup (\cup_{t>0} \exp(t, U)),$$

where $p \in X$ and U is an open set of U_pX. Notice that the action on X of an element of Γ has a natural continuous extension on $\overline{X}$.

Given a set Ω in X, we write $\partial\Omega$ for the boundary of Ω in X, while $\Omega(\infty)$ will denote the intersection between $X(\infty)$ and the closure of Ω in $\overline{X}$.

The isometric transformations of X can be classified in terms of the so-called *displacement function* $X \ni p \mapsto d_X(p, \gamma(p))$. An isometry $\gamma : X \to X$ is called *elliptic* if it has at least one fixed point in X; *hyperbolic* if the displacement function attains its minimum in X, and such minimum is strictly positive; *parabolic* otherwise.

3.2. Closed geodesics and limit sets. If $M = X/\Gamma$ is a complete Riemannian surface, then the only elliptic element of Γ is the identity. Let G be a closed geodesic in M. Fix one of its lifts $\widetilde{G}$ in X and let $c : \mathbf{R} \to X$ be a parameterization of $\widetilde{G}$ by arc-length. There is one isometry $\gamma \in \Gamma$ such that $\gamma(c(0)) = c(T)$, where T is the length of G. Then $c(t+T) = c(t)$ for every $t \in \mathbf{R}$. (The proof goes as in the hyperbolic case; see [21, Theorem 9.6.2].) The isometry γ is hyperbolic (see [4, section 6]) and $\widetilde{G}$ is called an *axis* of γ. Actually, every hyperbolic isometry γ has at least one axis. (Given any point p at which the displacement function attains its minimum, the geodesic line between p and $\gamma(p)$ is indeed an axis.) If none of the half-planes bounded by $\widetilde{G}$ is flat, then, given a neighborhood U of $c(-\infty)$ and a neighborhood V of $c(+\infty)$ in $\overline{X}$, there exists $\bar{n} \in \mathbf{N}$ such that

$$\gamma^n\left(\overline{X} \setminus U\right) \subset V, \qquad \gamma^{-n}\left(\overline{X} \setminus V\right) \subset U, \tag{6}$$

for every $n \geq \bar{n}$, as proved by Ballmann in [3].

The existence of a closed geodesic on Riemannian surfaces is established by some classical results under very general assumptions. Lyusternik and Fet [16] proved that all complete compact Riemannian surfaces contain at least one closed geodesic. After that, Thorbergsson [31] extended the result to all complete, connected Riemannian surfaces neither homeomorphic to the plane nor to the cylinder. As a consequence, if Γ is not cyclic, then $M = X/\Gamma$ contains a closed geodesic.

DEFINITION 3.1. *Let Γ be a group of isometries of X. Fix $p \in X$ and consider $\overline{\Gamma(p)}$, the closure in $\overline{X}$ of the Γ-orbit of p. The set $\mathcal{L}(\Gamma) = \overline{\Gamma(p)} \cap X(\infty)$ is called the limit set of Γ.*

The definition of $\mathcal{L}(\Gamma)$ is actually independent of the choice of the point p. Fix, indeed, p, p' in X, and let $\gamma_n(p) \to z \in X(\infty)$ as $n \to \infty$, with $\{\gamma_n\}_{n\in\mathbf{N}} \subset \Gamma$. Notice that for every $n \in \mathbf{N}$, $d(\gamma_n(p), \gamma_n(p')) = d(p, p')$. Moreover, every sector S of X such that $z \in S(\infty)$ contains a subsector S' such that $z \in S'(\infty)$ and whose distance from the boundary of S is larger than $d(p, p')$. Therefore, in the cone topology of $\overline{X}$, $\gamma_n(p') \to z$ as $n \to \infty$.

DEFINITION 3.2. *Let X be a Hadamard surface and $M = X/\Gamma$ be a complete Riemannian surface. We say that M is of the first kind if $\mathcal{L}(\Gamma) = X(\infty)$; otherwise we say that M is of the second kind.*

Remark 3.3. If Γ is cyclic, then M is of the second kind. Indeed, if $\Gamma = \{\mathrm{Id}\}$, then $\mathcal{L}(\Gamma)$ is empty. If Γ is nontrivial, then it is generated by a hyperbolic or a parabolic isometry, denoted by γ. If γ is hyperbolic, then it translates at least one geodesic, hence $\mathcal{L}(\Gamma)$ is made of the two endpoints of such geodesic (just consider the orbit of a point on the geodesic). If γ, instead, is parabolic, then there exists a point $z \in X(\infty)$ which is invariant under the action of γ, together with the horospheres centered at z (see [4, Lemma 6.6]). Fix any point $x \in X$ and consider a horosphere U centered at z which contains x. Then $\mathcal{L}(\Gamma)$ is contained in $U(\infty)$. In both cases, M is not of the first kind. Therefore, if M is of the first kind, then it contains a closed geodesic.

DEFINITION 3.4. *Let M be a Riemannian surface. We say that its Gaussian curvature K is uniformly negative if* $\sup_M K < 0$.

If K is uniformly negative on X, then, for every two elements x, y of $X(\infty)$, there exists a parameterized geodesic $c : \mathbf{R} \to X$ such that $c(+\infty) = x$ and $c(-\infty) = y$. That is, using the terminology introduced by Eberlein and O'Neill in [10], X is a *visibility manifold*.

3.3. Geodesic coordinates. Geodesic coordinates can be globally defined on X. They depend on the choice of an element q of UX and are defined through the map

$$\begin{array}{rccl} \phi_q : & \mathbf{R}^2 & \longrightarrow & X, \\ & (x,y) & \longmapsto & \pi(e^{yf} \circ e^{\frac{\pi}{2} g} \circ e^{xf}(q)). \end{array} \tag{7}$$

Let $B : \mathbf{R}^2 \to \mathbf{R}$ be the solution of the system

$$\begin{cases} B_{yy} + KB = 0, \\ B(x,0) \equiv 1, \\ B_y(x,0) \equiv 0, \end{cases} \tag{8}$$

in which the index y appearing in B_y, B_{yy} stands for the partial differentiation with respect to y. Notice that since $K \leq 0$, B is globally defined and

$$B(x,y) \geq 1, \tag{9}$$

$$yB_y(x,y) \geq 0 \tag{10}$$

for every $(x,y) \in \mathbf{R}^2$. The coordinate expression for the metric m on X is given by

$$m(x,y) = B^2(x,y)dx^2 + dy^2.$$

(See, for instance, [13].) The unit bundle UX can be identified with

$$\left\{ \left(x, y, \frac{\cos\theta}{B(x,y)}, \sin\theta \right) \in \mathbf{R}^4 \,\middle|\, (x,y) \in \mathbf{R}^2,\ \theta \in \mathcal{S}^1 \right\}.$$

In the coordinates (x,y,θ), f and g are given by

$$f(x,y,\theta) = \left(\frac{\cos\theta}{B(x,y)}, \sin\theta, F(x,y)\cos\theta \right)^T, \qquad g(x,y,\theta) = (0,0,1)^T, \tag{11}$$

where

$$F(x,y) = \frac{B_y(x,y)}{B(x,y)}. \tag{12}$$

Equivalently, (D_ε^X) can be written as follows:

$$\dot{x} = \frac{\cos\theta}{B}, \tag{13}$$

$$\dot{y} = \sin\theta, \tag{14}$$

$$\dot{\theta} = u + F\cos\theta. \tag{15}$$

Notice that, as it can be easily deduced from (8), F satisfies the Cauchy problem

$$F_y = -K - F^2, \qquad F(x,0) \equiv 0. \tag{16}$$

If $M = X/\Gamma$ is a complete Riemannian surface, then, for any fixed $q \in N$, the map ϕ_q, defined as in (7), is a local diffeomorphism at every point of $\mathbf{R}^2$. The pullback of

the metric m through ϕ_q endows $\mathbf{R}^2$ with a Riemannian structure, which renders $\mathbf{R}^2$ isomorphic to X.

Remark 3.5. The previous formulation of Dubins' problem in coordinates still makes sense when only local, instead of global, geodesic coordinates are defined. Consider the case of a Riemannian surface M with possibly sign-varying curvature. Recall that a *half-plane* of M is a simply connected open subset of M bounded by a simple open complete geodesic. If M contains a half-plane H and if K is nonpositive on H, then H admits a system of geodesic coordinates. Indeed, for any $q \in N$ such that $\phi_q(\mathbf{R}\times[0,\infty)) = H$, $\mathbf{R}\times[0,\infty)$ can be endowed with the metric $\phi_q^* m$, and Dubins' problem on H can be described by equations (13)–(15). Similarly, if $q = (p, v) \in N$ and $a, b, r > 0$ are such that $K \leq 0$ on $B_M(p, r)$ and $a + b < r$, then the rectangle $[-a, a] \times [-b, b] \subset \mathbf{R}^2$, endowed with $\phi_q^* m$, is a nonpositively curved covering domain of a neighborhood of p. Again, system (13)–(15) describes Dubins' problem on such nonflat rectangle.

4. Statement of the results. We collect here the statements of the main results proved in the next section, in order to stress the interrelations between the proposed necessary and sufficient conditions for the UCC property to hold.

PROPOSITION 4.1. *Let M be a complete, connected Riemannian surface. The Dubins problem on M has the* UCC *property if at least one of the following conditions is satisfied:* (*a*) *M has finite area or* (*b*) *the geodesic flow on M is topologically transitive.*

THEOREM 4.2. *Let M be a complete, connected Riemannian surface with nonpositive Gaussian curvature K. Denote by X the universal covering of M. The Dubins problem on M has the* UCC *property if at least one of the following conditions is satisfied:* (i) *M is of the first kind or* (ii) *for every $r > 0$ and every sector S in X, $\sup_{p\in S} \int_{B_X(p,r)} K dA = 0$.*

Conditions which are both sufficient and necessary for UCC can be formulated under more restrictive hypotheses on K.

THEOREM 4.3. *Let M be a complete, connected Riemannian surface and assume that its Gaussian curvature K is uniformly negative. Then the Dubins problem on M has the* UCC *property if and only if M is of the first kind.*

THEOREM 4.4. *Let M be a complete, connected Riemannian surface and assume that its Gaussian curvature K is nonpositive and bounded. Assume, moreover, that M is of the second kind and denote by X its universal covering. The Dubins problem on M has the* UCC *property if and only if, for every $r > 0$, for every sector S in X, $\sup_{p\in S} \int_{B_X(p,r)} K dA = 0$.*

The results stated in Theorem 4.2 are proved in Propositions 5.9 and 5.10. For what concerns Theorem 4.3, it is a combination of Proposition 5.9 and Corollary 5.15. Finally, the proof of Theorem 4.4 is split into the ones of Proposition 5.10 and Corollary 5.18. For results concerning manifolds with possibly sign-varying curvature, we refer to the detail of sections 5.2, 5.3, and 5.4.

5. Proofs of the results.

5.1. Additional properties of Riemannian surfaces of nonpositive curvature. In this paragraph, we prove some geometric properties of Riemannian surfaces of nonpositive curvature, which are at the heart of the arguments of the theorems stated above. In particular, we show how certain properties of limit sets which hold in the hyperbolic case can be recovered in the nonconstant curvature case, under suitable nonflatness assumptions. To this extent we propose the following definition.

DEFINITION 5.1. *A sector S of a Hadamard surface X is said to be heavy if, for every sector S' contained in S, $\int_{S'} KdA = -\infty$.*

For example, as we will often use, if $r, \delta > 0$ exist such that $\int_{B_X(p,r)} KdA \leq -\delta$ for every $p \in S$, then S is heavy. Heavy sectors enjoy the following visibility-type property.

LEMMA 5.2. *Let S be a heavy sector of a Hadamard surface X. Then, for every z_1 in the interior of $S(\infty)$ and every $z_2 \in X(\infty)$ different from z_1, there exists a geodesic $c : \mathbf{R} \to X$ such that $c(-\infty) = z_1$ and $c(+\infty) = z_2$.*

The lemma is basically a reformulation of [4, Exercise (i), p. 57]. For sake of completeness, let us provide its proof. Consider a convex sector S' bounded by two rays having z_1 and z_2 as endpoints. Since z_1 belongs to the interior of $S(\infty)$, then S' contains a subsector which is a included in S. Therefore, by definition of heavy sector, $\int_{S'} KdA = -\infty$. Denote by x the vertex of S' and let p_n, q_n be two sequences of points in $\partial S'$, tending, respectively, to z_1, z_2, as n tends to infinity. Denote by T_n the geodesic triangle with vertices x, p_n, and q_n. The Gauss–Bonnet theorem implies that $\int_{T_n} KdA > -\pi$ for every n. Therefore, there exists a compact subset C of S' such that, possibly passing to a subsequence, the geodesic segment between p_n and q_n intersects C for every n. By compactness of $UC = \{q \in UX |\ \pi(q) \in C\}$, there exists $q \in UC$ such that the geodesic $s \mapsto \exp(s, q)$ connects, as required, z_1 and z_2.

Given a Hadamard surface X and a group Γ of isometries of X, let us denote by $\mathcal{A}(\Gamma)$ the closure in $X(\infty)$ of the set of endpoints of axis of hyperbolic isometries contained in Γ.

LEMMA 5.3. *Let X be a Hadamard surface and Γ a group of isometries of X. If $\mathcal{A}(\Gamma)$ is nonempty, then, for every heavy sector S of X, $\mathcal{L}(\Gamma) \cap S(\infty) = \mathcal{A}(\Gamma) \cap S(\infty)$.*

Proof. Clearly $\mathcal{A}(\Gamma) \subset \mathcal{L}(\Gamma)$. Fix an axis G of a hyperbolic transformation of Γ, a point z in $S(\infty)$, and assume that there exists a neighborhood J of z which does not contain any endpoint of complete geodesics of the type $\gamma(G)$, $\gamma \in \Gamma$. Let H be a half-plane of X such that $z \in H(\infty) \subset J$, whose existence is guaranteed by Lemma 5.2. Then, the entire Γ-orbit of G is contained in $X \setminus H$. Therefore, z is not a density point of the Γ-orbit of any point of G, i.e., $z \notin \mathcal{L}(\Gamma)$. □

The following lemma concerns the existence of half-planes in nonpositively curved Riemannian surfaces. Its role will be crucial while proving the noncontrollability results contained in section 5.3.

LEMMA 5.4. *Let M be a complete, connected Riemannian surface with nonpositive curvature and let X be its universal covering. Assume that M is of the second kind and that X contains a heavy sector S. Then the canonical projection $\Pi : X \to M$ embeds in M a half-plane of X, contained in S.*

Proof. We start the proof by noticing that if there exists an open, nonempty, connected subset U of $S(\infty) \setminus \mathcal{L}(\Gamma)$ such that $\gamma(U) \cap U = \emptyset$ for every $\gamma \in \Gamma \setminus \{\mathrm{Id}\}$, then every half-plane H bounded by U and a geodesic with endpoints in U is embedded by Π in M. Indeed, $H \cap \gamma(H) \neq \emptyset$ implies that $H(\infty) \cap (\gamma(H))(\infty) = H(\infty) \cap \gamma(H(\infty)) \neq \emptyset$ and γ must therefore be the identity. The existence of a geodesic with endpoints in U follows from Lemma 5.2.

We distinguish two cases, depending on whether Γ contains a hyperbolic isometry.

Assume first that $\Gamma \setminus \{\mathrm{Id}\}$ is made of parabolic isometries only. In particular, Γ is cyclic, since M contains no closed geodesic. Denote by γ a generator of Γ. The set of fixed points of γ on $X(\infty)$ is closed and nonempty. If it does not contain $S(\infty)$, then there exists an open, nonempty, connected subset U of $S(\infty)$ such that $\gamma^j(U) \cap U = \emptyset$ for every $j \in \mathbf{Z} \setminus \{0\}$. Then the lemma follows from the above remark. We are

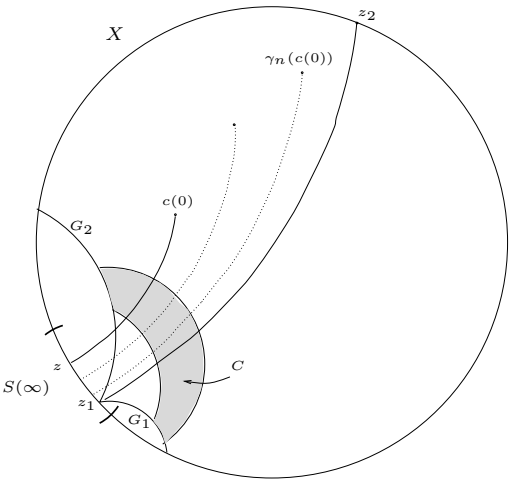


FIG. 2.

left to deal with the case in which every point of $S(\infty)$ is fixed by γ. Consider a parameterized geodesic $c : \mathbf{R} \to X$ such that $c(+\infty)$ and $c(-\infty)$ are both in $S(\infty)$. Since γ fixes the endpoints of c, then P_c, the set spanned by all the geodesics parallel to c, is invariant under the action of γ. The set P_c is isometric to $\mathbf{R} \times \Sigma$, where Σ is a closed geodesic line in X (see [4, Lemma 2.4]). By hypothesis, c bounds no flat half-plane; therefore Σ is bounded. Since γ induces a continuous transformation of Σ, there must exist at least one complete geodesic of X fixed by γ. Thus γ, which is a nonhyperbolic isometry of P_c, is the identity transformation, and the statement of the lemma is trivially true. (Notice that the argument, which follows the proof of Lemma 6.8 in [4], shows that γ has at most one fixed point in the interior of $S(\infty)$, unless $\Gamma = \{\mathrm{Id}\}$.)

Assume now that Γ contains at least one hyperbolic isometry. According to Lemma 5.3, then, $\mathcal{L}(\Gamma) \cap S(\infty) = \mathcal{A}(\Gamma) \cap S(\infty)$. Notice that by definition of heavy sector, for every complete geodesic G with at least one endpoint in the interior of $S(\infty)$, the half-planes bounded by G are not flat. Therefore, $\mathcal{L}(\Gamma) \cap S(\infty)$ has a nonempty interior, since otherwise (6) would imply that M is of the first kind. In particular, $S(\infty)$ is not contained in $\mathcal{L}(\Gamma)$. Fix $z \in S(\infty) \setminus \mathcal{L}(\Gamma)$. According to the remark at the beginning of the proof, it is enough to prove that z is an isolated point of the Γ-orbit of z in $X(\infty)$. To this extent, let us show that the Γ-orbit of z in $X(\infty)$ has no density point in the interior of $S(\infty) \setminus \mathcal{L}(\Gamma)$. Assume, by contradiction, that there exist z_1 in the interior of $S(\infty) \setminus \mathcal{L}(\Gamma)$ and a sequence γ_n of isometries in Γ such that $\gamma_n(z)$ tends to z_1 as n goes to infinity. Let $c : [0, \infty) \to X$ be a ray with $c(+\infty) = z$. Passing possibly to a subsequence, $\gamma_n(c(0))$ converges to a point z_2 of $\mathcal{L}(\Gamma)$. According to Lemma 5.2, there exist two geodesics G_1 and G_2, each of them having one endpoint in z_1, such that the region of X bounded by G_1 and G_2 contains $\gamma_n(c(0))$ for every n. Thus, there exists a compact subset C of X which intersects each ray $c_n(\cdot) = \gamma_n(c(\cdot))$, $n \in \mathbf{N}$ (see Figure 2). Therefore, we can fix a sequence t_n of positive real numbers such that for every $x \in X$, $c_n(t_n)$ has distance from x uniformly bounded with respect to n. Fix $x \in X$ and notice that

$$t_n = d(c_n(0), c_n(t_n)) \geq d(c_n(0), x) - d(x, c_n(t_n)),$$

which implies that t_n tends to infinity with n. Thus, $c(t_n)$ tends to z as n goes to infinity. On the other hand, $\gamma_n^{-1}(x)$ has distance from $c(t_n)$ uniformly bounded with respect to n. It follows that as n goes to infinity, $\gamma_n^{-1}(x)$ converges to z, which contradicts the fact that z does not belong to $\mathcal{L}(\Gamma)$. ∎

5.2. Controllability results. Let us begin the section with the following elementary remark. Consider a complete Riemannian surface M, with no restriction on the sign of its curvature K. Setting $t = 1$ and $T = 2$ in (5), we have that for every $q \in N$ and $\varepsilon > 0$, $e^f(q) \in \mathrm{Int}(A_q^2(M, \varepsilon))$. Fix q and let q' vary in $B_N(q, 2)$, which is a relatively compact neighborhood of q. The continuous dependence of $A_{q'}^2(M, \varepsilon)$ on q' implies that there exists $\rho(\varepsilon, q) \in (0, 1)$ such that for every $q' \in B_N(q, 2)$,

$$B_N(e^f(q'), \rho(\varepsilon, q)) \subset A_{q'}(M, \varepsilon).$$

Therefore,

$$q \in \mathrm{Int}(A_{q'}(M, \varepsilon)) \quad \text{if} \quad e^f(q') \in B(q, \rho(\varepsilon, q)). \tag{17}$$

Indeed, for every $q' \in N$ such that $e^f(q') \in B(q, \rho(\varepsilon, q))$, we have

$$d_N(q, q') \le d_N(q, e^f(q')) + d_N(q', e^f(q')) < 2.$$

LEMMA 5.5. *Let M be a complete, connected Riemannian surface. If $q \in N$ and $\varepsilon > 0$ exist such that $A_q(M, \varepsilon)$ has finite volume, then M has finite area and $A_q(M, \varepsilon) = N$.*

Proof. Fix $q \in N$ and $\varepsilon > 0$ such that $A_q = A_q(M, \varepsilon)$ has finite volume. We want to prove that ∂A_q is empty. Let, by contradiction, $\bar{q} \in \partial A_q$, and define $\bar{\rho} = \rho(\varepsilon, \bar{q}) > 0$, where the function ρ satisfies (17). A well-known theorem by Krener [14] states that any attainable set of a bracket-generating system is contained in the closure of its interior. Therefore, $V = A_q \cap B_N(\bar{q}, \bar{\rho})$ has a nonempty interior and, in particular, positive volume. Since e^f is a volume-preserving diffeomorphism of N (see, for instance, [23]) and A_q has finite volume, then the sets $e^{nf}(V)$, for $n \in \mathbf{N}$, cannot be pairwise disjoint, being $e^{nf}(V) \subset A_q$ for every $n \in \mathbf{N}$. Therefore, there exist $n_1 < n_2$ such that $e^{n_1 f}(V) \cap e^{n_2 f}(V)$ is nonempty. Equivalently, there exists a point q' in $e^{(n_2 - n_1 - 1)f}(V)$ whose image by e^f lies in V. Due to (17), the contradiction is reached. ∎

COROLLARY 5.6. *Let M be a complete, connected Riemannian surface with finite area. Then the Dubins problem on M has the* UCC *property.*

Notice that in the nonpositive curvature case, if M has finite area, then it is of the first kind (see, for instance, [3]). The converse, clearly, is not true in general. It is, however, when M is a finitely connected hyperbolic surface (see, for instance, [21]). In particular, we proved that the Dubins problem on a finitely connected hyperbolic surface of the first kind is UCC. Actually, first kind implies UCC in complete generality, as will be shown in Proposition 5.9.

Let us now prove the second part of Proposition 4.1. Recall that the geodesic flow on M is said to be *topologically transitive* if there exists $q \in N$ such that the orbit of q for the geodesic flow is dense in N.

PROPOSITION 5.7. *Let M be a complete, connected Riemannian surface such that the geodesic flow on M is topologically transitive. Then the Dubins problem on M has the* UCC *property.*

Proof. Fix $q_0 \in N$ such that

$$\mathcal{O}_{q_0} = \{e^{tf}(q_0) \mid t \in \mathbf{R}\}$$

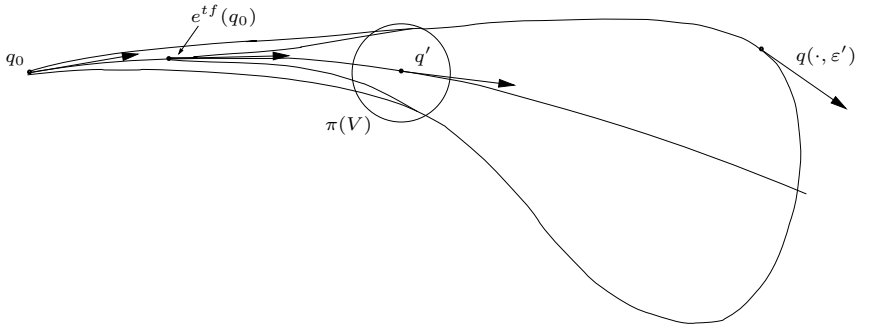


FIG. 3. *Construction of a teardrop trajectory from* $e^{tf}(q_0)$.

is dense in N. The property formulated in (17) implies that for every $\varepsilon > 0$,

$$N = A_{q_0}(M,\varepsilon) \cup \left(A_{q_0^-}(M,\varepsilon)\right)^-,$$

where $(A_{q_0^-}(M,\varepsilon))^-$ denotes the image of $A_{q_0^-}(M,\varepsilon)$ through the involutive diffeomorphism $q \mapsto q^-$. Hence, there exists a decreasing sequence ε_n, converging to zero as n tends to infinity, such that

$$q_0^- \in A_{q_0}(M,\varepsilon_n) \quad \text{for every } n \geq 1 \tag{18}$$

or

$$q_0 \in A_{q_0^-}(M,\varepsilon_n) \quad \text{for every } n \geq 1. \tag{19}$$

Without loss of generality, we can assume that (18) holds, since

$$\mathcal{O}_{q_0^-} = (\mathcal{O}_{q_0})^-$$

is dense in N, which means that the roles of q_0 and q_0^- are symmetric.

Since $\varepsilon \mapsto A_{q_0}(M,\varepsilon)$ is monotone with respect to the inclusion, then q_0 belongs to the set

$$W = \{q \in N \mid q^- \in A_q(M,\varepsilon) \text{ for every } \varepsilon > 0\}.$$

The UCC property for Dubins' problem on M is equivalent to the equality $N = W$ (see Remark 2.2). It is enough to show that $\mathcal{O}_{q_0} \subset W$, due to (5) and to the density of $\mathcal{O}_{q_0}$ in N.

This is trivial for what concerns the negative orbit of q_0. Fix now $\varepsilon, t > 0$. We have to show that there exists an ε-admissible trajectory connecting $e^{tf}(q_0)$ and $(e^{tf}(q_0))^-$. Notice that $A_{e^{tf}(q_0)}(M,\varepsilon)$ contains a neighborhood V of $q' = e^{(t+1)f}(q_0)$. For every $\varepsilon' > 0$, let $q(\cdot,\varepsilon') : [0, T_{\varepsilon'}] \to N$ be an ε'-admissible trajectory such that $q(0,\varepsilon') = q_0$ and $q(T_{\varepsilon'},\varepsilon') = q_0^-$. Then, as ε' goes to zero, we have that $T_{\varepsilon'}$ goes to infinity, while, for every $L > 0$, the restriction of $q(\cdot,\varepsilon')$ on $[0,L]$ (respectively, $[T_{\varepsilon'} - L, T_{\varepsilon'}]$) converges uniformly to $[0,L] \ni s \mapsto e^{sf}(q_0)$ (respectively, $[T_{\varepsilon'} - L, T_{\varepsilon'}] \ni s \mapsto e^{(s-T_{\varepsilon'})f}(q_0^-)$). In particular, for ε' small enough, $q(t+1,\varepsilon')$ and $q(T_{\varepsilon'} - t - 1,\varepsilon')^-$ belong to V (see Figure 3). Provided that $\varepsilon' \leq \varepsilon$, an ε-admissible trajectory steering $e^{tf}(q_0)$ to $(e^{tf}(q_0))^-$ can be obtained by gluing three ε-admissible subtrajectories, the first

going from $e^{tf}(q_0)$ to $q(t+1,\varepsilon')$, the second coinciding with the restriction of $q(\cdot,\varepsilon')$ on $[t+1, T_{\varepsilon'}-t-1]$, and the last connecting $q(T_{\varepsilon'}-t-1,\varepsilon')$ and $(e^{tf}(q_0))^-$. □

Recall that the *injectivity radius at a point p of a Riemannian manifold M* is defined as the least upper bound $i_p(M)$ of all $r>0$ such that the map $(t,q)\mapsto \exp(t,q)$, restricted to the punctured disk $(0,r)\times U_pM$, is injective. Before stating sufficient conditions for UCC on nonpositively curved Riemannian surfaces, we prove the following technical result.

LEMMA 5.8. *Let M be a Riemannian surface, $q=(p,v)\in N$, and $\varepsilon, R, \delta>0$. Assume that $R<i_p(M)$, that $K\le 0$ on $B_M(p,R)$, and that $\int_{B_M(p,R)}(-K)dA\le\delta$. There exist $R(\varepsilon),\delta(\varepsilon)>0$, depending only on ε, such that if $R\ge R(\varepsilon)$ and $\delta\le\delta(\varepsilon)$, then there exists an ε-admissible trajectory going from q to q^- and contained in $UB_M(p,R)$.*

Proof. Notice that it is enough to prove the existence of $\delta(\varepsilon)$ and $R(\varepsilon)$ in the case $\varepsilon=1$. Indeed, given $\varepsilon>0$, denote by M^ε the Riemannian surface $(M,\varepsilon m)$. Let, moreover, K^ε and dA^ε be, respectively, the curvature and the surface form on M^ε. Then $K^\varepsilon=\varepsilon K$, $dA=\varepsilon dA^\varepsilon$, and, for every $p\in M$ and $r>0$, $B_M(p,\varepsilon r)=B_{M^\varepsilon}(p,r)$. Thus,

$$\int_{B_M(p,\varepsilon r)}(-K)dA=\int_{B_{M^\varepsilon}(p,r)}(-K^\varepsilon)dA^\varepsilon.$$

The reduction to the case $\varepsilon=1$ is proved (with $\varepsilon R(\varepsilon)=R(1)$ and $\delta(\varepsilon)=\varepsilon\delta(1)$), since an ε-admissible trajectory in M is a 1-admissible trajectory in M^ε and $i_p(M)=\varepsilon i_p(M^\varepsilon)$.

Let $q=(p,v)$ and R be, respectively, a point of N and a positive constant such that $R<i_p(M)$ and $K\le 0$ on $B_M(p,R)$. Denote by Q the square $[-R/2,R/2]\times[-R/2,R/2]\subset\mathbf{R}^2$, endowed with the metric ϕ_q^*m. Recall that Dubins' problem on Q is described by (13)–(15).

Let $u:[0,T]\to[-1,1]$ be a 1-admissible control, and denote by $(x(\cdot),y(\cdot),\theta(\cdot))$ the coordinates of the corresponding trajectory starting from a given $(x_0,y_0,\theta_0)\in UQ$.

Assume that $(x(t),y(t),\theta(t))\in UQ$ and $\dot x(t)>0$ for every $t\in(0,T)$. We can define a map $\tau:[x_0,x(T)]\to[0,T]$ by means of the relation $x(\tau(\xi))=\xi$. Notice that τ is continuous, as well as the function $\eta:[x_0,x(T)]\to\mathbf{R}$ which maps ξ to $\eta(\xi)=y(\tau(\xi))$. Let Ω be the open region of Q defined by

$$\Omega=\cup_{\xi\in(x_0,x(T))}I(\eta(\xi)),$$

where, for every $l\in\mathbf{R}$, $I(l)$ denotes the open interval with 0 and l as boundary points.

Let (X,Y,Θ) be the solution of

$$\begin{cases}\dot X=\cos\Theta,\\ \dot Y=\sin\Theta,\\ \dot\Theta=u\end{cases}$$

with initial condition $(X(0),Y(0),\Theta(0))=(x_0,y_0,\theta_0)$. For every $t\in[0,T]$,

$$\begin{aligned}|\theta(t)-\Theta(t)|=\left|\int_0^t\cos(\theta(s))\frac{B_y(x(s),y(s))}{B(x(s),y(s))}ds\right| &= \left|\int_{x_0}^{x(t)}B_y(\xi,\eta(\xi))d\xi\right|\\ \le\int_{x_0}^{x(t)}d\xi\left(\int_{I(\eta(\xi))}B_{yy}(\xi,v)dv\right) &= \int_{x_0}^{x(t)}\int_{I(\eta(\xi))}-K(\xi,v)B(\xi,v)dv\,d\xi,\end{aligned}$$

where the last equality follows from (8). Since the surface element dA of Q is equal to $B(\xi, v)dv\,d\xi$, we have

$$|\theta(t) - \Theta(t)| \leq \int_Q (-K)dA. \tag{20}$$

Integrating the equations satisfied by $y(\cdot)$ and $Y(\cdot)$, we get, for every $t \in [0, T]$,

$$|y(t) - Y(t)| \leq \int_0^t |\theta(s) - \Theta(s)|ds \leq t \int_Q (-K)dA. \tag{21}$$

Similarly, we obtain

$$\begin{aligned}
|x(t) - X(t)| &= \left| \int_0^t \left(\frac{\cos(\theta(t))}{B(x(t), y(t))} - \cos(\Theta(t)) \right) dt \right| \\
&\leq \left| \int_0^t (\cos(\theta(t)) - \cos(\Theta(t)))\, dt \right| + \left| \int_{x_0}^{x(t)} (1 - B(\xi, \eta(\xi)))\, d\xi \right| \\
&\leq t \int_Q (-K)dA + \int_{x_0}^{x(t)} \left(\int_{I(\eta(\xi))} B_y(\xi, v)dv \right) d\xi \\
&= t \int_Q (-K)dA + \int_{x_0}^{x(t)} \left(\int_{I(\eta(\xi))} \left(\int_{I(v)} (-K(\xi, l)B(\xi, l))dl \right) dv \right) d\xi \\
&= t \int_Q (-K)dA + \int_{x_0}^{x(t)} \left(\int_{I(\eta(\xi))} |\eta(\xi) - l|(-K(\xi, l)B(\xi, l))dl \right) d\xi \\
&\leq \left(t + (x(t) - x_0)\frac{R}{2} \right) \int_Q (-K)dA \\
&\leq t \left(1 + \frac{R}{2} \right) \int_Q (-K)dA. \qquad (22)
\end{aligned}$$

Let

$$\delta = \int_Q (-K)dA. \tag{23}$$

Our aim is to show that under suitable assumptions on R and δ, there exists a 1-admissible trajectory contained in UQ which steers $(0, 0, 0)$ to $(0, 0, -\pi)$. As a model we consider the teardrop in the plane of size 1, whose projection is contained in a square of center $(0, 0)$ and side length 6.

Let $\alpha \in (0, \pi/2)$ and consider the trajectory $(x^+(\cdot, \alpha), y^+(\cdot, \alpha), \theta^+(\cdot, \alpha))$ in UQ with initial condition $(x^+(0, \alpha), y^+(0, \alpha), \theta^+(0, \alpha)) = (0, 0, 0)$ corresponding to the control

$$u^+(s, \alpha) = \begin{cases} 1 & \text{if} \quad 0 \leq s \leq \alpha, \\ -1 & \text{if} \quad s > \alpha. \end{cases}$$

Let $t^+(\alpha)$ be the first time t at which $\theta^+(t, \alpha) = -\pi/2$. The fact that a finite $t^+(\alpha)$ with such property exists, for δ and $1/R$ small enough, depends on the following considerations. First, notice that the trajectory is defined as long as its projection on the plane (x, y) stays in Q; thus, for a time $T_{\max}$ larger than $R/2$. If $\dot{x}^+(s, \alpha) > 0$ for

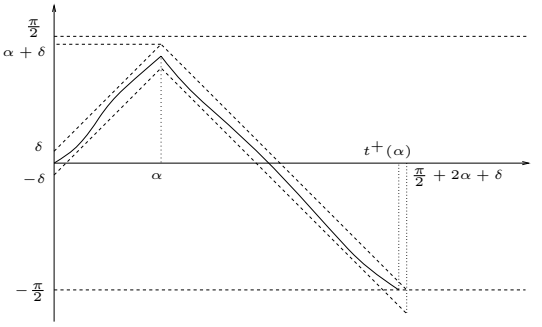


FIG. 4. *The graph of* $\theta^+(\cdot, \alpha)$.

every $s \in [0, t]$, then $|\theta^+(t, \alpha) - \int_0^t u^+(s, \alpha) ds| \leq \delta$, as it follows from (20). Assume that

$$\delta < \pi/2 - \alpha. \tag{24}$$

In particular, $\dot{x}^+(t, \alpha) > 0$ until $\theta^+(\cdot, \alpha)$ assumes the value $-\pi/2$ or up to time $T_{\max}$. If, moreover,

$$R \geq 2\pi + 2\alpha, \tag{25}$$

then $T_{\max} \geq R/2 > \pi/2 + 2\alpha + \delta$ and, since $\int_0^{\pi/2+2\alpha} u^+(s, \alpha) ds = -\pi/2$, $t^+(\alpha)$ is well defined and satisfies

$$\left| t^+(\alpha) - \frac{\pi}{2} - 2\alpha \right| \leq \delta. \tag{26}$$

(See Figure 4.) Let $u^-(\cdot, \alpha) = -u^+(\cdot, \alpha)$, define $(x^-(\cdot, \alpha), y^-(\cdot, \alpha), \theta^-(\cdot, \alpha))$ as the trajectory in UQ corresponding to the control $u^-(\cdot, \alpha)$ with initial condition

$$(x^-(0, \alpha), y^-(0, \alpha), \theta^-(0, \alpha)) = (0, 0, 0).$$

As above, if δ and R satisfy (24) and (25), then $t^-(\alpha)$, the first time t for which $\theta^-(t, \alpha) = \pi/2$, is well defined and the estimate

$$\left| t^-(\alpha) - \frac{\pi}{2} - 2\alpha \right| \leq \delta \tag{27}$$

holds true.

Fix $R = 3\pi$ and notice that, as a consequence, (25) is always satisfied. Let U be a neighborhood of $(\pi/3, \pi/3)$ compactly contained in $(0, \pi/2) \times (0, \pi/2)$. Thus there exists $\delta_0 > 0$ such that if $\delta \leq \delta_0$, then

$$\mathcal{F}(\alpha_0, \alpha_1) = (x^+(t^+(\alpha_0), \alpha_0), y^+(t^+(\alpha_0), \alpha_0)) - (x^-(t^-(\alpha_1), \alpha_1), y^-(t^-(\alpha_1), \alpha_1))$$

is well defined for every (α_0, α_1) in U, and

$$t^+(\alpha_0), t^-(\alpha_1) \leq \frac{3}{2}\pi.$$

Notice that δ_0 depends only on the choice of U. The quantity $|\mathcal{F}(\alpha_0, \alpha_1)|$ measures how far the two trajectories corresponding to the controls $u^+(\cdot, \alpha_0)$ and $u^-(\cdot, \alpha_1)$ are

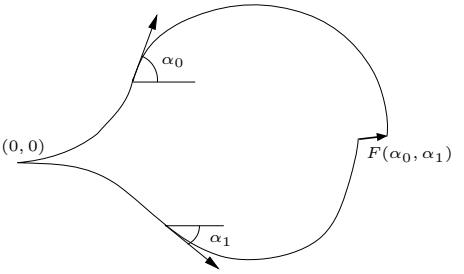


FIG. 5.

from matching their ends (see Figure 5). The lemma is proved if we show that the minimum of $|\mathcal{F}|$ on U is equal to zero.

For every $\alpha \in (0,\pi/2)$, let $(X^{\pm}(\cdot,\alpha), Y^{\pm}(\cdot,\alpha), \Theta^{\pm}(\cdot,\alpha))$ be the trajectories in $U\mathbf{R}^2 \cong \mathbf{R}^2 \times \mathcal{S}^1$ corresponding to the controls $u^{\pm}(\cdot,\alpha)$ and to the initial conditions $(X^{\pm}(0,\alpha), Y^{\pm}(0,\alpha), \Theta^{\pm}(0,\alpha)) = (0,0,0)$. For every (α_0,α_1) in U, let

$$\begin{aligned}\mathcal{G}(\alpha_0,\alpha_1) &= \left(X^{+}\left(\frac{\pi}{2}+2\alpha_0,\alpha_0\right), Y^{+}\left(\frac{\pi}{2}+2\alpha_0,\alpha_0\right)\right)\\ &\quad - \left(X^{-}\left(\frac{\pi}{2}+2\alpha_1,\alpha_1\right), Y^{-}\left(\frac{\pi}{2}+2\alpha_1,\alpha_1\right)\right)\\ &= 2(\sin\alpha_0 - \sin\alpha_1, -\cos\alpha_0 - \cos\alpha_1),\end{aligned}$$

which is the flat counterpart of $\mathcal{F}$. Note that, by construction, $\mathcal{G}(\pi/3,\pi/3) = (0,0)$. Moreover,

$$\det D\mathcal{G}(\pi/3,\pi/3) = \det\begin{pmatrix} \cos\dfrac{\pi}{3} & -\cos\dfrac{\pi}{3} \\ \sin\dfrac{\pi}{3} & \sin\dfrac{\pi}{3} \end{pmatrix} = \frac{\sqrt{3}}{2},$$

and so if $\mathcal{S}^1 \ni s \mapsto (\alpha_0(s),\alpha_1(s)) \in U$ is a simple closed loop encircling $(\pi/3,\pi/3)$ and being close enough to $(\pi/3,\pi/3)$, then $s \mapsto \mathcal{G}(\alpha_0(s),\alpha_1(s))$ encircles $(0,0)$. Fix a loop $(\alpha_0(\cdot),\alpha_1(\cdot))$ of such kind and let

$$r = \min_{s\in\mathcal{S}^1} d_{\mathbf{R}^2}((0,0),\mathcal{G}(\alpha_0(s),\alpha_1(s))) > 0.$$

We stress that r, by construction, is independent of the specific point p and of the manifold M. In fact, it depends only on the dynamics of the flat Dubins problem.

According to (21) and (22), we get, for every $(\alpha_0,\alpha_1) \in U$ and for $j = 0,1$,

$$\begin{aligned}|x^{\pm}(t^{\pm}(\alpha_j),\alpha_j) - X^{\pm}(t^{\pm}(\alpha_j),\alpha_j)| &\le t^{\pm}(\alpha_j)\left(1+\frac{R}{2}\right)\delta \quad \le \quad \frac{3}{2}\pi\left(1+\frac{3}{2}\pi\right)\delta,\\ |y^{\pm}(t^{\pm}(\alpha_j),\alpha_j) - Y^{\pm}(t^{\pm}(\alpha_j),\alpha_j)| &\le t^{\pm}(\alpha_j)\delta \quad \le \quad \frac{3}{2}\pi\delta.\end{aligned}$$

On the other hand, for for every $(\alpha_0,\alpha_1) \in U$ and $j = 0,1$,

$$\begin{aligned}|X^{\pm}(t^{\pm}(\alpha_j),\alpha_j) - X^{\pm}(\pi/2+2\alpha_j,\alpha_j)| &\le \delta,\\ |Y^{\pm}(t^{\pm}(\alpha_j),\alpha_j) - Y^{\pm}(\pi/2+2\alpha_j,\alpha_j)| &\le \delta,\end{aligned}$$

as it follows from (26) and (27). Therefore, by triangle inequality,

$$|\mathcal{F}(\alpha_0, \alpha_1) - \mathcal{G}(\alpha_0, \alpha_1)| \leq c_0 \delta$$

for every $(\alpha_0, \alpha_1) \in U$, where c_0 is a universal positive constant. If

$$\delta \leq \min\left\{\delta_0, \frac{r}{c_0}\right\}, \tag{28}$$

then it follows from standard degree theory considerations that there exists (α_0, α_1) in U such that $\mathcal{F}(\alpha_0, \alpha_1) = (0, 0)$. The lemma is proved, with $R(1) = 3\pi$ and $\delta(1)$ equal to the right-hand side of (28). □

From now on, let, for every complete Riemannian surface M, for every $p \in M$ and $r > 0$,

$$\Phi(p, r) = \int_{B_M(p,r)} K dA.$$

PROPOSITION 5.9. *Let M be of the first kind. Then the Dubins problem on M is* UCC.

Proof. Fix $q \in N$ and $\varepsilon > 0$. Motivated by Remark 2.2, we are going to prove the so-called weak controllability from q, i.e., that there exists $\hat{q} \in A(q, \varepsilon)$ such that $\hat{q}^- \in A(\hat{q}, \varepsilon)$.

Let $\tilde{q} \in UX$ be such that $\Pi_*(\tilde{q}) = q$ and define $\tilde{q}_1 = e^f(\tilde{q})$. Recall that, due to (5), $\tilde{q}_1$ belongs to the interior of $A_{\tilde{q}}(X, \varepsilon)$. Therefore, there exists $\rho > 0$ such that for every $\theta \in [-\rho, \rho]$, the ray $[0, \infty) \ni s \mapsto \exp(s, e^{\theta g}(\tilde{q}_1))$ is projected by Π in the interior of $\pi(A_q)$. Denote by S the sector of X spanned by all rays $\exp(\cdot, e^{\theta g}(\tilde{q}_1))$, $\theta \in [-\rho, \rho]$.

Consider the case in which there exists $p \in S$ such that $\Phi(p, R(\varepsilon)) \geq -\delta(\varepsilon)$, where $R(\cdot)$ and $\delta(\cdot)$ are the functions whose existence was established by Lemma 5.8. Notice that $i_p(X) = \infty$. Then, for every $\hat{q} \in U_pX$, we would have $\hat{q}^- \in A_{\hat{q}}(X, \varepsilon)$. Since $U_pX \cap A_{\tilde{q}}(X, \varepsilon) \neq \emptyset$, then the weak controllability from q would follow. Thus, we can assume that S is heavy.

Recall that M contains at least one closed geodesic G (see Remark 3.3). The set $\Pi^{-1}(G)$ is the union of infinitely many complete geodesics. Let $\widetilde{G}$ be one of these lifts of G. Fix one endpoint of $\widetilde{G}$ and denote it by z.

Lemma 5.3 implies that, without loss of generality, $z \in S(\infty)$ and $\widetilde{G}$ is not the boundary of a flat half-plane. Since $\widetilde{G}$ is the axis of a hyperbolic isometry contained in Γ, it follows from (6) that there exist components of $\Pi^{-1}(G)$ which are entirely contained in S. Thus, we can assume that $\widetilde{G} \subset S$ (see Figure 6).

Choose a parameterization $c : \mathbf{R} \to X$ of $\widetilde{G}$. A standard application of Topogonov theorem to the triangle $(\pi(\tilde{q}_1), c(t), c(0))$ shows that the angle between $\widetilde{G}$ and the segment from $\pi(\tilde{q}_1)$ to $c(t)$ goes to zero as t goes to plus or minus infinity (see [4, p. 34]).

Fix $p \in G$ and let $\sigma > 0$ be such that for every $q' \in U_pM$, $B_N(q', \sigma) \subset A^2_{e^{-f}(q')}(M, \varepsilon)$. There exists a ray from $\pi(\tilde{q})$, contained in S, which intersects $\widetilde{G}$ at a point which projects to p, with an angle smaller than σ, and at a distance from $\pi(\tilde{q})$ larger than one.

Therefore, there exists an admissible curve for $(\mathcal{D}^M_\varepsilon)$ which starts from the point-with-direction q and arrives tangentially on G. Moreover, the same procedure can be applied in order to arrive tangentially on G with the opposite orientation. Gluing

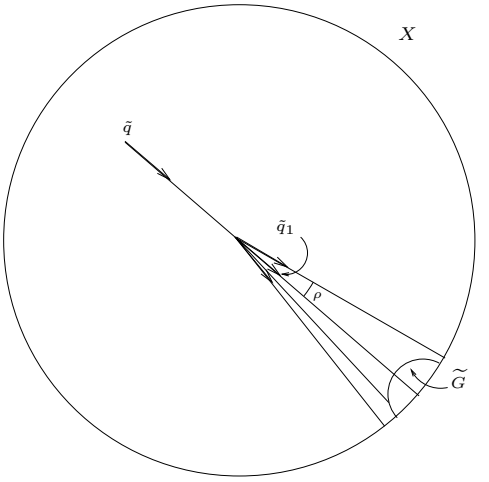


FIG. 6.

pieces of admissible trajectories, one obtains an admissible strategy which connects q with q^-. □

In the above proof we additionally gave an argument for the following proposition.

PROPOSITION 5.10. *Let M be a connected complete nonpositively curved Riemannian surface and denote by X its universal covering. If for every $r > 0$ and every sector S of X, $\sup_{p\in S} \Phi(p, r) = 0$, then the Dubins problem on M is* UCC.

Remark 5.11. If $\lim_{p\to\infty,\, p\in M} K(p) = 0$, then $\lim_{p\to\infty,\, p\in X} \Phi(p, r) = 0$ for every $r > 0$, and therefore the Dubins problem on M is UCC. This does not come unexpected. Indeed, it is a special case of [6, Theorem 5.14], where unrestricted complete controllability was proved for all complete, connected Riemannian surfaces whose curvature tends to zero at infinity.

COROLLARY 5.12. *If $K \leq 0$ and $\int_M K dA$ is finite, then the Dubins problem on M is* UCC.

Proof. Assume that the UCC property does not hold. In particular, M is of the second kind and, according to Proposition 5.10, there exist $r > 0$ and a sector S in X such that

$$\sup_{p\in S} \Phi(p, r) < 0, \tag{29}$$

where X denotes the universal covering of M. Then S is heavy and Lemma 5.4 guarantees the existence of a half-plane $H \subset S$ embedded by Π in M. Finally, $\int_M K dA = \int_H K dA = -\infty$. □

Remark 5.13. Let M be a complete, connected unbounded Riemannian surface and assume that $K \leq 0$ outside a compact subset C of M. Then, for every $p \in M \setminus C$, for every $q \in U_p M$, the square $[-d(p, C)/2, d(p, C)/2] \times [-d(p, C)/2, d(p, C)/2]$ can be endowed with the metric $\phi_q^* m$. Therefore, for every $r > 0$, on a complement of a compact subset of M, which depends on r, it is possible to define

$$\widetilde{\Phi}(p, r) = \int_{\tilde{B}(p,r)} K dA,$$

where $\tilde{B}(p, r)$ is the ball of center $(0, 0)$ and radius r contained in a sufficiently large square of the type $[-\rho, \rho] \times [-\rho, \rho]$ endowed with $\phi_q^* m$, for some $q \in U_p M$. The

definition makes sense, since the value of $\int_{\tilde{B}(p,r)} K dA$ does not depend on the choice of q. Clearly, if $i_p(M) \geq r$, then $\widetilde{\Phi}(p,r) = \Phi(p,r)$.

If $\lim_{p\to\infty} \widetilde{\Phi}(p,r) = 0$, then the Dubins problem on M is UCC. Indeed, it is a general fact that on an unbounded surface M every attainable set is unbounded (see Lemma 5.5). Then, for every $\varepsilon > 0$ and every $q \in N$, there exists $\hat{q} \in A_q(M,\varepsilon)$ such that $\widetilde{\Phi}(\pi(\hat{q}), R(\varepsilon))$ is well defined and it is larger than $-\delta(\varepsilon)$. Applying Lemma 5.8 to $(-R(\varepsilon), R(\varepsilon)) \times (-R(\varepsilon), R(\varepsilon))$, endowed with $\phi_{\hat{q}}^* m$, we obtain the desired weak controllability from q.

5.3. Noncontrollability results.

LEMMA 5.14. *Let M be a complete Riemannian surface. If a half-plane H in M exists, such that $\sup_H K < 0$, then the Dubins problem on M is not* UCC.

Proof. Let $a = |\sup_H K| > 0$. Fix a system of geodesic coordinates (x,y) on H such that H is parameterized by $\{(x,y) |\ x \in \mathbf{R}, y \geq 0\}$. Let us compare $F(x,y)$, defined as in (12), with

$$G(y) = \sqrt{a} \tanh\left(\sqrt{a} y\right),$$

the solution of (16) corresponding to $K \equiv -a$. If $(x,y) \in H$ and $F(x,y) = G(y)$, then $F_y(x,y) - G_y(y) = -K(x,y) - a \geq 0$. Therefore

$$F(x,y) \geq \sqrt{a} \tanh\left(\sqrt{a} y\right). \tag{30}$$

Let $q = (0,1,\pi/2) \in UH$. We claim that for ε small enough, the attainable set $A_q(M,\varepsilon)$ is contained in the interior of UH. Indeed, let $\varepsilon \in (0, \sqrt{a}\tanh(\sqrt{a}))$ and consider any admissible control $u : [0,\infty) \to [-\varepsilon,\varepsilon]$. Denote by $(x(\cdot), y(\cdot), \theta(\cdot))$ its corresponding trajectory such that $(x(0), y(0), \theta(0)) = q$. Define $\theta_0 \in (0, \pi/2)$ through the relation

$$\sqrt{a} \tanh\left(\sqrt{a}\right) \cos\theta_0 = \varepsilon \tag{31}$$

and let

$$T = \sup\{\tau > 0 \,|\, \pi - \theta_0 \leq \theta(\tau) \leq \theta_0\}.$$

Notice that $y(t) \geq 1$ for every $t \in [0,T)$. It follows from (30) that $F(x(t), y(t)) \geq \sqrt{a}\tanh(\sqrt{a})$ for every $t \in [0,T)$. Therefore, if T were finite, then

$$F(x(T), y(T)) \cos\theta_0 > \varepsilon,$$
$$F(x(T), y(T)) \cos(\pi - \theta_0) < -\varepsilon.$$

This would not be compatible with (15), the equation satisfied by $\theta(\cdot)$, and the definition of T. ∎

COROLLARY 5.15. *If K is uniformly negative and M is of the second kind, then the Dubins problem is not* UCC. *More precisely, if $0 < \varepsilon < |\sup_M K|$, then (D_ε^M) is not completely controllable.*

Proof. The first part of the statement follows from Lemma 5.4, which ensures that Π embeds a half-plane of X in M. The quantitative assertion is easily derived from the proof proposed for Lemma 5.14, where the point $q_1 = (0,1,\pi/2)$ can be replaced by $(0,y,\pi/2)$, with y arbitrarily large. ∎

LEMMA 5.16. *Let M be a complete Riemannian surface. Let $H \subset M$ be a half-plane such that K is nonpositive and bounded on H. Assume that there exists $r > 0$ such that $\sup_{p\in H} \Phi(p,r) < 0$. Then the Dubins problem on M is not* UCC.

Proof. Parameterize H over $\{(x,y)|\ x \in \mathbf{R}, y \geq 0\}$ with a system of geodesic coordinates. By contradiction, let us assume that the Dubins problem on M is UCC. Therefore, for every $\varepsilon > 0$, there exists a curve leaving from $(0,0)$ in the direction $(0,0,\pi/2)$ and coming back to ∂H in finite time, having geodesic curvature bounded by ε. In particular, there exists an ε-admissible trajectory $q = (x,y,\theta) : [0,T] \to N$ such that $q(0) = (0,0,\pi/2)$ and $q(T) = (x(T), y(T), k\pi)$, with $x(T) \in \mathbf{R}$, $y(T) > 0$, and $k \in \{0,1\}$.

Without loss of generality, $\theta(t) \in (0,\pi)$ for every $t \in [0,T)$. Hence, $t \mapsto y(t)$ is nonnegative and strictly increasing on $[0,T]$. Notice that T tends to infinity as ε goes to zero, since otherwise the Ascoli theorem would immediately lead to a contradiction.

The following claim will be repeatedly applied.

CLAIM 5.17. *For every $\rho, \tau > 0$, there exists $\varepsilon(\rho,\tau) > 0$ with the following property: let $\gamma : [0,\tau] \to H$ be a curve with geodesic curvature bounded by $\varepsilon(\rho,\tau)$, parameterized by arc-length and such that $\phi_{\dot\gamma(0)}([0,\tau] \times [-\rho,\rho])$ is contained in H; then, for every $t \in [0,\tau]$, $\operatorname{dist}(\gamma(t), \exp(t,\dot\gamma(0))) < \rho$.*

The claim can be obtained by a standard perturbation argument, as ε approaches zero, applied to the system (13)–(15). The use of geodesic coordinates is justified by the hypothesis that $\phi_{\dot\gamma(0)}([0,\tau] \times [-\rho,\rho]) \subset H$. Notice that ε can be chosen only depending on τ and ρ, because K is uniformly bounded from below.

Up to a global rescaling of M, as performed in the proof of Lemma 5.8, we can assume that $r = 1/4$. Assume from now on that $\varepsilon < \varepsilon(1/4, 1)$ and $T > 2$.

Let T_1 be the largest number between $T-2$ and the maximal time $t \in [0,T)$ such that $\theta(t) = \pi/2$. Let, moreover,

$$\Omega = \cup_{t \in (T_1,T)} \{x(t)\} \times (0, y(t)),$$

and D be the disk of center $p = \exp(1/2, \mathcal{R}q(T-1))$ and radius $1/4$. Assume that $\theta(T) = 0$ (symmetric arguments hold when $\theta(T) = \pi$). The Gauss–Bonnet theorem, applied to Ω, implies that

$$\theta(T_1) + \int_\Omega (-K) dA = -\int_{T_1}^T u(t) dt.$$

Since both $\theta(T_1)$ and $\int_\Omega (-K) dA$ are nonnegative, then

$$\max\left\{\theta(T_1), \int_\Omega (-K) dA\right\} \leq \left|\int_{T_1}^T u(t) dt\right| \leq 2\varepsilon. \tag{32}$$

A contradiction is reached if we show that for ε small enough, D is contained in Ω. It follows from (32) and the definition of T_1 that for $\varepsilon < \pi/4$, $T_1 = T-2$. Claim 5.17 implies that $y(1) \geq 3/4$. Thus, the y-component of p is larger than $y(T-1) - 1/2 > y(1) - 1/2 \geq 1/4$. Therefore, D does not intersect the line $\{(x,0)|\ x \in \mathbf{R}\}$. If D contained a point of the type $\pi(q(\tau))$, for $\tau \in [T-2, T]$, then the distance from p to the geodesic $s \mapsto \exp(s, q(T-1))$ would be smaller than

$$d(p, \pi(q(\tau))) + d(\pi(q(\tau)), \exp(\tau - T + 1, q(T-1))) < \frac{1}{2}.$$

(See Figure 7.) Therefore, drawing the minimizing segment from p to the considered geodesic, we would construct a triangle (having p and $\pi(q(T-1))$ as two of its vertices) with two orthogonal angles. This is impossible, due to the Gauss–Bonnet theorem.

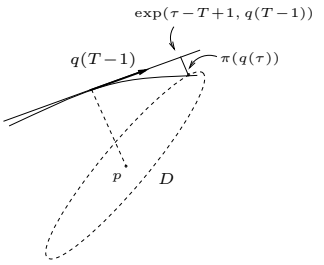


FIG. 7.

Assume now that $y \in [0, y(T)]$ exists such that $(x(T), y)$ belongs to D. Then the distance from $\exp(-1, q(T))$ to the geodesic $s \mapsto (x(T), s)$ is strictly smaller than one. Again, this leads to a contradiction.

The same argument shows that D does not intersect $\Sigma = \{\exp(s, \mathcal{R}q(T-2))|\ s \in J\}$, where J is the maximal interval containing zero such that $\exp(s, \mathcal{R}q(T-2)) \in H$ for every $s \in J$. Since $\theta(T-2)$ is contained in $(0, \pi/2)$, and Σ cannot intersect $\{(x(T-2), y)|\ y \geq 0\}$ twice, then the stated inclusion of D in Ω follows. □

From Lemma 5.4, one easily obtains the next corollary.

COROLLARY 5.18. *Let M be a complete, connected Riemannian surface. Assume that K is nonpositive and bounded on M and that M is of the second kind. Let X be the universal covering of M. Assume that $\sup_{p\in S} \Phi(p, r) < 0$ for some sector S of X and some $r > 0$. Then the Dubins problem on M is not* UCC.

5.4. Riemannian surfaces with nonpositive curvature outside a compact set. In this section, let us focus on Riemannian surfaces with sign-varying curvature, allowing K to be positive on a bounded set. Recall that a subset U of a finitely connected complete Riemannian surface M is called a *Riemannian half-cylinder* if it is diffeomorphic to $\mathcal{S}^1 \times [0, \infty)$. With a Riemannian half-cylinder U, we can associate its curvature at infinity $K_\infty(U) = -\int_U K dA - k(\partial U)$, where $k(\partial U)$ denotes the integral of the geodesic curvature of ∂U. The Cohn-Vossen theorem [8, 27] guarantees that $K_\infty(U) \geq 0$. When $K_\infty(U)$ is strictly positive, U is called a *strict Riemannian half-cylinder*. Notice that if U' is a Riemannian half-cylinder homotopic to U, then $K_\infty(U) = K_\infty(U')$. In particular, U' is strict if and only if U is.

We can prove the following.

PROPOSITION 5.19. *Let M be a finitely connected complete Riemannian surface such that K is bounded and nonpositive outside a compact set. Assume, moreover, that every Riemannian half-cylinder in M is strict. Then the Dubins problem on M is* UCC *if and only if, for every half-plane H contained in M and every $r > 0$, $\sup_{p\in H} \Phi(p, r) = 0$. In particular, the Dubins problem on M is* (i) UCC *if $\int_M K dA > -\infty$;* (ii) *not* UCC *if there exists a strict Riemannian half-cylinder $U \subset M$ such that $\sup_{p\in U} \Phi(p, r) < 0$ for some $r > 0$.*

We do not provide a complete argument, since Proposition 5.19 basically follows from what precedes and from the results of Shioya [27]. Let us just sketch the main points of the proof.

Shioya proved that if U is a strict Riemannian half-cylinder, then, through every point of U far enough from ∂U, passes a complete geodesic $c : \mathbf{R} \to M$, contained in U and which has a finite number $n(U)$ of self-intersections (only depending on U). In addition, $n(U) = 0$ if $\int_U K dA = -\infty$, which means that U contains a half-plane of M. When $n(U) \geq 1$, the self-intersections of c are *regular* in the sense defined in

Chapter 1 of [27]. Such notion of regularity guarantees that up to reorientation of c, the restriction of $\phi_{\dot c(0)}$ to $\mathbf{R}\times[0,+\infty)$ takes values in U and each of its preimages is finite, with cardinality bounded by $n(U)+1$. Therefore, $\phi_{\dot c(0)}(\mathbf{R}\times[0,+\infty))$ lifts to a nonpositively curved half-plane.

If there exist $r>0$ and a half-plane H contained in M such that $\sup_{p\in H}\Phi(p,r)<0$, then the Dubins problem on M is not UCC (Proposition 5.16).

Fix $q\in N$ and $\varepsilon>0$ and assume that, for every half-plane H contained in M and every $r>0$, $\sup_{p\in H}\Phi(p,r)=0$. Recall that $A_q(M,\varepsilon)$ is unbounded. If $\int_U KdA>-\infty$ for some Riemannian half-cylinder U, then, for every $r>0$, $\lim_{p\in U,p\to\infty}\widetilde{\Phi}(p,r)=0$. (For the definition of $\widetilde{\Phi}$, see Remark 5.13.) If $A_q(M,\varepsilon)$ contains a point $\hat q$ such that $\widetilde{\Phi}(\pi(\hat q),R(\varepsilon))=0$, then the weak controllability from q is proved. Otherwise, there must exist a Riemannian half-cylinder U such that $\int_U KdA=-\infty$ and $\pi(A_q(M,\varepsilon))\cap U$ is unbounded. Thus, there exist a parameterized complete geodesic $c(\cdot)$ of the type described above and $\hat q\in\pi^{-1}(c(0))\cap A(q,\varepsilon)$ pointing inside the half-plane $\phi_{\dot c(0)}(\mathbf{R}\times[0,+\infty))$. From the usual reasoning based on (5), $\phi_{\dot c(0)}(\mathbf{R}\times[0,+\infty))\cap\pi(A_q(M,\varepsilon))$ contains a sector S. If $\lim_{p\in S,\,p\to\infty}\Phi(p,R(\varepsilon))=0$, then the weak controllability from q follows. Otherwise, $\int_S KdA=-\infty$. From a visibility property analogous to Lemma 5.2, it follows that S contains a half-plane of M. The weak controllability from q follows.

Remark 5.20. The argument above shows that when $\sup_{p\in H}\Phi(p,r)=0$ for every half-plane H contained in M and every $r>0$, UCC holds even in the case in which K is unbounded. Analogously, Lemma 5.14 implies that, if K is uniformly negative (possibly unbounded) on a strict Riemannian half-cylinder, then the Dubins problem on M is not UCC.

6. Remarks on optimal control. Let M be any complete oriented Riemannian surface, with no assumption on the sign of K. Denote by $\langle\lambda,\cdot\rangle$ the action of a covector $\lambda\in T_q^*N$ on T_qN.

Fix $q_1=(p_1,v_1)$, $q_2=(p_2,v_2)\in N$ and assume that $q_2\in A_{q_1}(M,\varepsilon)$. Then, there exists an ε-admissible curve $q:[0,T]\to N$ such that $q(0)=q_1$, $q(T)=q_2$, and for which T is minimal. According to the Pontryagin maximum principle, the time-optimality of $q(\cdot)$ implies that there exist an absolutely continuous curve $\lambda:[0,T]\to T^*N$ and a constant $c\geq 0$ such that for almost every $t\in[0,T]$, $\lambda(t)\in T^*_{q(t)}N\setminus\{0\}$ and the function

$$H_w(\lambda)=\langle\lambda,f(q)\rangle+w\,\langle\lambda,g(q)\rangle\,,\quad \lambda\in T_q^\star N,\quad q\in N,\quad w\in[-\varepsilon,\varepsilon],$$

verifies

$$H_{u(t)}(\lambda(t))=\max_{w\in[-\varepsilon,\varepsilon]}H_w(\lambda(t))\;=\;\langle\lambda,f(q)\rangle+\varepsilon|\langle\lambda(t),g(q(t))\rangle|\;\equiv\;c, \tag{33}$$

$$\dot\lambda(t)=\overrightarrow{H_{u(t)}}(\lambda(t)), \tag{34}$$

where, for every $w\in[-\varepsilon,\varepsilon]$, $\overrightarrow{H_w}$ denotes the Hamiltonian vector field associated with $H_w:T^*N\to\mathbf{R}$.

A simple computation shows that

$$[f\pm g,[f,g]]=-K\,g\pm f.$$

In particular, $(f,g,[f,g])$, $(g,[f,g],[f+g,[f,g]])$, and $(g,[f,g],[f-g,[f,g]])$ are three moving frames in N. Therefore, the results by Schättler [24, 25] (see also [2])

imply that $q(\cdot)$ is piecewise smooth. A subinterval I of $[0,T]$ is called a *bang arc* if, up to a modification on a set of measure zero, u is constant and equal to ε or $-\varepsilon$ along I, while it is called a *singular arc* if $u|_I$ is smooth (again, up to a modification on a set of measure zero) but $|u|$ is not constantly equal to ε on I. Schättler's results give restrictions on the structure of time-optimal controls, stating that for every $q \in N$, there exists a neighborhood U of q such that every time-optimal trajectory of (D_ε) contained in U is the concatenation of three bang arcs or of a bang, a singular, and a bang arc (where some arc possibly has zero length).

In order to show that every time-optimal trajectory follows locally Dubins' pattern, let us prove that every singular arc corresponds to a geodesic of M, i.e., that $u = 0$ almost everywhere along singular arcs.

The covector lift $\lambda(\cdot)$ of a time-optimal trajectory $q : [0,t] \to N$ can be represented through the three real-valued functions:

$$\begin{aligned}\varphi_1(t) &= \langle \lambda(t), f(q(t)) \rangle ,\\ \varphi_2(t) &= \langle \lambda(t), g(q(t)) \rangle ,\\ \varphi_3(t) &= \langle \lambda(t), [f,g](q(t)) \rangle .\end{aligned}$$

Note that, according to (33),

$$u(t) = \operatorname{sign}(\varphi_2(t))\, \varepsilon \tag{35}$$

for almost every t such that $\varphi_2(t) \neq 0$ and that

$$|\varphi_1(t)| + |\varphi_2(t)| + |\varphi_3(t)| \neq 0 \tag{36}$$

for every t, since $\lambda(t) \neq 0$ and $f,\ g$, and $[f,g]$ are everywhere linearly independent.

Equation (34) implies that φ_1, φ_2, φ_3 satisfy the system of differential equations

$$\dot\varphi_1(t) = -u(t)\varphi_3(t), \tag{37}$$

$$\dot\varphi_2(t) = \varphi_3(t), \tag{38}$$

$$\dot\varphi_3(t) = u(t)\varphi_1(t) - K(q(t))\varphi_2(t). \tag{39}$$

Let $I \subset [0,T]$ be a singular arc. Modify, if necessary, u on a set of measure zero in order to render it smooth on I. By definition of singular arc, $\varphi_2|_{I'} \equiv 0$ on some subinterval I' of I. From (38), we have that $\varphi_3 = 0$ on I'. Plugging the information in (37) and (39), we get that φ_1 is constant and that $\varphi_1(t)u(t) = 0$ for every $t \in I'$. Due to (36), $\varphi_1 \neq 0$, and thus $u = 0$ along I'. Since u is smooth on I, it follows that $\varphi_2 = u = 0$ on I.

We proved that every singular arc of a time-optimal trajectory of (D_ε) (independently of the sign of K) corresponds to a geodesic segment and that time-optimal trajectories follow locally Dubins' pattern. The following proposition provides further restrictions on the structure of time-optimal trajectories under the hypothesis that K is uniformly negative.

PROPOSITION 6.1. *If K is uniformly negative and ε is small enough, then every time-optimal trajectory $q : [0,T] \to N$ of (D_ε) is the concatenation of a bang, a singular, and a bang arc (some of which possibly having zero length). In particular, $q(\cdot)$ follows Dubins' pattern.*

Proof. Assume that

$$K \leq -\varepsilon^2. \tag{40}$$

From (33) and (35), we get that $u\,\varphi_1 = \operatorname{sign}(\varphi_2)\,\varepsilon\,c - \varepsilon^2\varphi_2$ and so, for almost every $t \in [0,T]$,

$$\ddot{\varphi}_2(t) = -(K(q(t)) + \varepsilon^2)\varphi_2(t) + \operatorname{sign}(\varphi_2(t))\,\varepsilon\,c. \tag{41}$$

Therefore, $\ddot{\varphi}_2\varphi_2 \geq 0$ almost everywhere. If τ is such that $\varphi_2(\tau) = 0$ and $\varphi_2 > 0$ on a left neighborhood of τ, then $\varphi_2 > 0$ on $(\tau, T]$. Symmetric arguments show that the set $\{t \in [0,T] \mid \varphi_2(t) \neq 0\}$ has at most two connected components and that $q(\cdot)$ is the concatenation of a bang, a singular, and a bang arc. □

Acknowledgment. It is a pleasure for us to thank P. Pansu for his seminal suggestion and his plentiful advice.

REFERENCES

[1] A. A. Agrachev and Y. L. Sachkov *Control Theory from the Geometric Viewpoint*, Encyclopaedia Math. Sci., 87, Springer-Verlag, Berlin, 2004.
[2] A. A. Agrachev and M. Sigalotti, *On the local structure of optimal trajectories in* $\mathbf{R}^3$, SIAM J. Control Optim., 42 (2003), pp. 513–531.
[3] W. Ballmann, *Axial isometries of manifolds of non-positive curvature*, Math. Ann., 259 (1982), pp. 131–144.
[4] W. Ballmann, M. Gromov, and V. Schroeder, *Manifolds of Nonpositive Curvature*, Progr. Math., 61, Birkhäuser, Boston, 1985.
[5] U. Boscain and Y. Chitour, *Time optimal synthesis for left-invariant control systems on* $SO(3)$, SIAM J. Control Optim., to appear.
[6] Y. Chitour, *Applied and Theoretical Aspects of the Controllability of Nonholonomic Control System*, Ph.D. Thesis, Rutgers University, New Brunswick, NJ, 1996.
[7] Y. Chitour and M. Sigalotti, *Dubins' problem on surfaces.* I. *Nonnegative curvature*, J. Geom. Anal., 15 (2005), pp. 565–587.
[8] S. Cohn-Vossen, *Kürzeste Wege und Totalkrümmung auf Flächen*, Compos. Math., 2 (1935), pp. 69–133.
[9] L. E. Dubins, *On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents*, Amer. J. Math., 79 (1957), pp. 497–516.
[10] P. Eberlein and B. O'Neill, *Visibility manifolds*, Pacific J. Math., 46 (1973), pp. 45–110.
[11] A. Huber, *On subharmonic functions and differential geometry in the large*, Comment. Math. Helv., 32 (1957), pp. 13–72.
[12] V. Jurdjevic, *Geometric Control Theory*, Cambridge Stud. Adv. Math. 52, Cambridge University Press, Cambridge, UK, 1997.
[13] W. Klingenberg, *A Course in Differential Geometry*, Graduate Texts in Math. 51, Springer-Verlag, New York, Heidelberg, 1978.
[14] A. Krener, *A generalization of Chow's theorem and the bang-bang theorem to non-linear control problems*, SIAM J. Control, 12 (1974), pp. 43–51.
[15] C. Lobry, *Contrôlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.
[16] L. A. Lyusternik and A. I. Fet, *Variational problems on closed manifolds*, Dokl. Akad. Nauk SSSR (N.S.), 81 (1951), pp. 17–18 (in Russian).
[17] D. Mittenhuber, *Dubins' problem in the hyperbolic plane using the open disc model*, in Geometric Control and Non-Holonomic Mechanics (Mexico City, 1996), CMS Conf. Proc. 25, Amer. Math. Soc., Providence, RI, 1998, pp. 115–152.
[18] D. Mittenhuber *Dubins' problem is intrinsically three-dimensional*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 1–22.
[19] F. Monroy–Perez, *Three-dimensional non-Euclidean Dubins' problem*, in Geometric Control and Non-Holonomic Mechanics (Mexico City, 1996), CMS Conf. Proc. 25, Amer. Math. Soc., Providence, RI, 1998, pp. 153–181.
[20] R. N. Murray, Z. X. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1994.
[21] J. C. Ratcliffe, *Foundations of Hyperbolic Manifolds*, Graduate Texts in Math. 149, Springer-Verlag, New York, 1994.
[22] J. A. Reeds and L. A. Shepp, *Optimal paths for a car that goes both forwards and backwards*, Pacific J. Math., 145 (1990), pp. 367–393.

[23] T. Sakai, *Riemannian Geometry*, Transl. Math. Monogr. 149, Amer. Math. Soc., Providence, RI, 1996.

[24] H. Schättler, *On the local structure of time-optimal bang-bang trajectories in* $\mathbf{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186–204.

[25] H. Schättler *The local structure of time-optimal trajectories in dimension three under generic conditions*, SIAM J. Control Optim., 26 (1988), pp. 899–918.

[26] U. Serres, *On the curvature of two-dimensional optimal control systems and Zermelo's navigation problem*, J. Math. Sci. to appear.

[27] T. Shioya, *Behavior of distant maximal geodesics in finitely connected complete* 2*-dimensional Riemannian manifolds*, Mem. Amer. Math. Soc., 108, Providence, RI, 1994.

[28] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, Brandeis University, Waltham, MA, 1970.

[29] H. J. Sussmann, *Shortest* 3*-dimensional paths with a prescribed curvature bound*, in Proc. IEEE CDC (New Orleans, 1995), IEEE Publications, New York, 1995, pp. 3306–3312.

[30] H. J. Sussmann and G. Tang, *Shortest Paths for the Reeds-Shepp Car: A Worked Out Example of the Use of Geometric Techniques in Nonlinear Optimal Control*, Rutgers Center for Systems and Control Technical Report 91-10, Rutgers University, New Brunswick, 1991.

[31] G. Thorbergsson, *Closed geodesics on non-compact Riemannian manifolds*, Math. Z., 159 (1978), pp. 249–258.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 483–495

OPTIMAL CONTROL OF FAVORABLE GAMES WITH EXPECTED LOSS CONSTRAINT*

JOHANNES LEITNER†

Abstract. Using a forward-backward iterative method, we maximize the expected profit of a gambling strategy for an arbitrary finite sequence of independent favorable gambles with payoffs in L^1 under an expected loss constraint. The solution is explicit up to solving a finite sequence of (deterministic) optimization problems over $\mathbf{R}^2_+$.

Key words. stochastic control, constraint portfolio optimization, gambling theory, favorable games, inequalities for stochastic processes, gain-loss ratio

AMS subject classifications. 91A60, 91A15, 91A50, 93E20

DOI. 10.1137/050623231

1. Introduction.

Discrete gambling theory and stochastic games have been studied intensively for a long time. See the classic book [9] or, e.g., [20] for a modern introduction. A special case of stochastic games are the so-called favorable games; see [9] (see [24] for unfavorable games). Optimal gambling strategies (optimal with respect to expected wealth or the probability of reaching a goal) for favorable games were found in [6]. Kulldorff [18] also considers discrete time favorable games with a finite time limit and obtains the corresponding continuous time result as a limiting case. For the discrete time model, Kelly [17] found optimal growth rate strategies. The corresponding continuous time equivalent was solved by Heath et al. [15]. Furthermore, they found expected time optimal strategies with which to reach a goal for the continuous time model.

We consider the problem of maximizing the expected profit, i.e., expected change of wealth relative to some initial capital, in a sequence of independent gambles with return $V \in L^1$ under an expected loss (EL) constraint on $E[V_-]$. Maximizing the probability of reaching a certain wealth goal in a stochastic game is a good strategy for "how to gamble if you must" [9]. In contrast to this, the approach we are considering provides a good strategy for a gambler who is tempted by a favorable game but is risk-aware enough to impose a risk constraint by means of a bound on the EL relative to the initial capital. Clearly, depending on the type of gamble, in general that does not exclude arbitrarily high losses.

The EL is a special case of the so-called lower partial nth-moment ($LPM_n, n \geq 0$) risk measures studied, e.g., in [1], [3], [4], [11], [16], [23]. However, it is crucial in our approach to use the EL rather than the expected shortfall below some other wealth level or performance bound. The idea of measuring utility with respect to relative changes of wealth rather than absolute wealth goes back to [21].

Note also that EL is positively homogeneous, monotone, and convex, but not translation invariant, and hence is not a coherent risk measure; see [2]. It is rather a

*Received by the editors January 25, 2005; accepted for publication (in revised form) November 28, 2005; published electronically April 21, 2006. Support for this work was provided by the Austrian Science Foundation (FWF) under the Wittgenstein-Preis Program Z36-MAT.
http://www.siam.org/journals/sicon/45-2/62323.html
†Research Unit for Financial and Actuarial Mathematics, Institute for Mathematical Economics, Vienna University of Technology, Wiedner Hauptstraße 8-10/105-1, A-1040 Vienna, Austria (jleitner@fam.tuwien.ac.at).

loss measure than a risk measure. However, it satisfies convexity, a generally accepted property of a risk measure, which should encourage risk mitigation by diversification. See [12] for the notion of convex risk measures.

Trivially, the problem of maximizing expected return under an EL constraint is equivalent to the problem of minimizing EL under an expected return constraint. There is a third way to solve these constrained problems: Equivalently we can also consider the problem of maximizing the ratio of expected returns over EL! Our objective is to demonstrate that this latter, now *unconstrained* optimization problem can be solved *explicitly* in our restricted discrete time gambling framework. Unlike in the case of the discrete time unconstrained expected return (utility) optimization problem, the solution cannot be found by a backward iteration, which would lead to a discrete time Bellman-type equation. The problem does not decouple completely and a combined forward-backward iteration method has to be applied. Unfortunately, so far we did not succeed in applying our method to a more general model, such as a finite state Markovian gambling model. However, see [19] for an abstract duality result that allows one to determine the lowest upper bound of the ratio of expected returns over EL as the optimal value of a dual optimization problem over the set of all corresponding equivalent martingale measures for a general market model.

Optimal strategies for the expected profit over the EL ratio can in general, be unique at best up to a strictly positive scaling factor. This can actually be very convenient: While the ratio captures the trade-off, a gambler has to make between upside chances, measured by expected profits, and downside risk, measured by EL, the scaling factor allows the gambler to introduce an additional constraint, e.g., for bounded gambles a maximal loss constraint, a bound on the probability of wealth falling below a certain level, or a bound on the expected shortfall below some benchmark. This could be used to compensate for the fact that ruin is a priori not possible in our model. Considering that small or moderate losses are painful but qualitatively different from high losses causing irreversible bankruptcy, this feature of the proposed ratio might help a gambler (investor) to decide on a strategy.

For example, consider a coin-tossing game: Let $(S_i)_{1\le i\le N}$ be an i.i.d. sequence of random variables with $P(S_i=1)=1-P(S_i=-1)>\frac{1}{2}$ for all $1\le i\le N, N\ge 2$. Let S_i describe the payoff of a game at time i after having bet one currency unit. Denote by $V_i, 1\le i\le N$, the total change of wealth relative to some initial capital at time i after having bet on $S_j, 1\le j\le i$, according to some (nonanticipating) strategy. We are going to show that the following strategy leads to an optimal expected profit over the EL ratio $\frac{E[V_N]}{E[(V_N)_-]}$: Betting, say, one currency unit on S_1, V_1 equals S_1. It is then optimal to bet $|V_1|$ currency units on S_2 so that $V_2=0$ if $S_1\ne S_2$ and $V_2=2S_1$ if $S_1=S_2$. Interestingly, it is optimal to stop gambling (and not optimal to continue gambling) if $V_2=0$ holds; even so, S_3 is again a favorable game for $N\ge 3$. After having won twice or lost twice, it is optimal to keep doubling the bet on S_i to $|V_{i-1}|=2^{i-2}$ at each time $2<i\le N$ until either $S_i\ne S_1$ and we stop there (in which case $V_i=\cdots=V_N=0$) or $i=N$ and $S_1=\cdots=S_N$, in which case $V_N=2^{N-1}S_1$ holds. By a simple calculation, we find $\frac{E[V_N]}{E[(V_N)_-]}=(\frac{P(S_1=1)}{P(S_1=-1)})^N-1$ for the optimal ratio. By appropriate scaling of the initial bet, a bound on the maximal loss can be guaranteed.

In section 2 the model is specified and a stopping result is shown, which will allow us to split up the optimization problem into local backward-iterative optimization problems. This local problem is considered in section 3. Optimal values and gambling strategies are given in section 4. In section 5 we will give some easy examples, e.g., tossing of a coin. In section 6 we give some results for the multidimensional case.

2. Independent gambles. Let $\mathbf{\Omega} = (\Omega, (\mathcal{F}_t)_{0\le t\le T}, \mathcal{F}, P), 0 < T < \infty$, be a filtered probability space in discrete time. Let $S = (S_t)_{0\le t\le T}$ be an adapted $\mathbf{R}$-valued process on $\mathbf{\Omega}$ with $S_0 = 0$ and such that $S_t \in L^1$ is independent of $\mathcal{F}_{t-1}$ for all $1 \le t \le T$. Write $H \in \mathcal{F}_t$ if H is a random variable measurable with respect to $\mathcal{F}_t$.

Define $\mathbf{V}$ as the set of all adapted $\mathbf{R}$-valued processes $V = (V_t)_{0\le t\le T}$ on $\mathbf{\Omega}$ which are of the following form: $V_0 = 0$ and $V_t = \sum_{s=0}^{t-1} H_s S_{s+1}, 0 < t \le T$, where $H_s \in \mathcal{F}_s$ is such that $H_s S_{s+1} \in L^1, 0 \le s < T$. For $0 \le t \le T$, define $\mathbf{V}^t := \{V^t | V \in \mathbf{V}\}$, where $V^t = (V_u^t)_{0\le u\le T}$ is the process V stopped at t: $V_u^t := V_{u\wedge t}, 0 \le u \le T$. Set $\mathcal{V}^t := \{V_t | V \in \mathbf{V}^t\} \subseteq L^1$. $\mathcal{V}^t$ is to be interpreted as the set of all attainable wealth changes relative to some initial capital. $\mathcal{V}^t$ is called *arbitrage-free* if $P(V < 0) = 0$ for $V \in \mathcal{V}^t$ implies $V = 0$. It is well known that $\mathcal{V}^t$ is arbitrage-free for all $0 < t \le T$ iff $S_t \ne 0$ implies $P(S_t < 0), P(S_t > 0) > 0$ for all $0 < t \le T$; see [7] (and [8] for a general version of this result). In particular, $E[V_-] = 0$ then implies $V = 0$ for all $V \in \mathcal{V}^t$. By replacing S_t with $-S_t$ if $E[S_t] < 0$, we can assume $E[S_t] \ge 0$ for all $0 < t \le T$. For simplicity (but essentially without loss of generality), we can assume $S_t \ne 0$ for all $0 < t \le T$.

We always assume that $\mathcal{V} := \mathcal{V}^T$ is arbitrage-free. $\mathcal{V}$ is called *favorable* if there exists a $V \in \mathcal{V}$ with $E[V] > 0$. Set $\alpha(V) := \frac{E[V]}{E[V_-]}$ for $V \in \mathcal{V} \setminus \{0\}$ and $\alpha(0) = 0$. Consider the following optimization problems:

$$\bar{\alpha}^t := \sup_{V\in\mathcal{V}^t} \alpha(V), \quad 0 < t \le T. \tag{2.1}$$

We have $\bar{\alpha}^t \ge 0$ and $\bar{\alpha}^t = 0$ iff for all $V \in \mathcal{V}^t$, $E[V] = 0$ holds. Clearly, $\bar{\alpha}^t \le \bar{\alpha}^{\bar{t}}$ for all $0 < t \le \bar{t} \le T$. The optimization problem (2.1) will be solved by a forward-backward iterative method. The case $T = 1$ is trivial and we assume from now on that $1 < T < \infty$. Set also $S := S_T$.

The proofs of our main results, Theorem 4.1 and Theorem 4.2, are based on three steps. Not being in a Hamilton–Jacobi–Bellman setting, we cannot directly solve the optimization problem (2.1) by a backward-iterative method; i.e., we cannot apply the *dynamic programming principle*.

The first step is to transform problem (2.1) into a nested optimization problem over $V \in \mathcal{V}^{T-1}$ and a local optimization problem at time $T-1$ given V with optimal value α_V. This is done in the next subsection. The trick is to observe that an optimal strategy that returns to zero can be stopped and remains optimal.

The second step is to solve the local problem. This is quite technical but elementary and is done in section 3. The trick here is to observe that an optimal strategy can be chosen to assume only two values, depending on the sign of the value process at time $0 < t < T$. These two values are determined as the solution of a deterministic problem over $\mathbf{R}_+^2$. We also have to handle some marginal cases, e.g., if $E[S_T] = 0$. Moreover, it is possible that even if the games $S_t, 1 \le t < \tilde{t}$, are all favorable up to time $\tilde{t} \le T$, it is still optimal not to start gambling before time $\tilde{t}$ (this is where the dynamic programming principle breaks down).

The last step is to observe that the optimal value α_V of the local problem depends only on $\alpha(V)$. This allows us to then iterate the local optimization backwards in time. This is done in the proof of Theorem 4.1.

2.1. Stopping at zero. Let us start with a very simple result that will be used repeatedly without further mention.

LEMMA 2.1. *Let a, A, b, B be real numbers with $A, A + B > 0$ and $B \ne 0$; then $\frac{a+b}{A+B} \ge \frac{a}{A}$ iff $\frac{b}{B} \ge \frac{a}{A}$ and $B > 0$, or $\frac{b}{B} \le \frac{a}{A}$ and $B < 0$.*

The following result means that an optimal strategy can be stopped when the process of wealth changes returns to 0 at time $T-1$ and remains optimal.

PROPOSITION 2.2. *The following assertion holds:*

$$\bar{\alpha}^T = \max\left(\sup_{V\in\mathcal{V}^{T-1}} \sup_{\substack{HS\in L^1, H\in\mathcal{F}_{T-1}\\ HS=HS\mathbf{1}_{\{V\neq 0\}}}} \alpha(V+HS), \alpha(S)\right). \tag{2.2}$$

Proof. Let $V\in\mathcal{V}^{T-1}$ and $H\in\mathcal{F}_{T-1}$ such that $HS\in L^1$. Note that $\bar{V} := V + \mathbf{1}_{\{V\neq 0\}}HS \in \mathcal{V}^T$ and $W := \mathbf{1}_{\{V=0\}}HS \in \mathcal{V}^T$. Hence for all $x,y\in\mathbf{R}_+$, $x\bar{V}+yW\in\mathcal{V}^T$ holds and we have $\alpha(x\bar{V}+yW) = \frac{xE[\bar{V}]+yE[W]}{xE[\bar{V}_-]+yE[W_-]}$. Considering this as a function of (x,y) and by partial differentiation with respect to x and y at $x=y=1$, one checks that $\alpha(\bar{V})\neq\alpha(W)$ implies either $\alpha(\bar{V})>\alpha(\bar{V}+W)$ or $\alpha(W)>\alpha(\bar{V}+W)$. Since S is independent of $\mathcal{F}_{T-1}$, we have $\alpha(W)=\alpha(S)$ for $\mathbf{1}_{\{V=0\}}H\geq 0$ and $\mathbf{1}_{\{V=0\}}H\neq 0$, implying the assertion. ∎

Instead of solving the nested optimization problem in Proposition 2.2, it is more convenient to solve the following related problem: Fix a $V\in\mathcal{V}^{T-1}\setminus\{0\}$ and denote by $\mathcal{H}_V$ the set of $\mathcal{F}_{T-1}$-measurable random variables H on $\mathbf{\Omega}$ such that $VHS\in L^1$. For $\lambda\in\mathbf{R}$ and $H\in\mathcal{H}_V$, set $V_{\lambda,H} := V(\lambda+HS)$. Note that $V_{\lambda,H}\in\mathcal{V}$ for all $(\lambda,H)\in\mathbf{R}\times\mathcal{H}_V$ and consider the optimization problem

$$\alpha_V := \sup_{\lambda\in\mathbf{R}, H\in\mathcal{H}_V} \alpha(V_{\lambda,H}). \tag{2.3}$$

We have $\alpha_V=0$ iff $E[S]=E[V]=0$. Set $\alpha_V := \alpha(S)$ for $V=0$ and note that $\alpha_V\geq\alpha(S)$ for all $V\in\mathcal{V}^{T-1}$: For $V\in\mathcal{V}^{T-1}\setminus\{0\}$ we have $H := V^{-1}\mathbf{1}_{\{V\neq 0\}}\in\mathcal{H}_V$, $V_{0,H}\in\mathcal{V}$, since H is $\mathcal{F}_{T-1}$-measurable, and $\alpha(V_{0,H}) = \frac{E[S\mathbf{1}_{\{V\neq 0\}}]}{E[S_-\mathbf{1}_{\{V\neq 0\}}]} = \alpha(S)$ since S is independent of $\mathcal{F}_{T-1}$.

LEMMA 2.3. $\bar{\alpha}^T = \sup_{V\in\mathcal{V}^{T-1}}\alpha_V$.

Proof. Note first that $\alpha_V\leq\bar{\alpha}^T$ holds since $\alpha_V=\alpha(S)$ for $V=0$ and for $V\in\mathcal{V}^{T-1}\setminus\{0\}$, $V_{\lambda,H}\in\mathcal{V}$ for all $(\lambda,H)\in\mathbf{R}\times\mathcal{H}_V$. Since $\alpha_V\geq\sup_{\substack{HS\in L^1, H\in\mathcal{F}_{T-1}\\ HS=HS\mathbf{1}_{\{V\neq 0\}}}}\alpha(V+HS)$ and $\alpha_V\geq\alpha(S)$, the assertion now follows from Proposition 2.2. ∎

3. The local problem. The local optimization problem (2.3) will be solved using a necessary and sufficient 1st order condition for optimality of the ratio α.

3.1. The 1st order optimality condition. Let $u:\mathbf{R}\to\mathbf{R}$ be a function such that $U(V) := E[u(V)]\in\mathbf{R}$ for all $V\in L^1$. For $V,Z\in L^1$, the right directional derivative in direction Z at V of U is defined as $\partial_Z^+U(V) := \lim_{\epsilon\searrow 0}\frac{U(V+\epsilon Z)-U(V)}{\epsilon}$ if the limit exists; e.g., for convex u the limit always exists in $[-\infty,\infty)$.

LEMMA 3.1. *For $V,Z\in L^1$ we have $\partial_Z^+E[V_-] = E[\mathbf{1}_{\{V=0\}}Z_-] - E[\mathbf{1}_{\{V<0\}}Z]$, and for $E[V_-]>0$,*

$$\partial_Z^+\alpha(V) = \frac{E[Z]-\alpha(V)(E[\mathbf{1}_{\{V=0\}}Z_-]-E[\mathbf{1}_{\{V<0\}}Z])}{E[V_-]}$$

holds.

Proof. Note that $\frac{(x+\epsilon y)_- - x_-}{\epsilon}$ decreases to, resp., $-y$, y_-, 0 for $x<0$, $x=0$, $x>0$, as $\epsilon\searrow 0$. The first assertion follows from the monotone convergence theorem. The remaining assertions follow from the chain rule. ∎

The ratio α is in general not concave, but shares same of the properties of concave functionals.

PROPOSITION 3.2. *Let $\mathcal{V}$ be a linear subspace of L^1 such that $E[V_-] > 0$ for all $V \in \mathcal{V} \setminus \{0\}$; i.e., $\mathcal{V}$ is arbitrage-free. Then $\hat{V} \in \mathcal{V} \setminus \{0\}$ is a global maximum for α, i.e., $\alpha(V) \le \alpha(\hat{V})$ for all $V \in \mathcal{V}$, iff $E[\hat{V}] \ge 0$ and the local 1st order optimality condition $\partial_Z^+ \alpha(\hat{V}) \le 0$ holds for all $Z \in \mathcal{V}$.*

Proof. The clearly necessary condition $E[\hat{V}] \ge 0$ and $\partial_Z^+ \alpha(\hat{V}) \le 0$ for all $Z \in \mathcal{V}$ holds iff $E[\hat{V}] \ge 0$ and $E[Z] \le \alpha(\hat{V}) \partial_Z^+ E[\hat{V}_-]$. If $E[\hat{V}] = 0$, this implies $E[Z] = 0 = \alpha(Z)$ for all $Z \in \mathcal{V}$. If $E[\hat{V}] > 0$, this condition, together with convexity of $E[(\cdot)_-]$, implies $\alpha(\hat{V} + \epsilon Z) \le \frac{E[\hat{V}] + \epsilon E[Z]}{E[\hat{V}_-] + \epsilon \partial_Z^+ E[\hat{V}_-]} \le \alpha(\hat{V})$ for all $\epsilon \ge 0$ if $\partial_Z^+ E[\hat{V}_-] \ge 0$, resp., for all $0 \le \epsilon < -\frac{E[\hat{V}]}{E[Z]} \le -\frac{E[\hat{V}_-]}{\partial_Z^+ E[\hat{V}_-]}$ otherwise, and hence $\hat{V}$ is a local maximum for α. Assume $\alpha(\tilde{V}) > \alpha(\hat{V})$ for a $\tilde{V} \in \mathcal{V}$. We can also assume $E[\tilde{V}] = E[\hat{V}]$. From $\partial_Z^+ \alpha(\hat{V}) \le 0$ for $Z = \tilde{V} - \hat{V}$ it follows that $\partial_Z^+ E[\hat{V}_-] \ge 0$, and hence $\alpha(\hat{V} + \epsilon Z) \le \alpha(\hat{V})$ for all $\epsilon \ge 0$, a contradiction. □

3.2. Simple strategies. The scaling properties of α suggest that we look for solutions (λ, H) of (2.3), where $H = \hat{f}\mathbf{1}_{\{V>0\}} - \hat{g}\mathbf{1}_{\{V<0\}}$ for $\hat{f}, \hat{g} \in \mathbf{R}$. Note that then

$$\alpha(V_{1,H}) = \frac{E[V_+](1 + \hat{f}E[S]) + E[V_-](-1 + \hat{g}E[S])}{E[V_+]F(\hat{f}) + E[V_-]G(\hat{g})}, \tag{3.1}$$

where F, G are defined for $h \in \mathbf{R}$ by $F(h) := E[(1+hS)_-]$ and $G(h) := E[(-1+hS)_-]$.

Define $u^+(h) := \lim_{\epsilon \searrow 0} \frac{u(h+\epsilon)-u(h)}{\epsilon}$ and $u^-(h) := \lim_{\epsilon \searrow 0} \frac{u(h)-u(h-\epsilon)}{\epsilon}$ for a convex function u defined in a neighborhood of $h \in \mathbf{R}$. Set $\bar{S} := \operatorname{esssup} S \in (0, \infty]$ and $\underline{S} := \operatorname{essinf} S \in [-\infty, 0)$. F and G play a crucial role in solving problems (2.1) and (2.3). We collect some of their simple properties for later reference.

LEMMA 3.3. *The following assertions hold:*

1. *F and G are convex nonnegative functions on $\mathbf{R}$, G is strictly positive, $F^{\pm}, G^{\pm}$ are nondecreasing, F^+, G^+ are right-continuous, F^-, G^- are left-continuous, and $\lim_{\epsilon \searrow 0} F^+(h-\epsilon) = F^-(h) \le F^+(h)$, $\lim_{\epsilon \searrow 0} G^+(h-\epsilon) = G^-(h) \le G^+(h)$ holds for all $h \in \mathbf{R}$.*
2. *For all $h \in [0, \infty)$, $F^+(h) = -E[\mathbf{1}_{\{1+hS \le 0\}} S]$, $F^-(h) = -E[\mathbf{1}_{\{1+hS<0\}} S]$ and $G^+(h) = -E[\mathbf{1}_{\{-1+hS<0\}} S]$, $G^-(h) = -E[\mathbf{1}_{\{-1+hS \le 0\}} S]$ holds.*
3. *For all $h \in (-\infty, 0]$, $F^+(h) = -E[\mathbf{1}_{\{1+hS<0\}} S]$, $F^-(h) = -E[\mathbf{1}_{\{1+hS \le 0\}} S]$ and $G^+(h) = -E[\mathbf{1}_{\{-1+hS \le 0\}} S]$, $G^-(h) = -E[\mathbf{1}_{\{-1+hS<0\}} S]$ holds.*
4. *We have $\lim_{h \to \infty} F^+(h) = E[S_-]$, $\lim_{h \to -\infty} F^+(h) = -E[S_+]$, $F^+(h) = 0$ if $h \in [-\bar{S}^{-1}, -\underline{S}^{-1}) \cup \{0\}$ and $F^+(h) \ne 0$ if $h \notin F^{-1}(0) = [-\bar{S}^{-1}, -\underline{S}^{-1}]$ ($\frac{1}{\pm\infty} := 0$).*
5. *We have $\lim_{h \to \infty} G^+(h) = E[S_-]$, $\lim_{h \to -\infty} G^+(h) = -E[S_+]$, $G^+(h) = -E[S]$ if $h \in [\underline{S}^{-1}, \bar{S}^{-1}) \cup \{0\}$ and $G^+(h) \ne -E[S]$ if $h \notin G^{-1}(1) = [\underline{S}^{-1}, \bar{S}^{-1}]$.*

Proof. The assertions easily follow from the convexity of the function $x \mapsto x_-$. Monotonous convergence of difference quotients for convex functions, together with the monotonous convergence theorem, allows us to calculate the one-sided derivatives of F, G. □

The following key auxiliary result will allow us to replace general strategies $(\lambda, H) \in \mathbf{R} \times \mathcal{H}_V$ in the optimization problem (2.3) with simpler ones which assume only two values.

LEMMA 3.4. *For all $V \in \mathcal{V}^{T-1} \setminus \{0\}$, the optimization problem (2.3) admits a solution $(\hat\lambda, \hat H)$, where $\hat\lambda \in \{-1,0,1\}$ and $\hat H = \hat f \mathbf{1}_{\{V>0\}} - \hat g \mathbf{1}_{\{V<0\}}$ for constants $\hat f, \hat g \in \mathbf{R}$.*

Proof. The trick of the proof is to apply 1st order optimality conditions to finite-dimensional approximations of the original optimization problem. Let $(\lambda^n, H^n)_{n\geq 1}$ be a sequence in $\mathbf{R}\times\mathcal{H}_V$ such that $\lim_{n\to\infty}\alpha(V_{\lambda^n,H^n}) = \alpha_V$. Approximating each H^n by a sequence $(H^n_k)_{k\geq 1}$ of finite step functions in $\mathbf{R}\times\mathcal{H}_V$, we easily find, using continuity of α on $\mathbf{R}\times\mathcal{H}_V\setminus\{0\}$, a sequence $(H^n_{k_n})_{n\geq 1}$ with $\lim_{n\to\infty}\alpha(V_{\lambda^n,H^n_{k_n}}) = \alpha_V$. Note that we can assume $H^n_{k_n} = \mathbf{1}_{\{V\neq 0\}}H^n_{k_n}$. Setting $\mathcal{F}^n := \sigma(\mathbf{1}_{\{V=0\}}, \mathbf{1}_{\{V>0\}}, H^i_{k_i}, 1\leq i\leq n)$, we have constructed an increasing sequence of finite σ-algebras $\mathcal{F}^n \subseteq \mathcal{F}_{T-1}$ such that $\{V=0\}$ is an atom in $\mathcal{F}^n$ and with $\{V<0\}, \{V>0\} \in \mathcal{F}^n$ for all $n\geq 1$. Denote by $\mathcal{H}^n$ the $\mathcal{F}^n$-measurable elements of L^∞. Since for $n\geq 1$, $\alpha(V_{\lambda^n,H^n}) = \alpha(V_{x\lambda^n,xH^n})$ for all $x\in(0,\infty)$, the finite-dimensional optimization problem $\sup_{\lambda\in\mathbf{R},H\in\mathcal{H}^n}\alpha(V_{\lambda,H})$ admits, by continuity of α on $\mathbf{R}\times\mathcal{H}^n\setminus\{0\}$, a solution on a compact sphere. Therefore, by scaling,

$$\alpha^n_V := \sup_{\lambda\in\{-1,0,1\},H\in\mathcal{H}^n}\alpha(V_{\lambda,H}) \tag{3.2}$$

admits for all $n\geq 1$ a solution $(\hat\lambda,\hat H)\in\{-1,0,1\}\times\mathcal{H}^n$ with $\alpha(V_{\hat\lambda,\hat H}) =: \alpha^n_V \geq 0$ and such that $\lim_{n\to\infty}\alpha^n_V = \alpha_V$. We are going to show that $\hat H$ can be chosen to be constant on $\{V>0\}$ and on $\{V<0\}$ and to vanish on $\{V=0\}$. This implies that $\alpha_V = \alpha^n_V = \alpha(V_{\hat\lambda,\hat H})$ holds for all $n\geq 1$, and the assertion will be proved.

If $\alpha_V = 0$, then $\alpha^n_V = 0$ for all $n\geq 1$ is equivalent to $E[S] = E[V] = 0 = \alpha_V$, and we can choose $(\hat\lambda,\hat H) = (0,0)$ for all $n\geq 1$. Assume $\alpha_V > 0$ and $N\in\mathbf{N}$ such that $\alpha^n_V > 0$ for all $n\geq N$. Fix $n\geq N$. We have to distinguish two cases. First, assume $\hat\lambda = 0$. We then have

$$\alpha(V_{\hat\lambda,\hat H}) = \frac{(E[(V\hat H)_+] - E[(V\hat H)_-])E[S]}{E[(V\hat H)_+]E[S_-] + E[(V\hat H)_-]E[S_+]} \leq \frac{E[S]}{E[S_-]},$$

and hence we can choose $\hat H = \mathbf{1}_{\{V>0\}} - \mathbf{1}_{\{V<0\}}$ so that $\alpha(V_{0,\hat H}) = \frac{E[S]}{E[S_-]}$. Assume now $\hat\lambda\in\{-1,1\}$. Let $(A_m)_{0\leq m\leq M}$ be the sequence of nonempty atoms in $\mathcal{F}^n$; i.e., for $0\leq m,\tilde m\leq M$, $P(A_m)>0$ and $A_m\cap A_{\tilde m} = \emptyset$ iff $m\neq\tilde m$ and $\mathcal{F}^n = \sigma(A_m, 1\leq m\leq M)$. By assumption, $M\geq 2$ and, without loss of generality, we can assume that $A_0 = \{V=0\}$ and that there exists a $1\leq\tilde M<M$ such that $\bigcup_{m=1}^{\tilde M}A_m = \{\hat V>0\}$ and $\bigcup_{m=\tilde M+1}^{M}A_m = \{\hat V<0\}$, where $\hat V := \hat\lambda V$. Note that $V_{\lambda,H} = (\lambda V)_{1,\lambda H}$ for $\lambda\in\{-1,1\}$. Hence for $\hat\lambda\hat H = \sum_{i=1}^{\tilde M}\mathbf{1}_{A_i}h_i - \sum_{i=\tilde M+1}^{M}\mathbf{1}_{A_i}h_i$, where $h_i\in\mathbf{R}, 1\leq i\leq M$, we have

$$\alpha(V_{\hat\lambda,\hat H}) = \frac{\sum_{i=1}^{\tilde M}E[\hat V_+\mathbf{1}_{A_i}](1+h_iE[S]) + \sum_{i=\tilde M+1}^{M}E[\hat V_-\mathbf{1}_{A_i}](-1+h_iE[S])}{\sum_{i=1}^{\tilde M}E[\hat V_+\mathbf{1}_{A_i}]F(h_i) + \sum_{i=\tilde M+1}^{M}E[\hat V_-\mathbf{1}_{A_i}]G(h_i)}.$$

The necessary 1st order optimality conditions, $\frac{\partial^+}{\partial h_i}\alpha(V_{\hat\lambda,\hat H})\leq 0$ and $\frac{\partial^-}{\partial h_i}\alpha(V_{\hat\lambda,\hat H})\leq 0$ for $1\leq i\leq M$, imply $\alpha^n_V F^-(h_i)\leq E[S]\leq\alpha^n_V F^+(h_i)$ for all $1\leq i\leq\tilde M$ and $\alpha^n_V G^-(h_i)\leq E[S]\leq\alpha^n_V G^+(h_i)$ for all $\tilde M<i\leq M$. Define $I := \{h\in\mathbf{R}|\ \alpha^n_V F^-(h)\leq E[S]\leq\alpha^n_V F^+(h)\}$ and $\tilde I := \{h\in\mathbf{R}|\ \alpha^n_V G^-(h)\leq E[S]\leq\alpha^n_V G^+(h)\}$. By Lemma 3.3, we have $I = [\underline f,\bar f]$ and $\tilde I = [\underline g,\bar g]$ for constants $-\infty<\underline f\leq\bar f\leq\infty$ and $-\infty<\underline g\leq$

$\bar{g} \leq \infty$, where "∞]" is to be interpreted as "∞)" for $\bar{f} = \infty$, resp., $\bar{g} = \infty$. Clearly, if $\underline{f} = \bar{f}$, resp., if $\underline{g} = \bar{g}$, then $\hat{H}$ is constant on $\{\hat{V} > 0\}$, resp., on $\{\hat{V} < 0\}$. It follows from straightforward calculations, that for $\underline{f} < \bar{f}$, resp., $\underline{g} < \bar{g}$, $h_i \in I, 1 \leq i \leq \tilde{M}$, resp., $h_i \in \tilde{I}, \tilde{M} + 1 \leq i \leq M$, can be chosen to be equal to $\hat{f} \in I$, resp., $\hat{g} \in \tilde{I}$. □

We can now give a first characterization of solutions (λ, H) of problem (2.3) in the case $\lambda \in \{-1, 1\}$. Note, however, that at this point it is not clear under which conditions there exists such a solution. We will see in Proposition 3.8 that such solutions do not exist in all cases.

PROPOSITION 3.5. *Let $V \in \mathcal{V}^{T-1} \setminus \{0\}$. $(\hat{\lambda}, \hat{\lambda}\hat{H}) \in \{-1, 1\} \times \mathcal{H}_V$ solves (2.3) iff $E[\hat{V}] \geq 0$ for $\hat{V} := \hat{\lambda} V$ and the following inequalities hold:*

$$\alpha_V F^-(\hat{H}) \leq E[S] \leq \alpha_V F^+(\hat{H}) \quad \text{a.s. on } \{\hat{V} > 0\}, \tag{3.3}$$

$$\alpha_V G^-(\hat{H}) \leq E[S] \leq \alpha_V G^+(\hat{H}) \quad \text{a.s. on } \{\hat{V} < 0\}. \tag{3.4}$$

Proof. Note that the assertion holds if $\alpha_V = 0$. Assume $\alpha_V > 0$. From the proof of Lemma 3.4, we know that $(\hat{\lambda}, \hat{\lambda}\hat{H}) \in \{-1, 1\} \times \mathcal{H}_V$ solves (2.3) if (3.3) and (3.4) hold. Conversely , given a solution $(\hat{\lambda}, \hat{\lambda}\hat{H}) \in \{-1, 1\} \times \mathcal{H}_V$ for (2.3), it is not difficult to show that (3.3) and (3.4) hold by calculating right-derivatives at $\hat{H}$ in directions $\mathbf{1}_A$ for measurable sets A, where (3.3) or (3.4) does not hold, and applying the 1st order optimality conditions from Proposition 3.2.

It remains to show that $E[\hat{V}] \geq 0$. This is clearly the case if $E[S] = 0$. Assume $E[S] > 0$. From the proof of Lemma 3.4 and by Lemma 3.3(4), (5), we know that if $(\hat{\lambda}, \hat{\lambda}\hat{H})$ solves (2.3); then we can find $\hat{f}, \hat{g} > 0$ such that $(1, \hat{f}\mathbf{1}_{\{\hat{V}>0\}} - \hat{g}\mathbf{1}_{\{\hat{V}<0\}})$ solves (2.3) for $\hat{V}$ and $F(\hat{f}) > 0$. Since

$$\frac{1 + \hat{f}E[S]}{F(\hat{f})} = \frac{\hat{f}^{-1} + E[S]}{F(\hat{f})\hat{f}^{-1}} \geq \frac{\hat{f}^{-1} + E[S]}{F^-(\hat{f})} > \frac{E[S]}{F^-(\bar{f})} \geq \alpha_V$$

by convexity of F, we find $\frac{-1+\hat{g}E[S]}{G(\hat{g})} < \alpha_V$, and, for $x, y > 0$,

$$\frac{x(1 + \hat{f}E[S]) + y(-1 + \hat{g}E[S])}{xF(\hat{f}) + yG(\hat{g})} \geq \frac{y(1 + \hat{f}E[S]) + x(-1 + \hat{g}E[S])}{yF(\hat{f}) + xG(\hat{g})}$$

iff $x \geq y$, and hence $E[\hat{V}_+] \geq E[\hat{V}_-]$. □

In order to study the existence of solutions (λ, H) of problem (2.3) with $\lambda \in \{-1, 1\}$, we must study several marginal cases which require separate analysis. This will be performed in the following two subsections before we consider the generic case.

3.3. Marginal cases. For $V \in \mathcal{V}^{T-1} \setminus \{0\}$, denote by $\hat{\mathcal{V}}_V$ the set of all $(\lambda, H) \in \{-1, 0, 1\} \times L^\infty(\mathcal{F}_{T-1})$ such that $\alpha_V = \alpha(V_{\lambda,H})$.

Let us first deal with the case $E[S] = 0$. The following result essentially means that one cannot improve a strategy by gambling if $E[S] = 0$.

PROPOSITION 3.6. *Let $V \in \mathcal{V}^{T-1} \setminus \{0\}$ and let $(\lambda, H) \in \{-1, 0, 1\} \times L^\infty(\mathcal{F}_{T-1})$. The following assertions hold:*

1. $\alpha_V = 0 \Leftrightarrow E[S] = E[V] = 0 \Leftrightarrow \hat{\mathcal{V}}_V = \{-1, 0, 1\} \times L^\infty(\mathcal{F}_{T-1})$.
2. $\alpha_V = \max(\alpha(V), \alpha(-V))$ *iff* $E[S] = 0$. *If* $\alpha_V > 0$ *and* $E[S] = 0$, *then* $(\lambda, H) \in \hat{\mathcal{V}}_V$ *iff* $\lambda \in \{-1, 1\}$, $E[\lambda V] > 0$ *and* $F(H) = 0$ *a.s. on* $\{\lambda V > 0\}$, *and* $G(H) = 1$ *a.s. on* $\{\lambda V < 0\}$. *Furthermore,* $(\mathrm{sgn}E[V], 0) \in \hat{\mathcal{V}}_V$.

Proof. Assertion (1) is easy to see, and, clearly, (2) holds if $\alpha_V = 0$. Assume $\alpha_V > 0$. $E[S] = 0$ implies $G \geq 1$ since G is convex, $G(0) = 1$, and, by Lemma 3.3,

$G^{\pm}(0) = 0$ holds. Hence, for all $H \in L^\infty(\mathcal{F}_{T-1})$ we have $\alpha(V_{0,H}) = 0$. Thus $\max(\alpha(V_{1,H}), \alpha(V_{-1,H})) \le \max(\alpha(V), \alpha(-V)) = \max(\alpha(V_{1,0}), \alpha(V_{-1,0}))$, and $\alpha_V = \max(\alpha(V), \alpha(-V))$ follows. Conversely, without loss of generality we can assume $E[V] > 0$ and $\alpha_V = \alpha(V) > 0$. $E[S] > 0$ implies $G^+(0) < 0$, and hence we have $G(h) < 1$ for small $h > 0$ and $\alpha(V(1 - h\mathbf{1}_{\{V<0\}}S)) > \alpha(V)$. This contradiction proves the first part of assertion (2). The second part follows directly from Proposition 3.5 and Lemma 3.3(4), (5). □

3.4. Waiting for a good chance. It might happen that α_V for $V \in \mathcal{V}^{T-1} \setminus \{0\}$ equals $\alpha(S)$. This means that gambling up to time $T-1$ with return V is as good as waiting until time $T-1$ and gambling just once on S. The following results characterize this situation.

LEMMA 3.7. *Let $V \in \mathcal{V}^{T-1} \setminus \{0\}$ and set $V_H := V + HS$ for $H \in \mathcal{F}_{T-1}$ such that $HS \in L^1$. We have*

$$\partial^+_{V_H}\alpha(S) = \frac{E[V] - \alpha(S)\big(P(S=0)E[V_-] - P(S<0)E[V]\big)}{E[S_-]}. \tag{3.5}$$

Furthermore, $\partial^+_{V_H}\alpha(S) \le 0$ for all $H \in \mathcal{F}_{T-1}$ with $HS \in L^1$ iff the following condition is satisfied:

$$\alpha(V) \le \frac{\alpha(S)P(S=0)}{1 + \alpha(S)P(S<0)}. \tag{3.6}$$

Proof. By Lemma 3.1, (3.5) follows from

$$\begin{aligned}\partial^+_{V_H}E[S_-] &= E[\mathbf{1}_{\{S=0\}}V_-] - E[\mathbf{1}_{\{S<0\}}(V + HS)] \\ &= P(S=0)E[V_-] - P(S<0)E[V] + E[H]E[S_-].\end{aligned}$$

We find $\partial^+_{V_H}\alpha(S) \le 0$ iff $\alpha(V) - \alpha(S)\big(P(S=0) - P(S<0)\alpha(V)\big) \le 0$, which is equivalent to $\alpha(V) \le \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$. □

PROPOSITION 3.8. *Let $V \in \mathcal{V}^{T-1} \setminus \{0\}$. $\alpha_V = \alpha(S)$ is equivalent to the inequality*

$$\max(\alpha(V), \alpha(-V)) \le \frac{\alpha(S)P(S=0)}{1 + \alpha(S)P(S<0)}. \tag{3.7}$$

If $\alpha_V = \alpha(S) > 0$, then the following assertions hold:

1. *$(0, H)$ with $H \in \mathcal{H}_V$ solves problem (2.3) iff $H \ge 0$ a.s. on $\{V > 0\}$, $H \le 0$ a.s. on $\{V < 0\}$, and $H\mathbf{1}_{\{V \ne 0\}} \ne 0$.*
2. *$(\hat\lambda, \hat\lambda\hat H) \in \{-1, 1\} \times \mathcal{H}_V$ solves problem (2.3) iff the following conditions hold:*
   (a) *Equality holds in (3.7).*
   (b) *$\bar s_0 := \sup\{h | P(0 < S < h) = 0\} > 0$ and $\underline{s}_0 := \inf\{h | P(h < S < 0) = 0\} < 0$. If $E[V] = 0$, then S is bounded away from 0.*
   (c) *$E[\hat V] \ge 0$ for $\hat V := \hat\lambda V$ and a.s. $\hat H \in [-\underline{s}_0^{-1}, \infty)$ a.s. on $\{\hat V > 0\}$, resp., $\hat H \in [\bar s_0^{-1}, \infty)$ a.s. on $\{\hat V < 0\}$.*

   In particular, if (a) and (b) are satisfied, then there exists a solution $(\hat\lambda, \hat\lambda\hat H) \in \{-1, 1\} \times \mathcal{H}_V$ to problem (2.3).

Proof. The equivalence of $\alpha_V = \alpha(S)$ and (3.7) follows from Proposition 3.2 and Lemma 3.7. Assume $\alpha_V = \alpha(S) > 0$. For $H \in \mathcal{H}_V$ we have

$$\alpha(V_{0,H}) = \frac{(E[(VH)_+] - E[(VH)_-])E[S]}{E[(VH)_+]E[S_-] + E[(VH)_-]E[S_+]} \le \alpha(S)$$

and equality holds iff $(VH)_- = 0$ and $(VH)_+ \neq 0$, proving assertion (1). In order to prove assertion (2), assume now $(\hat\lambda, \hat\lambda\hat H) \in \{-1,1\}\times\mathcal{H}_V$ to solve (2.3). By Proposition 3.5 and inequalities (3.3) and (3.4), we find $E[\hat V] \geq 0$, $F^-(\hat H) \leq E[S_-] \leq F^+(\hat H)$ a.s. on $\{\hat V > 0\}$, and $G^-(\hat H) \leq E[S_-] \leq G^+(\hat H)$ a.s. on $\{\hat V < 0\}$. In particular by Lemma 3.3, F^+, resp., G^+, assumes their maximum $E[S_-]$ on $\mathbf{R}$. Condition (c) and the first part of condition (b) follow now since $F^+(h) = E[\mathbf{1}_{\{1+hS<0\}}S_-] = E[S_-]$ iff $P(-h^{-1} < S < 0) = 0$, and $G^+(h) = E[\mathbf{1}_{\{-1+hS<0\}}S_-] = E[S_-]$ iff $P(0 < S < h^{-1}) = 0$. For $f \in (-\underline{s}_0^{-1}, \infty)$ and $g \in (\bar s_0^{-1}, \infty)$, we find $F(f) = fE[S_-] - P(S<0)$, resp., $G(g) = P(S \leq 0) + gE[S_-]$, and calculate for $\bar V := \hat V(1+(f\mathbf{1}_{\{\hat V>0\}} + g\mathbf{1}_{\{\hat V>0\}})S)$

$$\alpha(S) = \alpha_V = \alpha(\bar V) = \frac{E[\hat V] + (E[\hat V_+]f + E[\hat V_-]g)E[S]}{E[\hat V_+]F(f) + E[\hat V_-]G(g)}$$

$$= \frac{E[\hat V] + (E[\hat V_+]f + E[\hat V_-]g)E[S]}{(E[\hat V_+]f + E[\hat V_-])E[S_-] - E[\hat V]P(S<0) + E[\hat V_-]P(S=0)},$$

which is equivalent to $\alpha(\hat V) = \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$, proving (a). $E[V] = 0$ now implies $P(S=0) = 0$, and hence the second part of (b) holds. The converse implication easily follows by again using Proposition 3.5. □

3.5. The generic case. We can now deal with the generic case.

PROPOSITION 3.9. *Let $V \in \mathcal{V}^{T-1}\setminus\{0\}$. If $\alpha_V > \alpha(S)$, then $(\hat\lambda, \hat\lambda\hat H) \in \{-1,0,1\}\times\mathcal{H}_V$ solves problem* (2.3) *iff $\hat\lambda \in \{-1,1\}$, $E[\hat\lambda V] > 0$, and inequalities* (3.3) *and* (3.4) *hold with $\hat V := \hat\lambda V$. Furthermore, problem* (2.3) *admits a solution.*

Proof. By Lemma 3.4 we know that there exists a solution to problem (2.3), and, from the proof of Proposition 3.8(1), we know that necessarily $\hat\lambda \in \{-1,1\}$. The assertion follows now from Proposition 3.5. □

4. Optimal gambling. We can now solve problem (2.1). Denote by $\hat{\mathcal{V}}^t$ the set of solutions for problem (2.1) for $0 < t \leq T$. We introduce the following auxiliary function $R = R_S$:

$$R(x,f,g) := \frac{(x+1)(1+fE[S]) + (-1+gE[S])}{(x+1)F(f) + G(g)}, \quad x,f,g \in \mathbf{R}_+. \tag{4.1}$$

We summarize our results in the following two theorems.

THEOREM 4.1. *The optimization problem* (2.1) *admits a solution for all $0 < T < \infty$. There exists a $V \in \mathcal{V}^{T-1}$ such that $\alpha_V = \bar\alpha^T$ and $\alpha_V = \bar\alpha^T$ holds for $V \in \mathcal{V}^{T-1}$ iff $\alpha(V) = \bar\alpha^{T-1}$. Furthermore, for $1 < T < \infty$, $\bar\alpha^T$ is characterized by the following statements:*

1. *$\bar\alpha^T = 0$ iff $\bar\alpha^{T-1} = 0$ and $E[S] = 0$.*
2. *$\bar\alpha^T \geq \bar\alpha^{T-1}$, and equality holds iff $E[S] = 0$.*
3. *$\bar\alpha^T \geq \alpha(S)$, and equality holds iff $\bar\alpha^{T-1} \leq \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$.*
4. *$\bar\alpha^T = \sup_{f,g\geq 0} R(\bar\alpha^{T-1}, f, g)$. Furthermore, $R(\bar\alpha^{T-1},\cdot,\cdot)$ assumes its maximum on $\mathbf{R}_+^2$ iff one of the following statements holds:*
   (a) *$\bar\alpha^{T-1} > \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$.*
   (b) *$\bar\alpha^{T-1} = \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$ and if $E[S] > 0$, then in addition, $\underline{s}_0 < 0 < \bar s_0$.*

   Under this condition, $\bar\alpha^T = R(\bar\alpha^{T-1}, \hat f, \hat g)$ for $\hat f, \hat g \geq 0$ iff the following inequalities hold:

$$\bar\alpha^T F^-(\hat f) \leq E[S] \leq \bar\alpha^T F^+(\hat f), \tag{4.2}$$

$$\bar\alpha^T G^-(\hat g) \leq E[S] \leq \bar\alpha^T G^+(\hat g). \tag{4.3}$$

Proof. Existence of a solution $V_{\lambda,H}$, $(\lambda,H)\in\{-1,0,1\}\times\mathcal{H}_V$, where $V\in\mathcal{V}^{T-1}$, to problem (2.1) will be proved by induction. Clearly, there exists such a solution for $T=1$. Let $T>1$ and assume existence of solutions for all $1\le t<T$. In particular $\bar\alpha^{T-1}<\infty$. The assertion follows now from Lemma 2.3 and Lemma 3.4.

Assertion (1) and $\bar\alpha^T\ge\bar\alpha^{T-1}$ are easy to see. If $\bar\alpha^T=\bar\alpha^{T-1}>0$, then by the first part of the proof there exists a $V\in\mathcal{V}^{T-1}$ such that $\alpha_V=\bar\alpha^T=\alpha(V)$. By Proposition 3.6(2), we find $E[S]=0$, and, conversely, $E[S]=0$ implies $\alpha_V=\alpha(V)$ for all $V\in\mathcal{V}^{T-1}$ with $E[V]\ge0$, and hence $\bar\alpha^T=\sup_{V\in\mathcal{V}^{T-1}}\alpha_V=\sup_{V\in\mathcal{V}^{T-1}}\alpha(V)=\bar\alpha^{T-1}$. Clearly, $\bar\alpha^T\ge\alpha(S)$, and assertion (3) follows from Proposition 3.8. For the proof of assertion (4), we have to distinguish three cases. First, assume $\bar\alpha^{T-1}>\frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$. Then there exists a $V\in\mathcal{V}^{T-1}$ with $E[V]>0$, and, for all such V, there exist constants $f_V,g_V\ge0$ such that

$$\begin{aligned}\alpha_V&=\frac{E[V_+](1+f_VE[S])+E[V_-](-1+g_VE[S])}{E[V_+]F(f_V)+E[V_-]G(g_V)}\\&=\frac{(\alpha(V)+1)(1+f_VE[S])+(-1+g_VE[S])}{(\alpha(V)+1)F(f_V)+G(g_V)}\\&=R(\alpha(V),f_V,g_V)=\max_{f,g\ge0}R(\alpha(V),f,g);\end{aligned}$$

i.e., α_V and f_V,g_V depend only on $\alpha(V)$. We write $\hat f\big(\alpha(V)\big)$, resp., $\hat g\big(\alpha(V)\big)$ for f_V, resp., g_V. As we have seen at the end of the proof of Proposition 3.5, $\frac{1+f_VE[S]}{F(f_V)}>\alpha_V$, which implies $R(x,\hat f(\alpha(V)),\hat g(\alpha(V)))$ to be increasing in $x>0$, and hence $\bar\alpha^T=\sup_{V\in\mathcal{V}^{T-1}}\alpha_V=\sup_{V\in\hat{\mathcal{V}}^{T-1}}\alpha_V=R(\bar\alpha^{T-1},\hat f,\hat g)$, where $\hat f,\hat g\ge0$ are such that $\bar V:=V(1+(\hat f\mathbf{1}_{\{V>0\}}-\hat g\mathbf{1}_{\{V<0\}})S)$ solves problems (2.3) and (2.1) for $V\in\hat{\mathcal{V}}^{T-1}$, i.e., $\alpha(V)=\bar\alpha^{T-1}$.

For the second case, assume $\bar\alpha^{T-1}=\frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$. Then $\bar\alpha^T=\alpha(S)$, and the first part of assertion (4) follows since $\lim_{h\to\infty}\frac{F(h)}{h}=\lim_{h\to\infty}\frac{G(h)}{h}=E[S_-]$ implies $\lim_{h\to\infty}R(x,h,h)=\alpha(S)$ for all $x\ge0$ and $R(\bar\alpha^{T-1},f,g)\le\bar\alpha^T$ for all $h_+,h_-\ge0$. By Proposition 3.8(2), existence of optimal $\hat f,\hat g\ge0$ follows for $\alpha(S)>0$. For $\alpha(S)=0$, we have $\bar\alpha^T=\bar\alpha^{T-1}=0$, and the assertion is easy to see. Finally, for $\bar\alpha^{T-1}<\frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$, we clearly have $\alpha(S)>0$, and, again by Proposition 3.8(2a), there cannot exist an optimal pair $\hat f,\hat g$. The last part of assertion (4) follows from inequalities (3.3) and (3.4). ∎

Set for $E[S]>0$, ${}_{T-1}\hat{\mathcal{V}}^T:=\{HS\in L^1\setminus\{0\}|\,H\in\mathcal{F}_{T-1},H\ge0\}$, resp., ${}_{T-1}\hat{\mathcal{V}}^T:=\{HS\in L^1|\,H\in\mathcal{F}_{T-1}\}$ for $E[S]=0$. Concerning the set of solutions for problem (2.1), we have the following result.

THEOREM 4.2. *For $1<T<\infty$, $\hat{\mathcal{V}}^T$ is characterized by the following statements:*

1. *If $\bar\alpha^{T-1}=0$ and $E[S]=0$, then $\hat{\mathcal{V}}^T=\mathcal{V}^T$.*
2. *If $\bar\alpha^{T-1}=0$, $E[S]>0$, and S is not bounded away from 0, then $\hat{\mathcal{V}}^T={}_{T-1}\hat{\mathcal{V}}^T$.*
3. *If $\bar\alpha^{T-1}=0$, $E[S]>0$, and S is bounded away from 0, then*

$$\begin{aligned}\hat{\mathcal{V}}^T=\Big\{&V(1+HS)+HS\mathbf{1}_{\{V=0\}}\big|V\in\mathcal{V}^{T-1},H\in[-\underline{s}_0^{-1},\infty)\ \text{a.s. on}\ \{V>0\},\\&H\in[\bar s_0^{-1},\infty)\ \text{a.s. on}\ \{V<0\},HS\mathbf{1}_{\{V=0\}}\in{}_{T-1}\hat{\mathcal{V}}^T\Big\}.\end{aligned}\tag{4.4}$$

4. *If $\bar\alpha^{T-1}=\frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}>0$ and $\bar s_0\underline{s}_0=0$, then $\hat{\mathcal{V}}^T={}_{T-1}\hat{\mathcal{V}}^T$.*

5. *If* $\bar{\alpha}^{T-1} = \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)} > 0$ *and* $\underline{s}_0 < 0 < \bar{s}_0$, *then*

$$\hat{\mathcal{V}}^T = \Big\{ V(1+HS) + HS\mathbf{1}_{\{V=0\}} \big| V \in \hat{\mathcal{V}}^{T-1}, H \in [-\underline{s}_0^{-1}, \infty) \text{ a.s. on } \{V>0\},$$

$$H \in [\bar{s}_0^{-1}, \infty) \text{ a.s. on } \{V<0\}, HS\mathbf{1}_{\{V=0\}} \in {}_{T-1}\hat{\mathcal{V}}^T \cup \{0\} \Big\} \cup {}_{T-1}\hat{\mathcal{V}}^T. \tag{4.5}$$

6. *If* $0 < \bar{\alpha}^{T-1} < \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$, *then* $\hat{\mathcal{V}}^T = {}_{T-1}\hat{\mathcal{V}}^T$.
7. *If* $\bar{\alpha}^{T-1} > \frac{\alpha(S)P(S=0)}{1+\alpha(S)P(S<0)}$ *and* $E[S] > 0$, *then*

$$\hat{\mathcal{V}}^T = \Big\{ V(1+HS) \big| V \in \hat{\mathcal{V}}^{T-1}, \bar{\alpha}^T F^-(H) \le E[S] \le \bar{\alpha}^T F^+(H) \text{ a.s. on } \{V>0\},$$

$$\bar{\alpha}^T G^-(H) \le E[S] \le \bar{\alpha}^T G^+(H) \text{ a.s. on } \{V<0\} \Big\}. \tag{4.6}$$

In particular, $H \ge \underline{S}^{-1} \ge 0$ *a.s. on* $\{V>0\}$ *and* $H \ge \bar{S}^{-1} \ge 0$ *a.s. on* $\{V<0\}$ *for* $V(1+HS) \in \hat{\mathcal{V}}^T$.

8. *If* $\bar{\alpha}^{T-1} > 0$ *and* $E[S] = 0$, *then*

$$\hat{\mathcal{V}}^T = \Big\{ V(1+HS) \big| V \in \hat{\mathcal{V}}^{T-1}, H \in [-\bar{S}^{-1}, -\underline{S}^{-1}] \text{ a.s. on } \{V>0\},$$

$$H \in [\underline{S}^{-1}, \bar{S}^{-1}] \text{ a.s. on } \{V<0\} \Big\}. \tag{4.7}$$

Proof. The proof is an immediate consequence of our considerations. First, note that the eight cases above are mutually exclusive and form a complete list of all possible cases. Assertion (1) follows from Proposition 3.6. Assertions (2)–(6) follow from Proposition 3.8 and Theorem 4.1. Assertions (7) and (8) follow from Proposition 3.9 and Theorem 4.1. □

5. Examples. Note that inequalities (4.2) and (4.3) are in general not easily solved explicitly since $\bar{\alpha}^T$ is part of the solution to be found.

The following result is an easy consequence of Theorem 4.1 and Theorem 4.2.

COROLLARY 5.1. *Assume* $(S_t)_{1\le t\le T}, T < \infty$, *is an i.i.d. sequence of random variables* $S_t \sim S$ *with* $E[S] > 0$ *and* $P(S<0) > 0$. *Then the following statements hold:*

1. $\bar{\alpha}^1 = \alpha(S)$, *and* $\bar{\alpha}^t$ *strictly increases in* $1 \le t \le T$.
2. *For* $1 < t \le T$, $\bar{\alpha}^t = \max_{f,g\ge 0} R(\bar{\alpha}^{t-1}, f, g)$ *and* $\hat{f}_t, \hat{g}_t \ge 0$ *are optimal iff the following inequalities hold:*

$$\bar{\alpha}^t F^-(\hat{f}_t) \le E[S] \le \bar{\alpha}^t F^+(\hat{f}_t), \tag{5.1}$$

$$\bar{\alpha}^t G^-(\hat{g}_t) \le E[S] \le \bar{\alpha}^t G^+(\hat{g}_t). \tag{5.2}$$

3. *The sequence* $(\hat{f}_t, \hat{g}_t)_{1<t\le T}$ *of optimal pairs can be chosen to be nonincreasing in* $1 < t \le T$. *If* S *admits a density with respect to the Lebesgue-measure on* $\mathbf{R}$, *then* $\hat{f}_t, \hat{g}_t$ *are strictly decreasing and satisfy the following equation:*

$$F'(\hat{f}_t) = G'(\hat{g}_t) = \frac{E[S]}{\bar{\alpha}^t}. \tag{5.3}$$

4. *For* $1 < t \le T$,

$$\hat{\mathcal{V}}^t = \Big\{ V(1+HS) \big| V \in \hat{\mathcal{V}}^{t-1}, \bar{\alpha}^t F^-(H) \le E[S] \le \bar{\alpha}^t F^+(H) \text{ a.s. on } \{V>0\},$$

$$\bar{\alpha}^t G^-(H) \le E[S] \le \bar{\alpha}^t G^+(H) \text{ a.s. on } \{V<0\} \Big\}. \tag{5.4}$$

Note. If S admits a density f_S, then it is easy to check that $F'(f) = G'(g)$ iff $\int_{-\frac{1}{f}}^{\frac{1}{g}} u f_S(u)du = 0$. This can facilitate the problem of solving (5.3) considerably.

5.1. Optimization for random walks. Assume that S satisfies $S \in \{\underline{S}, 0, \bar{S}\}$ almost surely, where $-\infty < \underline{S} < 0 < \bar{S} < \infty$, $P(S = \underline{S}), P(S = \bar{S}) > 0$, and $E[S] > 0$. Note that in this case we have $F^+(h) = F^-(h) = 0$ for all $0 \le h < -\underline{S}^{-1}$ and $F^+(h) = E[S_-]$ for all $h \ge -\underline{S}^{-1}$. Furthermore, $G^+(h) = G^-(h) = -E[S]$ for all $0 \le h < \bar{S}^{-1}$ and $G^+(h) = E[S_-]$ for all $h \ge \bar{S}^{-1}$. Since $\bar{\alpha}^T \ge \alpha(S)$, we find $(\hat{f}, \hat{g}) := (-\underline{S}^{-1}, \bar{S}^{-1})$ to solve $R_S(\alpha, \hat{f}, \hat{g}) = \sup_{f,g\ge 0} R_S(\alpha, f, g)$ for all $\alpha \ge 0$. By Corollary 5.1, we find the optimal strategy described in the introduction. Note also that $F(\hat{f}) = 0$ and $G(\hat{g}) = 1 - \hat{g}E[S]$ so that $R_S(\alpha, \hat{f}, \hat{g}) + 1 = (\alpha + 1)\frac{1+\hat{f}E[S]}{1-\hat{g}E[S]}$ follows by a simple calculation. We find for $T > 1$, $\bar{\alpha}^T = (\bar{\alpha}^{T-1} + 1)\frac{1+\hat{f}E[S]}{1-\hat{g}E[S]} - 1$. In the i.i.d. case ($S_t \sim S$ for all $1 \le t \le T$), we find $\bar{\alpha}^T = (\frac{1+\hat{f}E[S]}{1-\hat{g}E[S]})^T - 1$ for all $T \ge 1$. For $P(S \in \{-1, 1\}) = 1$, this simplifies to $\bar{\alpha}^T = (\frac{P(S=1)}{P(S=-1)})^T - 1$, as claimed in the introduction. It easily follows from Proposition 2.2 and Corollary 5.1(1) that it is not optimal to continue gambling after reaching a zero wealth change. It is easy to check that the corresponding $\bar{\alpha}$ diverges to infinity when we try to approximate Brownian motion by a sequence of such random walk models.

6. Some remarks on the multidimensional case. For a random variable X with c.d.f. F_X, define the right-continuous inverse q_X of F_X by $q_X(s) := \inf\{t \in \mathbf{R} |\, F_X(t) > s\}$, $s \in (0, 1)$. For random variables $X, Y \in L^\infty$, we say X stochastically dominates Y in 2nd order ($Y \preceq^2 X$) if

$$\int_{-\infty}^{t} F_X(s)ds \le \int_{-\infty}^{t} F_Y(s)ds \ \forall\, t \in \mathbf{R}. \tag{6.1}$$

Recall the following result; see, e.g., [12, Thm. 2.58].

THEOREM 6.1. *For $X, Y \in L^\infty$, the following conditions are equivalent:*

1. $Y \preceq^2 X$.
2. $E[U(Y)] \le E[U(X)]$ *for all nondecreasing concave functions U defined on* $\mathbf{R}$.
3. *For all* $c \in \mathbf{R}$, $E[(c + X)_-] \le E[(c + Y)_-]$.

LEMMA 6.2. *Assume $E[S] = E[\tilde{S}]$ and $\tilde{S} \preceq^2 S$ for an $\mathcal{F}_T$-measurable $\tilde{S} \in L^1$, independent of $\mathcal{F}_{T-1}$. Then $R_S(\alpha, f, g) \ge R_{\tilde{S}}(\alpha, f, g)$ for all $\alpha, f, g \ge 0$. In particular, α_V calculated with respect to S is greater than or equal to α_V calculated with respect to $\tilde{S}$ for all $V \in \mathcal{V}^{T-1}$.*

Proof. The result follows from Theorem 6.1, which implies $E[(\pm 1 + hS)_-] = hE[(\frac{\pm 1}{h} + S)_-] \le hE[(\frac{\pm 1}{h} + \tilde{S})_-] = E[(\pm 1 + h\tilde{S})_-]$ for all $h > 0$. □

Let at time $T - 1$ gambles $S^i, i \in I \ne \emptyset$ be available, where $S^i \in L^1(\mathcal{F}_T)$ is independent of $\mathcal{F}_{T-1}$ and $E[S^i] \ge 0$ for all $i \in I$. Without loss of generality, we can assume $E[S^i] \in \{0, 1\}$ for all $i \in I$. Assume that there exist $\preceq^2$-maximal elements in $\{S_i |\, i \in I\}$; i.e., assume that there exists an $\bar{I} \subseteq I$ such that for all $i \in I$ with $E[S^i] > 0$, there exists an $\bar{i} \in \bar{I}$ such that $S^i \preceq^2 S^{\bar{i}}$.

Lemma 6.2 immediately implies the following result.

PROPOSITION 6.3. *In the above setting, we have*

$$\alpha_I := \sup_{i\in I} \sup_{f,g\ge 0} R_{S^i}(\alpha, f, g) = \sup_{i\in \bar{I}} \sup_{f,g\ge 0} R_{S^i}(\alpha, f, g). \tag{6.2}$$

In particular, if there exists an $\bar{i} \in I$ such that $E[S^{\bar{i}}] > 0$ and $S^i \preceq^2 S^{\bar{i}}$ for all $i \in I$ with $E[S^i] > 0$, then $\alpha_I = \sup_{f,g\ge 0} R_{S^{\bar{i}}}(\alpha, f, g)$.

This result allows us, e.g., to deal with the classical CAPM (capital asset pricing model; see [22]) or its (symmetric) α-stable generalizations ($1 < \alpha \leq 2$; see [10]) by reducing the multidimensional case to the one-dimensional case.

Acknowledgments. I am grateful to the anonymous referees for their helpful comments. I sincerely thank Walter Schachermayer for his hospitality.

REFERENCES

[1] J. Ang, *A note on the E, SL portfolio selection model*, J. Financial Quant. Anal., 10 (1975), pp. 849–857.
[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.
[3] V. S. Bawa, *Safety first, stochastic dominance and optimal portfolio choice*, J. Financial Quant. Anal., 13 (1978), pp. 255–271.
[4] V. S. Bawa and E. B. Lindenberg, *Capital market equilibrium in a mean-lower partial moment framework*, J. Financial Econ., 5 (1977), pp. 189–200.
[5] A. Bernardo and O. Ledoit, *Gain, loss and asset pricing*, J. Political Economy, 108 (2000), pp. 144–172.
[6] L. Breiman, *Optimal gambling systems for favorable games*, in Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley, 1961, pp. 65–78.
[7] R. C. Dalang, A. Morton, and W. Willinger, *Equivalent martingale measures and no-arbitrage in stochastic securities market models*, Stoch. Stoch. Reports, 29 (1989), pp. 185–202.
[8] F. Delbaen and W. Schachermayer, *A general version of the fundamental theorem of asset pricing*, Math. Ann., 300 (1994), pp. 463–520.
[9] L. E. Dubins and L. J. Savage, *Inequalities for Stochastic Processes: How to Gamble if You Must*, Dover, New York, 1976.
[10] E. F. Fama, *Risk, return, and equilibrium*, J. Political Economy, 79 (1971), pp. 30–55.
[11] P. Fishburn, *Mean-risk analysis with risk associated with below-target return*, Amer. Econ. Rev., 67 (1977), pp. 116–126.
[12] H. Föllmer and A. Schied, *Stochastic Finance*, de Gruyter, Berlin, 2002.
[13] H. Föllmer and P. Leukert, *Quantile hedging*, Finance Stoch., 3 (1999), pp. 251–273.
[14] H. Föllmer and P. Leukert, *Efficient hedging: Cost versus shortfall risk*, Finance Stoch., 4 (2000), pp. 117–146.
[15] D. Heath, S. Orey, V. Pestien, and W. Sudderth, *Minimizing or maximizing the expected time to reach zero*, SIAM J. Control Optim., 25 (1987), pp. 195–205.
[16] W. W. Hogan and J. M. Warren, *Computation of efficient boundary in the E-S portfolio selection model*, J. Financial Quant. Anal., 7 (1972), pp. 1881–1896.
[17] J. L. Kelly, *A new interpretation of information rate*, Bell System Tech. J., 35 (1956), pp. 917–926.
[18] M. Kulldorff, *Optimal control of favorable games with a time limit*, SIAM J. Control Optim., 31 (1993), pp. 52–69.
[19] J. Leitner, *Optimal portfolios with expected loss constraints and shortfall risk optimal martingale measures*, Statist. Decisions, 23 (2005), pp. 49–66.
[20] A. P. Maitra and W. D. Sudderth, *Discrete Gambling and Stochastic Games*, Springer, New York, 1996.
[21] H. Markowitz, *The utility of wealth*, J. Political Economy, 60 (1952), pp. 151–158.
[22] H. M. Markowitz, *Portfolio Selection*, Wiley, New York, 1959.
[23] B. K. Stone, *A general class of three-parameter risk measures*, J. Finance, 28 (1973), pp. 675–685.
[24] W. D. Sudderth and A. Weerasinghe, *Controlling a process to a goal in finite time*, Math. Oper. Res., 14 (1989), pp. 400–409.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 496–518

MIXED ZERO-SUM STOCHASTIC DIFFERENTIAL GAME AND AMERICAN GAME OPTIONS*

S. HAMADÈNE†

Abstract. In this paper we solve the mixed zero-sum stochastic differential game problem in the general case. The main tool is the notion of a local solution of backward stochastic differential equations (BSDEs) with two reflecting barriers. As an application we deal with the American game options.

Key words. mixed zero-sum stochastic differential game, reflected backward SDEs, Dynkin games, American game options

AMS subject classifications. 91A15, 91B74, 60G40, 91A60

DOI. 10.1137/S036301290444280X

1. Introduction.

Suppose the dynamics of a stochastic system is described by a functional differential equation of the form

$$dx_t = f(t, x., u_t, v_t)dt + \Sigma(t, x.)dB_t,\ t \leq T, \text{ and } x_0 = \bar{x}_0 \in R^m.$$

Here the stochastic processes $(u_t)_{t\leq T}$ and $(v_t)_{t\leq T}$ are adapted and stand for, respectively, the intervention functions of two agents c_1 and c_2 on that system. Additionally, the agents are allowed to exit from the control of the system when they decide, i.e., c_1 (resp., c_2) can stop controlling at a stopping time τ (resp., σ). The actions of the agents are not free and their advantages are antagonistic, which means that between c_1 and c_2 there is a payoff $J(u, \tau; v, \sigma)$ which is a cost for c_1 and a reward for c_2.

The problem we are interested in is finding a saddle-point for the payoff functional $J(u, \tau; v, \sigma)$, i.e., a quadruple $(u^*, \tau^*; v^*, \sigma^*)$ which satisfies that for any (u, τ) and (v, σ), we have

$$J(u^*, \tau^*; v, \sigma) \leq J(u^*, \tau^*; v^*, \sigma^*) \leq J(u, \tau; v^*, \sigma^*).$$

This property of $(u^*, \tau^*; v^*, \sigma^*)$ means that when the agents c_1 and c_2 control with (u^*, τ^*) and (v^*, σ^*), respectively, it is advantageous for both to keep the same control. Otherwise, if one of them decides unilaterally to change a strategy, it is penalized if the other keeps its initial one. Actually, when c_1 chooses (u^*, τ^*) as a control action and keeps it, then the best choice for c_2 is to act with (v^*, σ^*). If c_2 deviates, it would get a lesser amount (than what it would obtain if it acts with (v^*, σ^*)). Conversely if c_2 chooses to control with (v^*, σ^*), then the best choice for c_1 is to act with (u^*, τ^*); otherwise c_1 will pay a greater amount.

This problem is a zero-sum game of mixed type because it combines control and stopping.

Mixed game problems are linked to backward stochastic differential equations (BSDEs) with two reflecting barriers, which were first introduced by Cvitanic and Karatzas in [6] (see also [19] and [16]). In [16], Hamadène and Lepeltier showed that if the BSDE with two reflecting barriers associated with the mixed game has a

*Received by the editors April 8, 2004; accepted for publication (in revised form) December 5, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/sicon/45-2/44280.html
†Laboratoire de Statistique et Processus, Université du Maine, 72085 Le Mans Cedex 9, France (hamadene@univ-lemans.fr).

solution, then the game has a value and a saddle-point. The expression of the saddle-point is given explicitly. However, solutions for BSDEs with two reflecting barriers are obtained mainly under two conditions. The first one is the well-known Mokobodski's hypothesis, which requires the existence of a difference of two nonnegative supermartingales between the barriers [6], [16]. The second one is a regularity assumption on one of the barriers which, roughly speaking, should be a semimartingale [6], [19]. For general barriers the solution may not exist. Therefore the mixed game problem remains open in the general case, i.e., without assuming either Mokobodski's hypothesis or regularity of one of the barriers.

Thus the main objective of this paper is to solve this problem in its general setting. We show that, in order to solve the mixed game problem, it is enough to solve locally its associated double obstacle reflected BSDE. Therefore we prove that the game has a saddle-point and a value function which is also continuous.

A particular case of the mixed game problem is the well-known Dynkin game. Actually the latter corresponds to the case when the agents have no control actions but just stopping. The Dynkin game has been widely studied in the past [1, 2, 4, 5, 10, 23, 25, 26, 29]. However, to our knowledge, the most general paper on this subject is the one by Lepeltier and Maingueneau [25]. They proved, without assuming Mokobodski's hypothesis, the existence of the value function for the game which is a right continuous with left limits process. So as a by-product of our result on mixed games, we deduce that, in the case of Brownian filtration, the value function of the Dynkin game is continuous. Additionally, we give a convenient characterization of the saddle-point.

Another setting of Dynkin games that deserves to be mentioned is the one with mixed strategies. This consists of the relaxation of the stopping times. They are replaced by either nondecreasing processes or processes whose sections are stopping times. For more details on this setting see, e.g., the paper by Touzi and Vieille [30] for the case of finite duration T and the one by Laraki and Solan [24] for the case when $T = \infty$. In this framework the required properties for the processes that define the payoff of the game are somehow less stringent than in the usual setting.

Finally, let us point out that when the controllers are not allowed to stop the game (i.e., $\sigma = \tau = T$), this problem turns into the well-known zero-sum stochastic differential game, which was studied by many authors, e.g., Elliott [11], Davis and Elliott [8], and Hamadène and Lepeltier [17], [18].

In the second part of this paper we deal with American game options, which we briefly introduce. In a financial market let us assume we have two agents, a broker c_1 and a trader c_2. The trader pays a premium to the broker and has the right to ask for a payment (a "contingent claim" in the usual language of financial markets), provided by the broker, when he decides to exercise within a period of time $[0, T]$. Thus when c_2 decides to exercise at time σ, she gets an amount equal to L_σ. The difference with the standard American options is that the broker is allowed to cancel the contract which binds her to the trader. In that case he pays a money penalty. Indeed, if c_1 decides to cancel the contract at time τ, she pays c_2 an amount equal to $L_\tau + \alpha_\tau$ (α_τ stands for the money penalty). If c_1 and c_2 agree to stop the option at τ, then the reward for c_2 is equal to $L_\tau + \alpha'_\tau$ with $0 \leq \alpha'_t \leq \alpha_t$.

This problem is typically a Dynkin game. As usual, the main issues are as follows:

(i) What is the fair price of the option at $t \in [0, T]$?

(ii) Does a hedging strategy for the broker exist?

(iii) When it is optimal for the broker (resp., trader) to cancel (resp., exercise) the option?

Cvitanic and Ma [7] and Kifer [20] dealt with American game options. Both papers showed that the value of the option is given by Y_0, where $(Y_t)_{t\leq T}$ is the value function of the associated Dynkin game. In addition, in [20], the author showed the existence of a hedge which replicates the option for the broker in a complete financial market. However, Kifer, whose work is based heavily on the results of [25], dealt with a less general setting than ours. Additionally, in our work we obtain more regularity properties for the model. In [7] the authors make use of BSDEs but consider only specific processes $(L_t)_{t\leq T}$ and $(\alpha_t)_{t\leq T}$, namely, Itô processes.

In this paper we show that the value of the option at any $t \in [0,T]$ is given by $e^{rt}Y_t$, where r is the credit rate of the nonrisky asset and $(Y_t)_{t\leq T}$ is the continuous *local* solution for a specific BSDE with two reflecting barriers. Moreover a hedging strategy after t for the broker, against the game option, exists. We do not require any restriction on the processes $(L_t)_{t\leq T}$ and $(\alpha_t)_{t\leq T}$.

This paper is organized as follows. In section 2, we present the mixed game problem, while section 3 is related to BSDEs with two reflecting barriers. We give the main theorem, which allows us to solve the mixed zero-sum stochastic differential game problem. The proof is given in the appendix. In section 4, we deal with the mixed game problem and show the existence of the value and of a saddle-point for the game. The expression of the latter is given. Moreover we consider an example for which we give a numerical result. Finally, in the last section we study the American game options in a complete financial market. We give the fair price of the option at any time $t \in [0,T]$ and show the existence of a hedging strategy after t. Once again we deal with an example and give some numerical results.

2. Setting of the mixed zero-sum stochastic differential game problem. Throughout this paper $(\Omega, \mathcal{F}, P)$ is a fixed probability space on which is defined a standard m-dimensional Brownian motion $B = (B_t)_{t\leq T}$, whose natural filtration is $(F_t^0 := \sigma\{B_s, s \leq t\})_{t\leq T}$. We denote by $(F_t)_{t\leq T}$ the completed filtration of $(F_t^0)_{t\leq T}$ with the P-null sets of $\mathcal{F}$; therefore $(F_t)_{t\leq T}$ satisfies the usual conditions, i.e., it is right continuous and complete. Moreover

- $\mathcal{P}$ is the σ-algebra on $[0,T]\times\Omega$ of F_t-progressively measurable sets;
- for any stopping time ν, $\mathcal{T}_\nu$ is the set of F_t-measurable stopping times τ such that P-a.s. $\nu \leq \tau \leq T$; $\mathcal{T}_0$ will be simply denoted $\mathcal{T}$;
- $\mathcal{H}^{2,k}$ is the set of $\mathcal{P}$-measurable processes $w = (w_t)_{t\leq T}$, R^k-valued and square integrable w.r.t. $dt \otimes dP$;
- $\mathcal{S}^2$ is the set of $\mathcal{P}$-measurable and continuous processes $w' = (\omega'_t)_{t\leq T}$ such that $E[\sup_{t\leq T} |w'_t|^2] < \infty$.

We now introduce the mixed stochastic differential game problem.

Let $\mathcal{C}$ be the space of continuous functions $\bar{x}$ from $[0,T]$ into R^m and let $\bar{\mathcal{F}}_t$ be the σ-field of $\mathcal{C}$ generated by $\{\Pi_s : \bar{x} \in \mathcal{C} \mapsto \bar{x}_s,\ s \leq t\}$. By $\mathcal{D}$ we denote the σ-field of $[0,T]\times\mathcal{C}$ consisting of all the subsets D, which have the property that the section of D at time t is in $\bar{\mathcal{F}}_t$ and the section of D at $\bar{x}$ is Lebesgue measurable for any $\bar{x}$ (see, e.g., [11] for more details).

Let us consider a mapping $\Sigma :\ (t,\bar{x}) \in [0,T]\times\mathcal{C} \mapsto \Sigma(t,\bar{x}) \in R^{m\times m}$ satisfying the following:

$$\begin{cases} \text{(1.1) } \Sigma \text{ is } \mathcal{D}\text{-measurable.} \\ \text{(1.2) There exists a constant } C \text{ such that } |\Sigma(t,\bar{x}) - \Sigma(t,\bar{x}')| \leq C\|\bar{x} - \bar{x}'\|_t \text{ and} \\ \qquad |\Sigma(t,\bar{x})| \leq C(1 + \|\bar{x}\|_t), \text{ where for any } \bar{x} \in \mathcal{C} \text{ and } t \leq T, \|\bar{x}\|_t = \sup_{s\leq t} |\bar{x}_s|. \\ \text{(1.3) For any } (t,\bar{x}) \in [0,T]\times\mathcal{C}, \text{ the matrix } \Sigma(t,\bar{x}) \text{ is invertible and } |\Sigma^{-1}(t,\bar{x})| \leq \\ \qquad C(1 + \|\bar{x}\|_t^m) \text{ for some constants } C \text{ and } m \geq 0. \end{cases}$$

Let $x_0 \in R^m$ and let $x = (x_t)_{t\le T}$ be the solution of the following standard functional differential equation:

$$x_t = x_0 + \int_0^t \Sigma(s, x.)dB_s,\ t \le T; \tag{1}$$

here we emphasize that Σ depends not only on x_t but also on the path of the process x up to time t. The process $(x_t)_{t\le T}$ exists, since Σ satisfies (1.1)–(1.3) (see, e.g., [28, p. 375]).

Let A (resp., B) be a compact metric space and $\mathcal{U}$ (resp., $\mathcal{V}$) be the space of $\mathcal{P}$-measurable processes $u := (u_t)_{t\le T}$ (resp., $v := (v_t)_{t\le T}$) with values in A (resp., B). Let $f : [0,T] \times \mathcal{C} \times A \times B \to R^m$ be such that

$$\begin{cases} \text{(1.4) there exists a constant } C \text{ such that } |f(t,\bar{x},\tilde{u},\tilde{v})| \le C(1 + \|\bar{x}\|_t) \text{ for any} \\ \qquad t, \bar{x}, \tilde{u}, \text{ and } \tilde{v}; \\ \text{(1.5) for any } (t,\bar{x}) \in [0,T] \times \mathcal{C},\ \text{the mapping } (\tilde{u},\tilde{v}) \mapsto f(t,\bar{x},\tilde{u},\tilde{v}) \text{ is continuous;} \\ \text{(1.6) for any } (\tilde{u},\tilde{v}) \in A \times B,\ \text{the function } (t,\bar{x}) \mapsto f(t,\bar{x},\tilde{u},\tilde{v}) \text{ is } \mathcal{D}\text{-measurable.} \end{cases}$$

For $(u,v) \in \mathcal{U} \times \mathcal{V}$, let $\Gamma^{u,v} := (\Gamma_t^{u,v})_{t\le T}$ be the local martingale defined by

$$d\Gamma_t^{u,v} = \Gamma_t^{u,v}\Sigma^{-1}(t,x.)f(t,x.,u_t,v_t)dB_t, t \le T,\ \text{and } \Gamma_0^{u,v} = 1.$$

As Σ and f verify (1.1), (1.3), (1.4), and (1.6), the measure $P^{u,v}$ defined by $P^{u,v}(A) = E[\Gamma_T^{u,v} 1_A]$ for $A \in \mathcal{F}$ is actually a probability on $(\Omega, \mathcal{F})$ (see, e.g., [13]). Additionally, thanks to Girsanov's theorem (see, e.g., [21, p. 191] or [28, p. 331]), under the probability $P^{u,v}$, the process $dB_t^{u,v} = dB_t - \Sigma^{-1}(t,x.)f(t,x.,u_t,v_t)dt$, $t \le T$, is a Brownian motion and $(x_t)_{t\le T}$ satisfies

$$x_t = x_0 + \int_0^t f(s,x.,u_s,v_s)ds + \int_0^t \Sigma(s,x.)dB_s^{u,v},\ t \le T, \tag{2}$$

which means that $(x_t)_{t\le T}$ is a weak solution for the stochastic functional differential equation (2).

The process $x = (x_t)_{t\le T}$ stands for the dynamics of the system when it is not controlled. On that system, two agents c_1 and c_2 intervene. A control action for c_1 (resp., c_2) is a process $u = (u_t)_{t\le T}$ (resp., $v = (v_t)_{t\le T}$) which belongs to $\mathcal{U}$ (resp., $\mathcal{V}$). Thereby $\mathcal{U}$ (resp., $\mathcal{V}$) is called *the set of admissible controls* for c_1 (resp., c_2). When c_1 and c_2 act with, respectively, u and v, the law of the dynamics of the system is the same as the one of x under $P^{u,v}$. It turns out that the interventions of the agents give rise to a drift $f(t,x.,u_t,v_t)$ in the dynamics.

As previously mentioned, an intervention strategy for an agent combines control and stopping. Therefore a strategy for c_1 (resp., c_2) is a pair (u,τ) (resp., (v,σ)), where u (resp., v) is an admissible control and τ (resp., σ) a stopping time. It implies that the agents keep controlling until one of them decides to stop and the system is actually stopped at $\tau \wedge \sigma$. So assume that c_1 (resp., c_2) chooses (u,τ) (resp., (v,σ)) as a strategy of intervention which is not free and generates a payoff, which is a cost for c_1 and a reward for c_2. The expression of that payoff is given by

$$J(u,\tau;v,\sigma) = E^{u,v}\left[\int_0^{\tau\wedge\sigma} h(s,x.,u_s,v_s)ds + U_\tau 1_{[\tau<\sigma]} + L_\sigma 1_{[\sigma<\tau<T]} + Q_\tau 1_{[\tau=\sigma<T]} + \xi 1_{[\tau=\nu=T]}\right], \tag{3}$$

where

$$\begin{cases} (1.7)\ h : [0,T]\times\mathcal{C}\times A\times B \to R^+ \text{ is } \mathcal{D}\otimes\mathcal{B}(A\times B)/\mathcal{B}(R^+)\text{-measurable and for any } (t,\bar{x})\in[0,T]\times\mathcal{C},\ \text{the mapping } h(t,\bar{x},.,.):\ (\tilde{u},\tilde{v})\in A\times B\mapsto h(t,\bar{x},\tilde{u},\tilde{v}) \text{ is continuous. In addition there exists a constant } C \text{ such that for any } (t,\bar{x},\tilde{u},\tilde{v})\in[0,T]\times\mathcal{C}\times A\times B, |h(t,\bar{x},\tilde{u},\tilde{v})|\le C(1+\|\bar{x}\|_t). \text{ The process } (h(t,x.(\omega),u_t,v_t))_{t\le T} \text{ stands for the instantaneous payoff between } c_1 \text{ and } c_2,\\ (1.8)\ U=(U_t)_{t\le T}, Q=(Q_t)_{t\le T},\ \text{and } L=(L_t)_{t\le T} \text{ are processes of } \mathcal{S}^2 \text{ such that } L_t\le Q_t\le U_t, \forall\, t\le T,\\ (1.9)\ \xi\in L^2(\Omega,F_T,dP;R) \text{ and satisfies P-}a.s.\ L_T\le\xi\le U_T. \end{cases}$$

The meaning of $J(u,\tau;v,\sigma)$ is as follows:

(i) When the two agents decide to stop at the same time $\tau=\sigma<T$, c_2 obtains from c_1 an amount equal to $\int_0^\tau h(s,x,u_s,v_s)ds+Q_\tau 1_{[\tau=\sigma<T]}+\xi 1_{[\tau=\sigma=T]}$.

(ii) If c_1 is the first to decide to stop controlling at $\tau<T$, he pays c_2 the amount $\int_0^\tau h(s,x,u_s,v_s)ds+U_\tau$.

(iii) If c_2 is the first to decide to stop controlling at $\sigma<T$, he gets from c_1 the amount $\int_0^\sigma h(s,x,u_s,v_s)ds+L_\sigma$.

The difference U_t-L_t is the money penalty that c_1 pays for the unilateral decision to exit from the system, while Q_t-L_t is a money penalty for c_1, which corresponds to the situation when both agree to stop controlling before T.

The problem we are interested in is finding two intervention strategies (u^*,τ^*) and (v^*,σ^*), respectively, for c_1 and c_2 such that for any (u,τ) and (v,σ), we have

$$J(u^*,\tau^*;v,\sigma)\le J(u^*,\tau^*;v^*,\sigma^*)\le J(u,\tau;v^*,\sigma^*).$$

The quadruple $(u^*,\tau^*;v^*,\sigma^*)$ is called a *saddle-point* for the zero-sum mixed stochastic differential game.

The particular case when $h\equiv 0$ and $f\equiv 0$ corresponds to the well-known Dynkin game studied by many authors; see, e.g., [1, 2, 4, 5, 6, 10, 23, 25, 26, 29].

3. Connection with BSDEs with two reflecting barriers. Let $\varphi:[0,T]\times\Omega\times R^{1+m}\to R$ satisfy the following conditions:

$$\begin{cases} (2.1)\ \forall (y,z)\in R^{1+m},\ \text{the process } (\varphi(t,\omega,y,z))_{t\le T} \text{ is } \mathcal{P}\text{-measurable and the process } (\varphi(t,\omega,0,0))_{t\le T} \text{ belongs to } \mathcal{H}^{2,1}.\\ (2.2)\ \varphi \text{ is uniformly Lipschitz with respect to } (y,z);\ \text{i.e., there exists a constant } C \text{ such that } |\varphi(t,\omega,y,z)-\varphi(t,\omega,y',z')|\le C(|y-y'|+|z-z'|)\ \forall t,y,y',z, \text{ and } z'. \end{cases}$$

A solution for the double obstacle reflected BSDE, whose coefficient (resp., terminal value; resp., lower and upper barriers) is φ (resp., ξ; resp., L and U), is a quadruple $(\bar{Y}_t,\bar{Z}_t,\bar{K}_t^+,\bar{K}_t^-)_{t\le T}$ of $\mathcal{P}$-measurable processes with values in $R^{1+m+1+1}$ such that

$$(4)\quad \begin{cases} \bar{Y},\bar{K}^\pm\in\mathcal{S}^2,\ \bar{Z}\in\mathcal{H}^{2,m};\ \bar{K}^\pm \text{ are nondecreasing processes } (\bar{K}_0^\pm=0);\\ \bar{Y}_t=\xi+\displaystyle\int_t^T\varphi(s,\bar{Y}_s,\bar{Z}_s)ds+(\bar{K}_T^+-\bar{K}_t^+)-(\bar{K}_T^--\bar{K}_t^-)-\int_t^T\bar{Z}_s dB_s\ \forall t\le T;\\ L_t\le\bar{Y}_t\le U_t\ \forall t\le T\ \text{ and }\ \displaystyle\int_0^T(U_s-\bar{Y}_s)d\bar{K}_s^-=\int_0^T(L_s-\bar{Y}_s)d\bar{K}_s^+=0. \end{cases}$$

For general processes L and U, (4) may not have a solution. Actually if $L\equiv U$, and L is not a semimartingale, then obviously there is no solution. As shown in

[16], in order to solve the mixed game problem, it is sufficient to solve its associated BSDE with two reflecting barriers, which is of type (4) with a specific drift φ. Indeed, φ is just the minmax of the *Hamiltonian* function of the game. However, a careful examination of the proof of Theorem 4 in [16] shows that it is unnecessary to solve the associated BSDE in $[0,T]$, but just in $[0,\theta]$, where θ is an appropriate stopping time. In a way, we need to solve (4) *locally* and not globally. This observation is the key point of this work. Theorem 3.1 below provides a local solution for the above reflected BSDE, which enables us to solve the mixed game problem in its general setting, i.e., without assuming either Mokobodski's hypothesis or regularity of the barriers. The proof of this theorem, given in the appendix, is based on a penalization scheme of a BSDE with one reflecting barrier and concatenation.

THEOREM 3.1. *There exists a $\mathcal{P}$-measurable lower semicontinuous process $(Y_t)_{t\le T}$ such that*

(i) $Y_T = \xi$;

(ii) *for any stopping time ν, there exist another stopping time $\theta_\nu \ge \nu$, a process $(Z_s^\nu)_{s\ge 0} \in \mathcal{H}^{2,m}$, and two nondecreasing processes $(K^{\nu,\pm})_{s\ge 0} \in \mathcal{S}^2$ $(K_\nu^{\nu,\pm} = 0)$ such that the quadruple $(Y, Z^\nu, K^{\nu,\pm})$ satisfies P-a.s.*

$$
(5)\quad \begin{cases} \forall s \in [\nu,\theta_\nu], Y_s = Y_{\theta_\nu} + \displaystyle\int_s^{\theta_\nu} \varphi(s,Y_s,Z_s^\nu)ds + (K_{\theta_\nu}^{\nu,+} - K_s^{\nu,+}) \\ \qquad\qquad - (K_{\theta_\nu}^{\nu,-} - K_s^{\nu,-}) - \displaystyle\int_s^{\theta_\nu} Z_s^\nu dB_s; \\ \forall s \in [\nu,\theta_\nu], L_s \le Y_s \le U_s \text{ and } \displaystyle\int_\nu^{\theta_\nu} (Y_s - L_s)dK_s^{\nu,+} = \int_\nu^{\theta_\nu} (Y_s - U_s)dK_s^{\nu,-} = 0; \\ Y \text{ is continuous on } [\nu,\theta_\nu]; \end{cases}
$$

(iii) *if $\tau_\nu^* := \inf\{s \ge \nu, Y_s = U_s\} \wedge T$ and $\sigma_\nu^* := \inf\{s \ge 0, Y_s = L_s\} \wedge T$, then $\tau_\nu^* \vee \sigma_\nu^* \le \theta_\nu$.*

4. Resolution of the mixed zero-sum stochastic differential game. We focus once again on the mixed zero-sum stochastic differential game introduced in section 2. We show it has a value and a saddle-point; the expression of the latter is given. To do so we use the notion of local solutions for BSDEs with two reflecting barriers introduced by Theorem 3.1.

Thus let us set $H(t,\bar{x},z,u,v) = z\Sigma^{-1}(t,\bar{x})f(t,\bar{x},u,v) + h(t,\bar{x},u,v)$ for any $(t,\bar{x},z,u,v) \in [0,T]\times\mathcal{C}\times R^m \times A \times B$. The function H is called the *Hamiltonian* associated with the game. We now make the following assumption:

[H] We assume that

$$
\begin{cases} \text{(i)} & \text{the function } \Sigma^{-1}.f \text{ is bounded;} \\ \text{(ii)} & \inf_{u\in A}\sup_{v\in B} H(t,\bar{x},z,u,v) = \sup_{v\in B}\inf_{u\in A} H(t,\bar{x},z,u,v) \text{ for any } (t,\bar{x},z) \in \\ & [0,T]\times\mathcal{C}\times R^m. \end{cases}
$$

Assumption [H](ii), called the *Isaacs condition*, is quite natural in zero-sum differential games (see, e.g., [8, 11, 16]). We have the following result.

PROPOSITION 4.1. *There exist two $\mathcal{D}\otimes\mathcal{B}(R^m)$-measurable functions u^* and v^* from $[0,T]\times\mathcal{C}\times R^m$ into A and B, respectively, such that*

(i) *$(u^*(t,\bar{x},z),v^*(t,\bar{x},z))$ is a saddle-point for H, i.e., for any $t,\bar{x},z,u,v$ we have*

$$H(t,\bar{x},z,u^*(t,\bar{x},z),v)\le H(t,\bar{x},z,(u^*,v^*)(t,\bar{x},z))\le H(t,\bar{x},z,u,v^*(t,\bar{x},z));$$

(ii) *the function $z\mapsto H(t,\bar{x},z,(u^*,v^*)(t,\bar{x},z))$ is uniformly Lipschitz.*

Proof. The function $(t,\bar{x},u,v)\mapsto\Sigma^{-1}(t,\bar{x}).f(t,\bar{x},u,v)$ (resp., $h(t,\bar{x},u,v)$) is $\mathcal{D}\otimes\mathcal{B}(A\times B)$-measurable. In addition, the functions f and h are continuous w.r.t. (u,v) and the sets A and B are compact. Therefore, according to Benes' selection theorem [3], there exist two $\mathcal{D}\otimes\mathcal{B}(R^m)$-measurable functions u^* and v^* from $[0,T]\times\mathcal{C}\times R^m$ into A and B, respectively, such that

$$\inf_{u\in A}\sup_{v\in B}H(t,\bar{x},z,u,v)=\sup_{v\in B}H(t,\bar{x},z,u^*(t,\bar{x},z),v)$$

and

$$\sup_{v\in B}\inf_{u\in A}H(t,\bar{x},z,u,v)=\inf_{u\in A}H(t,\bar{x},z,u,v^*(t,\bar{x},z)).$$

It follows that

$$\inf_{u\in A}H(t,\bar{x},z,u,v^*(t,\bar{x},z))\le H(t,\bar{x},z,(u^*,v^*)(t,\bar{x},z))\le\sup_{v\in B}H(t,\bar{x},z,u^*(t,\bar{x},z),v).$$

Then taking into account [H](ii), we obtain

$$\inf_{u\in A}H(t,\bar{x},z,u,v^*(t,\bar{x},z))=H(t,\bar{x},z,(u^*,v^*)(t,\bar{x},z))=\sup_{v\in B}H(t,\bar{x},z,u^*(t,\bar{x},z),v).$$

Henceforth $(u^*,v^*)(t,\bar{x},z)$ is a saddle-point for the function $(u,v)\longmapsto H(t,\bar{x},z,u,v)$.

Next, for any $z,z'\in R^m$ we have

$$\begin{aligned}&\left|\inf_{u\in A}\sup_{v\in B}H(t,\bar{x},z,u,v)-\inf_{u\in A}\sup_{v\in B}H(t,\bar{x},z',u,v)\right|\\&\le\sup_{u\in A}\sup_{v\in B}|H(t,\bar{x},z,u,v)-H(t,\bar{x},z',u,v)|\le\|\Sigma^{-1}.f\|.|z-z'|.\end{aligned}$$

The result follows because of the boundedness of $\Sigma^{-1}.f$. □

For $(t,\omega,z)\in[0,T]\times\Omega\times R^m$, let us set

$$H^*(t,\omega,z)=H(t,x.(\omega),z,(u^*,v^*)(t,x.(\omega),z)),$$

where x is the solution of (1). Let $(Y_t)_{t\le T}$ be the process constructed as in Theorem 3.1, but associated with (H^*,ξ,L,U) (note that H^* does not depend on y). Using once again Theorem 3.1 with $\nu=0$, let $(Z_t,K_t^+,K_t^-)_{t\le T}$ be the triple of processes and θ the stopping time such that

$$(6)\quad\begin{cases}\text{(i) } Z\in\mathcal{H}^{2,m},K^\pm\in\mathcal{S}^2\text{ nondecreasing }(K_0^\pm=0);\\ \text{(ii) }\forall t\le\theta,Y_t=Y_\theta+\displaystyle\int_t^\theta H^*(s,\omega,Z_s)ds+(K_\theta^+-K_t^+)\\ \qquad\qquad-(K_\theta^--K_t^-)-\displaystyle\int_t^\theta Z_sdB_s\,;\,Y_T=\xi;\\ \text{(iii) }\forall t\le\theta,L_t\le Y_t\le U_t\text{ and }\displaystyle\int_0^\theta(Y_s-L_s)dK_s^+=\int_0^\theta(Y_s-U_s)dK_s^-=0;\\ \text{(iv) }Y\text{ is continuous on }[0,\theta].\end{cases}$$

Let $\tau^* := \inf\{s \geq 0, Y_s = U_s\} \wedge T$ and $\sigma^* := \inf\{s \geq 0, Y_s = L_s\} \wedge T$; then we know that $\tau^* \vee \sigma^* \leq \theta$. On the other hand, if $u^* := (u^*(t,x,Z_t))_{t\leq T}$ and $v^* := (v^*(t,x,Z_t))_{t\leq T}$, then the controls u^* and v^* belong to $\mathcal{U}$ and $\mathcal{V}$, respectively, since $u^*(t,\bar{x},z)$ and $v^*(t,\bar{x},z)$ are $\mathcal{D}\otimes\mathcal{B}(R^m)$-measurable. We are ready to give the main result of this paper.

THEOREM 4.2. *The strategy $(\tau^*,u^*;\sigma^*,v^*)$ is a saddle-point for the mixed zero-sum stochastic differential game; i.e., it satisfies*

$$J(u^*,\tau^*;v,\sigma) \leq J(u^*,\tau^*;v^*,\sigma^*) = Y_0 \leq J(u,\tau;v^*,\sigma^*)\ \forall\,(v,\sigma)\ \textit{and}\ (u,\tau).$$

Additionally, Y_0 is the value of the game, i.e.,

$$Y_0 = \inf_{u\in\mathcal{U};\tau\in\mathcal{T}}\sup_{v\in\mathcal{V};\sigma\in\mathcal{T}} J(u,\tau;v,\sigma) = \sup_{v\in\mathcal{V};\sigma\in\mathcal{T}}\inf_{u\in\mathcal{U};\tau\in\mathcal{T}} J(u,\tau;v,\sigma).$$

Proof. Let us show that $Y_0 = J(\tau^*,u^*;\sigma^*,v^*)$. First recall that $\tau^*\vee\sigma^* \leq \theta$ and that Y is continuous on $[0,\theta]$. In addition K^+ and K^- are continuous on $[0,\theta]$ and move only when Y reaches either L or U; thus $K^+_{\tau^*\wedge\sigma^*} = K^-_{\tau^*\wedge\sigma^*} = 0$. It follows that

$$\begin{aligned} Y_0 &= Y_{\tau^*_\wedge\sigma^*} + K^+_{\tau^*\wedge\sigma^*} - K^-_{\tau^*\wedge\sigma^*} + \int_0^{\tau^*\wedge\sigma^*} H(r,x,Z_r,(u^*,v^*)(r,x,Z_r))dr \\ &\qquad - \int_0^{\tau^*\wedge\sigma^*} Z_r dB_r \\ &= Y_{\tau^*_\wedge\sigma^*} + \int_0^{\tau^*\wedge\sigma^*} h(r,x,(u^*,v^*)(r,x,Z_r))dr - \int_0^{\tau^*\wedge\sigma^*} Z_r dB_r^{u^*,v^*}. \end{aligned}$$

As $(\int_0^t Z_r dB_r^{u^*,v^*})_{s\leq T}$ is an (F_t,P^{u^*,v^*})-martingale, taking expectation we get

$$Y_0 = E^{u^*,v^*}[Y_0] = E^{u^*,v^*}\left[Y_{\tau^*_\wedge\sigma^*} + \int_0^{\tau^*\wedge\sigma^*} h(r,x(\omega),Z_r,(u^*,v^*)(r,x(\omega),Z_r))dr\right]$$

because Y_0 is F_0-measurable, and hence deterministic. Now P-a.s., and also P^{u^*,v^*}-a.s. (since they are equivalent probabilities); we have

$$Y_{\tau^*\wedge\sigma^*} = Y_{\tau^*}1_{[\tau^*<\sigma^*]} + Y_{\sigma^*}1_{[\sigma^*<\tau^*]} + Y_{\sigma^*}1_{[\sigma^*=\tau^*<T]} + \xi 1_{[\tau^*=\sigma^*=T]}.$$

But $Y_{\tau^*}1_{[\tau^*<\sigma^*]} = Y_{\tau^*}1_{[\tau^*<\sigma^*]}.1_{[\theta<T]} + Y_{\tau^*}1_{[\tau^*<\sigma^*]}.1_{[\theta=T]}$. On $[\theta<T]$, $\tau^*\leq\theta<T$, and Y is continuous on $[0,\theta]$; therefore $Y_{\tau^*} = U_{\tau^*}$. On the other hand, on $[\theta=T]$, Y is continuous on $[0,T]$ and $Y_{\tau^*}1_{[\tau^*<\sigma^*]} = U_{\tau^*}1_{[\tau^*<\sigma^*]}$, since in this case $\tau^*<\sigma^*$ implies $\tau^*<T$. It follows that $Y_{\tau^*}1_{[\tau^*<\sigma^*]} = U_{\tau^*}1_{[\tau^*<\sigma^*]}$. In the same way we have $Y_{\sigma^*}1_{[\sigma^*<\tau^*]} = L_{\sigma^*}1_{[\sigma^*<\tau^*]}$. Finally, $Y_{\sigma^*}1_{[\sigma^*=\tau^*<T]} = L_{\sigma^*}1_{[\sigma^*=\tau^*<T]} = U_{\tau^*}1_{[\sigma^*=\tau^*<T]} = Q_{\tau^*}1_{[\sigma^*=\tau^*<T]}$ due to $L\leq Q\leq U$. Thus we have

$$Y_{\tau^*\wedge\sigma^*} = U_{\tau^*}1_{[\tau^*<\sigma^*]} + L_{\sigma^*}1_{[\sigma^*<\tau^*]} + Q_{\sigma^*}1_{[\sigma^*=\tau^*<T]} + \xi 1_{[\tau^*=\sigma^*=T]}\ \ P\text{- and } P^{u^*,v^*}\text{-a.s.},$$

and then

$$\begin{aligned} Y_0 &= E^{u^*,v^*}\left[\int_0^{\tau^*\wedge\sigma^*} h(r,x,(u^*,v^*)(r,x,Z_r))dr + U_{\tau^*}1_{[\tau^*<\sigma^*]} + L_{\sigma^*}1_{[\sigma^*<\tau^*]}\right. \\ &\qquad\qquad \left. + Q_{\sigma^*}1_{[\sigma^*=\tau^*<T]} + \xi 1_{[\tau^*=\sigma^*=T]}\right] \\ &= J(u^*,\tau^*;v^*,\sigma^*). \end{aligned}$$

Next let $v=(v_t)_{t\leq T}$ be an admissible control for c_2 and let $\sigma\in\mathcal{T}$. Let us show that $Y_0\geq J(u^*,\tau^*;v,\sigma)$. Once again since $\tau^*\wedge\sigma\leq\theta$ (thanks to $\tau^*\leq\theta$), we have

$$\begin{aligned}Y_0 &= Y_{\tau^*\wedge\sigma}+K^+_{\tau^*\wedge\sigma}+\int_0^{\tau^*\wedge\sigma}H(r,x,Z_r,(u^*,v^*)(r,x,Z_r))dr-\int_0^{\tau^*\wedge\sigma}Z_rdB_r\\ &= Y_{\tau^*\wedge\sigma}+\int_0^{\tau^*\wedge\sigma}\{H(r,x,Z_r,(u^*,v^*)(r,x,Z_r))-H(r,x,Z_r,u^*(r,x,Z_r),v_r)\}dr\\ &\quad+\int_0^{\tau^*\wedge\sigma}h(r,x,u^*(r,x,Z_r),v_r)dr-\int_0^{\tau^*\wedge\sigma}Z_rdB^{u^*,v}_r+K^+_{\tau^*\wedge\sigma}.\end{aligned}$$

Here, without $\tau^*\leq\theta$ we would not have been able to write this equality since in that case we were not sure whether $\tau^*\wedge\sigma\leq\theta$. As $K^+_{\tau^*\wedge\sigma}\geq 0$ and $H(r,x,Z_r,(u^*,v^*)(r,x,Z_r))-H(r,x,Z_r,u^*(r,x,Z_r),v_r)\geq 0$, then

$$Y_0\geq Y_{\tau^*\wedge\sigma}+\int_0^{\tau^*\wedge\sigma}h(r,x,Z_r,u^*(r,x,Z_r),v_r)dr-\int_0^{\tau^*\wedge\sigma}Z_rdB^{u^*,v}_r.$$

Once more $(\int_0^s Z_rdB^{u^*,v}_r)_{s\leq T}$ is an $(F_s,P^{u^*,v})$-martingale; then taking the expectation with respect to $P^{u^*,v}$ and taking into account the fact that Y_0 is deterministic, we obtain

$$Y_0=E^{u^*,v}[Y_0]\geq E^{u^*,v}\left[Y_{\tau^*\wedge\sigma}+\int_0^{\tau^*\wedge\sigma}h(r,x,u^*(r,x,Z_s),v_r)dr\right].$$

But once again

$$Y_{\tau^*\wedge\sigma}\quad=Y_{\tau^*}1_{[\tau^*<\sigma]}+Y_\sigma 1_{[\sigma<\tau^*]}+Y_{\tau^*}1_{[\tau^*=\sigma<T]}+\xi 1_{[\tau^*=\sigma=T]},$$

and $Y_\sigma 1_{[\sigma<\tau^*]}\geq L_\sigma 1_{[\sigma<\tau^*]}$, $Y_{\tau^*}1_{[\tau^*<\sigma]}=U_{\tau^*}1_{[\tau^*<\sigma]}$. Additionally, $Y_{\tau^*}1_{[\tau^*=\sigma<T]}=U_{\tau^*}1_{[\tau^*=\sigma<T]}\geq Q_{\tau^*}1_{[\tau^*=\sigma<T]}$. Therefore we have

$$\begin{aligned}Y_0 &\geq E^{u^*,v}\left[\int_0^{\tau^*\wedge\sigma}h(r,x,u^*(r,x,Z_r),v_r)dr+U_{\tau^*}1_{[\tau^*<\sigma]}+L_\sigma 1_{[\sigma<\tau^*]}\right.\\ &\qquad\left.+Q_{\tau^*}1_{[\tau^*=\sigma<T]}+\xi 1_{[\tau^*=\sigma=T]}\right]\\ &= J(\tau^*,u^*;\sigma,v)\end{aligned}$$

and $Y_0=J(u^*,\tau^*;v^*,\sigma^*)\geq J(u^*,\tau^*;v,\sigma)$. In the same way we can show that $Y_0=J(u^*,\tau^*;v^*,\sigma^*)\leq J(u,\tau;v^*,\sigma^*)$ for any $\tau\in\mathcal{T}$ and u an admissible control for c_1. It follows that $(u^*,\tau^*;v^*,\sigma^*)$ is a saddle-point for the game.

Finally, let us show that the value of the game is Y_0. We proved that

$$J(u^*,\tau^*;v,\sigma)\leq Y_0=J(u^*,\tau^*;v^*,\sigma^*)\leq J(u,\tau;v^*,\sigma^*)$$

for any $(u,v)\in\mathcal{U}\times\mathcal{V}$ and $\tau,\sigma\in\mathcal{T}$. Thereby

$$Y_0\leq\inf_{u\in\mathcal{U},\tau\in\mathcal{T}}J(u,\tau;v^*,\sigma^*)\leq\sup_{v\in\mathcal{V},\sigma\in\mathcal{T}}\inf_{u\in\mathcal{U},\tau\in\mathcal{T}}J(u,\tau;v,\sigma).$$

On the other hand

$$Y_0 \geq \sup_{v\in\mathcal{V},\sigma\in\mathcal{T}} J(u^*,\tau^*;v,\sigma) \geq \inf_{u\in\mathcal{U},\tau\in\mathcal{T}} \sup_{v\in\mathcal{V},\sigma\in\mathcal{T}} J(u,\tau;v,\sigma).$$

Now due to the inequality

$$\inf_{u\in\mathcal{U},\tau\in\mathcal{T}} \sup_{v\in\mathcal{V},\sigma\in\mathcal{T}} J(u,\tau;v,\sigma) \geq \sup_{v\in\mathcal{V},\sigma\in\mathcal{T}} \inf_{u\in\mathcal{U},\tau\in\mathcal{T}} J(u,\tau;v,\sigma)$$

we have

$$Y_0 = \inf_{u\in\mathcal{U},\tau\in\mathcal{T}} \sup_{v\in\mathcal{V},\sigma\in\mathcal{T}} J(u,\tau;v,\sigma) = \sup_{v\in\mathcal{V},\sigma\in\mathcal{T}} \inf_{u\in\mathcal{U},\tau\in\mathcal{T}} J(u,\tau;v,\sigma).$$

The proof is now complete. □

Let $t\in[0,T]$ be fixed; for τ, $\sigma\in\mathcal{T}_t$, and $(u,v)\in\mathcal{U}\times\mathcal{V}$, we define the conditional payoff after t as follows:

$$J_t(u,\tau;v,\sigma) := E^{u,v}\left[\int_t^{\tau\wedge\sigma} h(s,x,u_s,v_s)ds + U_\tau 1_{[\tau<\sigma]} + L_\sigma 1_{[\sigma<\tau<T]} + Q_\tau 1_{[\tau=\sigma<T]} + \xi 1_{[\tau=\nu=T]}|F_t\right].$$

In the previous result we characterized Y_0 as the value of the game. However, more can be said about the process $(Y_t)_{t\leq T}$. Actually we have the following proposition.

PROPOSITION 4.3. *The process $(Y_t)_{t\leq T}$ is continuous, and for any stopping time ν we have*

$$\begin{aligned} Y_\nu &= \operatorname{ess\,inf} f_{u\in\mathcal{U};\tau\in\mathcal{T}_\nu} \operatorname{ess\,sup}_{v\in\mathcal{V};\sigma\in\mathcal{T}_\nu} J_\nu(u,\tau;v,\sigma) \\ &= \operatorname{ess\,sup}_{v\in\mathcal{V};\sigma\in\mathcal{T}_\nu} \operatorname{ess\,inf} f_{u\in\mathcal{U};\tau\in\mathcal{T}_\nu} J_\nu(u,\tau;v,\sigma); \end{aligned} \tag{7}$$

i.e., the process $(Y_t)_{t\leq T}$ is the value function of the game.

Proof. Let ν be a stopping time. Let θ_ν ($\geq\nu$) be the stopping time with which $(Z_s^\nu)_{s\in[0,T]}\in\mathcal{H}^{2,m}$ and $(K_s^{\nu,\pm})_{s\in[0,T]}\in\mathcal{S}^2$ nondecreasing ($K_\nu^{\nu,\pm}=0$) are such that

$$\begin{cases} \forall s\in[\nu,\theta_\nu], Y_s = Y_{\theta_\nu} + \displaystyle\int_s^{\theta_\nu} H^*(r,Z_r^\nu)dr + (K_{\theta_\nu}^{\nu,+} - K_s^{\nu,+}) \\ \qquad\qquad - (K_{\theta_\nu}^{\nu,-} - K_s^{\nu,-}) - \displaystyle\int_s^{\theta_\nu} Z_r^\nu dB_r,\ Y_T=\xi, \\ \forall s\in[\nu,\theta_\nu], L_s\leq Y_s\leq U_s \ \text{ and } \displaystyle\int_\nu^{\theta_\nu}(Y_r-L_r)dK_r^{\nu,+} = \int_\nu^{\theta_\nu}(Y_r-U_r)dK_r^{\nu,-}=0, \\ Y \text{ is continuous on } [\nu,\theta_\nu]. \end{cases} \tag{8}$$

The existence of θ_ν, Z^ν, and $K^{\nu,\pm}$ stems from Theorem 3.1. If $\tau_\nu^* = \inf\{s\geq\nu, Y_s=U_s\}\wedge T$ and $\sigma_\nu^* = \inf\{s\geq\nu, Y_s=L_s\}\wedge T$, then P-a.s., $\tau_\nu^*\vee\sigma_\nu^*\leq\theta_\nu$. For $s\geq 0$, let $u_s^* = u^*(s,x,Z_s^\nu)$ and $v_s^* = v^*(s,x,Z_s^\nu)$; then u^* and v^* are admissible controls for c_1 and c_2, respectively. Along the same lines as in the proof of Theorem 4.2 we can show that $Y_\nu = J_\nu(u^*,\tau_\nu^*;v^*,\sigma_\nu^*)$. On the other hand if τ, $\sigma\in\mathcal{T}_\nu$ and $(u,v)\in\mathcal{U}\times\mathcal{V}$, then

$$P\text{-a.s. } Y_\nu\leq J_\nu(u,\tau;v^*,\sigma_\nu^*) \text{ and } Y_\nu\geq J_\nu(u^*,\tau_\nu^*;v,\sigma),$$

from which we can infer relation (7). We now focus on the continuity of $(Y_t)_{t\leq T}$. In the construction of Y we used an increasing scheme, but it is possible to use a

decreasing scheme by making the penalization at (11) on the barrier U. Therefore there exists an upper semicontinuous process $\tilde{Y}$ such that for any stopping time ν equality (7) is satisfied with $\tilde{Y}_\nu$ instead of Y_ν. It follows that $Y_\nu = \tilde{Y}_\nu$, P-a.s. for any stopping time ν. The optional section theorem (see [9, p. 220]) implies that P-a.s. for any $t \leq T$, $Y_t = \tilde{Y}_t$. Therefore the process Y is both upper and lower semicontinuous; i.e., it is continuous. □

Remark 4.4. (i) The value function of the zero-sum mixed game does not depend on the values $Q_s, s < T$, of the process Q.

(ii) In the case when $h \equiv 0$ and $f \equiv 0$, the problem is just the well-known zero-sum Dynkin game. Our results, Theorem 4.2 and Proposition 4.3, imply the game has a value function and a saddle-point when L and U are merely continuous processes. We do not require further conditions of Mokobodski type or regularity of the barriers. In addition, the value function is continuous and the saddle-point τ^*, σ^* is characterized as the first times when we have $Y = U$ and $Y = L$, respectively. Those properties of Dynkin games were not known. Actually in [25], for example, the authors show just that the value function is right continuous with left limits and that a saddle-point exists for bounded processes L and U and when $Q = L$. Additionally, a characterization such as ours of the saddle-point is not obtained.

Example 4.5. Let us consider the following example:

$$\begin{aligned} &dx_t = x_t dB_t,\ t \leq 1,\ \text{and } x_0 = 1;\\ &A = [0,1],\ B = [-1,1],\ f(t,x,u,v) = x_t(u-v),\\ &\text{and } h(t,x,u,v) = \min\{|x_t|, 2\} + \frac{1}{2}(u^2+v^2). \end{aligned}$$

Therefore $H(t,x,p,u,v) = p(u-v)+\min\{|x_t|,2\}+\frac{1}{2}(u^2+v^2)$, and obviously the Isaacs condition is satisfied with $u^*(p) = 1_{[p<-1]} - p1_{[-1\leq p\leq 0]}$ and $v^*(p) = 1_{[p<0]} - 1_{[p\geq 0]}$ for any $p \in R$. It follows that

$$\begin{aligned} H^*(t,x,p) := H(t,x,p,u^*(p),v^*(p)) &= pu^*(p) - pv^*(p) + \min\{|x_t|,2\}\\ &\quad + \tfrac{1}{2}(u^*(p)^2 + v^*(p)^2)\\ &= p1_{[p<-1]} - p^2 1_{[-1\leq p\leq 0]} + |p| + \min\{|x_t|,2\}\\ &\quad + \tfrac{1}{2}\{1 + 1_{[p<-1]} + p^2 1_{[-1\leq p\leq 0]}\}. \end{aligned}$$

Assume that for $t \leq 1$, $U_t = \min(2, x_t)$, $L_t = \min(1, \frac{1}{3}x_t)$, $\xi = \min(2, x_1)$, and let $(Y_t)_{t\leq 1}$ be the value function of the associated mixed stochastic differential game. Figure 1 shows a simulation of a path of the process Y and an evaluation of Y_0.

5. Application in finance. Let $\mathcal{M}$ be a financial market where, for the sake of simplicity, we have a bond $(C_t)_{t\leq T}$ whose evolution is

$$dC_t = rC_t dt, \quad t \leq T, \quad C_0 > 0,$$

and one risky asset, whose price $(S_t)_{t\leq T}$ is given by

$$dS_t = S_t(\mu dt + \nu dB_t), \quad t \leq T, \quad S_0 > 0.$$

Let $t \in [0,T]$ be fixed and let X be a nonnegative F_t-measurable random variable. A self-financing portfolio after t, whose endowment at t is X, is a $\mathcal{P}$-measurable process $\pi = (\beta_s, \gamma_s)_{s\in[t,T]}$ with values in R^2 such that, P-a.s.,

(i) $\displaystyle\int_t^T \{|\beta_s| + (\gamma_s S_s)^2\} ds < \infty,$

(ii) if $A_s^{\pi,X} = \beta_s.C_s + \gamma_s.S_s$, then $A_s^{\pi,X} = X + \displaystyle\int_t^s \beta_u dC_u + \int_t^s \gamma_u dS_u \ \ \forall s \in [t,T].$

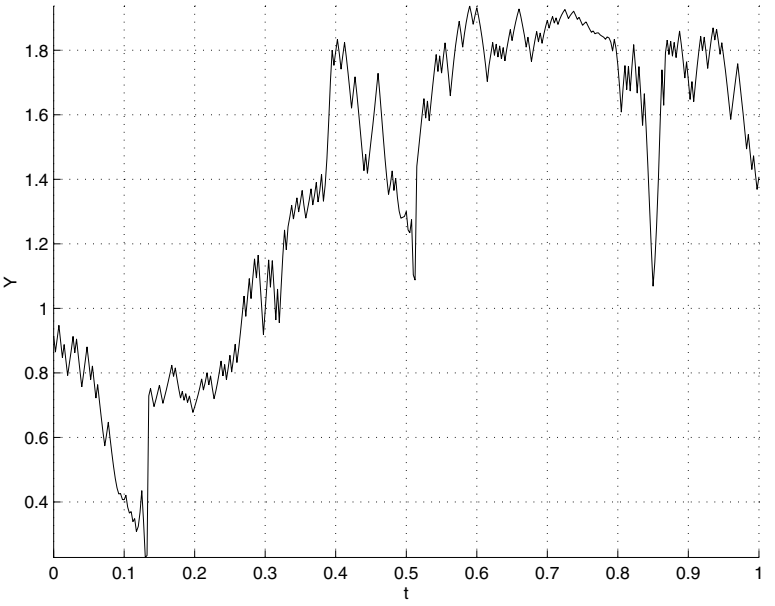


FIG. 1. *Trajectory of the value function* $(Y_t)_{t\leq 1}$ $(Y_0 = 0.9142)$.

Let P^* be the unique probability on $(\Omega, \mathcal{F})$ under which the actualized price of the risky asset is a martingale, i.e.,

$$\frac{dP^*}{dP} := \exp\left\{-\nu^{-1}(\mu - r)B_T - \frac{1}{2}(\nu^{-1}(\mu - r))^2 T\right\}.$$

Therefore under P^*, by Girsanov's theorem (see [21, p. 191] or [28, p. 331]), the process $(W_t = B_t + \nu^{-1}(\mu - r)t)_{t\leq T}$ is an (F_t, P^*)-Brownian motion.

An American game or Israeli option is a contract between a broker c_1 and a trader c_2 who are, respectively, the seller and the buyer of the option. The trader pays an initial amount (the price of the option) which guarantees a payment of $(L_t)_{t\leq T}$ (a "contingent claim" in the usual language of financial markets). The trader can exercise whenever he decides before the maturity T of the option. Thus, if the trader decides to exercise at τ, he gets the amount L_τ. On the other hand, the broker is allowed to cancel the contract which binds him to c_2. Therefore if he chooses σ as the contract cancellation time, he pays the amount U_σ equal to at least the one that would earn c_2 if he decided to exercise at σ, i.e., $U_\sigma \geq L_\sigma$. The difference $U_\sigma - L_\sigma$ is the premium that the broker pays for his decision to cancel the contract. If c_1 and c_2 decide together to stop the contract at the same time τ, then c_2 gets a reward equal to $Q_\tau 1_{[\tau<T]} + \xi 1_{[\tau=T]}$.

In [7], the authors consider this problem, but with payoffs L and U as Itô processes (they also suppose that $Q \equiv L$). Additionally, in [20] Kifer studied this problem in detail. However, since his work is based on the paper by Lepeltier and Maingueneau [25], he was able to consider only the case when $Q = U$. Our framework is more general and we have deduced more properties of the solution of the problem (e.g., continuity of the fair price, characterization of the optimal strategy) than Kifer did.

Let us now consider the probability space $(\Omega, \mathcal{F}, P^*)$ endowed with the Brownian motion $(W_t)_{t\leq T}$. Let $(\tilde{Y})_{t\leq T}$ be the continuous process (through Proposition 4.3) such

that for any $t \le T$, there exist a stopping time $\theta_t \ge t$ and a process $(\tilde{Z}_s)_{s\in[t,T]}$ of $\mathcal{H}^{2,m}$ and $(\tilde{K}_s^+, \tilde{K}_s^-)_{s\in[t,T]}$, which are a pair of nondecreasing processes of $\mathcal{S}^2$ $((\tilde{K}_t^+, \tilde{K}_t^-) = (0,0))$, such that

$$\begin{cases} \forall s \le T, e^{-rs}L_s \le \tilde{Y}_s \le e^{-rs}U_s \text{ and } \tilde{Y}_T = e^{-rT}\xi; \\ \tilde{Y}_s = \tilde{Y}_{\theta_t} + (\tilde{K}_{\theta_t}^+ - \tilde{K}_s^+) - (\tilde{K}_{\theta_t}^- - \tilde{K}_s^-) - \displaystyle\int_s^{\theta_t} \tilde{Z}_r dW_r \ \forall s \in [t, \theta_t], \\ \displaystyle\int_t^{\theta_t} (e^{-ru}U_u - \tilde{Y}_u) d\tilde{K}_u^- = \int_t^{\theta_t} (e^{-ru}L_u - \tilde{Y}_u) d\tilde{K}_u^+ = 0. \end{cases}$$

According to Proposition 4.3, if we set $\tau_t^* := \inf\{s \ge t, \tilde{Y}_s = e^{-rs}U_s\} \wedge T$ and $\sigma_t^* := \inf\{s \ge t, \tilde{Y}_s = e^{-rs}L_s\} \wedge T$, then we have

$$\forall\ \tau, \sigma \in \mathcal{T}_t,\ \bar{\Gamma}_t(\tau_t^*, \sigma) \le \tilde{Y}_t = \bar{\Gamma}_t(\tau_t^*, \sigma_t^*) \le \bar{\Gamma}_t(\tau, \sigma_t^*), \tag{9}$$

where $\bar{\Gamma}_t(\tau,\sigma) = E^*[e^{-r\tau}U_\tau 1_{[\tau<\sigma]} + e^{-r\sigma}L_\sigma 1_{[\sigma<\tau]} + e^{-r\tau}Q_\tau 1_{[\tau=\sigma<T]} + e^{-rT}\xi 1_{[\tau=\sigma=T]}|F_t]$.

A hedge for the broker against the game option after t is a pair (π, τ), where π is a self-financing portfolio after t, whose endowment at t is X and τ is a stopping time $\ge t$, satisfying

$$A_{s\wedge\tau}^{\pi,X} \ge R(s,\tau) := U_\tau 1_{[\tau<s]} + L_s 1_{[s<\tau]} + Q_s 1_{[s=\tau<T]} + \xi 1_{[s=\tau=T]}\ \forall s \in [t,T] P\text{-a.s.}$$

Since the broker is allowed to cancel the option he does not need to hedge with a portfolio up to maturity T but only up to its cancellation time. This is the main reason why a hedging strategy for the broker has two components, a portfolio and a stopping time. The quantity $R(s,\tau)$ is the amount that the broker has to pay if the option is exercised at s or cancelled at τ.

In [20], Kifer showed that the value (the fair price) of the option at $t = 0$ is $\tilde{Y}_0$ and a hedge exists. We show that the fair price of the option at time t is $e^{rt}\tilde{Y}_t$ and a hedge after t exists also.

First, as usual (see, e.g., [14], [22]), we define the value of the option at t by V_t, where $(V_t)_{t\le T}$ is an *rcll* (right continuous with left limits) process which satisfies $\forall t \le T$, P-a.s.

$V_t :=$ ess inf$\{X \ge 0,\ F_t$-measurable such that there exists a hedge (π,τ) after t against the game option, where π is a self-financing portfolio after t whose endowment at t is $X\}$.

Then we have the following theorem.

THEOREM 5.1. *P-a.s., for any $t \le T$, $V_t = e^{rt}\tilde{Y}_t$. Moreover a hedge after t against the option exists.*

Proof. Let $t \le T$ be fixed and let (π,τ) be a hedge after t against the game option. Therefore $\tau \ge t$, and $\pi = (\beta_s, \gamma_s)_{s\in[t,T]}$ is a self-financing portfolio whose value at t is X, satisfying $A_{s\wedge\tau}^{\pi,X} \ge R(s,\tau)\ \forall s \in [t,T]$. But

$$e^{-rs\wedge\tau}A_{s\wedge\tau}^{\pi,X} = e^{-rt}X + \nu\int_t^{s\wedge\tau} \gamma_u S_u e^{-ru} dW_u \ge e^{-rs\wedge\tau}R(s,\tau)\ \forall s \in [t,T]. \tag{10}$$

Let σ be a stopping time $\ge t$. Putting $s = \sigma$ and taking the conditional expectation in (10) we have

$$e^{-rt}X \ge E^*[e^{-r\tau\wedge\sigma}R(\sigma,\tau)|F_t].$$

Therefore

$$\begin{aligned} e^{-rt}X &\geq \text{ess sup}_{\sigma\geq t}E^*[e^{-r\tau\wedge\sigma}R(\sigma,\tau)|F_t] \\ &\geq \text{ess inf}_{\tau\geq t}\text{ess sup}_{\sigma\geq t}E^*[e^{-r\tau\wedge\sigma}R(\sigma,\tau)|F_t] = \tilde{Y}_t \end{aligned}$$

because of (9) and Proposition 4.3. It follows that $V_t \geq e^{rt}\tilde{Y}_t$. Let us now show the reverse inequality. Since $\tau_t^* \leq \theta_t$, then for any $s \geq t$ we have

$$\tilde{Y}_{s\wedge\tau_t^*} = \tilde{Y}_t - \tilde{K}^+_{s\wedge\tau_t^*} + \int_t^{s\wedge\tau_t^*} \tilde{Z}_u dW_u.$$

On the other hand, since $\tilde{Y}_{\tau_t^*}1_{[\tau_t^*<T]} = e^{-r\tau_t^*}U_{\tau_t^*}1_{[\tau_t^*<T]} \geq e^{-r\tau_t^*}Q_{\tau_t^*}1_{[\tau_t^*<T]}$, then

$$\begin{aligned} \tilde{Y}_{s\wedge\tau_t^*} &= \tilde{Y}_s 1_{[s<\tau_t^*]} + \tilde{Y}_{\tau_t^*}1_{[\tau_t^*<s]} + \tilde{Y}_{\tau_t^*}1_{[s=\tau_t^*<T]} + \xi 1_{[s=\tau_t^*=T]} \\ &\geq e^{-rs}L_s 1_{[s<\tau_t^*]} + e^{-r\tau_t^*}U_{\tau_t^*}1_{[\tau_t^*<s]} + e^{-r\tau_t^*}Q_{\tau_t^*}1_{[s=\tau_t^*<T]} + \xi 1_{[s=\tau_t^*=T]} \\ &= e^{-rs\wedge\tau_t^*}R(s,\tau_t^*). \end{aligned}$$

Henceforth we have

$$\tilde{Y}_t + \int_t^{s\wedge\tau_t^*} \tilde{Z}_u dW_u \geq e^{-rs\wedge\tau_t^*}R(s,\tau_t^*)\ \forall s \in [t,T].$$

Next for $s \in [t,T]$ let us set $\tilde{M}_s := e^{rs}(\tilde{Y}_t + \int_t^{s\wedge\tau_t^*} \tilde{Z}_u dW_u)$, $\gamma_s = e^{rs}\tilde{Z}_s(\nu S_s)^{-1}1_{[s\leq\tau_t^*]}$, and $\beta_s = (\tilde{M}_s - \gamma_s S_s)C_s^{-1}$. Therefore for any $s \in [t,T]$ we obviously have $\tilde{M}_s =$

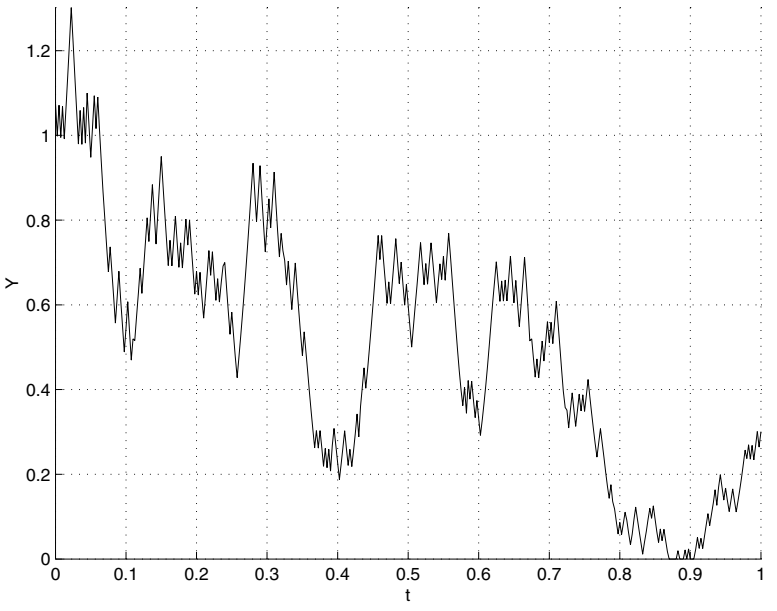


FIG. 2. $S_0 = 2$, $r = \mu = \nu = 1$, $T = 1$, $L_t = (S_t - 1)^+$, $U_t = 1.1(S_t - 1)^+$, *and* $\xi = 1.05(S_1 - 1)^+$.

$\beta_s C_s + \gamma_s S_s$. Additionally, using Itô's formula we obtain

$$\forall s \in [t,T],\ \tilde{M}_s = e^{rt}\tilde{Y}_t + \int_t^s e^{ru}\tilde{Z}_u 1_{[u\le\tau_t^*]}dW_u + \int_t^s r\tilde{M}_u du$$
$$= e^{rt}\tilde{Y}_t + \int_t^s \beta_u dC_u + \int_t^s \gamma_u dS_u.$$

It follows that $(\beta_s,\gamma_s)_{s\in[t,T]}$ is a self-financing portfolio whose value at t is $e^{rt}\tilde{Y}_t$. Since $\tilde{M}_{s\wedge\tau_t^*} \ge R(s,\tau_t^*)$ for any $s\in[t,T]$, then $((\beta_s,\gamma_s)_{s\in[t,T]},\tau_t^*)$ is a hedge against the game option. It implies that $e^{rt}\tilde{Y}_t \ge V_t$ and then $e^{rt}\tilde{Y}_t = V_t$ P-a.s. As $\tilde{Y}$ is continuous and V is $rcll$, those processes are undistinguishable, i.e., P-a.s. for any $t\le T$, $V_t = e^{rt}\tilde{Y}_t$. This ends the proof. □

We finally present a numerical result of an example. Let us take $S_0 = 2$, $r = \mu = \nu = 1$, $T = 1$, $L_t = (S_t - 1)^+$, $U_t = 1.1(S_t-1)^+$, and finally $\xi = 1.05(S_1-1)^+$. Figure 2 shows a simulation of a path of the process $(\tilde{Y}_t = V_t e^{-t})_{t\le 1}$ ($V_0 = 1.0746$).

Appendix. The proof of Theorem 3.1, divided into five steps, is given for the case when $\nu = 0$. If ν is not equal to 0, it just needs a slight modification.

For $n\ge 0$, let $\varphi_n : [0,T]\times\Omega\times R^{1+m}\to R$ such that $\varphi_n(t,\omega,y,z) = \varphi(t,\omega,y,z) + n(L_t - y)^+$. The function φ_n is uniformly Lipschitz with respect to (y,z). According to the result of El-Karoui et al. [12, Theorem 5.2], there exists a process $(Y^n, Z^n, K^{-,n})$ solution of the BSDE with one reflecting barrier associated with (φ_n,ξ,U), i.e.,

(11)
$$\begin{cases} Y^n \in \mathcal{S}^2,\ Z^n\in\mathcal{H}^{2,m};\ K^{-,n} \text{ is a continuous nondecreasing process with } K_0^{-,n} = 0;\\ Y_t^n = \xi + \displaystyle\int_t^T \varphi(s,Y_s^n,Z_s^n)ds + n\int_t^T (L_s - Y_s^n)^+ds - (K_T^{-,n} - K_t^{-,n}) - \int_t^T Z_s^n dB_s;\\ Y_t^n \le U_t\ \forall t\le T \ \text{ and } \displaystyle\int_0^T (U_s - Y_s^n)dK_s^{-,n} = 0. \end{cases}$$

As $\varphi^n \le \varphi^{n+1}$, thus a comparison theorem (see, e.g., [12, Theorem 4.1]) implies that for any $n\ge 0$, $Y^n \le Y^{n+1} \le U$. Let ν be a stopping time, let $n\ge 0$, and $\delta_\nu^n = \inf\{s\ge\nu, Y_s^n = U_s\}\wedge T$. Since $Y^n\le Y^{n+1}$ we have $\delta_\nu^{n+1}\le\delta_\nu^n$ and the sequence of stopping times $(\delta_\nu^n)_{n\ge0}$ converges to another stopping time $\delta_\nu := \lim_{n\to\infty}\delta_\nu^n$. In what follows, δ_0^n (resp., δ_0) will be simply denoted δ^n (resp., δ).

First let us give the following result.

PROPOSITION A.1. *There exists a process $Y := (Y_t)_{t\le T}$ such that P-a.s. for any $t\le T$, $Y_t^n\to Y_t$ as $n\to\infty$ and $(Y^n)_{n\ge0}\to Y$ in $\mathcal{H}^{2,1}$. Moreover Y is lower semicontinuous and satisfies*

$$\text{P-a.s. } \forall t\le T, L_t\le Y_t\le U_t, \text{ and } Y_T = \xi.$$

Proof. Since P-a.s. for any $n\ge0$ we have $Y^n\le Y^{n+1}\le U$ and Y^n is continuous, there exists a $\mathcal{P}$-measurable lower semicontinuous process $(Y_t)_{t\le T}$ such that P-a.s. for any $t\le T$, $Y_t^n\to Y_t$ as $n\to\infty$, $Y_t\le U_t$, and $Y_T = \xi$. On the other hand, the dominated convergence theorem implies that $(Y^n)_{n\ge0}\to Y$ in $\mathcal{H}^{2,1}$.

We want to show that P-a.s. for any $t\le T$, $Y_t\ge L_t$. It is enough to show that P-a.s., $Y_\nu\ge L_\nu$ (ν is a "whatever" stopping time); then the result follows from the optional section theorem [9, p. 220].

For any $n \geq 0$, the process Y^n satisfies (see, e.g., [12])

$$Y_\nu^n = \operatorname{ess\,inf}_{\tau \in \mathcal{T}_\nu} E\left[\int_\nu^\tau \varphi(s, Y_s^n, Z_s^n)ds + n\int_\nu^\tau (L_s - Y_s^n)^+ ds + U_\tau 1_{[\tau<T]} + \xi 1_{[\tau=T]} | F_\nu\right]$$

and

$$Y_\nu^n = E\left[\int_\nu^{\delta_\nu^n} \varphi(s, Y_s^n, Z_s^n)ds + n\int_\nu^{\delta_\nu^n} (L_s - Y_s^n)^+ ds + U_{\delta_\nu^n} 1_{[\delta_\nu^n<T]} + \xi 1_{[\delta_\nu^n=T]} | F_\nu\right].$$

Therefore we have

(12)
$$Y_\nu^n \geq E\left[-\int_\nu^{\delta_\nu^n} |\varphi(s, Y_s^n, Z_s^n)|ds + n\int_\nu^{\delta_\nu^n} (L_s - Y_s^n)^+ ds + U_{\delta_\nu^n} 1_{[\delta_\nu^n<T]} + \xi 1_{[\delta_\nu^n=T]} | F_\nu\right].$$

Since the process $K^{-,n}$ moves only when Y^n reaches the barrier U, then for any $s \leq T$ from (11) we have

$$Y_{(s\vee\nu)\wedge\delta_\nu^n}^n = Y_{\delta_\nu^n}^n + \int_{(s\vee\nu)\wedge\delta_\nu^n}^{\delta_\nu^n} \varphi(r, Y_r^n, Z_r^n)dr + n\int_{(s\vee\nu)\wedge\delta_\nu^n}^{\delta_\nu^n} (L_r - Y_r^n)^+ dr - \int_{(s\vee\nu)\wedge\delta_\nu^n}^{\delta_\nu^n} Z_r^n dB_r.$$

Therefore basic calculations (see, e.g., [12]) imply the existence of a constant C, which does not depend on n, such that

$$E\left[\int_\nu^{\delta_\nu^n} |Z_r^n|^2 dr\right] \leq C \ \forall n \geq 0, \tag{13}$$

since $Y_{\delta_\nu^n}^n$ is uniformly bounded in $L^2(\Omega, F, P)$. As $Y^0 \leq Y^n \leq U$, φ is Lipschitz, and taking into account (13), we deduce from (12) that

$$\int_\nu^{\delta_\nu} (L_r - Y_r)^+ dr = 0, \text{P-a.s.} \tag{14}$$

On the other hand, for any $s \leq T$ we have, due to $\delta_\nu \leq \delta_\nu^n$,

$$Y_{(s\vee\nu)\wedge\delta_\nu}^n = Y_{\delta_\nu}^n + \int_{(s\vee\nu)\wedge\delta_\nu}^{\delta_\nu} \varphi(r, Y_r^n, Z_r^n)dr + n\int_{(s\vee\nu)\wedge\delta_\nu}^{\delta_\nu} (L_r - Y_r^n)^+ dr - \int_{(s\vee\nu)\wedge\delta_\nu}^{\delta_\nu} Z_r^n dB_r,$$

which implies

$$Y_{(s\vee\nu)\wedge\delta_\nu}^n = Y_\nu^n - \int_\nu^{(s\vee\nu)\wedge\delta_\nu} \varphi(r, Y_r^n, Z_r^n)dr - n\int_\nu^{(s\vee\nu)\wedge\delta_\nu} (L_r - Y_r^n)^+ dr + \int_\nu^{(s\vee\nu)\wedge\delta_\nu} Z_r^n dB_r \ \forall s \leq T.$$

From $Y^n \leq Y^{n+1}$ and (13), using a result of Peng [27, Theorem 2.1] we deduce that the process $(Y_{(s\vee\nu)\wedge\delta_\nu})_{s\leq T}$ is right continuous, and hence Y is right continuous on

$[\nu,\delta_\nu]$. Therefore on the set $\{\nu<\delta_\nu\}$, we obtain from (14) that $Y_s\geq L_s$ for $s\in[\nu,\delta_\nu[$ and $Y_\nu\geq L_\nu$. The set $\{\nu=\delta_\nu\}$ is F_ν-measurable, and thus from (12) we have

(15)
$$1_{\{\nu=\delta_\nu\}}Y^n_\nu\geq E\left[1_{\{\nu=\delta_\nu\}}\left\{-\int_\nu^{\delta^n_\nu}|\varphi(s,Y^n_s,Z^n_s)|ds+U_{\delta^n_\nu}1_{[\delta^n_\nu<T]}+\xi 1_{[\delta^n_\nu=T]}\right\}|F_\nu\right].$$

But

$$E\left[1_{\{\nu=\delta_\nu\}}\int_\nu^{\delta^n_\nu}|\varphi(s,Y^n_s,Z^n_s)|ds\right]\to 0 \text{ as } n\to\infty.$$

Then taking the limit in (15), if necessary extracting a subsequence, we obtain

$$1_{\{\nu=\delta_\nu\}}(Y_\nu-U_{\delta_\nu}1_{[\delta_\nu<T]}-\xi 1_{[\delta_\nu=T]})\geq 0$$

which implies $1_{\{\nu=\delta_\nu\}}(Y_\nu-L_\nu)\geq 0$. It follows that P-*a.s.* $Y_\nu\geq L_\nu$ and the optional section theorem yields the desired result. □

On $[0,\delta]$ the process Y satisfies a BSDE with one reflecting lower barrier. Actually we have the following lemma.

LEMMA A.2. *There exist two $\mathcal{P}$-measurable processes $(\tilde K^+_s)_{s\leq T}$ and $(\tilde Z_s)_{s\leq T}$ with values, respectively, in R^+ and R^m such that P-a.s.*

$$(16)\quad\begin{cases}E\left[\int_0^T|\tilde Z_s|^2ds\right]<\infty;\ \tilde K^+ \text{ is continuous nondecreasing and } \tilde K^+_0=0;\\ Y_s=Y_\delta+\int_s^\delta\varphi(u,Y_u,\tilde Z_u)du+(\tilde K^+_\delta-\tilde K^+_s)-\int_s^\delta\tilde Z_u dB_u\ \forall s\in[0,\delta];\\ \int_0^\delta(Y_s-L_s)d\tilde K^+_s=0.\end{cases}$$

Proof. We know that the process $K^{-,n}$ of (11) moves only when Y^n reaches the barrier U; therefore for any $n\geq 0$ and $s\in[0,\delta]$ we have

$$Y^n_s=Y^n_\delta+\int_s^\delta\varphi(u,Y^n_u,Z^n_u)du+n\int_s^\delta(L_u-Y^n_u)^+du-\int_s^\delta Z^n_u dB_u$$

since $\delta\leq\delta^n$. As $\lim_{n\to\infty}E[|Y^n_\delta-Y_\delta|^2]=0$ and $Y_\delta\geq L_\delta$, the sequences

$$((Z^n_s1_{[s\leq\delta]})_{s\leq T})_{n\geq 0}\text{ and }\left(\left(n\int_0^{s\wedge\delta}(L_u-Y^n_u)^+dr\right)_{s\leq T}\right)_{n\geq 0}$$

converge, respectively, in $\mathcal{H}^{2,m}$ and $\mathcal{S}^2$ to $\tilde Z$ and $\tilde K^+$. In addition we have $\int_0^\delta(Y_u-L_u)d\tilde K^+_u=0$. The proof of those claims is the same as the proof of Theorem 5.2 in [12] in the part related to the penalization method (6th paragraph). The minor changes are that the terminal time is a bounded stopping time and that the terminal value Y^n_δ depends on n. But we have $\lim_n\to\infty Y^n_\delta=Y_\delta\geq L_\delta$. Therefore those changes are irrelevant. It follows that $(Y,\tilde Z,\tilde K^+)$ satisfies (16). □

Next for $n\geq 0$ and ν a stopping time let $\theta^n_\nu=\inf\{s\geq\delta_\nu,Y^n_s\leq L_s\}\wedge T$. Since $Y^n\leq Y^{n+1}$, then $\theta^n_\nu\leq\theta^{n+1}_\nu$; therefore $\theta^n_\nu\to\theta_\nu$ as $n\to\infty$ and θ_ν is also a stopping time. In what follows, θ_0 will be simply denoted θ.

Let us deal with the convergence of $Y^n_{t\wedge\theta^n}$ for $t \in [\delta, \theta]$.

PROPOSITION A.3. *P-a.s. we have*

(i) *for any* $t \in [\delta, \theta[$, $Y^n_{t\wedge\theta^n} \to Y_t$ *as* $n \to \infty$;

(ii) *on the set* $[Y_\delta = L_\delta]$ *(resp.,* $[Y_\delta > L_\delta]$*),* $Y^n_{\theta^n} \to Y_\theta$ *(resp.,* $L_\theta 1_{[\theta<T]} + \xi 1_{[\theta=T]}$*) as* $n \to \infty$.

Proof. (i) Let ω be fixed. If $t \in [\delta(\omega), \theta(\omega)[$, then there exists an $n_0(\omega)$ such that for $n \geq n_0$ we have $\theta^n \geq t$ and $Y^n_{t\wedge\theta^n} = Y^n_t \to Y_t$. Let us focus on (ii). If $\omega \in [Y_\delta = L_\delta]$, one has $\theta^n(\omega) = \delta(\omega)$ and $\theta(\omega) = \delta(\omega)$, since $Y^n \leq Y$. It follows that $Y^n_{\theta^n}(\omega) = Y^n_\delta(\omega) = Y^n_\theta(\omega) \to Y_\theta(\omega) = Y_\delta(\omega)$. Assume instead $\omega \in [Y_\delta > L_\delta]$. As $Y^n \nearrow Y$, there exists $n_1(\omega)$ such that for $n \geq n_1(\omega)$ we have $Y^n_\delta(\omega) > L_\delta(\omega)$. The continuity of Y^n and L implies that for $n \geq n_1$, $\theta^n(\omega) > \delta(\omega)$. Let us now write

$$Y^n_{\theta^n} = Y^n_{\theta^n} 1_{[\theta<T]} + Y^n_{\theta^n} 1_{[\theta=T]}.$$

We note that for $n \geq n_1$, $Y^n_{\theta^n} 1_{[\theta<T]} = L_{\theta^n} 1_{[\theta<T]} \to L_\theta 1_{[\theta<T]}$, while $Y^n_{\theta^n} 1_{[\theta=T]} \to \xi 1_{[\theta=T]}$ as $n \to \infty$. To see the latter, suppose there exists $k(\omega)$ such that $Y^k(\omega)$ does not reach $L(\omega)$ after $\delta(\omega)$. Then for $n \geq k$, we have $\theta_n(\omega) = T$ since $Y^k \leq Y^n$; therefore $Y^n_{\theta^n} 1_{[\theta=T]} \to \xi 1_{[\theta=T]}$. Otherwise, for any $n \geq 0$, $Y^n(\omega)$ reaches $L(\omega)$ after $\delta(\omega)$; this implies $\theta(\omega) = T$ and necessarily $L_T(\omega) = \xi(\omega)$, and thus at least for $n \geq n_1$, $Y^n_{\theta^n} 1_{[\theta=T]} = L_{\theta^n} 1_{[\theta=T]} \to \xi 1_{[\theta=T]}$.

It follows that for $\omega \in [Y_\delta > L_\delta]$, $Y^n_{\theta^n} \to L_\theta 1_{[\theta<T]} + \xi 1_{[\theta=T]}$, as $n \to \infty$. □

On $[\delta, \theta]$, the process Y satisfies a BSDE with one upper reflecting barrier. Specifically we have the following proposition.

PROPOSITION A.4.

(i) *P-a.s.* $Y_\theta 1_{[\theta<T]} = L_\theta 1_{[\theta<T]}$.

(ii) *There exist two processes* $\tilde{Z}^1 \in \mathcal{H}^{2,m}$ *and a nondecreasing* $\tilde{K}^{1-} \in \mathcal{S}^2$ *with* $\tilde{K}^{1-}_\delta = 0$ *such that*

$$\text{(17)} \quad \begin{cases} \forall t \in [\delta, \theta],\ Y_t = Y_\theta + \displaystyle\int_t^\theta \varphi(s, Y_s, \tilde{Z}^1_s) ds - (\tilde{K}^{1-}_\theta - \tilde{K}^{1-}_t) - \int_t^\theta \tilde{Z}^1_s dB_s; \\ \displaystyle\int_\delta^\theta (Y_s - U_s) d\tilde{K}^{1-}_s = 0. \end{cases}$$

(iii) *P-a.s.* $Y_\delta 1_{[\delta<T]} = U_\delta 1_{[\delta<T]}$.

Proof. Let us set $\zeta := \lim_{n\to\infty} Y^n_{\theta^n} = Y_\theta 1_{[Y_\delta = L_\delta]} + (L_\theta 1_{[\theta<T]} + \xi 1_{[\theta=T]}) 1_{[Y_\delta > L_\delta]}$ and let us consider a $(Y', \tilde{Z}', \tilde{K}'^-)$ solution of the BSDE with one upper barrier U on $[0, \theta]$, i.e.,

$$\text{(18)} \quad \begin{cases} Y',\ \tilde{K}'^- \in \mathcal{S}^2, \tilde{Z}' \in \mathcal{H}^{2,m}; \tilde{K}'^- \text{ is nondecreasing and } \tilde{K}'^-_0 = 0; \\ \forall t \leq \theta,\ Y'_t = \zeta + \displaystyle\int_t^\theta \varphi(s, Y_s, \tilde{Z}'_s) ds - (\tilde{K}'^-_\theta - \tilde{K}'^-_t) - \int_t^\theta \tilde{Z}'_s dB_s; \\ \forall t \leq \theta, Y'_t \leq U_t \text{ and } \displaystyle\int_0^\theta (Y'_s - U_s) d\tilde{K}'^-_s = 0. \end{cases}$$

Let us emphasize that in the arguments of φ we have Y and not Y'.

To begin, we show that P-a.s. $\forall t \leq T, Y^n_{(t\vee\delta)\wedge\theta^n} \to Y'_{(t\vee\delta)\wedge\theta}$ as $n \to \infty$. From (11), for $n \geq 0$ and $t \in [\delta, \theta]$ we have

$$Y^n_{t\wedge\theta^n} = Y^n_{\theta^n} + \int_t^\theta 1_{[s\leq\theta^n]} \varphi(s, Y^n_s, Z^n_s) ds - (K^{-,n}_{\theta^n} - K^{-,n}_{t\wedge\theta^n}) - \int_t^\theta 1_{[s\leq\theta^n]} Z^n_s dB_s$$

since either $\theta^n = \delta$ or $Y_t^n \geq L_t$ for $t \in [\delta, \theta^n]$. Using Itô's formula for $t \in [\delta, \theta]$ we get

$$\begin{aligned}
&(Y_{t\wedge\theta^n}^n - Y_t')^2 + \int_t^\theta |Z_s^n 1_{[s\leq\theta^n]} - \tilde{Z}_s'|^2 ds = (Y_{\theta^n}^n - \zeta)^2 \\
&+ 2\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')\{1_{[s\leq\theta^n]}\varphi(s, Y_s^n, Z_s^n) \\
&- \varphi(s, Y_s, \tilde{Z}_s')\}ds - 2\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(dK_{s\wedge\theta^n}^{-,n} - d\tilde{K}_s'^{-}) \\
&- 2\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(1_{[s\leq\theta^n]}Z_s^n - \tilde{Z}_s')dB_s.
\end{aligned}$$

But for $t \in [\delta, \theta^n]$,

$$\begin{aligned}
&\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(dK_{s\wedge\theta^n}^{-,n} - d\tilde{K}_s'^{-}) \\
&= \int_t^{\theta^n} (Y_s^n - Y_s')(dK_s^{-,n} - d\tilde{K}_s'^{-}) - \int_{\theta^n}^\theta (Y_{\theta^n}^n - Y_s')d\tilde{K}_s'^{-} \\
&\geq -\int_{\theta^n}^\theta (Y_{\theta^n}^n - Y_s')d\tilde{K}_s'^{-} \geq -2\sup_{s\leq\theta}|U_s|(\tilde{K}_\theta'^{-} - \tilde{K}_{\theta^n}'^{-})
\end{aligned}$$

due to $\int_t^{\theta^n}(Y_s^n - Y_s')(dK_s^{-,n} - d\tilde{K}_s'^{-}) \geq 0$. On the other hand, for $t \in [\theta^n, \theta]$,

$$\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(dK_{s\wedge\theta^n}^{-,n} - d\tilde{K}_s'^{-}) = -\int_t^\theta (Y_{\theta^n}^n - Y_s')d\tilde{K}_s'^{-} \geq -2\sup_{s\leq\theta}|U_s|(\tilde{K}_\theta'^{-} - \tilde{K}_{\theta^n}'^{-}).$$

Therefore for any $t \in [\delta, \theta]$ we have

$$\begin{aligned}
&(Y_{t\wedge\theta^n}^n - Y_t')^2 + \int_t^\theta |Z_s^n 1_{[s\leq\theta^n]} - \tilde{Z}_s'|^2 ds \leq (Y_{\theta^n}^n - \zeta)^2 + 4\sup_{s\leq\theta}|U_s|(\tilde{K}_\theta'^{-} - \tilde{K}_{\theta^n}'^{-}) \\
&+ 2\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(1_{[s\leq\theta^n]}\varphi(s, Y_s^n, Z_s^n) - \varphi(s, Y_s, \tilde{Z}_s'))ds \qquad (19) \\
&- 2\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(1_{[s\leq\theta^n]}Z_s^n - \tilde{Z}_s')dB_s.
\end{aligned}$$

For $t \in [\delta, \theta]$, the following estimate holds true:

$$\begin{aligned}
&\int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')(1_{[s\leq\theta^n]}\varphi(s, Y_s^n, Z_s^n) - \varphi(s, Y_s, \tilde{Z}_s'))ds \\
&= \int_t^\theta (Y_{s\wedge\theta^n}^n - Y_s')[1_{[s\leq\theta^n]}(\varphi(s, Y_s^n, Z_s^n) - \varphi(s, Y_s, \tilde{Z}_s')) + (1_{[s\leq\theta^n]} - 1)\varphi(s, Y_s, \tilde{Z}_s')]ds \\
&\leq \int_t^\theta \{C1_{[s\leq\theta^n]}|Y_{s\wedge\theta^n}^n - Y_s'|(|Y_s^n - Y_s| + |Z_s^n - \tilde{Z}_s'|) \\
&\qquad + |1_{[s\leq\theta^n]} - 1||\varphi(s, Y_s, \tilde{Z}_s')||Y_{s\wedge\theta^n}^n - Y_s'|\}ds \\
&\leq \int_t^\theta \left\{C|Y_{s\wedge\theta^n}^n - Y_s'|^2 + C|Y_{s\wedge\theta^n}^n - Y_s|^2 + \frac{1}{4}|Z_s^n 1_{[s\leq\theta^n]} - \tilde{Z}_s'|^2\right\} ds \\
&\qquad + \int_t^\theta |1_{[s\leq\theta^n]} - 1||\varphi(s, Y_s, \tilde{Z}_s')||Y_{s\wedge\theta^n}^n - Y_s'|ds,
\end{aligned}$$

where C is a constant independent of n, which may change from one line to another. Substituting in (19) and replacing t by $(t \vee \delta) \wedge \theta$ for $t \in [0,T]$, we obtain

$$
\begin{aligned}
&(Y^n_{(t\vee\delta)\wedge\theta^n} - Y'_{(t\vee\delta)\wedge\theta})^2 + \frac{3}{4}\int_{(t\vee\delta)\wedge\theta}^{\theta} |Z^n_s 1_{[s\le\theta^n]} - \tilde{Z}'_s|^2 ds \\
&\le (Y^n_{\theta^n} - \zeta)^2 + 4\sup_{s\le\theta}|U_s|(\tilde{K}'^-_\theta - \tilde{K}'^-_{\theta^n}) \\
&+ \int_{(t\vee\delta)\wedge\theta}^{\theta} \{C|Y^n_{s\wedge\theta^n} - Y'_s|^2 + C|Y^n_{s\wedge\theta^n} - Y_s|^2\} ds \\
&+ \int_{\delta}^{\theta} |1_{[s\le\theta^n]} - 1||\varphi(s, Y_s, \tilde{Z}'_s)||Y^n_{s\wedge\theta^n} - Y'_s| ds \\
&\quad -2\int_{(t\vee\delta)\wedge\theta}^{\theta} (Y^n_{s\wedge\theta^n} - Y'_s)(1_{[s\le\theta^n]} Z^n_s - \tilde{Z}'_s) dB_s.
\end{aligned} \tag{20}
$$

Taking expectations in *bhs* (both hand sides) and using Gronwall's inequality [15, p. 195] we obtain

$$\forall t \le T,\ E[(Y^n_{(t\vee\delta)\wedge\theta^n} - Y'_{(t\vee\delta)\wedge\theta})^2] \to 0 \text{ as } n \to \infty,$$

since

$$E[(Y^n_{\theta^n}-\zeta)^2 + 2\sup_{s\le\theta}|U_s|(\tilde{K}'^-_\theta - \tilde{K}'^-_{\theta^n}) + \int_\delta^\theta |Y^n_{s\wedge\theta^n} - Y_s|^2 + |1_{[s\le\theta^n]} - 1||\varphi(s, Y_s, \tilde{Z}'_s)|^2 ds] \to 0$$

as $n \to \infty$. Finally in (20), taking the supremum in t, expectations in *bhs*, and using first the Burkholder–Davis–Gundy inequality (see [21, p. 166] or [28, p. 160]) and then the inequality $|a.b| \le \epsilon|a|^2 + \frac{1}{\epsilon}|b|^2$ for any $a, b \in R^k$ and $\epsilon > 0$, we get

$$E\left[\sup_{t\le T} |Y^n_{(t\vee\delta)\wedge\theta^n} - Y'_{(t\vee\delta)\wedge\theta}|^2\right] \to 0 \text{ as } n \to \infty.$$

According to Proposition A.3, for any $t \le T$, $\lim_n \to \infty Y^n_{(t\vee\delta)\wedge\theta^n} = Y_{(t\vee\delta)} 1_{[t<\theta]} + \zeta 1_{[t\ge\theta]}$; therefore uniform convergence implies that the process $(Y_{(t\vee\delta)} 1_{[t<\theta]} + \zeta 1_{[t\ge\theta]})_{t\le T}$ is continuous, and for any $t \le T$ we have $Y'_{(t\vee\delta)\wedge\theta} = Y_{(t\vee\delta)} 1_{[t<\theta]} + \zeta 1_{[t\ge\theta]}$. It follows that for any $t \in [\delta, \theta]$, $Y'_t = Y_t 1_{[t<\theta]} + \zeta 1_{[t=\theta]}$.

Let us show (i). If $\delta = \theta < T$, then for any $n \ge 0$, $\delta = \theta^n = \theta$ and $Y^n_{\theta^n} = Y^n_\delta \le L_\delta$; therefore taking the limit we obtain $Y_\theta = L_\theta$ since $Y_\theta \ge L_\theta$ (Proposition A.1).

Next, suppose that $\delta < \theta < T$ (then $L_\delta < Y_\delta$), and let $(t_n)_{n\ge0}$ be a sequence of real numbers which converges increasingly to θ. Then $\lim_n \to \infty Y_{t_n} = \zeta = L_\theta$ since $\theta < T$ (thanks to A.3(ii)), but the process Y is lower semicontinuous, and thus $Y_\theta \le \lim_n \to \infty Y_{t_n} = L_\theta$. Henceforth we have also $Y_\theta = L_\theta$.

We now focus on (ii). Since $Y_\theta 1_{[\theta<T]} = L_\theta 1_{[\theta<T]}$, then $\zeta = Y_\theta$, and for any $t \in [\delta, \theta]$, $Y'_t = Y_t$. For $t \le T$, let us set $\tilde{Z}^1_t = \tilde{Z}'_t 1_{[\delta\le t\le\theta]}$ and $\tilde{K}^{1-}_t = \tilde{K}'^-_{(t\vee\delta)\wedge\theta} - \tilde{K}'^-_\delta$; therefore the triple $(Y, \tilde{Z}^1, \tilde{K}^{1-})$ satisfies (17). As a by-product we have also the continuity of Y on $[\delta, \theta]$.

Finally, it remains to show that $Y_\delta 1_{[\delta<T]} = U_\delta 1_{[\delta<T]}$. For $\omega \in [Y_\delta > L_\delta] \cap [\delta < T]$ we have $\theta(\omega) > \delta(\omega)$, and since Y is continuous on $[\delta, \theta]$ then $Y_\delta(\omega) =$

$\lim_{n\to\infty} Y_{\delta^n}(\omega) \geq \limsup_{n\to\infty} Y^n_{\delta^n}(\omega) = \lim_{n\to\infty} U_{\delta^n}(\omega) = U_\delta(\omega)$ which implies $Y_\delta(\omega) = U_\delta(\omega)$.

Let us now show that $Y_\delta = U_\delta$ on $[Y_\delta = L_\delta] \cap [\delta < T]$.

For $n \geq 0$, let $\nu^n = \inf\{s \geq \delta, Y^n_s = L_s\} \wedge T$. On $[Y_\delta = L_\delta]$, $\delta \leq \nu^n \leq \delta^n$ for any $n \geq 0$; therefore $\lim_{n\to\infty} \nu^n = \delta$. From (11), for any $n \geq 0$ we have

$$\begin{aligned} 1_{[Y_\delta = L_\delta]} Y^n_{\nu^n} &\geq \left(Y^n_{\delta^n} + \int_{\nu^n}^{\delta^n} \varphi(u, Y^n_u, Z^n_u)du - \int_{\nu^n}^{\delta^n} Z^n_u dB_u \right) 1_{[Y_\delta = L_\delta]} \\ &\geq Y^n_{\delta^n} 1_{[Y_\delta = L_\delta]} + \int_{\nu^n}^{\delta^n} 1_{[Y_\delta = L_\delta]} \varphi(u, Y^n_u, Z^n_u)du - \int_{\nu^n}^{\delta^n} 1_{[Y_\delta = L_\delta]} Z^n_u dB_u \end{aligned}$$

since $n(L_u - Y^n_u)^+ \geq 0$ and $K^{-,n}_{\delta^n} = K^{-,n}_{\nu^n}$ on $[Y_\delta = L_\delta]$. Using (13), we deduce that $E[\int_{\nu^n}^{\delta^n} 1_{[Y_\delta = L_\delta]} \varphi(u, Y^n_u, Z^n_u)du] \to 0$ as $n \to \infty$ since on $[Y_\delta = L_\delta]$, the sequence $(\nu^n - \delta^n)_{n\geq 0} \to 0$, and thus $\liminf_{n\to\infty} E[1_{[Y_\delta = L_\delta]}(Y^n_{\nu^n} - Y^n_{\delta^n})] \geq 0$. But P-a.s. we have

$$1_{[Y_\delta = L_\delta]}(Y^n_{\nu^n} - Y^n_{\delta^n}) = 1_{[Y_\delta = L_\delta]}(Y^n_{\nu^n} - Y^n_{\delta^n}) 1_{[\delta < T]} \to 1_{[Y_\delta = L_\delta]}(L_\delta - U_\delta) 1_{[\delta < T]},$$

and hence using the dominated convergence theorem we obtain

$$\begin{aligned} \liminf_{n\to\infty} E[1_{[Y_\delta = L_\delta]}(Y^n_{\nu^n} - Y^n_{\delta^n})] &= \lim_{n\to\infty} E[1_{[Y_\delta = L_\delta]}(Y^n_{\nu^n} - Y^n_{\delta^n})] \\ &= E[1_{[Y_\delta = L_\delta]}(L_\delta - U_\delta) 1_{[\delta < T]}] \geq 0. \end{aligned}$$

It follows that $1_{[Y_\delta = L_\delta]}(L_\delta - U_\delta) 1_{[\delta < T]} = 0$. □

Remark A.5. The motivation of (iii) is related to the mixed zero-sum stochastic differential game. Loosely speaking, we need to affirm that the process Y reaches U before θ.

We finally give the main result of this section.

For $t \leq T$, let us set $Z_t = \tilde{Z}_t 1_{[t\leq\delta]} + \tilde{Z}^1_t 1_{[\delta<t\leq\theta]}$, $K^+_t = \tilde{K}^+_{t\wedge\delta}$, and $K^-_t = \tilde{K}^{1-}_t$. We can easily see that $Z \in \mathcal{H}^{2,m}$ and that $K^\pm$ are nondecreasing processes belonging to $\mathcal{S}^2$ and verify $K^\pm_0 = 0$. Additionally, we have the following theorem.

THEOREM A.6. *The processes Y, Z, and $K^\pm$ satisfy*

$$(21) \quad \begin{cases} \text{(i) } \forall t \leq \theta, Y_t = Y_\theta + \displaystyle\int_t^\theta \varphi(s, Y_s, Z_s)ds + (K^+_\theta - K^+_t) - (K^-_\theta - K^-_t) \\ \qquad\qquad - \displaystyle\int_t^\theta Z_s dB_s \,;\, Y_T = \xi; \\ \text{(ii) } \forall t \leq \theta, L_t \leq Y_t \leq U_t \;\; \textit{and} \;\; \displaystyle\int_0^\theta (Y_s - L_s)dK^+_s = \int_0^\theta (Y_s - U_s)dK^-_s = 0; \\ \text{(iii) } Y \textit{ is continuous on } [0, \theta]. \end{cases}$$

In addition if $\tau^* = \inf\{s \geq 0, Y_s = U_s\} \wedge T$ and $\sigma^* = \inf\{s \geq 0, Y_s = L_s\} \wedge T$, then P-a.s. $\tau^* \vee \sigma^* \leq \theta$.

Proof. (ii) For any $t \leq T$ we know that $L_t \leq Y_t \leq U_t$ and $Y_T = \xi$. By Lemma A.2, $\int_0^\theta (Y_s - L_s)dK^+_s = \int_0^\delta (Y_s - L_s)dK^+_s = 0$. On the other hand, $\int_0^\theta (Y_s - U_s)dK^-_s = \int_\delta^\theta (Y_s - U_s)dK^-_s = 0$, because of A.4(ii) and the definition of K^-.

(i) If $t \in [\delta, \theta]$, then the equation is satisfied because of A.4(ii) being $K^+_\theta - K^+_t = 0$. Hence for $t \in [0, \delta]$,

$$Y_t = Y_\delta + \int_t^\delta \varphi(s, Y_s, \tilde{Z}_s)ds + (\tilde{K}^+_\delta - \tilde{K}^+_t) - \int_t^\delta \tilde{Z}_s dB_s.$$

On the other hand thanks to A.4(ii) we have

$$Y_\delta = Y_\theta + \int_\delta^\theta \varphi(s, Y_s, \tilde{Z}_s^1)ds - (\tilde{K}_\theta^{1-} - \tilde{K}_\delta^{1-}) - \int_\delta^\theta \tilde{Z}_s^1 dB_s.$$

We plug the latter into the previous one to prove (i), since $K_\delta^+ = K_\theta^+$ and $\tilde{K}_\delta^{1-} = \tilde{K}_t^{1-} = 0$ for any $t \leq \delta$.

Point (iii) is an immediate consequence of (ii).

Finally, let us deal with the last point. If $\theta < T$, the inequality stems from the continuity of Y on $[0, \theta]$ and A.4(i),(iii). If $\theta = T$, this is obvious. □

The proof of Theorem 3.1 for $\nu = 0$ is a direct consequence of the results established in Propositions A.1, A.3, and A.4, Lemma A.2, Remark A.5, and Theorem A.6. When $\nu \neq 0$, in the same way as before, we can construct a solution on the interval $[\nu, \theta_\nu]$, where θ_ν is a stopping time $\geq \nu$, which makes the claim of Theorem 3.1 valid.

Acknowledgments. The author thanks the referees for their helpful comments.

REFERENCES

[1] M. Alariot-Nazaret, *Jeux de Dynkin*, Thèse de doctorat en mathématiques, Université de Franche-Comté, Besançon, France, 1982.

[2] M. Alariot, J. P. Lepeltier, and B. Marchal, *Jeux de Dynkin*, 2nd Bad Honnef Workshop on Stochastic Processes, Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, 1982., pp. 23–32.

[3] V. E. Beneš, *Existence of optimal strategies based on specified information, for a class of stochastic decision problems*, SIAM J. Control, 8 (1970), pp. 179–188.

[4] A. Bensoussan and A. Friedman, *Non-linear variational inequalities and differential games with stopping times*, J. Funct. Anal., 16 (1974), pp. 305–352.

[5] J. M. Bismut, *Sur un problème de Dynkin*, Z. Wahrsch. Verw. Geb., 39 (1977), pp. 31–53.

[6] J. Cvitanic and I. Karatzas, *Backward SDEs with reflection and Dynkin games*, Ann. Probab., 24 (1996), pp. 2024–2056.

[7] J. Cvitanic and J. Ma, *Reflected forward-backward SDEs and obstacle problems with boundary conditions*, J. Appl. Math. Stochastic Anal., 14 (2001), pp. 113–138.

[8] R. J. Elliott and M. H. Davis, *Optimal play in a stochastic differential game*, SIAM J. Control Optim., 19 (1981), pp. 543–554.

[9] C. Dellacherie and P. A. Meyer, *Probabilités et Potentiel*, Hermann, Paris, 1975.

[10] E. B. Dynkin and A. A. Yuschkevitch, *Theorems and Problems in Markov Processes*, Plenum Press, New York, 1968.

[11] R. Elliott, *The existence of value in stochastic differential games*, SIAM J. Control Optim., 14 (1976), pp. 85–94.

[12] N. El-Karoui, C. Kapoudjian, E. Pardoux, S. Peng, and M. C. Quenez, *Reflected solutions of backward SDEs and related obstacle problems for PDEs*, Ann. Probab., 25 (1997), pp. 702–737.

[13] N. El-Karoui and S. Hamadène, *BSDEs and risk-sensitive control, zero-sum and nonzero-sum game problems of stochastic functional differential equations*, Stochastic Process Appl., 107 (2003), pp. 145–169.

[14] N. El-Karoui, E. Pardoux, and M.-C. Quenez, *Reflected backward SDEs and American options*, in Numerical Methods in Finance, L. Robers and D. Talay, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 215–231.

[15] W. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[16] S. Hamadène and J.-P. Lepeltier, *Reflected BSDEs and mixed game problem*, Stochastic Process Appl., 85 (2000), pp. 177–188.

[17] S. Hamadène and J. P. Lepeltier, *Backward equations, stochastic control and zero-sum stochastic differential games*, Stochastics Stochastic Rep., 54 (1995), pp. 221–231.

[18] S. Hamadène and J. P. Lepeltier, *Zero-sum stochastic differential games and backward equations*, Systems Control Lett., 24 (1995), pp. 259–263.

[19] S. Hamadène, J. P. Lepeltier, and A. Matoussi, *Double barrier backward SDEs with continuous coefficient*, in Backward Stochastic Differential Equations, Pitman Res. Notes Math. Ser. 364, Longman, Harlow, 1997, pp. 115–128.
[20] Y. Kifer, *Game options*, Finance Stoch., 4 (2000), pp. 443–463.
[21] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer, New York, 1991.
[22] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Appl. Math. 39, Springer, New York, 1998.
[23] I. Karatzas and H. Wang, *Connections between bounded variation control and Dynkin games*, in Optimal Control and Partial Differential Equations (Volume in honor of A. Bensoussan), J. L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, Amsterdam, 2001, pp. 363–373.
[24] R. Laraki and E. Solan, *The value of zero-sum stopping games in continuous time*, SIAM J. Control Optim., 43 (2005), pp. 1913–1922.
[25] J. P. Lepeltier and M. A. Maingueneau, *Le Jeu de Dynkin en Théorie Générale Sans l'Hypothecèse de Mokobodski*, Stochastics, 13 (1984), pp. 25–44.
[26] H. Morimoto, *Dynkin games and martingale methods*, Stochastics, 13 (1984), pp. 213–228.
[27] S. Peng, *Monotonic limit theorem of BSDE and nonlinear decomposition theorem of Doob-Meyer's type*, Probab. Theory Related Fields, 113 (1999), pp. 473–499.
[28] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Springer, Berlin, 1999.
[29] L. Stettner, *Zero-sum Markov games with stopping and impulsive strategies*, Appl. Math. Optim., 9 (1982/83), pp. 1–24.
[30] N. Touzi and N. Vieille, *Continuous-time Dynkin games with mixed strategies*, SIAM J. Control Optim., 41 (2002), pp. 1073–1088.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 519–531

OPTIMAL MEASUREMENT STRATEGY FOR NONLINEAR FILTERING*

N. U. AHMED† AND C. D. CHARALAMBOUS‡

Abstract. In this paper we consider the question of an optimal measurement strategy for nonlinear filtering. Given a set of measurement and observation options, the problem faced by a designer is to choose the option that gives the best nonlinear filter. We develop the necessary conditions of optimality whereby one can determine the best filter.

Key words. nonlinear filtering, conditional density, optimization

AMS subject classification. 93E11

DOI. 10.1137/040610428

1. Introduction.

Throughout this paper, we use the filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, where the filtration $\mathcal{F}_t, t \geq 0$, is an increasing family of subsigma algebras of the sigma algebra $\mathcal{F}$, which are right continuous having left limits. All the random processes appearing in the equations are assumed to be $\mathcal{F}_t$ adapted.

Consider the system governed by stochastic differential equations of the form

$$\begin{aligned} dx(t) &= f(t, x(t))dt + \sigma(t, x(t))dW(t), \qquad x(0) = \xi, \qquad t \geq 0, \\ dy(t) &= h(t, x(t))dt + \sigma_0(t)dV(t), \qquad\quad y(0) = 0, \qquad t \geq 0, \end{aligned} \tag{1}$$

where the state $x(t) \in R^n$ is the (unobserved) process to be estimated and $y(t) \in R^m$ is the (measurement or) observation process. The processes $\{W, V\}$ are $\{R^d, R^m\}$ valued independent standard Brownian motions; $f : R_0 \times R^n \equiv [0, \infty) \times R^n \to R^n$ and $\sigma : R_0 \times R^n \to M(n \times d)$; the class of $n \times d$ matrices (with entries real) and $h : R_0 \times R^n \to R^m$ and σ_0 is an $m \times m$ matrix valued function; and x_0 is an $\mathcal{F}_0$ measurable vector independent of the Wiener processes W and V.

Let φ be any bounded measurable function on R^n with values in R. The classical problem of nonlinear filtering [8] is to estimate $\varphi(x(t))$ given the history of y up to time t. It is well known that the best (unbiased minimum variance) filter is given by the conditional expectation. That is,

$$\hat{\varphi}(t) = E\{\varphi(x(t))|\mathcal{F}_t^y\} = \int_{R^n} \varphi(\xi) Q_t^y(d\xi), \tag{2}$$

where $\mathcal{F}_t^y \subset \mathcal{F}_t$ denotes the smallest sigma algebra induced by the history of the process y up to time t and Q_t^y is the conditional probability measure of $x(t)$ given this history. In other words, for any Borel measurable set $\Gamma \subset R^n$,

$$Q_t^y(\Gamma) = Prob.\{x(t) \in \Gamma|\mathcal{F}_t^y\}.$$

*Received by the editors June 23, 2004; accepted for publication (in revised form) December 5, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/sicon/45-2/61042.html
†School of Information Technology and Engineering and Department of Mathematics, University of Ottawa, Ottawa, Ontario, Canada (ahmed@site.uottawa.ca).
‡Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus (chadcha@ucy.ac.cy); also with the School of Information Technology and Engineering, University of Ottawa, Ottawa, Ontario, K1N 6N5, Canada.

For any given measurement strategy determined by the function h, it is well known that the probability measure valued process $Q_t^y(\cdot) \equiv Q_t^h(\cdot)$ is governed by the Kushner–Stratonovich equation, which is a nonlinear stochastic partial differential equation [7].

Thus, expression (2) can be written equivalently as

$$\hat{\varphi}^h(t) = E\{\varphi(x(t))|\mathcal{F}_t^y\} = \int_{R^n} \varphi(\xi) Q_t^h(d\xi), \tag{3}$$

emphasizing the dependence of the estimate on the measurement operator. In other words, the estimate strongly depends on the measurement strategy, or measurement devices and techniques, chosen entirely by h that appears in the observation equation. Thus, one may have the option of choosing this from a set of feasible strategies. Let $M \equiv B(R_0 \times R^n, R^m)$ denote the space of bounded Borel measurable maps. Furnished with the sup norm topology, we have

$$\| h \|_0 \equiv \sup\{\| h(t,x) \|, (t,x) \in R_0 \times R^n\},$$

which is a Banach space. Note that the inclusion map $M \hookrightarrow L_\infty(I \times R^n, R^m)$ is continuous. Let M_0 be a closed convex possibly bounded subset of M.

Problem. Our basic problem is to find a measurement strategy $h^o \in M_0$ that produces the best estimate out of all the estimates $\{\hat{\varphi}^h(t), h \in M_0\}$.

This requires a quantitative measure that distinguishes one from the rest. The most suitable measure for this is the entropy. Thus if we can find an admissible h^o that minimizes the entropy associated with the conditional probability measure $Q_t^h, t \geq 0$, we have the best estimator (out of all the best unbiased minimum variance estimates). Entropy measures have been employed in many applications of information theory, an anthology of which is found in [1]. In addition, entropy measures are used in control systems (see, for example, [2]). An application of entropy in estimation and tracking problems is found in [3, 4], while minimax control problems using relative entropy as a payoff are found in [13, 14]. Alternative methods addressing similar problems are found in [11, 12].

Some notation. Let H denote the Hilbert space $L_2(R^n)$ and $\mathcal{H} \equiv H^m$ the Cartesian product of m copies of the Hilbert space H, equipped with the natural inner product and norm. For any pair of Hilbert spaces $\{X, Y\}$, we use $\mathcal{L}(X, Y)$ to denote the Banach space of bounded linear operators from X to Y. In the case when $X = Y$, $\mathcal{L}(X, X) \equiv \mathcal{L}(X)$.

The rest of the paper is organized as follows. In section 2 we present the mathematical formulation of a general optimization problem, which we specialize to our original problem, which motivated this paper in the first place. We develop the necessary conditions of optimality. In section 3 we deal with the original problem. We conclude the paper with some comments and an elementary (one-dimensional) example for illustration.

2. Formulation of the optimization problem. In this section we formulate a general optimization problem and present the necessary conditions of optimality.

If one is willing to work with unnormalized bounded positive measures, one has the unnormalized conditional density equation [5, 6, 9], which is linear. This equation is given by

$$d\mu_t = A^*(t)\mu_t dt + B(h\mu_t) \cdot dy, \quad t \geq 0, \quad \mu_0 = \nu, \tag{4}$$

where A^* is the adjoint of the infinitesimal generator A of the Markov process x, B is simply the multiplicative linear operator

$$B(h\mu) = (\Gamma(t)h)\mu, \quad \text{with} \quad \Gamma(t) = (\sigma_0(t)\sigma_0^*(t))^{-1},$$

and $x \cdot y = (x, y)$ for $x, y \in R^m$. The relationship between Q and μ is simply the normalization factor $\mu_t(1) = \mu_t(R^n)$ which means $Q_t = \mu_t/\mu_t(1)$. Given the initial measure μ_0, one can prove the existence and uniqueness of the solution of this equation in the space of countably additive bounded positive measures [10]. However, if $\mu_0 = \nu$ is absolutely continuous with respect to the Lebesgue measure, with density $q_0 \in H \equiv L_2(R^n)$, then the measure equation (4) is equivalent to the density equation

$$d\rho_t = A^*(t)\rho_t dt + B(h\rho_t) \,.\, dy, \quad t \geq 0, \quad \rho_0 = q_0, \tag{5}$$

and, further, under certain assumptions on the coefficients of the operator A and h, one can prove that, for any finite interval $I = [0, T]$, the evolution equation (5) has a unique solution ρ which belongs to $C(I, H) \cap L_2(I, V) \cap L_\infty(I, H)$ with probability one, where $V = H^1$ is the standard Sobolev space. For more details see [5]. In view of the Girsanov theorem used in the original derivation of (5), it is known that one can replace the observation with a Brownian motion $\hat{W}$ which, without loss of generality, we assume to be a standard Brownian motion. Thus from now on we write (5) as

$$d\rho_t = A^*(t)\rho_t dt + B(h\rho_t) \,.\, d\hat{W}, \quad t \geq 0, \quad \rho_0 = q_0, \tag{6}$$

and formulate the optimization problem using this equation. Let $(\hat{\mathcal{F}}_t, \uparrow, t \geq 0)$ denote the smallest family of increasing subsigma algebras of the sigma algebra $(\mathcal{F}_t, \uparrow, t \geq 0)$ with respect to which $\{\hat{W}(s), 0 \leq s \leq t\}$ is measurable for all $t \geq 0$. Let X denote any Hilbert space and $L_2^e(I, X)$ the Hilbert space of $\hat{\mathcal{F}}_t$ adapted X valued random processes defined on I having square integrable second moments. Let $\bar{R}_0 \equiv [0, \infty]$, $\ell : I \times H \to \bar{R}_0$, and $\Psi : H \to \bar{R}_0$. The objective functional is given by

$$J(h) \equiv E\left\{ \int_0^T \ell(t, \rho_t)dt + \Psi(\rho_T) \right\}. \tag{7}$$

The problem involves finding an $h^o \in M_0$ that minimizes the functional (7) subject to the dynamic constraint (6). In order to develop necessary conditions of optimality for this problem, we need some basic results.

In particular, we present the following lemma which proves the continuous dependence and differentiability of the solution with respect to $h \in M$.

Without loss of generality, we assume that the operator A, and hence A^*, is time invariant. For the time varying case, the semigroup is replaced with the transition operator.

LEMMA 1. *Consider the system* (6) *with initial condition* $\rho_0 = q_0 \in H$. *Suppose* A^* *generates a* C_0 *semigroup* $S(t), t \geq 0$, *in* H *and the operator* B *is the mutiplication operator mapping* H *to* $\mathcal{H} \equiv H^m$, *and let* M_0 *be a closed bounded subset of* M. *Then the solution* ρ *is continuously dependent on* $h \in M_0$ *and once Gâteaux differentiable. Further, the Gâteaux differential of* ρ *at* h^o *in the direction* $h - h^o$, *denoted by* p^o, *has the following properties:* $p^o \in L_2^e(I, H)$, *with sample paths belonging to* $C(I, H)$ *almost surely, and it satisfies the evolution equation*

$$dp^o = A^*p^o dt + B(h^o p^o) \cdot d\hat{W}(t) + B((h - h^o)\rho^o) \cdot d\hat{W}(t), \quad t \in I, \quad p_0^o = 0. \tag{8}$$

Proof. Since $S(t)$ is a semigroup in H, it is clear that there exists a number $K > 0$, possibly dependent on T, such that

$$\sup\{\| S(t) \| \equiv \| S(t) \|_{\mathcal{L}(H)}, t \in I\} \leq K.$$

Let $h^n \in M_0$, and let ρ^n be the corresponding solution of (6). Suppose $h^n \xrightarrow{s} h^o$ in the Banach space M and ρ^o is the solution corresponding to h^0. We show that $\rho^n \xrightarrow{s} \rho^o$ in $L_2^e(I, H)$. Clearly the solutions $\{\rho^n, \rho^0\}$ are mild and given by the solutions of the integral equations

$$\rho^n(t) = S(t)\rho_0 + \int_0^t S(t-s)B(h^n\rho^n) \cdot d\hat{W},$$
$$\rho^o(t) = S(t)\rho_0 + \int_0^t S(t-s)B(h^o\rho^o) \cdot d\hat{W}, \qquad t \in I.$$

Subtracting the second equation from the first, we obtain

$$\rho^n(t) - \rho^o(t) = \int_0^t S(t-r)B(h^n(\rho^n - \rho^o)) \cdot d\hat{W} + \int_0^t S(t-r)B((h^n - h^o)\rho^o) \cdot d\hat{W}, \qquad t \in I. \tag{9}$$

Since $\Gamma(t)$ is bounded and M_0 is a bounded subset of M, there exist constants $c_0, c_1 > 0$ such that

$$E \| \rho^n(t) - \rho^o(t) \|_H^2 \leq 2K^2c_o^2 \int_0^t E \| \rho^n(r) - \rho^o(r) \|_H^2 \, dr + 2K^2c_1^2 \int_0^t E \| (h^n - h^o)\rho^o(r) \|_{\mathcal{H}}^2 \, dr. \tag{10}$$

Since the set M_0 is a bounded subset of M, there exists a number $0 < m_0 < \infty$ such that

$$\sup\{\| h \|_{B(I \times R^n, R^m)}, h \in M_0\} \leq m_0.$$

Hence $c_0 = c_1 m_0$. Using the Gronwall inequality, it follows from inequality (10) that

$$E \| \rho^n(t) - \rho^o(t) \|_H^2 \leq c_3 \int_0^T E \| (h^n - h^o)\rho^o(r) \|_{\mathcal{H}}^2 \, dr \quad \forall \, t \in I, \tag{11}$$

where

$$c_3 = 2K^2c_1^2 \exp\{2K^2c_0^2T\}.$$

Now by virtue of the Lebesgue dominated convergence theorem, it follows from (11) that

$$\rho^n \xrightarrow{s} \rho^o \quad \text{in} \quad L_2^e(I, H).$$

This proves continuous dependence. Next, we prove Gâteaux differentiability. Let $h, h^o \in M_0$. Since M_0 is a closed convex set, the function $h^\varepsilon \equiv h^o + \varepsilon(h - h^o) \in M_0$

also for $\varepsilon \in [0,1]$. Let ρ^o denote the solution of (6) corresponding to h^o and ρ^ε the solution corresponding to h^ε. Define

$$p^\varepsilon \equiv \{(\rho^\varepsilon - \rho^o)/\varepsilon\}.$$

Then one can easily verify that p^ε satisfies the evolution equation (in the mild sense)

$$dp^\varepsilon = A^* p^\varepsilon dt + B(h^o p^\varepsilon) \cdot d\hat{W} + B((h-h^o)\rho^\varepsilon) \cdot d\hat{W}, \qquad p^\varepsilon(0) = 0. \tag{12}$$

Using the (Markov) semigroup, we observe that p^ε must be the solution of the integral equation

$$\begin{aligned} p^\varepsilon(t) = &\int_0^t S(t-r)B(h^o p^\varepsilon(r)) \cdot d\hat{W}(r) \\ &+ \int_0^t S(t-r)B((h-h^o)\rho^\varepsilon(r)) \cdot d\hat{W}(r), \qquad t \in I. \end{aligned} \tag{13}$$

It follows from continuous dependence of solution with respect to h, as proved in the first part of this lemma, that $\rho^\varepsilon \xrightarrow{s} \rho^o$ in $L_2^e(I,H)$ as $\varepsilon \to 0$. This implies that the second integral on the right side of (13) also converges strongly. Thus letting $\varepsilon \downarrow 0$, it follows from the above equation that

$$p^\varepsilon \xrightarrow{s} p^o \quad \text{in} \quad L_2^e(I,H).$$

In other words, p^o is the mild solution of the evolution equation

$$dp^o = A^* p^o dt + B(h^o p^o) \cdot d\hat{W}(t) + B((h-h^o)\rho^o) \cdot d\hat{W}(t), \quad p^o(0) = 0, \quad t \in I, \tag{14}$$

which uniquely defines the Gâteaux differential of ρ at h^o in the direction $(h-h^o)$. This completes the proof. □

Remark 1. Since $H \equiv L_2(R^n)$ is a Hilbert space, both $T(t)$ and its adjoint $S(t) = T^*(t)$ are strongly continuous bounded linear operators on H. See [6] or any other book on semigroup theory.

Remark 2. Note that the semigroup $S(t), t \geq 0$, is the adjoint of the Markov semigroup $T(t), t \geq 0$, defined by

$$(T(t)\varphi)(\xi) \equiv \int_{R^n} P_t(\xi, dy)\varphi(y),$$

where $P_t(\xi, dy)$ denotes the transition probability measure. Strictly speaking, these semigroups are the continuous extensions of the Markov semigroups from the space of bounded measurable functions $B(R^n)$ to H. Further, they are positive semigroups in the sense that $(T(t)\phi)(\xi) \geq 0$ given that $\phi(\xi) \geq 0$. Since a Markov semigroup is a contraction semigroup, the number K used in Lemmas 1 and 2 can be taken as 1.

Now we are prepared to consider the optimization problem (6)–(7) with the constraints as specified there. Since we are going to consider entropy functionals later and these functionals are concave, we shall not impose any convexity condition on our objective functional. This is compensated by assuming that M_0 is a closed bounded subset of the Banach space $M = B(I \times R^n, R^m)$, implying the boundedness of the solution set as proved in the following lemma.

LEMMA 2. *Suppose A^* and the set $M_0 \subset M$ satisfy the assumptions of Lemma 1. Then, for any $q_0 \in H$ and for every $h \in M_0$, (6) has a unique mild solution*

$\rho^h \in L_2^e(I,H)$. *Further, the solution set* $\mathcal{S} \equiv \{\rho^h, h \in M_0\}$ *is a bounded subset of* $L_2^e(I,H) \cap L_\infty^e(I,H)$.

Proof. Recall that by definition, a mild solution ρ^h must satisfy the integral equation

$$\rho^h(t) = S(t)q_0 + \int_0^t S(t-r)B((h\rho^h)(r)) \cdot d\hat{W}(r), \quad t \in I. \tag{15}$$

We will see that existence and uniqueness of a solution of this integral equation follow from the fact that, for any given $h \in M_0$, $\rho \longrightarrow B(h\rho)$ is a bounded linear (hence Lipschitz) map in H. Indeed, since $q_0 \in H$ and $h \in M_0$, it is easy to verify that the operator G, defined by

$$(G\rho)(t) \equiv S(t)q_0 + \int_0^t S(t-r)(B(h\rho))(r) \cdot d\hat{W}(r), \quad t \in I,$$

maps $L_2^e(I,H)$ into itself. This follows from the inequality

$$\| G\rho \|^2_{L_2^e(I,H)} \leq 2K^2 \big\{ T|q_0|_H^2 + \beta_h^2 \, \| \rho \|^2_{L_2^e(I,H)} \big\},$$

where $\beta_h \equiv \sup\{|\Gamma(t)h(t,x)|_{R^m}, (t,x) \in I \times R^n\}$ and $K \equiv \sup\{\| S(t) \|_{\mathcal{L}(H)}, t \in I\}$. On $L_2^e(I,H)$, we can introduce the metric d given by

$$d^2(\rho,q) \equiv \sup\{E|\rho(s) - q(s)|_H^2, s \in I\}.$$

Clearly, with respect to this metric topology, $L_2^e(I,H)$ is a complete metric space. For $t \in I$, define $d_t^2(\rho,q) \equiv \sup\{E|\rho(r) - q(r)|_H^2, 0 \leq r \leq t\}$. Then using the operator G as defined above, it is easy to verify that

$$d_t^2(G\rho, Gq) \leq K^2\beta_h^2 \int_0^t d_s^2(\rho,q)ds, \quad t \in I. \tag{16}$$

By successive substitution of (16) into itself, after n iterations we obtain

$$d_t^2(G^n\rho, G^nq) \leq ((K^2\beta_h^2 t)^n/n!)d_t^2(\rho,q), \quad t \in I.$$

Since $t \longrightarrow d_t$ is a nondecreasing function, it follows from this that

$$d^2(G^n\rho, G^nq) \leq ((K^2\beta_h^2 T)^n/n!)d^2(\rho,q).$$

Thus for n sufficiently large, the nth iterate G^n of the operator G is a contraction on the metric space $(L_2^e(I,H), d)$, and hence by Banach fixed point theorem, G^n, and hence G itself, has a unique fixed point which we denote by ρ^h. Thus $\rho^h \in L_2^e(I,H)$ is the unique solution of the integral equation (15), and hence, by definition, this is the unique mild solution of (6). Next we verify that the solution set is bounded. Taking the H-norm, it follows from (15) that

$$E \, \| \rho^h(t) \|^2 \leq 2K^2 E \, \| q_0 \|^2 + 2K^2 \int_0^t E \, \| B(h\rho^h) \|^2_{\mathcal{H}} \, dr, \quad t \in I.$$

Since the set $M_0 \subset B(I \times R^n, R^m)$ is bounded, $c_0 \equiv \{\beta_h, h \in M_0\} < \infty$. Thus it follows from this that

$$E \, \| \rho^h(t) \|^2 \leq 2K^2 E \, \| \rho_0 \|^2 + 2K^2 c_0^2 \int_0^t E \, \| \rho^h(r) \|^2 \, dr, \quad t \in I. \tag{17}$$

Applying the Gronwall lemma, we arrive at the estimate

$$E\,\|\,\rho^h(t)\,\|^2 \le \big(2K^2\exp\big(2K^2c_0^2T\big)\big)E\,\|\,\rho_0\,\|^2, \quad t\in I, \quad \forall\ h\in M_0. \tag{18}$$

Since this is true for all $t\in I$, and for all $h\in M_0$, the solution set $\mathcal{S}$, corresponding to the measurement constraints M_0, is a bounded subset of $L_2^e(I,H)\cap L_\infty^e(I,H)$. This completes the proof. □

Remark 3. It is not clear to us at this time if the solution set $\mathcal{S}$ is closed in $L_2^e(I,H)$. Our conjecture is based on Mazur's theorem which states that a convex set in a locally convex topological space is closed if and only if it is weakly closed, and the fact that the space $B(I\times R^n,R^m)$ is continuously embedded in $L_\infty(I\times R^n,R^m)$ and hence M_0, considered as a subset of $L_\infty(I\times R^n,R^m)$, is a weak star compact subset of the space $L_\infty(I\times R^n,R^m)$.

Here we are interested only in the optimality conditions. Now we are prepared to present the necessary conditions of optimality. This is the main result of this paper.

THEOREM 3. *Suppose the cost integrand ℓ and the terminal cost Ψ of the objective functional* (7) *are once Gâteaux differentiable on H, ℓ is measurable in t on I, and the set of admissible measurement strategies M_0 is a closed bounded convex subset of $M\equiv B(I\times R^n,R^m)$. Then, for a pair $\{h^o,\rho^o\}$, satisfying the evolution equation*

$$d\rho_t^o = A^*\rho_t^o dt + B(h^o\rho_t^o)\,.\,d\hat{W}, \quad t\ge 0, \quad \rho_0 = q_0, \tag{19}$$

to be optimal, it is necessary that there exists a $\psi\in L_2^e(I,H)$ satisfying the adjoint equation

$$d\psi = -A\psi dt + |B(h^o)|^2\psi dt + \ell_\rho(t,\rho^o(t))dt - B(h^o\psi)\cdot d\hat{W}, \tag{20}$$

$$\psi(T) = -\Psi_\rho(\rho_T^o) \tag{21}$$

and the inequality

$$E\int_I (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\ dt \ge 0 \quad \forall\ h\in M_0. \tag{22}$$

Proof. The proof is based on variational arguments. Let $h^o\in M_0$ be the optimal strategy and ρ^0 the solution of the evolution equation (6) corresponding to $h=h^o$. Let $h\in M_0$ be any other element. Since M_0 is a closed convex set, it is evident that

$$h^\varepsilon \equiv h^o + \varepsilon(h-h^o)\in M_0 \quad \forall\ \varepsilon\in[0,1].$$

Let $\rho^\varepsilon\in L_2^e(I,H)$, having sample paths in $C(I,H)$ almost surely, be the solution of (6) corresponding to h^ε replacing h. By optimality of h^o, it is clear that

$$J(h^\varepsilon)\ge J(h^o) \quad \forall\ \varepsilon\in[0,1]. \tag{23}$$

By our assumption, both ℓ and Ψ are Gâteaux differentiable on H and it follows from Lemma 1 that the solutions are once Gâteaux differentiable. Thus the objective functional J is also once Gâteaux differentiable. Denote the Gâteaux differential of J at h^o in the direction $h-h^o$ by $dJ(h^o,h-h^o)$. By letting $\varepsilon\downarrow 0$ to the difference quotient

$$(1/\varepsilon)[J(h^\varepsilon)-J(h^o)]$$

and using Lemma 1, inequality (23), and Fubini's theorem, we arrive at

$$dJ(h^o, h-h^o) = E\left\{\int_I \langle p^o(t), \ell_\rho(t,\rho^o(t))\rangle_H + \langle p^o(T), \Psi_\rho(\rho^o_T)\rangle_H\right\} \geq 0, \tag{24}$$

where we have used ℓ_ρ, Ψ_ρ to denote the first Fréchet derivatives of ℓ and Ψ on H. For h^o to be optimal, it is necessary that this inequality hold for all $h \in M_0$. Recall that p^o is the solution of the variational equation

$$dp^o = A^*p^o dt + B(h^o p^o)\cdot d\hat{W}(t) + B((h-h^o)\rho^o)\cdot d\hat{W}(t), \qquad p^o(0)=0, \quad t\in I. \tag{25}$$

Since p^o has continuous sample paths with probability one, the terminal values are well defined random variables with values in H and have finite second moments. Clearly the map

$$p^o \to dJ(h^o, h-h^o)$$

is a continuous linear map from $L_2^e(I,H)$ to R. Further, it is also clear from the variational equation (25) that the map

$$B((h-h^o)\rho^o) \to p^o$$

is a continuous linear map from $L_2^e(I,H)$ to itself. More precisely, this is a map from the martingale

$$\int_0^t B((h-h^o)\rho^o)\cdot d\hat{W} \to p^o.$$

Hence the composition map

$$B((h-h^o)\rho^o) \to dJ(h^o, h-h^o)$$

is a continuous linear map. Thus by the standard representation theorem for continuous functionals on Hilbert spaces and classical martingale representation theory, there exists a $\psi \in L_2^e(I,H)$ (having almost surely continuous sample paths) determining the intensity of the martingale, $M_t \equiv \int_0^t B(h^o\psi)d\hat{W}, t\geq 0$, dual to the first martingale term of the variational equation (14), such that

$$\begin{aligned} dJ(h^o, h-h^o) &\equiv E\left\{\int_I \langle p^o(t), \ell_\rho(t,\rho^o(t))\rangle_H dt + \langle p^o(T), \Psi_\rho(\rho^o_T)\rangle_H\right\} \\ &= E\int_0^T (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\, dt. \end{aligned} \tag{26}$$

It follows from (24) and (26) that inequality (22) is necessary for h^o to be optimal. It remains to characterize the element ψ. Using the Itô differential of the functional (p^o,ψ) and the variational equation (14), we obtain

$$\begin{aligned} d(p^o,\psi) &= (dp^o,\psi) + (p^o, d\psi) + \langle dp^o, d\psi\rangle \\ &= (A^*p^o + B(h^op^o)\cdot d\hat{W} + B((h-h^o)\rho^o)\cdot d\hat{W}, \psi) + (p^o, d\psi) + \langle dp^o, d\psi\rangle \\ &= (p^o, d\psi + A\psi dt - |B(h^o)|^2\psi dt + B(h^o\psi)\cdot d\hat{W}) \\ &\quad - (B((h-h^o)\rho^o\psi)\cdot d\hat{W}) - (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\, dt, \end{aligned} \tag{27}$$

where we have used the bracket $\langle \cdot, \cdot \rangle$ to denote the quadratic variation term. The formal application of the Itô differential rule can be justified rigorously by Yosida approximation of the unbounded operator A^* by a sequence of bounded operators $A_n^* \equiv nA^*R(n, A^*)$, where $R(\lambda, A^*)$ denotes the resolvent of the operator A^* for $\lambda \in \rho(A^*)$, the resolvent set of A^*. This off course requires some limiting arguments in the process. This is a standard technique used in infinite-dimensional problems with unbounded operators. We do not go into detail here; the interested reader may consult any book on semigroup theory (e.g., [2]). Returning to our problem, by integrating (27) over the interval $I = [0, T]$ and taking the expectation, we obtain

$$E(p^0(T), \psi(T)) = E\Big\{ \int_I (p^o(t), d\psi + A\psi dt - |B(h^o)|^2 \psi dt + B(h^o\psi) \cdot d\hat{W}(t)) \\ - \int (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\, dt \Big\}. \tag{28}$$

Choosing ψ, existence of which is already proved, as any solution of the equation

$$d\psi + A\psi dt - |B(h^o)|^2 \psi dt + B(h^o\psi) \cdot d\hat{W}(t) = \ell_\rho(t, \rho^o(t))dt, \quad t \in I, \tag{29}$$

we find that (28) reduces to

$$-E(p^0(T), \psi(T)) + E\int_I (p^o(t), \ell_\rho(t, \rho^o(t)))\, dt = E\int_I (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\, dt. \tag{30}$$

As it stands, no initial or boundary conditions have been specified so far for (29). Choosing the boundary condition as

$$\psi(T) = -\Psi_\rho(\rho_T^o), \tag{31}$$

(29) determines ψ uniquely. As a result of this choice, (30) takes the form

$$E(p^0(T), \Psi_\rho(\rho^o(T))) + E\int_I (p^o(t), \ell_\rho(t, \rho^o(t)))dt \\ = E\int_I (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\, dt, \tag{32}$$

which coincides with expression (26) as required. In other words, ψ characterized by the adjoint equation (29) with the boundary condition (31) satisfies the identity (26). Combining (24) and (32), we obtain the inequality

$$dJ(h^0, h-h^o) = E\int_0^T (B((h-h^o)\rho^o), B(h^o\psi))_{\mathcal{H}}\, dt \geq 0. \tag{33}$$

Observe that we have now derived all the necessary conditions. Equations (29) and (31) give the adjoint equations (20), (21) as stated in the theorem. Since inequality (33) holds for any $h \in M_0$, we have the necessary inequality (22). This completes the proof. □

Implementation of the necessary conditions. The necessary conditions given by (19)–(22), like the celebrated Pontryagin minimum principle, help to isolate candidates from the set M_0 which contain the optimal. Thus these conditions can be used to construct an optimal measurement strategy h for the measurement channel (see (1)). Necessary conditions for a risk-sensitive payoff in which (6) includes a potential can be derived as well [15].

3. Application to observer design. In this section we apply some results of the previous section to design problems. For illustration we conclude with an elementary example. Our objective is to determine the best observer from a given family of options. The fundamental question is, What should be the appropriate measure that distinguishes one or more members being the best from the given family of options? Before we state the criteria, recall that the mutual information, associated with the random element $x(t)$ and the history of observation $\mathcal{Y}_0^t$ (dependent on the history of x), denoted by $I(x(t), \mathcal{Y}_0^t)$, is given by

$$I(x(t), \mathcal{Y}_0^t) = H(x(t)) - H(x(t)|\mathcal{Y}_0^t), \quad t \geq 0,$$

where $H(x(t))$ denotes the entropy of the random element $x(t)$, and $H(x(t)|\mathcal{Y}_0^t)$ is the conditional entropy of $x(t)$ given the measurement history $\mathcal{Y}_0^t$ up to time t. Since the measurement history depends on the unknown process x, it is natural to choose a measurement strategy that maximizes the mutual information. But since in the filtering or estimation problem considered here nothing can be done to modify the source x, this is equivalent to minimizing the conditional entropy $H(x(t)|\mathcal{Y}_0^t)$. This argument suggests the first of two possible criteria as described below. A related, although alternative formulation, is found in [13].

Entropy functional. In order to reduce uncertainty in our estimates, it seems to be appropriate to minimize the weighted entropy functional given by

$$\Psi(q) = -\int_{R^n} \{q(x) \log q(x)\} \phi(x) dx, \tag{34}$$

where ϕ is a nonnegative continuous bounded function having compact support in R^n. By setting $\phi \equiv 1$, we obtain the classical entropy. In any case, the running cost here is zero, that is, $\ell \equiv 0$. Thus the cost functional is given by

$$J(h) \equiv E\{\Psi(\rho_T^h)\}, \tag{35}$$

where ρ^h denotes the solution of the filtering equation (6) corresponding to the choice of $h \in M_0$. In this case, the adjoint equation is given by

$$\begin{aligned} -d\psi &= A\psi dt - |B(h^o)|^2 \psi dt + B(h^o\psi) \cdot d\hat{W}(t), \qquad t \in I, \\ \psi(T) &= -(1 + \log \rho_T^o)\phi. \end{aligned} \tag{36}$$

Remark 4. In some communication problems, the entropy functional may have to be maximized. Applications requiring source coding, those requiring the jamming of opponents' information exchange capabilities, etc. are some examples in which one wants to maximize uncertainty (information). In this situation, the necessary conditions of optimality (19), (20), (21) remain valid and the inequality (22) is reversed.

Relative entropy. Another possible measure is given by the relative entropy, relative to a given density ν, similar to [14]. Suppose an ideal measure or density $\nu(x), x \in R^n$, is given. Then the objective functional may be taken to be one that measures the closeness of the chosen ρ^h from the desired one. This requires that the elements of the solution set $\mathcal{S} \equiv \{\rho^h, h \in M_0\}$, or a closed subset of this, be absolutely continuous with respect to the desired density ν. Thus we may consider the relative entropy

$$\begin{aligned} J(h) \equiv H(\rho_T^h|\nu) &= \int_{R^n} \log\big(\rho_T^h(x)/\nu(x)\big) \rho_T^h(x) dx \\ &= \int_{R^n} \big\{\rho_T^h(x)/\nu(x)\big\} \big\{\log\big(\rho_T^h(x)/\nu(x)\big)\big\} \nu(x) dx \end{aligned} \tag{37}$$

as the objective functional. The problem involves choosing an $h^o \in M_0$ that minimizes the relative entropy

$$J(h) \equiv H(\rho_T^h|\nu)$$

over the admissible set M_0. Again the adjoint system is given by (36) with the terminal condition given by

$$\psi(T) = -(1 + \log(\rho_T^h/\nu)).$$

Remark 5. Often engineers have a very specific and physically implementable and relatively inexpensive family of measurement options that determine the set $M_{00} \subset M_0$. For example, M_{00} may be described by a family of neural networks

$$M_{00} \equiv \left\{ h \in M_0 : h(t,x) = \sum_{k=1}^{p} \alpha_k(t)\varphi_k(x), x \in R^n \right\},$$

where $\{\varphi_k : R^n \longrightarrow R^m\}$ are the basis functions of the neural network. Note that currently, neural networks are built by the use of a whole class of continuous as well as discontinuous elementary functions, and therefore are able to generate a wide variety of bounded measurable functions. Once a structure is put in, all that the engineer needs to do is choose the weights $\{\alpha_k, k = 1, 2, \ldots, p\}$ optimally. Clearly this set is determined only by a finite number of time dependent functions taking values from a given set $\Lambda \subset R^p$. In fact, one may wish to simplify further and choose constant weights. This can be done relatively easily by the use of the general algorithm given in the previous section. Implementation of an element from M_0 may be expensive and technically difficult. Let $h^* \in M_0$ be an ideal choice and ρ^* the corresponding ideal density. Falling back to the inexpensive and physically implementable class M_{00}, one wishes to make a choice that comes as close as possible to the ideal. Here, one may use the relative entropy $H(\rho^h|\rho^*)$ as the objective functional and choose an h from the set M_{00} that minimizes the objective functional

$$J(h) \equiv H(\rho^h|\rho^*).$$

This is precisely the problem stated above.

Remark 6. Note that we have used unnormalized densities, and thus the entropies and the relative entropies are not the true entropies according to strict definition. In any case, they are related to the normalized densities as follows. Let $e_n(\rho)$ denote the normalized entropy and $e(\rho)$ the unnormalized entropy corresponding to any density ρ. One can easily verify that they are related by the identity

$$e(\rho) = \|\rho\| \, (e_n(\rho) - \log \|\rho\|).$$

Similarly one can verify that the unnormalized relative entropy is related to the normalized by the identity

$$H(\rho|\nu) = H_n(\rho|\nu) + \|\rho\| \log(\|\rho\| \, / \, \|\nu\|),$$

where $\|\rho\| \equiv \int_{R^n} \rho(x)dx$.

An elementary example. Here we present an elementary example to illustrate the basic procedure. Consider a noisy signal $\{x(t), t \geq 0\}$ (vector or scalar valued) which is a Markov diffusion process, in particular an Itô process, with the corresponding infinitesimal generator A. Then there exist a pair of functions $\{b, a\}$ such that the differential generator A and its formal adjoint A^* are given by

$$(A\varphi)(x) \equiv (1/2)a(x)(\partial^2/\partial x^2)\varphi + b(x)(\partial/\partial x)\varphi, \quad x \in R,$$
$$(A^*\varphi)(x) \equiv (1/2)(\partial^2/\partial x^2)(a\varphi) - (\partial/\partial x)(b\varphi), \quad x \in R.$$

Let the measurement process be given by the scalar valued signal $\{y(t), t \in I\}$ generated by a noisy measurement device given by

$$dy = \beta \tanh(F(t)x(t))dt + \sigma_0 dV, \quad \sigma_0 \not\equiv 0.$$

The function $\{F(t), t \geq 0\}$ represents an on-off sampler given by

$$F(t) = \begin{cases} 1 & \text{for } t \in I_i, \quad i = 0, 2, 4, 6, \ldots, \\ 0 & \text{for } t \in I_i, \quad i = 1, 3, 5, \ldots, \end{cases}$$

where the intervals $\{I_i\}$ are mutually disjoint, each having a fixed length with the union covering the positive half of the real line, $\cup_{i=0}^{\infty} I_i = R_+ = [0, \infty)$. The hyperbolic tangent is used to represent a nonlinear (saturating) amplifier and β is the amplifier gain with admissible values $\beta \in [0, \alpha] \equiv \Lambda$. The objective is to find the best amplifier gain $\beta^* \in \Lambda$ that minimizes the terminal entropy given by expression (35). One possible way to compute the optimal solution is by the following algorithm. We start with any arbitrary element $\beta^0 \in \Lambda$ and solve for ρ^0 the density equation

$$d\rho = A^*\rho dt + (1/\sigma_0^2)(h^0\rho) \cdot d\hat{W}, \quad \rho(0) = q_0,$$
$$h^0(t, x) = \beta^0 \tanh(F(t)x).$$

Using the pair (β^0, ρ^0) in the adjoint system

$$-d\psi = A\psi dt - (1/\sigma^0)^4(h^0)^2\psi dt + (1/\sigma^0)^2(h^0\psi) \cdot d\hat{W}$$
$$\psi(T) = -(1 + \log \rho_T^0),$$

we obtain ψ^0. The triple $\{\beta^0, \rho^0, \psi^0\}$ is then used to compute the gradient $\tilde{D}J(\beta^0)$ as follows:

$$\tilde{D}J(\beta^0) \equiv \int_{I \times R} \{E(\rho^0\psi^0)(1/\sigma^0)^4\beta^0(\tanh^2(F(t)x)\}dxdt.$$

Using this gradient, define the next best value for β as follows:

$$\beta^1 = \beta^0 - \varepsilon\tilde{D}J(\beta^0) \quad \text{for } \varepsilon > 0 \quad \text{so that } \beta^1 \in \Lambda,$$

giving

$$h^1(t, x) = \beta^1 \tanh(F(t)x).$$

Using this new value, the iteration starts all over again from step one, giving ρ^1 leading to the pair $\{\beta^1, \rho^1\}$. This is then used for solving the adjoint system yielding ψ^1. The triple $\{\beta^1, \rho^1, \psi^1\}$ is used to compute the gradient $\tilde{D}J(\beta^1)$. These steps produce two sequences $\{\beta^k, J(\beta^k) = J(h^k)\}$. Comparing the difference $|J(\beta^{k+1}) - J(\beta^k)|$ with the specified tolerance, we see that the iteration is stopped if satisfied or continued if not.

Acknowledgments. The authors would like to express their thanks to the anonymous reviewers for their valuable comments.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley, New York, 1991.

[2] G. N. Saridis, *Entropy formulation of optimal and adaptive control*, IEEE Trans. Automat. Control, 33 (1988), pp. 713–721.

[3] G. N. Saridis, *Stochastic Processes, Estimation, and Control: The Entropy Approach*, John Wiley, New York, 1995.

[4] H. Yue and H. Wang, *Minimum entropy control of closed-loop tracking errors for dynamic stochastic systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 118–122.

[5] N. U. Ahmed, *Linear and Nonlinear Filtering for Scientists and Engineers*, World Scientific, Singapore, 1999.

[6] N. U. Ahmed, *Semigroup Theory with Applications to Systems and Control*, Longman Scientific and Technical, Harlow, UK; copublished in the United States with John Wiley, New York, 1991.

[7] H. J. Kushner, *Nonlinear filtering: The exact dynamical equations satisfied by the conditional modes*, IEEE Trans. Automat. Control, 12 (1967), pp. 262–267.

[8] R. S. Bucy, *Nonlinear filtering theory*, IEEE Trans. Automat. Control, 10 (1965), p. 198.

[9] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Geb., 11 (1969), pp. 230–243.

[10] N. U. Ahmed, *Existence and uniqueness of measure valued solutions for Zakai equation*, Publ. Math. Debrecen, 49 (1996), pp. 251–264.

[11] B. M. Miller and W. J. Runggaldier, *Optimization of observations: A stochastic control approach*, SIAM J. Control Optim., 35 (1997), pp. 1030–1052.

[12] V. Krishnamurthy, *Algorithm for optimal scheduling and management of hidden Markov model sensors*, IEEE Trans. Signal Process., 50 (2002), pp. 1382–1397.

[13] C. D. Charalambous and F. Rezaei, *Optimization of stochastic uncertain systems: Large deviations and robustness for partially observable diffusions*, in Proceedings of the American Control Conference (Boston, June 30–July 2, 2004), IEEE, Los Alamitos, CA, pp. 3152–3169.

[14] C. D. Charalmbous, F. Rezaei, and N. U. Ahmed, *Optimization of stochastic uncertain systems: Entropy rate functionals, Cramer's theorem, minimax games and robustness*, SIAM J. Control Optim., submitted.

[15] C. D. Charalambous and J. L. Hibey, *Minimum principle for partially observable nonlinear risk-sensitive control problems using measured-valued decompositions*, Stochastic Stochastics Rep., 57 (1996), pp. 247–288.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 532–547

SHAPE OPTIMIZATION OF AN ELLIPTIC EQUATION IN AN EXTERIOR DOMAIN*

GENGSHENG WANG†, LIJUAN WANG‡, AND DONGHUI YANG†

Abstract. This work concerns the shape optimization problem governed by some elliptic equations in exterior domains. The existence of optimal solutions is obtained and the so-called $\overset{\wedge}{\Gamma}$-property for a family $\mathcal{O}$ of open subsets is established; i.e., if $\Omega_m, \Omega \in \mathcal{O}$ are such that $\Omega_m \to \Omega$ in the Hausdorff metric, then for any $K \subset\subset R^N \setminus \overline{\Omega}$, there exists $m(K) > 0$ such that for all $m \geq m(K), K \subset\subset R^N \setminus \overline{\Omega}_m$.

Key words. shape optimization, exterior domain, elliptic equation

AMS subject classifications. 49Q10, 65K10

DOI. 10.1137/050625394

1. Introduction.

Let $D \subset R^N, N \geq 1$, be a given bounded and open subset. We shall study the existence of the following shape optimization problem:

$$(P) \qquad \text{Min}_{\Omega \in \mathcal{O}} \int_{R^N \setminus \overline{\Omega}} J(x, u_\Omega, \ldots, D^l u_\Omega)\, dx$$

subject to

$$(1.1) \qquad \begin{cases} Au_\Omega = f \quad \text{in } R^N \setminus \overline{\Omega}, \\ u_\Omega \in W_0^{l,p}(R^N \setminus \overline{\Omega}). \end{cases}$$

Here $f \in L^q(R^N)$ is a given function with $\frac{1}{p} + \frac{1}{q} = 1$, $1 < p, q < \infty$; $l \geq 1$ is an integer; $\mathcal{O}$ is a certain family of open subsets contained in D, which will be defined in section 2; J is a functional; and A is a nonlinear elliptic operator, which will be made precise in section 3.

There have been many works on the existence theory for shape optimization problems governed by elliptic equations (cf. [3], [4], [5], [7], [9], [10], [11], [13], [14], and [15]). In all these works, the domains in which the elliptic equations are defined are bounded. To the best of our knowledge, the shape optimization problems governed by elliptic equations defined on unbounded domains have not yet been studied. The problem (P) studied in this work is a shape optimization problem governed by elliptic equations on exterior domains. This work is inspired by [14], where the following shape optimization problem was studied:

$$(P_1) \qquad \text{Min}_{\Omega \in \mathcal{O}} \int_{\Omega} J(x, u_\Omega, \ldots, D^l u_\Omega)\, dx$$

*Received by the editors February 26, 2005; accepted for publication (in revised form) December 3, 2005; published electronically April 21, 2006. This work was supported by National Natural Science Foundation of China grants 10401041 and 10471053 and by the Grant of Key Laboratory–Optimal Control and Discrete Mathematics of Hubei Province.

http://www.siam.org/journals/sicon/45-2/62539.html

†Department of Mathematics, Wuhan University, Wuhan, Hubei, 430072, People's Republic of China (wanggs@public.wh.hb.cn, dongfyang@yahoo.com.cn).

‡Corresponding author. Department of Mathematics, Wuhan University, Wuhan, Hubei, 430072, People's Republic of China (hepeijie@public.wh.hb.cn).

subject to

$$\begin{cases} Au_\Omega = f \quad \text{in } \Omega, \\ u_\Omega \in W_0^{l,p}(\Omega). \end{cases}$$

Like most shape optimization problems governed by elliptic equations defined on bounded domains, the existence of optimal shapes for problem (P_1) can be implied by the following (cf. [14], [11]):

(i) If $\{\Omega_m\}_{m=1}^\infty \subset \mathcal{O}$, then there is a subsequence $\{\Omega_{m_k}\}_{k=1}^\infty$ of $\{\Omega_m\}_{m=1}^\infty$ such that $\Omega_{m_k} \to \Omega$ in the Hausdorff metric and $\Omega \in \mathcal{O}$.

(ii) If $\{\Omega_m\}_{m=1}^\infty \subset \mathcal{O}$ and $\Omega \in \mathcal{O}$ are such that $\Omega_m \to \Omega$ in the Hausdorff metric, then for any open subset K with $K \subset\subset \Omega$, i.e., $\overline{K} \subset \Omega$, there is an integer $m(K) > 0$ such that for all $m \geq m(K), K \subset\subset \Omega_m$ (this property is called the Γ-property for $\mathcal{O}$).

(iii) If $\{\Omega_m\}_{m=1}^\infty \subset \mathcal{O}$ and $\Omega \in \mathcal{O}$ are such that $\Omega_m \to \Omega$ in the Hausdorff metric, then $\chi_{\Omega_m} \to \chi_\Omega$ a.e. in R^N.

(iv) If $u \in W^{l,p}(R^N)$ and $u = 0$ a.e. in $R^N \setminus \Omega$ for some $\Omega \in \mathcal{O}$, then $u \in W_0^{l,p}(\Omega)$.

However, conditions (i)–(iv) cannot guarantee the existence of the optimal solutions for problem (P) since the equations in problem (P) are defined on the exterior domains. In this paper, we observe that the existence of the optimal solutions for problem (P) can be implied by (i), (ii), (iv) and the following:

(iii)′ If $\{\Omega_m\}_{m=1}^\infty \subset \mathcal{O}$ and $\Omega \in \mathcal{O}$ are such that $\Omega_m \to \Omega$ in the Hausdorff metric, then for any open subset K with $K \subset\subset R^N \setminus \overline{\Omega}$, there exists an integer $m(K) > 0$ such that for all $m \geq m(K), K \subset\subset R^N \setminus \overline{\Omega}_m$. (This property is called the $\overset{\wedge}{\Gamma}$-property for $\mathcal{O}$.)

Moreover, we establish the $\overset{\wedge}{\Gamma}$-property for $\mathcal{O}$. We must mention that the Γ-property holds for any family of open subsets contained in D, but the $\overset{\wedge}{\Gamma}$-property holds only for certain families of open subsets contained in D, which will be shown in section 2.

2. $\overset{\wedge}{\Gamma}$-property for $\mathcal{O}$. In this section, we shall introduce the family $\mathcal{O}$, which is the same as that in [14], and prove the $\overset{\wedge}{\Gamma}$-property for $\mathcal{O}$.

Let $\Omega \subset R^N$ be a bounded and open subset. We say that Ω is of class $\mathcal{C}$ if the following conditions (2.1), (2.2), and (2.3) are fulfilled. There is a family $\mathcal{F}_\Omega$ of real valued continuous functions g defined on $\overline{S(0, k_\Omega)}$, where $S(0, k_\Omega) \subset R^{N-1}$ is the open ball centered at 0 and is of radius $k_\Omega > 0$ such that

$$\partial\Omega = \bigcup_{g \in \mathcal{F}_\Omega} \{R_g(s, 0) + o_g + g(s)y_g; s \in S(0, k_\Omega)\}. \tag{2.1}$$

Here $o_g \in \partial\Omega$ is a certain point, which gives the center of the local system of axes, and $R_g : R^N \to R^N$ is a certain unitary rotation operator such that $R_g(0, 0, \ldots, 0, 1) = y_g$, where the unit vector y_g gives the local vertical direction. Moreover, for $g \in \mathcal{F}_\Omega$, there is $a_\Omega > 0$ such that

$$R_g(s, 0) + o_g + (g(s) - t)y_g \in \Omega \text{ for all } t \in (0, a_\Omega), \quad s \in S(0, k_\Omega), \tag{2.2}$$

and

$$R_g(s, 0) + o_g + (g(s) + t)y_g \in R^N \setminus \overline{\Omega} \text{ for all } t \in (0, a_\Omega), \quad s \in S(0, k_\Omega). \tag{2.3}$$

Roughly speaking, $\partial\Omega$ is locally the graph of a continuous function satisfying the segment property, and Ω is under the graph.

By the Lebesgue lemma, it follows from (2.1) that there is an $r_\Omega \in (0, k_\Omega)$ such that the restricted local charts defined on $\overline{S(0, r_\Omega)}$ also give a covering of $\partial\Omega$:

$$\partial\Omega = \bigcup_{g\in\mathcal{F}_\Omega} \{R_g(s,0) + o_g + g(s)y_g; s \in \overline{S(0,r_\Omega)}\}. \tag{2.4}$$

Let k_0, r_0, a_0 be positive constants with $r_0 < k_0$. We define $\mathcal{O}(k_0, r_0, a_0)$ to be the family of all open subsets Ω contained in D such that every $\Omega \in \mathcal{O}(k_0, r_0, a_0)$ is of class $\mathcal{C}$ with $k_\Omega \geq k_0, a_\Omega \geq a_0, r_\Omega \leq r_0$, where k_Ω, a_Ω, and r_Ω are given as in (2.1)–(2.4). In what follows, we will write $\mathcal{O}$ for $\mathcal{O}(k_0, r_0, a_0)$ for simplicity of notation. Moreover, we require that the family of functions $\mathcal{F} = \cup_{\Omega\in\mathcal{O}}\mathcal{F}_\Omega$ be equicontinuous and equibounded on $\overline{S(0,k_0)}$.

We define the usual topology on $\mathcal{O}$ by the Hausdorff–Pompeiu distance between the complementary sets, i.e.,

(2.5)
$$\rho(\Omega_1,\Omega_2) = \max\{\sup_{x\in\overline{D}\setminus\Omega_1} d(x, \overline{D}\setminus\Omega_2), \sup_{y\in\overline{D}\setminus\Omega_2} d(\overline{D}\setminus\Omega_1, y)\} \quad \text{for all } \Omega_1,\Omega_2\in\mathcal{O},$$

where $d(\cdot,\cdot)$ denotes the Euclidean metric in R^N. We denote by Hlim the limit in terms of (2.5).

By Theorem 1 of Chapter 3 in [11] and Theorem 2.1 in [14], the following result follows immediately.

LEMMA 2.1. *For any sequence $\{\Omega_n\}_{n=1}^\infty \subset \mathcal{O}$, there are subsequences $\{\Omega_{n_k}\}_{k=1}^\infty$ of $\{\Omega_n\}_{n=1}^\infty$ and $\Omega^* \in \mathcal{O}$ such that $\mathrm{Hlim}_{k\to\infty}\Omega_{n_k} = \Omega^*$.*

In [6] and [8], the Γ-property for $\mathcal{O}$ (indeed for any family of open subsets $\Omega \subset D$) was established.

LEMMA 2.2 (Γ-property for $\mathcal{O}$). *Assume that $\{\Omega_n\}_{n=1}^\infty \subset \mathcal{O}, \Omega_0 \in \mathcal{O}$, and $\Omega_0 = \mathrm{Hlim}\Omega_n$. Then for each open subset K satisfying $\overline{K} \subset \Omega_0$, there exists a positive integer n_K (depending on K) such that $\overline{K} \subset \Omega_n$ for all $n \geq n_K$.*

The next lemma plays a key role in this paper.

LEMMA 2.3 ($\overset{\wedge}{\Gamma}$-property for $\mathcal{O}$). *Assume that $\{\Omega_n\}_{n=1}^\infty \subset \mathcal{O}, \Omega_0 \in \mathcal{O}$, and $\Omega_0 = \mathrm{Hlim}\Omega_n$. Then for any open subset K satisfying $\overline{K} \subset R^N \setminus \overline{\Omega}_0$, there exists a positive integer n_K such that $\overline{K} \subset R^N \setminus \overline{\Omega}_n$ for $n \geq n_K$.*

Proof. Throughout the proof of Lemma 2.3, we shall use $B(x,r)$ and $S(y,r)$ to denote the closed ball with center at x and radius r in R^N and the open ball with center at y and radius r in R^{N-1}, respectively.

Let K be an open subset such that $\overline{K} \subset R^N \setminus \overline{\Omega}_0$. It suffices to prove that there is a neighborhood U_{Ω_0} of $\overline{\Omega_0}$ such that $U_{\Omega_0} \cap \overline{K} = \emptyset$ and $\overline{\Omega}_n \subset U_{\Omega_0}$ for n large enough.

We observe first that

$$\inf_{x\in\overline{K}, y\in\overline{\Omega}_0} d(x,y) \equiv c_0 > 0 \tag{2.6}$$

and there are $\varepsilon_0 > 0$ and a positive integer n_0 such that

$$\left\{x \in \Omega_n : d(x, \partial\Omega_n) \geq \frac{\varepsilon_0}{2}\right\} \neq \emptyset \text{ for } n \geq n_0. \tag{2.7}$$

Indeed, since $\Omega_0 \in \mathcal{O}, \Omega_0 \neq \emptyset$, there exist $x_0 \in \Omega_0$ and $\varepsilon_0 > 0$ such that $B(x_0, \varepsilon_0) \subset \Omega_0$. Hence by Lemma 2.2, there is a positive integer n_0 such that for all $n \geq n_0$, $B(x_0, \varepsilon_0) \subset \Omega_n$. Hence $x_0 \in \{x \in \Omega_n : d(x, \partial\Omega_n) \geq \frac{\varepsilon_0}{2}\}$, and thus (2.7) holds.

Now we take $\varepsilon > 0$ such that $\varepsilon < \frac{1}{2}\min\{\varepsilon_0, a_0, k_0 - r_0, \frac{c_0}{12}\}$. By the equicontinuity of the family $\cup_{\Omega\in\mathcal{O}}\mathcal{F}_\Omega$, there is $0 < \delta(\varepsilon) < \frac{\varepsilon}{2}$ such that

$$|g(s) - g(\widetilde{s})| < \frac{\varepsilon}{2} \quad \text{for all } g \in \mathcal{F}_\Omega \text{ and } \Omega \in \mathcal{O}, \tag{2.8}$$
$$\text{for all } s, \widetilde{s} \in S(0, k_0) \text{ and } |s - \widetilde{s}\,|_{R^{N-1}} \le \delta(\varepsilon).$$

Let $\Omega_n^1 = \{x \in \Omega_n : d(x, \partial\Omega_n) \ge \frac{\delta(\varepsilon)}{2}\}$ and $\Omega_n^2 = \{x \in \Omega_n : d(x, \partial\Omega_n) < \frac{\delta(\varepsilon)}{2}\}$. It is clear that by (2.7), $\Omega_n^1 \ne \emptyset$ for all $n \ge n_0$.

First we claim that

$$\Omega_n^1 \subset \overline{\Omega}_0 \subset \left\{x \in R^N : d(x, \overline{\Omega}_0) \le \frac{c_0}{12}\right\}. \tag{2.9}$$

In order to prove (2.9), we observe that

$$R^N \Big\backslash \left((\overline{D} \setminus \Omega_n) + \frac{\delta(\varepsilon)}{4}B(0,1)\right) \subset R^N \setminus (\overline{D} \setminus \Omega_0) \subset (R^N \setminus \overline{D}) \cup \overline{\Omega}_0. \tag{2.10}$$

Indeed, since $\Omega_0 = \mathrm{Hlim}\Omega_n$, we have that (see page 48 of [12])

$$\inf\{\lambda \ge 0 | \overline{D} \setminus \Omega_n \subset (\overline{D} \setminus \Omega_0) + \lambda B(0,1), \overline{D} \setminus \Omega_0 \subset (\overline{D} \setminus \Omega_n) + \lambda B(0,1)\} \to 0 \text{ as } n \to \infty.$$

Thus there is a positive integer $n_0(\varepsilon) \ge n_0$ such that for all $n \ge n_0(\varepsilon)$,

$$\inf\{\lambda \ge 0 | \overline{D} \setminus \Omega_n \subset (\overline{D} \setminus \Omega_0) + \lambda B(0,1), \overline{D} \setminus \Omega_0 \subset (\overline{D} \setminus \Omega_n) + \lambda B(0,1)\} \le \frac{\delta(\varepsilon)}{8},$$

and then for each $n \ge n_0(\varepsilon)$, there is a λ_n with $0 \le \lambda_n \le \frac{\delta(\varepsilon)}{4}$ such that

$$\overline{D} \setminus \Omega_0 \subset (\overline{D} \setminus \Omega_n) + \lambda_n B(0,1) \subset (\overline{D} \setminus \Omega_n) + \frac{\delta(\varepsilon)}{4}B(0,1),$$

from which the first inclusion in (2.10) follows. The second inclusion in (2.10) is obvious.

Then we prove that

$$\Omega_n^1 \subset R^N \Big\backslash \left((\overline{D} \setminus \Omega_n) + \frac{\delta(\varepsilon)}{4}B(0,1)\right) \text{ for all } n \ge n_0(\varepsilon). \tag{2.11}$$

By contradiction, we assume that there exists $x_n \in \Omega_n^1$ and $x_n \in (\overline{D}\setminus\Omega_n) + \frac{\delta(\varepsilon)}{4}B(0,1)$. It follows from the latter that

$$x_n = y_n + \frac{\delta(\varepsilon)}{4}z_n \text{ for some } y_n \in \overline{D} \setminus \Omega_n \text{ and } z_n \in B(0,1).$$

Hence $d(x_n, y_n) \le \frac{\delta(\varepsilon)}{4}$. By Lemma A in the appendix, there exists $w_n \in \partial\Omega_n$ satisfying $d(x_n, w_n) \le d(x_n, y_n) \le \frac{\delta(\varepsilon)}{4}$, which leads to a contradiction and implies (2.11) as desired.

Notice that $\Omega_n \subset D$ for all $n \ge 1$. Equation (2.9) follows immediately from (2.10) and (2.11).

Next we claim that

$$\Omega_n^2 \subset \left\{x \in R^N : \ d(x, \overline{\Omega}_0) \le \frac{c_0}{12}\right\}. \tag{2.12}$$

Let $n \geq n_0(\varepsilon)$ be arbitrary but fixed and let $x \in \Omega_n^2$. It suffices to find $z \in \overline{\Omega}_0$ such that $d(x,z) \leq \frac{c_0}{12}$.

To this end, we notice that there exists $y \in \partial\Omega_n$ such that

$$d(x,y) = d(x,\partial\Omega_n) < \frac{\delta(\varepsilon)}{2}. \tag{2.13}$$

By (2.4), there is a $g \in \mathcal{F}_{\Omega_n}$ and an $s_n \in \overline{S(0,r_n)}$ such that

$$y = R_g(s_n,0) + o_g + g(s_n)y_g.$$

By (2.8), we get that

$$-\frac{\varepsilon}{2} < g(s) - g(s_n) < \frac{\varepsilon}{2} \quad \text{for all } s \in S(0,k_0) \cap \overline{S\left(s_n, \frac{\delta(\varepsilon)}{2}\right)},$$

which, together with (2.2), implies that

$$\begin{aligned}
&\left\{R_g(s,0) + o_g + (g(s_n) - t)y_g \in R^N : \ s \in S(0,k_0) \cap \overline{S\left(s_n, \frac{\delta(\varepsilon)}{2}\right)}, \right. \\
&\qquad\qquad \left. t \in \left[\frac{\varepsilon}{2}, a_0 - \frac{\varepsilon}{2}\right]\right\} \\
&= \left\{R_g(s,0) + o_g + g(s)y_g - (g(s) - g(s_n) + t)y_g \in R^N : \right. \\
&\qquad\qquad \left. s \in S(0,k_0) \cap \overline{S\left(s_n, \frac{\delta(\varepsilon)}{2}\right)}, \ t \in \left[\frac{\varepsilon}{2}, a_0 - \frac{\varepsilon}{2}\right]\right\} \\
&\subset \left\{R_g(s,0) + o_g + (g(s) - t)y_g \in R^N : \ s \in S(0,k_0) \cap \overline{S\left(s_n, \frac{\delta(\varepsilon)}{2}\right)}, t \in (0,a_0)\right\} \\
&\subset \Omega_n.
\end{aligned} \tag{2.14}$$

Since $s_n \in \overline{S(0,r_n)}$, $r_n \leq r_0 < k_0$, and $0 < 2\delta(\varepsilon) < \varepsilon < \frac{k_0 - r_0}{2}$, we get

$$\overline{S\left(s_n, \frac{\delta(\varepsilon)}{2}\right)} \subset S(0,k_0).$$

Then by (2.14), we have that

$$\left\{R_g(s,0) + o_g + (g(s_n) - t)y_g \in R^N : \ s \in \overline{S\left(s_n, \frac{\delta(\varepsilon)}{2}\right)}, t \in \left[\frac{\varepsilon}{2}, a_0 - \frac{\varepsilon}{2}\right]\right\} \subset \Omega_n. \tag{2.15}$$

Now we take $z = R_g(s_n,0) + o_g + (g(s_n) - \varepsilon)y_g$. Since $0 < 2\delta(\varepsilon) < \varepsilon < \frac{a_0}{2}$, it follows from (2.15) that

$$z \in \Omega_n, \ \ d(z,\partial\Omega_n) \geq \min\left\{\frac{\delta(\varepsilon)}{2}, \frac{\varepsilon}{2}, a_0 - \frac{3\varepsilon}{2}\right\} = \frac{\delta(\varepsilon)}{2} \tag{2.16}$$

and

$$d(y,z) = \varepsilon. \tag{2.17}$$

By (2.13) and (2.17), we have that

$$d(x,z) \le d(x,y) + d(y,z) \le 2\varepsilon \le \frac{c_0}{12}.$$

By (2.9) and (2.16), we get $z \in \overline{\Omega}_0$. Thus we have proved (2.12).

By (2.9) and (2.12), we get that

$$\overline{\Omega}_n \subset \left\{ x \in R^N : d(x, \overline{\Omega}_0) \le \frac{c_0}{12} \right\}.$$

Take $U_{\Omega_0} = \left\{ x \in R^N : d(x, \overline{\Omega}_0) < \frac{c_0}{6} \right\}$. Then by (2.6), we have that

$$U_{\Omega_0} \cap \overline{K} = \emptyset \text{ and } \overline{\Omega}_n \subset U_{\Omega_0}.$$

This completes the proof. □

Finally, we shall show by a counterexample that the $\overset{\wedge}{\Gamma}$-property may not hold for some family $\widetilde{O}$ of open subsets $\Omega \subset D$, even though each Ω in $\widetilde{O}$ is of class $\mathcal{C}$.

Example. Let

$$D = \{(x,y) \in R^2 : |x| < 2, |y| < 2\}, \quad \widetilde{\mathcal{O}} = \{\Omega \subset D : \Omega \text{ is open and of class } \mathcal{C}\},$$

$$\Omega_0 = \left\{ (x,y) \in R^2 : 0 < y < 1 - |x|, \frac{1}{2} < |x| < 1 \right\} \cup \left\{ (x,y) \in R^2 : |x| \le \frac{1}{2}, 0 < y < \frac{1}{2} \right\},$$

$$\Omega_n = \Omega_0 \cup \left\{ (x,y) \in R^2 : \frac{1}{2} \le y < 1 - n|x|, |x| < \frac{1}{2n} \right\}, \qquad n \ge 1,$$

and

$$K = \left\{ (x,y) \in R^2 : |x| < \frac{3}{2}, |y| < 1 \right\} \Big\backslash \left\{ (x,y) \in R^2 : |x| \le \frac{5}{4}, |y| \le \frac{3}{4} \right\}.$$

One can easily check that

(i) $\Omega_0 \in \widetilde{\mathcal{O}}$ and $\Omega_n \in \widetilde{\mathcal{O}}, n = 1, 2, \ldots$;

(ii) $\Omega_0 = \text{Hlim}\Omega_n$;

(iii) $\overline{K} \subset R^N \setminus \overline{\Omega}_0$, but $\overline{K} \cap \overline{\Omega}_n \ne \emptyset$ for any $n \ge 1$.

Thus $\widetilde{\mathcal{O}}$ has no $\overset{\wedge}{\Gamma}$-property.

Figures 1–3 show the regions Ω_0, Ω_n, and $\overline{K}$.

3. Existence of optimal shape for problem (*P*). In this section, we shall first introduce the elliptic operator A and cost functional J in problem (P) given in section 1, and then prove the existence of an optimal shape for problem (P).

Let A be a differential operator from $W^{l,p}(R^N)$ to $W^{-l,q}(R^N)$ defined by

$$Az = \sum_{|\alpha| \le l} (-1)^{|\alpha|} D^\alpha A_\alpha(x, z, \ldots, D^l z), \quad x \in R^N. \tag{3.1}$$

Here α is a multi-index such that $|\alpha| \le l$ and $A_\alpha : R^N \times R^T \to R$ (T is the number of all partial derivatives in R^N from the order 0 to the order l) satisfies the following:

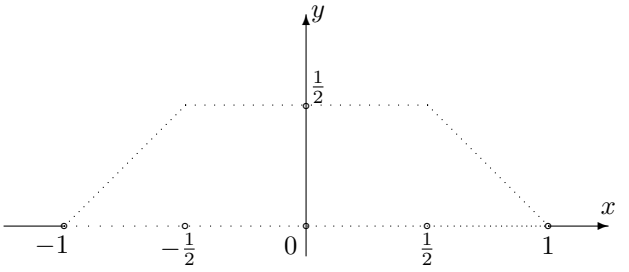


FIG. 1. Ω_0 *is the region surrounded by imaginary lines.*

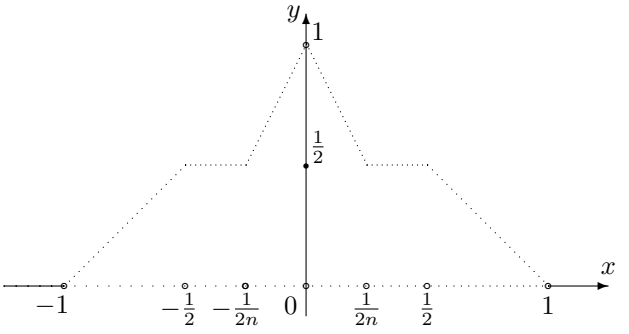


FIG. 2. Ω_n *is the region surrounded by imaginary lines.*

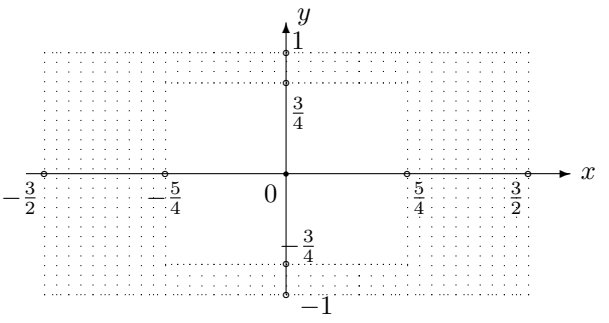


FIG. 3. K *is the region occupied by imaginary lines.*

(H_1) For each multi-index α, A_α is measurable in $x \in R^N$ and continuous in the other variables (denoted by $\xi \in R^T$) as follows:

$$|A_\alpha(x,\xi)| \le c_1(|\xi|_{R^T}^{p-1} + \mu(x)), \quad x \in R^N, \quad \xi \in R^T,$$

where $\mu \in L^q(R^N)$ is a given function, $\xi = (\xi_\alpha)_{|\alpha|\le l}$, and c_1 is a positive constant independent of α.

(H_2) There exists a positive constant c_2 such that for almost all $x \in R^N$ and all $\xi, \eta \in R^T$,

$$\sum_{|\alpha|\le l} (A_\alpha(x,\xi) - A_\alpha(x,\eta))(\xi_\alpha - \eta_\alpha) \ge c_2|\xi - \eta|_{R^T}^p.$$

(H_3) There exists a positive constant c_3 such that for almost all $x \in R^N$ and all $\xi \in R^T$,

$$\sum_{|\alpha|\le l} A_\alpha(x,\xi)\xi_\alpha \ge c_3\Big(\sum_{|\alpha|=l} |\xi_\alpha|^p + |\xi_0|^p\Big),$$

where $\xi_0 \equiv \xi_\alpha$ with $\alpha = 0$ is the component of ξ corresponding to the 0 order partial derivative, and $\xi_\alpha, |\alpha| = l$, are the components of ξ corresponding to the highest order partial derivatives.

Let $\mathcal{O}$ be given as in section 2 and let $\Omega \in \mathcal{O}$. We say $u_\Omega \in W_0^{l,p}(R^N \setminus \overline{\Omega})$ is a weak solution of (1.1) if

$$\sum_{|\alpha|\le l} \int_{R^N\setminus\overline{\Omega}} A_\alpha(x, u_\Omega, \dots, D^l u_\Omega)\cdot D^\alpha v\, dx = \int_{R^N\setminus\overline{\Omega}} f\cdot v\, dx \tag{3.2}$$

for all $v \in W_0^{l,p}(R^N \setminus \overline{\Omega})$.

THEOREM 3.1. *Under assumptions* $(H_1)-(H_3)$, (1.1) *has a unique weak solution for each open and bounded subset* $\Omega \subset R^N$.

Proof. It follows from (H_2) that for almost all $x \in R^N$ and all $\xi, \eta \in R^T$,

$$\sum_{|\alpha|\le l} (A_\alpha(x,\xi) - A_\alpha(x,\eta))(\xi_\alpha - \eta_\alpha) \ge 0. \tag{3.3}$$

Now we shall show that

$$A : W_0^{l,p}(R^N \setminus \overline{\Omega}) \to W^{-l,q}(R^N \setminus \overline{\Omega}) \text{ is coercive.} \tag{3.4}$$

Indeed, it follows from (H_3) that

$$\begin{aligned}
&\frac{\langle Au, u\rangle_{W_0^{l,p}(R^N\setminus\overline{\Omega}),W^{-l,q}(R^N\setminus\overline{\Omega})}}{\|u\|_{W_0^{l,p}(R^N\setminus\overline{\Omega})}}\\
&\ge \frac{c_3\Big(\sum_{|\alpha|=l}\|D^\alpha u\|^p_{L^p(R^N\setminus\overline{\Omega})} + \|u\|^p_{L^p(R^N\setminus\overline{\Omega})}\Big)}{\|u\|_{W_0^{l,p}(R^N\setminus\overline{\Omega})}}\\
&\ge C\|u\|^{p-1}_{W_0^{l,p}(R^N\setminus\overline{\Omega})} \to +\infty \qquad \text{as } \|u\|_{W_0^{l,p}(R^N\setminus\overline{\Omega})} \to +\infty,
\end{aligned}$$

which implies (3.4). Here and throughout the proof of Theorem 3.1, C denotes several positive constants.

Then by (H_1), (3.3), (3.4), and using the same arguments as those in Proposition 1.2 of Chapter 2 in [1], we infer that (1.1) has at least one weak solution.

Next, we prove the uniqueness. Suppose that u_1 and u_2 are two weak solutions of (1.1); then by (H_2) and (3.2), we obtain that

$$\begin{aligned}
0 &= \sum_{|\alpha|\le l} \int_{R^N\setminus\overline{\Omega}} [A_\alpha(x, u_1, \dots, D^l u_1) - A_\alpha(x, u_2, \dots, D^l u_2)]\cdot[D^\alpha u_1 - D^\alpha u_2]\, dx\\
&\ge C\|u_1 - u_2\|^p_{W_0^{l,p}(R^N\setminus\overline{\Omega})},
\end{aligned}$$

which indicates that $u_1 = u_2$. This completes the proof. □

Next we suppose that the function $J(x,\xi): R^N \times R^T \to R$ satisfies the following:
(H_4) J is measurable in $x \in R^N$ and continuous in $\xi \in R^T$,

$$0 \le J(x,\xi) \le c_4(|\xi|^p_{R^T} + \eta(x)), \quad x \in R^N, \ \ \xi \in R^T,$$

with $c_4 > 0$ and $\eta \in L^1(R^N)$.

The following lemma, which was provided in [14], will be used in our proof of the existence of problem (P).

LEMMA 3.2. *Let Ω be a bounded domain of class $\mathcal{C}$. If $z \in W^{l,p}(R^N)$ and $z = 0$ a.e. in $R^N \setminus \Omega$, then $z \in W^{l,p}_0(\Omega)$.*

The main result of the paper is as follows.

THEOREM 3.3. *The shape optimization problem (P) has at least one solution.*

Proof. Throughout the proof of Theorem 3.3, we shall use spt$u$ and $U(x,r)$ to denote the support of u and the open ball in R^N with center at x and radius r, respectively.

Let $d = \mathrm{Min}_{\Omega\in\mathcal{O}} \int_{R^N\setminus\overline{\Omega}} J(x, u_\Omega, \dots, D^l u_\Omega)\, dx$. It is obvious that $d > -\infty$. Then there exists a sequence $\{\Omega_n\}_{n=1}^\infty \subset \mathcal{O}$ such that

$$d = \lim_{n\to\infty} \int_{R^N\setminus\overline{\Omega}_n} J(x, u_n, \dots, D^l u_n)\, dx, \tag{3.5}$$

where $u_n \equiv u_{\Omega_n}$ is the weak solution of (1.1) in $R^N \setminus \overline{\Omega}_n$. By Lemma 2.1, there exist a subsequence of $\{\Omega_n\}_{n=1}^\infty$, still denoted by itself, and $\Omega^* \in \mathcal{O}$, such that $\Omega^* = \mathrm{Hlim}\Omega_n$.

By taking $v = u_n, \Omega = \Omega_n$ in (3.2) and by (H_3), we get that

$$\sum_{|\alpha|=l} \|D^\alpha u_n\|^p_{L^p(R^N\setminus\overline{\Omega}_n)} + \|u_n\|^p_{L^p(R^N\setminus\overline{\Omega}_n)} \le \frac{1}{2}\|u_n\|^p_{L^p(R^N\setminus\overline{\Omega}_n)} + C\|f\|^q_{L^q(R^N)},$$

which implies that

$$\|u_n\|_{W^{l,p}_0(R^N\setminus\overline{\Omega}_n)} \le C;$$

here and throughout the proof of Theorem 3.3, C denotes several positive constants independent of n. Let

$$\hat{u}_n(x) = \begin{cases} u_n(x) & \text{in } R^N \setminus \overline{\Omega}_n, \\ 0 & \text{in } \overline{\Omega}_n; \end{cases} \tag{3.6}$$

then $\{\hat{u}_n\}_{n=1}^\infty$ is bounded in $W^{l,p}(R^N)$. Hence there exists a subsequence of $\{\hat{u}_n\}$, still denoted by itself, such that

$$\hat{u}_n \to \hat{u} \ \text{ weakly in } W^{l,p}(R^N) \text{ and strongly in } L^p(\overline{D}) \tag{3.7}$$

for some $\hat{u} \in W^{l,p}(R^N)$.

We observe that

$$\hat{u}\,(x) = 0 \text{ a.e. in } \Omega^*. \tag{3.8}$$

Indeed, for any open subset K satisfying $\overline{K} \subset \Omega^*$, by Lemma 2.2, there is a positive integer n_K, such that $\overline{K} \subset \Omega_n \subset D$ for all $n \ge n_K$. Thus

$$\int_K |\,\hat{u}\,(x)|^p\, dx = \lim_{n\to\infty} \int_K |\hat{u}_n(x)|^p\, dx \le \overline{\lim_{n\to\infty}} \int_{\Omega_n} |\hat{u}_n(x)|^p\, dx = 0,$$

which implies that $\hat{u}(x) = 0$ a.e. in K. Since $K \subset \overline{K} \subset \Omega^*$ is arbitrary, (3.8) follows.

We claim first that

$$\hat{u}(x) \in W_0^{l,p}(R^N \setminus \overline{\Omega}^*). \tag{3.9}$$

It suffices to construct a sequence of functions in $W_0^{l,p}(R^N \setminus \overline{\Omega}^*)$ converging to $\hat{u}$ in the $W^{l,p}(R^N \setminus \overline{\Omega}^*)$-norm.

To this end, we let $\eta_{\frac{1}{4}}(x) = 4^N \eta(4x)$, where $\eta : R^N \to R$ is defined by

$$\eta(x) = \begin{cases} c \cdot \exp\left(\frac{1}{|x|^2 - 1}\right) & \text{if } |x| < 1, \\ 0 & \text{if } |x| \geq 1 \end{cases}$$

with the constant c satisfying $\int_{R^N} \eta(x)\, dx = 1$, and let

$$\varphi_r(x) = \left(\eta_{\frac{1}{4}} * \chi_{U(0, r+\frac{1}{2})}\right)(x) = \int_{R^N} \eta_{\frac{1}{4}}(x - y) \chi_{U(0, r+\frac{1}{2})}(y)\, dy.$$

One can easily check that $\varphi_r(x) \in C_0^\infty(R^N)$,

(3.10)

$$|D^\alpha \varphi_r(x)| \leq C \text{ for all } |\alpha| \leq l \text{ and } x \in R^N, \quad \varphi_r(x) = \begin{cases} 1 & \text{in } \overline{U(0, r)}, \\ 0 & \text{in } R^N \setminus U(0, r + \frac{7}{8}); \end{cases}$$

here and throughout the proof of Theorem 3.3, C denotes several positive constants independent of r. Thus we have that

$$(\varphi_r \hat{u})(x) \in W^{l,p}(R^N).$$

Let $\widetilde{r}_0 > 0$ be such that $\overline{\Omega}^* \subset U(0, \frac{\widetilde{r}_0}{2})$. Then for all $r \geq \widetilde{r}_0$, the set $\widetilde{\Omega}_r = U(0, r + 1) \setminus \overline{\Omega}^*$ is of class $\mathcal{C}$ and $R^N \setminus \widetilde{\Omega}_r = \overline{\Omega}^* \cup (R^N \setminus U(0, r + 1))$.

By (3.8) and (3.10), we obtain that

$$(\varphi_r \hat{u})(x) = 0 \text{ a.e. in } R^N \setminus \widetilde{\Omega}_r \quad \text{for all } r \geq \widetilde{r}_0;$$

then by Lemma 3.2, $\varphi_r \hat{u} \in W_0^{l,p}(\widetilde{\Omega}_r)$, and thus

$$\varphi_r \hat{u} \in W_0^{l,p}(R^N \setminus \overline{\Omega}^*).$$

It is clear that $\varphi_r \hat{u} \to \hat{u}$ in $W^{l,p}(R^N \setminus \overline{\Omega}^*)$. Thus (3.9) holds.

Then we claim that

$$\begin{aligned} &\sum_{|\alpha| \leq l} \int_{R^N \setminus \overline{\Omega}^*} A_\alpha(x, \hat{u}, \ldots, D^l \hat{u}) \cdot D^\alpha v\, dx \\ &\quad = \int_{R^N \setminus \overline{\Omega}^*} f(x) \cdot v(x)\, dx \text{ for all } v(x) \in W_0^{l,p}(R^N \setminus \overline{\Omega}^*). \end{aligned} \tag{3.11}$$

It suffices to prove that

$$\sum_{|\alpha|\le l}\int_{R^N\setminus\overline{\Omega}^*} A_\alpha(x,\hat{u},\dots,D^l\,\hat{u})\cdot D^\alpha v\,dx$$
$$=\int_{R^N\setminus\overline{\Omega}^*} f(x)\cdot v(x)\,dx \quad \text{for all } v(x)\in C_0^\infty(R^N\setminus\overline{\Omega}^*),$$

i.e., for each $v(x)\in C_0^\infty(R^N\setminus\overline{\Omega}^*)$,

$$\sum_{|\alpha|\le l}\int_{\mathrm{spt}v} A_\alpha(x,\hat{u},\dots,D^l\,\hat{u})\cdot D^\alpha v\,dx=\int_{\mathrm{spt}v} f(x)\cdot v(x)\,dx. \tag{3.12}$$

We observe that

$$\sum_{|\alpha|\le l}\int_{\mathrm{spt}v} A_\alpha(x,\hat{u}_n,\dots,D^l\hat{u}_n)\cdot D^\alpha v\,dx=\int_{\mathrm{spt}v} f(x)\cdot v(x)\,dx. \tag{3.13}$$

Indeed, we set

$$\hat{v}=\begin{cases} v & \text{in } \mathrm{spt}v,\\ 0 & \text{in } R^N\setminus \mathrm{spt}v.\end{cases} \tag{3.14}$$

By Lemma 2.3, there exists a positive integer $n_0(v)$ such that

$$\mathrm{spt}\,\hat{v}=\mathrm{spt}v\subset R^N\setminus\overline{\Omega}_n \quad \text{for all } n\ge n_0(v). \tag{3.15}$$

Then for each $n\ge n_0(v)$, $\hat{v}\in C_0^\infty(R^N\setminus\overline{\Omega}_n)$. Thus by (3.2), we have that

$$\sum_{|\alpha|\le l}\int_{R^N\setminus\overline{\Omega}_n} A_\alpha(x,\hat{u}_n,\dots,D^l\hat{u}_n)\cdot D^\alpha\,\hat{v}\,dx=\int_{R^N\setminus\overline{\Omega}_n} f(x)\cdot\hat{v}\,dx,$$

which, together with (3.14) and (3.15), implies (3.13).

Now we shall prove that

$$A_\alpha(x,\hat{u}_n,\dots,D^l\hat{u}_n)\to A_\alpha(x,\hat{u},\dots,D^l\,\hat{u}) \text{ weakly in } L^q(\mathrm{spt}v) \text{ for } |\alpha|\le l. \tag{3.16}$$

Let $\varphi\in C_0^\infty(R^N\setminus\overline{\Omega}^*)$ be a positive function such that

$$\varphi=1 \quad \text{in } \mathrm{spt}v. \tag{3.17}$$

By Lemma 2.3, there exists a positive integer $n_1\ge n_0(v)$ such that

$$\mathrm{spt}\varphi\subset R^N\setminus\overline{\Omega}_n \quad \text{for all } n\ge n_1. \tag{3.18}$$

We set

$$\hat{\varphi}=\begin{cases}\varphi & \text{in } R^N\setminus\overline{\Omega}^*,\\ 0 & \text{in } \overline{\Omega}^*\end{cases} \tag{3.19}$$

and

(3.20)
$$\begin{aligned} I_n &\equiv \int_{R^N} \hat{\varphi} \sum_{|\alpha|\le l} (A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n) - A_\alpha(x, \hat{u}, \dots, D^l \hat{u})) \cdot (D^\alpha \hat{u}_n - D^\alpha \hat{u})\, dx \\ &= \int_{\mathrm{spt}\hat{\varphi}} \hat{\varphi} \sum_{|\alpha|\le l} (A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n) - A_\alpha(x, \hat{u}, \dots, D^l \hat{u})) \cdot (D^\alpha \hat{u}_n - D^\alpha \hat{u})\, dx \\ &\equiv I_n^1 + I_n^2 + I_n^3, \end{aligned}$$

where

$$I_n^1 \equiv \int_{\mathrm{spt}\hat{\varphi}} \sum_{|\alpha|\le l} A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n) \cdot D^\alpha [\hat{\varphi}\, (\hat{u}_n - \hat{u})]\, dx,$$

$$I_n^2 \equiv - \int_{\mathrm{spt}\hat{\varphi}} \hat{\varphi} \sum_{|\alpha|\le l} A_\alpha(x, \hat{u}, \dots, D^l \hat{u}) \cdot (D^\alpha \hat{u}_n - D^\alpha \hat{u})\, dx,$$

and

$$I_n^3 \equiv - \int_{\mathrm{spt}\hat{\varphi}} \sum_{|\alpha|\le l} A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n) \cdot [D^\alpha (\hat{\varphi}\, (\hat{u}_n - \hat{u})) - \hat{\varphi}\, (D^\alpha \hat{u}_n - D^\alpha \hat{u})]\, dx.$$

We claim that

$$I_n \to 0 \text{ as } n \to \infty. \tag{3.21}$$

Indeed, it follows from (3.2), (3.6), (3.7), (3.18), and (3.19) that

$$\begin{aligned} I_n^1 &= \int_{R^N \setminus \overline{\Omega}_n} \sum_{|\alpha|\le l} A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n) \cdot D^\alpha [\hat{\varphi}\, (\hat{u}_n - \hat{u})]\, dx \\ &= \int_{R^N \setminus \overline{\Omega}_n} f \cdot \hat{\varphi}\, (\hat{u}_n - \hat{u})\, dx = \int_{\mathrm{spt}\varphi} f \cdot \varphi (\hat{u}_n - \hat{u})\, dx \to 0 \ \text{ as } n \to \infty. \end{aligned} \tag{3.22}$$

Then by (H_1) and (3.7), we get that

$$\sum_{|\alpha|\le l} (\|A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n)\|_{L^q(R^N)} + \|A_\alpha(x, \hat{u}, \dots, D^l \hat{u})\|_{L^q(R^N)}) \le C, \tag{3.23}$$

which implies

$$I_n^2 \to 0 \ \text{ as } n \to \infty. \tag{3.24}$$

Now we deal with the term I_n^3. Since the term $D^\alpha(\hat{\varphi}\, (\hat{u}_n - \hat{u})) - \hat{\varphi}\, (D^\alpha \hat{u}_n - D^\alpha \hat{u})$ does not contain any derivative of order l of $\hat{u}_n - \hat{u}$ for any multi-index α with $|\alpha| \le l$, there exists a subsequence of $\{\hat{u}_n\}$, still denoted by itself, such that for all multi-indices α with $|\alpha| \le l$,

$$D^\alpha(\hat{\varphi}\, (\hat{u}_n - \hat{u})) - \hat{\varphi}\, (D^\alpha \hat{u}_n - D^\alpha \hat{u}) \to 0 \ \text{ strongly in } L^p(\mathrm{spt}\, \hat{\varphi}),$$

which, combined with (3.23), implies

$$I_n^3 \to 0 \ \text{ as } n \to \infty.$$

This together with (3.20), (3.22), and (3.24) proves (3.21).

By (H_2), (3.17), (3.19), and (3.21), we get that

$$\int_{\mathrm{spt}v} \sum_{|\alpha|\le l} |D^\alpha \hat{u}_n - D^\alpha \hat{u}|^p \, dx \le \int_{R^N} \hat{\varphi} \sum_{|\alpha|\le l} |D^\alpha \hat{u}_n - D^\alpha \hat{u}|^p \, dx \le C I_n \to 0.$$

Hence there exists a subsequence of $\{\hat{u}_n\}$, still denoted by itself, such that

$$D^\alpha \hat{u}_n \to D^\alpha \hat{u} \ \text{ a.e. in } \mathrm{spt}v \text{ for } |\alpha| \le l,$$

which, combined with (H_1), implies

$$A_\alpha(x, \hat{u}_n, \dots, D^l \hat{u}_n) \to A_\alpha(x, \hat{u}, \dots, D^l \hat{u}) \ \text{ a.e. in } \mathrm{spt}v \text{ for } |\alpha| \le l.$$

This together with (3.23) immediately shows (3.16).

Then by (3.16) we may pass to the limit for $n \to \infty$ in (3.13) to get (3.12).

Next we claim that

$$d \ge \int_{R^N \setminus \overline{\Omega}^*} J(x, \hat{u}, \dots, D^l \hat{u}) \, dx. \tag{3.25}$$

Let $R^N \setminus \overline{\Omega}^* = \cup_{j=1}^\infty G_j$, where $G_j, j = 1, \dots,$ are open and bounded subsets in $R^N \setminus \overline{\Omega}^*$ such that $\overline{G}_j \subset G_{j+1}$. By Lemma 2.3, for each j, there exists a positive integer n_j such that

$$\overline{G}_j \subset R^N \setminus \overline{\Omega}_n \text{ for all } n \ge n_j.$$

Thus

$$\underline{\lim}_{n\to\infty} \int_{R^N \setminus \overline{\Omega}_n} J(x, \hat{u}_n, \dots, D^l \hat{u}_n) \, dx \ge \underline{\lim}_{n\to\infty} \int_{G_j} J(x, \hat{u}_n, \dots, D^l \hat{u}_n) \, dx. \tag{3.26}$$

By the same arguments as above, we may prove that there exists a subsequence of $\{\hat{u}_n\}$, still denoted by itself, such that

$$J(x, \hat{u}_n, \dots, D^l \hat{u}_n) \to J(x, \hat{u}, \dots, D^l \hat{u}) \ \text{ a.e. in } G_j.$$

Then by (3.5) and (3.26) and using the Fatou lemma, we get that for each j,

$$d \ge \int_{G_j} J(x, \hat{u}, \dots, D^l \hat{u}) \, dx. \tag{3.27}$$

It is clear that

$$\lim_{j\to\infty} \chi_{G_j}(x) = \chi_{R^N \setminus \overline{\Omega}^*}(x), \tag{3.28}$$

where χ_{G_j} and $\chi_{R^N \setminus \overline{\Omega}^*}$ are the characteristic functions of G_j and $R^N \setminus \overline{\Omega}^*$, respectively.

Thus by (3.27) and (3.28) and using the Fatou lemma again, we get that

$$d \geq \underline{\lim}_{j\to\infty} \int_{G_j} J(x, \hat{u}, \ldots, D^l \hat{u})\, dx = \underline{\lim}_{j\to\infty} \int_{R^N \setminus \overline{\Omega}^*} \chi_{G_j} J(x, \hat{u}, \ldots, D^l \hat{u})\, dx$$
$$\geq \int_{R^N \setminus \overline{\Omega}^*} J(x, \hat{u}, \ldots, D^l \hat{u})\, dx.$$

This proves (3.25).

Finally, it follows from (3.9), (3.11), and (3.25) that Ω^* is a solution of problem (P), and this completes the proof. □

4. Application. In this section, we shall give some examples which are under the framework of problem (P).

Let $\Omega \subset R^N, N \geq 1$, be an open and bounded subset, let $\mathcal{O}$ be the family defined in section 2, and let $u_0 \in L^p(R^N)$ be a given observation function, where $p \geq 2$.

Example 4.1. Consider the following p-Laplace elliptic equation in exterior domain $R^N \setminus \overline{\Omega}$:

$$\begin{cases} -\mathrm{div}(|\nabla u|^{p-2}\nabla u) + k^2|u|^{p-2}u = f \quad \text{in } R^N \setminus \overline{\Omega}, \\ u \in W_0^{1,p}(R^N \setminus \overline{\Omega}), \end{cases} \tag{4.1}$$

where $k \neq 0$ is a constant and $f \in L^{\frac{p}{p-1}}(R^N)$ is a given function. By the definition of weak solution in section 3, $u_\Omega \in W_0^{1,p}(R^N \setminus \overline{\Omega})$ is a weak solution of (4.1) if

$$\int_{R^N \setminus \overline{\Omega}} |\nabla u_\Omega|^{p-2}\nabla u_\Omega \cdot \nabla v\, dx + k^2 \int_{R^N \setminus \overline{\Omega}} |u_\Omega|^{p-2}u_\Omega \cdot v\, dx = \int_{R^N \setminus \overline{\Omega}} f \cdot v\, dx \tag{4.2}$$

for all $v \in W_0^{1,p}(R^N \setminus \overline{\Omega})$.

Consider the following shape optimization problem:

$$\mathrm{Min}_{\Omega \in \mathcal{O}} \int_{R^N \setminus \overline{\Omega}} |u_\Omega - u_0|^p\, dx, \tag{P_2}$$

where u_Ω is the weak solution to (4.1). One can check that problem (P_2) is a special case of problem (P). As $p = 2$, in quantum mechanics, (4.1) is known as the Schrödinger equation and describes the steady state of a particle outside of a domain shaped by Ω (cf. [2]), and thus problem (P_2) may be understood to ask for the shape Ω from among the family $\mathcal{O}$ such that the corresponding steady state u of the particle distributed outside of Ω is "as close as possible" to a given state u_0.

Example 4.2. Consider the following semilinear elliptic equation in exterior domain $R^N \setminus \overline{\Omega}$:

$$\begin{cases} -\triangle u + g(u) = f \quad \text{in } R^N \setminus \overline{\Omega}, \\ u \in W_0^{1,2}(R^N \setminus \overline{\Omega}), \end{cases} \tag{4.3}$$

where $f \in L^2(R^N)$ is a given function and $g : R \to R$ satisfies the following hypotheses:

(i) $g : R \to R$ is continuous;

(ii) there are a positive constant c and a function $\mu \in L^2(R^N)$ such that for almost all $x \in R^N$ and all $y \in R$,

$$|g(y)| \leq c(|y| + \mu(x));$$

(iii) for all $y, z \in R, (g(y) - g(z))(y - z) \geq 0$;

(iv) there is a positive constant $\widetilde{c}$ such that for all $y \in R, g(y) \cdot y \geq \widetilde{c}\, y^2$.

By the definition of weak solution in section 3, $u_\Omega \in W_0^{1,2}(R^N \setminus \overline{\Omega})$ is a weak solution of (4.3) if

$$\int_{R^N \setminus \overline{\Omega}} \nabla u_\Omega \cdot \nabla v \, dx + \int_{R^N \setminus \overline{\Omega}} g(u_\Omega) \cdot v \, dx = \int_{R^N \setminus \overline{\Omega}} f \cdot v \, dx \tag{4.4}$$

for all $v \in W_0^{1,2}(R^N \setminus \overline{\Omega})$.

We consider the following shape optimization problem:

$$\text{Min}_{\Omega \in \mathcal{O}} \int_{R^N \setminus \overline{\Omega}} |u_\Omega - u_0|^2 \, dx, \tag{P_3}$$

where u_Ω is the weak solution of (4.3). It is clear that problem (P_3) is still under framework (P) after some simple modification.

Appendix.

LEMMA A. *Let $U \subset R^N, N \geq 1$, be an open subset. For any $x \in U$ and $y \in R^N \setminus \overline{U}$, there exists a constant t with $0 < t < 1$ such that $x + t(y - x) \in \partial U$.*

Proof. Since $x \in U$ and U is an open subset of R^N, there exists a constant $t_0 \in (0, 1)$ such that

$$x + t_0(y - x) \in U.$$

Let

$$t^* = \sup\{0 \leq t \leq 1 : \ x + t(y - x) \in U\}.$$

Then $t_0 \leq t^* \leq 1$.

Now we claim that

$$t^* \neq 1.$$

By contradiction, we assume that $t^* = 1$. Then there exists a sequence $\{t_n\}_{n=1}^{\infty}$ with $0 \leq t_n \leq 1$ and $x + t_n(y - x) \in U$ such that $t_n \to 1$ as $n \to \infty$. Thus $x + t_n(y - x) \to y \in \overline{U}$, which leads to a contradiction.

Next we claim that

$$x + t^*(y - x) \in \partial U.$$

By contradiction, we assume that $x + t^*(y - x) \notin \partial U$. Then

$$x + t^*(y - x) \in U$$

or

$$x + t^*(y - x) \in R^N \setminus \overline{U}.$$

If $x + t^*(y - x) \in U$, then there exists a positive constant $0 < \delta < 1 - t^*$ such that

$$x + (t^* + \delta)(y - x) \in U,$$

which contradicts the fact that $t^* = \sup\{0 \leq t \leq 1 : \ x + t(y - x) \in U\}$.

If $x + t^*(y - x) \in R^N \setminus \overline{U}$, then there exists a positive constant $0 < \widetilde{\delta} < t^*$ such that

$$x + t(y - x) \in R^N \setminus \overline{U} \text{ for all } \widetilde{\delta} \leq t \leq t^*,$$

which contradicts the fact that $t^* = \sup\{0 \leq t \leq 1 : x + t(y - x) \in U\}$.

This completes the proof. □

REFERENCES

[1] V. Barbu, *Nonlinear Semigroup and Differential Equations in Banach Spaces*, Noordhoff, The Netherlands, 1976.

[2] V. Barbu, *Partial Differential Equation and Boundary Value Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[3] D. Bucur and J. P. Zolesio, *N-dimensional shape optimization under capacitary constraints*, J. Differential Equations, 123 (1995), pp. 504–522.

[4] L. C. Berselli and P. Guasoni, *Some problems of shape optimization arising in stationary fluid motion*, Adv. Math. Sci. Appl., 14 (2004), pp. 279–293.

[5] D. Chenais, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–219.

[6] C. Dellacherie, *Analytical Sets, Capacities and Hausdorff Measures*, Lectures Notes in Appl. Math. 295, Springer-Verlag, Berlin, 1972 (in French).

[7] K. Kuratowski, *Introduction to Set Theory and Topology*, Pergamon Press, Oxford, UK, 1961.

[8] W. Liu and J. E. Rubio, *Local convergences and optimal shape design*, SIAM J. Control Optim., 30 (1992), pp. 49–62.

[9] Y. J. Li and H. Gao, *A domain optimization problem with boundary penalty cost functional*, J. Math. Anal. Appl., 301 (2005), pp. 170–186.

[10] P. Neittaanmaki and D. Tiba, *Existence and Approximation in Optimal Shape Design Problems*, Report 6/1999, University of Jyväskylä, Department of Mathematics Laboratory of Scientific Computing, Jyväskylä, Finland.

[11] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, 1984.

[12] R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory*, Press Syndicate of the University of Cambridge, Cambridge, UK, 1993.

[13] V. Sverak, *On optimal shape design*, J. Math. Pures Appl., 72 (1993), pp. 537–551.

[14] D. Tiba, *A property of Sobolev spaces and existence in optimal design*, Appl. Math. Optim., 47 (2003), pp. 45–58.

[15] X. Zhong, *On Nonhomogeneous Quasilinear Elliptic Equations*, Dissertation, Ann. Acad. Sci. Fenn. Math. Diss. 117, University of Jyväskylä, Jyväskylä, Finland, 1998.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 548–564

EXISTENCE OF REGULAR LAGRANGE MULTIPLIERS FOR A NONLINEAR ELLIPTIC OPTIMAL CONTROL PROBLEM WITH POINTWISE CONTROL-STATE CONSTRAINTS*

A. RÖSCH† AND F. TRÖLTZSCH‡

Abstract. A class of optimal control problems for semilinear elliptic equations with mixed control-state constraints is considered. The existence of bounded and measurable Lagrange multipliers is proven. As a particular application, the Lavrentiev-type regularization of pointwise state constraints is discussed. Here, the existence of associated regular multipliers is shown, too.

Key words. nonlinear programming, function space, optimal control, Lagrange multiplier, semilinear elliptic equation, mixed control-state constraint, pointwise state constraint, Lavrentiev-type regularization

AMS subject classifications. 49K20, 49N10, 49N15, 90C45

DOI. 10.1137/050625114

1. Introduction.

In this paper, we consider the optimal control problem

$$\min J(y,u) := \int_\Omega \psi(x,y(x),u(x))\,dx + \int_\Gamma \varphi(x,y(x))\,ds(x) \tag{1.1}$$

subject to the semilinear boundary value problem

$$\begin{aligned} (Ay)(x) + d(x,y(x)) &= \beta(x)u(x) && \text{in } \Omega,\\ \partial_\nu y(x) + b(x,y(x)) &= 0 && \text{on } \Gamma \end{aligned} \tag{1.2}$$

and to the mixed control-state constraints

$$\begin{aligned} 0 \le u(x) &\le c(x) + \gamma(x)y(x) && \text{a.e. in } \Omega,\\ u(x) &\le e(x) && \text{a.e. in } \Omega. \end{aligned} \tag{1.3}$$

We will investigate also another type of constraint that might be useful for applications. In this setting, A is a uniformly elliptic differential operator and Ω is a bounded domain of $\mathbb{R}^N$, $N \ge 2$, with boundary Γ and outward unit normal ν. Precise assumptions on and definitions of the quantities introduced above are formulated at the end of this section.

The investigation of optimal control problems of this type is interesting for different reasons. First of all, they exhibit nice theoretical properties. As we shall prove in this paper, the Lagrange multipliers associated with optimal solutions are quite regular functions, while they are measures for pure state constraints. This higher regularity allows for a better numerical analysis. For instance, second-order sufficient optimality conditions have been discussed in a fairly complete way [15], [14]. For pointwise state constraints, the associated theory is still partially open (cf. Casas et al. [7]) since it works only for a small dimension of Ω if the constraints are imposed

*Received by the editors February 23, 2005; accepted for publication (in revised form) December 17, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/sicon/45-2/62511.html

†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (arnd.roesch@oeaw.ac.at).

‡Technische Universität Berlin, Fakultät II – Mathematik und Naturwissenschaften, Str. des 17. Juni 136, D-10623 Berlin, Germany (troeltzsch@math.tu-berlin.de).

on the whole domain. Moreover, we mention the application of mixed constraints in the numerical analysis that is briefly addressed in section 7.

Constraints of this type are also interesting for problems in which the difference between control and state has to be bounded (problems of bottleneck type). An application to optimal heating has been given in [19].

Our main issue is the existence of bounded and measurable Lagrange multipliers associated with constraints (1.3). We shall consider the control u in $L^\infty(\Omega)$—this is the natural space obtained from the constraints. Therefore, the multipliers should belong to the dual space $L^\infty(\Omega)^*$, a space of Radon measures.

However, it is known from similar problems for linear and nonlinear parabolic problems that the Lagrange multipliers are more regular under natural assumptions [3], [4], [5], [18]. In fact, they can be constructed as functions of $L^2(\Omega)$ or even of $L^\infty(\Omega)$. Recently, the same result has been obtained for linear-quadratic elliptic problems by the convex duality theory [17]. The elliptic case is more difficult than the parabolic one due to the appearance of eigenvalues.

In principle, two ways are known for proving regularity. In [4], [5], the problems are linearized at the optimal solution and the Lagrange multipliers are obtained by the associated linear dual problem. Another technique was suggested later in [3]: An application of a Kuhn–Tucker theorem in Banach spaces delivers existence of Lagrange multipliers in $L^\infty(\Omega)^*$. In a second step, these multipliers are shown to be bounded functions. This step is fairly technical.

Here, we follow again the idea of linearization and application of duality theory. We think that this approach is more elementary, although not simple as well. Moreover, linearization and working with different norms and spaces seem to be interesting on their own. Therefore, we devote a section to this issue. Later, we shall embed problem (1.1)–(1.3) into a more general class of problems in function spaces and prove the regularity of Lagrange multipliers for this general class. The result for (1.1)–(1.3) is obtained as a conclusion.

Assumptions. $\Omega \subset \mathbb{R}^N$, $N \geq 2$, is a bounded Lipschitz domain. The functions $\psi = \psi(x, y, u) : \Omega \times \mathbb{R}^2 \to \mathbb{R}$, $d = d(x, y) : \Omega \times \mathbb{R} \to \mathbb{R}$, and $\varphi, b : \Gamma \times \mathbb{R} \to \mathbb{R}$ are measurable with respect to x for all fixed pairs (y, u) and differentiable with respect to y and u for all fixed x. They satisfy the following conditions on boundedness and Lipschitz continuity: There is a constant $K > 0$ and, for all $M > 0$, a constant $L(M) > 0$ exists such that

$$\left|\psi(x,0,0) + \frac{\partial\psi}{\partial y}(x,0,0) + \frac{\partial\psi}{\partial u}(x,0,0)\right| \leq K$$

$$|\psi(x,y_1,u_1) - \psi(x,y_2,u_2)| + \left|\frac{\partial\psi}{\partial y}(x,y_1,u_1) - \frac{\partial\psi}{\partial y}(x,y_2,u_2)\right|$$

$$+ \left|\frac{\partial\psi}{\partial u}(x,y_1,u_1) - \frac{\partial\psi}{\partial u}(x,y_2,u_2)\right| \leq L(M)\left\{|y_1 - y_2| + |u_1 - u_2|\right\}$$

for all u_i, y_i with $|u_i| + |y_i| \leq M$, $i = 1, 2$, and for a.a. $x \in \Omega$. The functions d, φ, b satisfy the same assumptions with respect to $x \in \Omega$ or $x \in \Gamma$.

The functions $\beta, \gamma, c, e : \Omega \to \mathbb{R}$ are bounded, measurable, and nonnegative. To avoid a trivial problem, we assume that $\beta \not\equiv 0$. A is the uniformly elliptic differential operator defined by

$$(Ay)(x) = -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}y(x)\right) + c_0(x)y(x)$$

with functions a_{ij} that belong to $C^{0,1}(\bar{\Omega})$, satisfy the condition $a_{ij}(x) = a_{ji}(x)$, and satisfy the condition of uniform ellipticity

$$\sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \geq \delta_0|\xi|^2 \qquad \text{a.e. on } \Omega$$

with some $\delta_0 > 0$ for all $\xi \in \mathbb{R}^N$. We assume $c_0(x) \geq 0$ a.e. on Ω and $\|c_0\|_{L^2(\Omega)} > 0$. It holds that $(\partial d/\partial y)(x, y) \geq 0$ and $(\partial b/\partial y)(x, y) \geq 0$ for all $y \in \mathbb{R}$ and almost all $x \in \Omega$ and Γ, respectively.

2. Lagrange multipliers and linearization. In (1.1)–(1.3), the state y is associated with u by a differentiable mapping G, $y = G(u)$, and thus y can be eliminated from the problem. Then $J(y, u) = J(G(u), u) = f(u)$, and the state y can also be eliminated in (1.3). Details are worked out in the next section. Finally, we arrive at a differentiable mathematical programming problem in Banach spaces that admits the following form:

$$\min f(u), \qquad g(u) \leq_K 0, \quad u \in C. \tag{2.1}$$

Here, $f : U \to \mathbb{R}$, $g : U \to Z$ are continuously Fréchet differentiable mappings defined in real Banach spaces U, Z. Moreover, $C \subset U$ is a convex set and $K \subset Z$ is a convex closed cone that defines a partial ordering $\leq_K$ in Z by $z \leq_K 0 \Leftrightarrow -z \in K$. In this sense, K can be considered as the nonnegative cone of Z, although the concrete meaning can be quite different.

Assume that $\bar{u}$ is a *local solution* of (2.1); i.e., for some $r > 0$ it holds that $f(\bar{u}) \leq f(u)$ for all $u \in U$ with $\|u - \bar{u}\|_U \leq r$, $g(u) \leq_K 0$, and $u \in C$. A Lagrange multiplier is an element of the dual cone K_+ of K,

$$K_+ = \{\mu \in Z^* : \mu(z) \geq 0 \quad \forall z \in K\},$$

where Z^* is the dual space of Z, the Banach space of all linear and continuous functionals on Z.

DEFINITION 2.1. *A* Lagrange multiplier *associated with $\bar{u}$ is an element $\mu \in K_+$ such that the* variational inequality

$$f'(\bar{u})(u - \bar{u}) + \mu(g'(\bar{u})(u - \bar{u})) \geq 0 \quad \forall u \in C \tag{2.2}$$

and the complementary slackness condition

$$\mu(g(\bar{u})) = 0 \tag{2.3}$$

hold true.

Remark. Equation (2.2) might be written as

$$\langle f'(\bar{u}) + g'(\bar{u})^*\mu, u - \bar{u}\rangle \geq 0 \quad \forall u \in C$$

with the adjoint operator $g'(\bar{u})^*$ and the pairing $\langle\cdot,\cdot\rangle$ between Z^* and Z. We avoid $g'(\bar{u})^*$ since it might be fairly complicated to find its concrete form in applications. Moreover, we aim at finding μ in a more "regular" subspace of Z^*.

In our applications, Z is of type $L^\infty(\Omega)$ and K is the associated nonnegative cone. It has a nonempty interior int K. Therefore, we are justified in requiring the following *constraint qualification* (regularity of $\bar{u}$):

There is a $u_0 \in C$ such that

$$-[g(\bar{u}) + g'(\bar{u})(u_0 - \bar{u})] \in \operatorname{int} K. \tag{2.4}$$

Under this assumption, a Lagrange multiplier $\mu \in Z^*$ exists, since (2.4) is equivalent to the well-known regularity condition of Zowe and Kurcyusz [23]; see also [13] and [20].

The condition of Zowe and Kurcyusz is sufficient for the existence of $\mu \in Z^*$ [23]. However, we shall apply another conclusion of (2.4) that was used in [23] as an auxiliary result: (2.4) permits us to view $\bar{u}$ as the solution of the linearized problem.

DEFINITION 2.2 (linearizing cone). *Let $\bar{z} \in K$ and $\bar{u} \in C$.*

$$C(\bar{u}) = \{v \in U : v = \lambda\,(u - \bar{u}),\ \lambda \geq 0,\ u \in C\},$$
$$K(\bar{z}) = \{w \in Z : w = \lambda\,(z - \bar{z}),\ \lambda \geq 0,\ z \in K\}.$$

The set

$$L(\bar{u}) = \{v \in U : v \in C(\bar{u}),\ g'(\bar{u})v \in -K(-g(\bar{u}))\}$$

is said to be the linearizing cone *of the feasible set at $\bar{u}$.*

LEMMA 2.3. *Under the regularity assumption* (2.4)*, it holds that*

$$f'(\bar{u})\,v \geq 0 \quad \forall v \in L(\bar{u}). \tag{2.5}$$

For this result of [23], we refer to [20, Theorem 1.2.1] and condition (2.2) of [20]. The reason for (2.5) is that (2.4) ensures $L(\bar{u}) \subset T(\bar{u})$, where $T(\bar{u})$ is the so-called tangent cone.

We rewrite (2.5) in a more convenient form. Obviously,

$$L(\bar{u}) = \{v \in U : v = \alpha\,(u - \bar{u}),\ g'(\bar{u})\,v + \beta\,g(\bar{u}) \leq_K 0,\ u \in C,\ \alpha \geq 0,\ \beta \geq 0\}.$$

Taking $\alpha = \beta = 1$, we obtain from (2.5)

$$f'(\bar{u})(u - \bar{u}) \geq 0 \tag{2.6}$$

for all $u \in C$ with

$$g(\bar{u}) + g'(\bar{u})(u - \bar{u}) \leq_K 0. \tag{2.7}$$

THEOREM 2.4. *If $\bar{u}$ is a local solution of problem* (2.1) *that satisfies the constraint qualification* (2.4)*, then $\bar{u}$ is a (global) solution of the* linearized programming problem

$$\min f'(\bar{u})\,u, \quad g(\bar{u}) + g'(\bar{u})(u - \bar{u}) \leq_K 0, \quad u \in C. \tag{2.8}$$

Proof. This follows from rewriting (2.6) in the form $f'(\bar{u})u \geq f'(\bar{u})\bar{u}$, together with (2.7). □

THEOREM 2.5. *Let the assumptions of Theorem* 2.4 *be satisfied. Then a Lagrange multiplier $\mu \in K_+$ associated with $\bar{u}$ exists. It is also a Lagrange multiplier for $\bar{u}$, considered as the solution of* (2.8)*. Conversely, let $\bar{u}$ solve* (2.8) *and let μ be any Lagrange multiplier for $\bar{u}$ in* (2.8)*. Then $\bar{u}$ and μ together satisfy the first-order necessary conditions for problem* (2.1)*.*

Proof. Define $\tilde{f}(u) = f'(\bar{u})u$, $\tilde{g}(u) = g(\bar{u}) + g'(\bar{u})(u - \bar{u})$. Then we have for all u,

$$\tilde{f}'(u) \equiv f'(\bar{u}), \quad \tilde{g}'(u) \equiv g'(\bar{u}), \quad \tilde{g}(\bar{u}) = g(\bar{u}).$$

The first statement of the theorem follows from [23], since the regularity condition (2.4) implies the one of Zowe and Kurcyusz. The converse is shown as follows: Let μ be any Lagrange multiplier associated with $\bar{u}$ as a solution of (2.8). By the associated complementarity condition and the relations above, we get $\mu(g(\bar{u})) = \mu(\tilde{g}(\bar{u})) = 0$. Moreover,

$$f'(\bar{u})(u-\bar{u}) + \mu(g'(\bar{u})(u-\bar{u})) = \tilde{f}'(\bar{u})(u-\bar{u}) + \mu(\tilde{g}'(\bar{u})(u-\bar{u})) \geq 0 \quad \forall u \in C.$$

Therefore, (2.2) and (2.3) are satisfied, and thus μ is a Lagrange multiplier for (2.1). □

In view of this, we may concentrate on the linearized problem (2.8) to find a Lagrange multiplier. Why is this helpful? There are two reasons: (2.8) can be considered in a space with coarser topology, where f and g are *not* differentiable. Moreover, we shall be able to use cones of nonnegative functions with empty interior such as $(L^2(\Omega))_+$.

3. Transformation and linearization of the control problem. In this section, we convert the elliptic optimal control problem (1.1)–(1.3) into a mathematical programming problem and consider its linearization, together with an associated dual problem.

THEOREM 3.1. *Under the general assumptions, for all $u \in L^p(\Omega)$, $p > N/2$, (1.2) has a unique solution $y = y(u) \in H^1(\Omega) \cap C(\bar{\Omega}) =: Y$. The associated control-to-state mapping $G : u \mapsto y$ is continuously Fréchet differentiable. Its derivative $G'(\bar{u})$ at $\bar{u} \in L^p(\Omega)$ is given by $G'(\bar{u})u = v$, where v solves the linearized equation*

$$\begin{aligned} (A\,v)(x) + d_y(x,\bar{y}(x))v(x) &= \beta(x)u(x) && \text{in } \Omega, \\ \partial_\nu v(x) + b_y(x,\bar{y}(x))v(x) &= 0 && \text{on } \Gamma, \end{aligned} \tag{3.1}$$

and $\bar{y} = G(\bar{u})$.

This result is standard. Existence of $y \in L^\infty(\Omega)$ follows for Lipschitz domains by a truncation technique and the Stampacchia method, as in Casas [6] (see also Alibert and Raymond [1]). In [6], the continuity of y was still shown for $C^{1,1}$-domains. In [1], continuity has been deduced from a result by Murthy and Stampacchia [12] that holds in Lipschitz domains. The reader might also consult the monograph [21], where these ideas are presented in detail in sections 4.1.3 and 7.1.2.

In what follows, we consider G and $G'(\bar{u})$ as operators with a range in $C(\bar{\Omega})$, although their range is actually contained also in $H^1(\Omega)$. Inserting $y = G(u)$ into J, the elliptic problem (1.1)–(1.3) is transformed into

$$\min f(u) := J(G(u), u) \tag{3.2}$$

subject to

$$\begin{aligned} 0 \;\leq\; u(x) &\leq c(x) + \gamma(x)(G(u))(x), \\ u(x) &\leq e(x) \end{aligned} \tag{3.3}$$

for almost all $x \in \Omega$. If we consider u as a function of $L^p(\Omega)$, then $0 \leq u(x) \leq e(x)$, and the assumption $e \in L^\infty(\Omega)$ guarantees $u \in L^\infty(\Omega)$. Therefore, from now on, we fix $U = L^\infty(\Omega)$, $Z = (L^\infty(\Omega))^2$, $K = (L^\infty(\Omega)_+)^2$, and we define $g : U \to Z$ by

$$(g(u))(x) = \begin{pmatrix} u(x) - c(x) - \gamma(x)(G(u))(x) \\ u(x) - e(x) \end{pmatrix}.$$

Moreover, we define $C = L^\infty(\Omega)_+$. Then (1.1)–(1.3) is equivalent to

$$\min f(u), \quad g(u) \leq_K 0, \quad u \in C. \tag{3.4}$$

Let $\bar{u}$ be a local solution of (3.4). "Local" is related to the topology of $L^\infty(\Omega)$. To satisfy the regularity condition, we must assume the existence of $u_0 \in L^\infty(\Omega)$ such that (2.4) is fulfilled. We require here a stronger condition and assume that $u_0 = 0$ satisfies (2.4). This means that $\varepsilon > 0$ exists such that

$$\begin{aligned} &\varepsilon \leq c(x) + \gamma(x)\bar{y}(x) - \gamma(x)(G'(\bar{u})\bar{u})(x), \\ &\varepsilon \leq e(x) \end{aligned} \tag{3.5}$$

is satisfied almost everywhere in Ω. This assumption appears to be natural in some sense: the smallest function—namely $u_0(x)$—should satisfy the upper bounds strictly. Since some monotonicity follows from the maximum principle for elliptic equations, this is reasonable. The next result is obtained as a direct conclusion of Theorem 2.4.

LEMMA 3.2. *If $\bar{u} \in L^\infty(\Omega)$ is a local solution of the optimal control problem* (1.1)–(1.3)*, and the regularity condition* (3.5) *is satisfied with some $\varepsilon > 0$, then $\bar{u}$ is also a solution of the linear optimal control problem*

$$\min f'(\bar{u})u$$

subject to $u \in L^\infty(\Omega)$ and

$$\begin{aligned} &u(x) \leq c(x) + \gamma(x)(G(\bar{u}))(x) + \gamma(x)(G'(\bar{u})(u - \bar{u}))(x), \\ &u(x) \leq e(x), \\ &u(x) \geq 0. \end{aligned} \tag{3.6}$$

Transforming this problem into an elliptic control problem, we arrive at

$$\begin{aligned} \min \Bigg\{ &\int_\Omega \psi_y(x, \bar{y}(x), \bar{u}(x))y(x)\,dx + \int_\Gamma \varphi_y(x, \bar{y}(x))y(x)\,ds(x) \\ &+ \int_\Omega \psi_u(x, \bar{y}(x), \bar{u}(x))u(x)\,dx \Bigg\} \end{aligned} \tag{3.7}$$

subject to the linearized equation (3.1) and

$$\begin{array}{rcccl} 0 & \leq & u(x) & \leq & c(x) + \gamma(x)\bar{y}(x) + \gamma(x)y(x) - \gamma(x)(G'(\bar{u})\bar{u})(x), \\ & & u(x) & \leq & e(x). \end{array} \tag{3.8}$$

We introduce the fixed part of the right-hand side by

$$\bar{c}(x) = c(x) + \gamma(x)\bar{y}(x) - \gamma(x)(G'(\bar{u})\bar{u})(x).$$

Moreover, we introduce an adjoint state $\hat{p} \in H^1(\Omega) \cap C(\bar{\Omega})$ by

$$\begin{array}{rcll} (A\hat{p})(x) + d_y(x, \bar{y}(x))\hat{p}(x) & = & \psi_y(x, \bar{y}(x), \bar{u}(x)) & \text{in } \Omega, \\ \partial_\nu \hat{p}(x) + b_y(x, \bar{y}(x))\hat{p}(x) & = & \varphi_y(x, \bar{y}(x)) & \text{on } \Gamma. \end{array} \tag{3.9}$$

Then it holds that

$$\int_\Omega \psi_y(x, \bar{y}(x), \bar{u}(x))y(x)\,dx + \int_\Gamma \varphi_y(x, \bar{y}(x))y(x)\,ds(x) = \int_\Omega \beta(x)\hat{p}(x)u(x)\,dx.$$

This well-known relation is easily obtained from the weak formulation of (3.1) and (3.9), respectively. With these notations, the linear programming problem (3.6) becomes equivalent to the *primal problem*

$$\min \int_\Omega a(x)u(x)\,dx \tag{3.10}$$

subject to

$$\begin{aligned} &u(x) \le \bar c(x) + (Su)(x),\\ &u(x) \le e(x),\\ &u(x) \ge 0 \end{aligned} \tag{3.11}$$

with

$$\begin{aligned} a &= \beta\,\hat p + \psi_u(\cdot, \bar y, \bar u),\\ Su &= \gamma\, G'(\bar u)\, u. \end{aligned}$$

By our assumptions, it holds that $a \in L^\infty(\Omega)$, and S is a linear continuous operator in $L^\infty(\Omega)$. The unknown function is $u \in L^\infty(\Omega)$. However, we also know that $\|Su\|_{L^2(\Omega)} \le c\|u\|_{L^2(\Omega)}$, and thus we can consider S as a linear continuous operator in $L^2(\Omega)$. Let us view S in this way and consider the linear programming problem (3.10)–(3.11) for $u \in L^2(\Omega)$ with all inequalities defined in $L^2(\Omega)$ as well. Notice that this does not change the feasible set, since $0 \le u \le e$ implies $u \in L^\infty(\Omega)$. Therefore, $\bar u$ solves the linear problem (3.11) in $L^2(\Omega)$.

4. The dual linear problem. Let us now establish the dual problem associated with the primal problem (3.10)–(3.11). In what follows, we denote by $(\cdot\,,\,\cdot)$ the inner product in $L^2(\Omega)$, and $S^* : L^2(\Omega) \to L^2(\Omega)$ denotes the adjoint operator to S. By Lagrange duality, the dual problem is obtained as follows: (3.10)–(3.11) is equivalent to

$$\min_{u\in L^2(\Omega)} \left\{ \max_{\mu\in(L^2(\Omega)_+)^3} \{(a,u) - (u,\mu_1) + (u - Su - \bar c, \mu_2) + (u - e, \mu_3)\} \right\},$$

where $\mu = (\mu_1, \mu_2, \mu_3)$. The dual problem is obtained by reversing the order of min and max,

$$\max_{\mu\in(L^2(\Omega)_+)^3} \left\{ (-\bar c, \mu_2) - (e, \mu_3) + \min_{u\in L^2(\Omega)} \{(a - \mu_1 + \mu_2 - S^*\mu_2 + \mu_3, u)\} \right\}.$$

The minimum is $-\infty$ if $0 \ne a - \mu_1 + \mu_2 - S^*\mu_2 + \mu_3$. This is meaningless for maximization, and hence we obtain the dual problem as

$$\left.\begin{aligned} &\max \quad -\int_\Omega (\bar c(x)\mu_2(x) + e(x)\mu_3(x))\,dx\\ &\text{subject to}\\ &\mu_2 + \mu_3 = -a + \mu_1 + S^*\mu_2\\ &\mu_i \ge 0, \quad i = 1,2,3. \end{aligned}\right\} \tag{4.1}$$

We show that (4.1) is solvable. Moreover, the problems (3.10)–(3.11) and (4.1) turn out to be in duality—they have the same optimal values. Preparing this discussion, we finish this section by computing S^*. This operator is defined by $(v\,,\,Su) = (S^*v\,,\,u)$

for all $u, v \in L^2(\Omega)$. We recall that $Su = \gamma G'(\bar{u})u = \gamma y$, where y is the solution of (3.1). Therefore,

$$(v\,,\,Su) = (v\,,\,\gamma y) = (\gamma v\,,\,y) = (\beta q\,,\,u),$$

where q is the solution of the adjoint equation

$$\begin{array}{rcll} (A\,q)(x) + d_y(x,\bar{y}(x))q(x) & = & \gamma(x)v(x) & \text{in } \Omega, \\ \partial_\nu q(x) + b_y(x,\bar{y}(x))q(x) & = & 0 & \text{on } \Gamma. \end{array} \tag{4.2}$$

This follows again in a standard way from the weak formulation of (3.1), (4.2). In view of this, S^* is given by

$$(S^*v)(x) = \beta(x)q(x), \tag{4.3}$$

where $q \in H^1(\Omega)$ is the solution of (4.2).

5. The duality relation. Here, we discuss the equality of the optimal values for the primal problem (3.10)–(3.11) and its dual (4.1).

THEOREM 5.1. *If $\bar{c}(x) \geq 0$ and $e(x) \geq 0$ hold a.e. in Ω, then the minimum of* (3.10)–(3.11) *is equal to the supremum of* (4.1).

Proof. We sketch the proof for convenience, since it is standard for continuous linear programming problems. We refer, for instance, to [8], [10], [22], [9], to the monograph [2] and—in the context of elliptic PDEs—to the recent paper [17].

(i) Define the convex cone E,

$$E = \{(\alpha, d_1, d_2) \in \mathbb{R} \times L^2(\Omega)^2 |\, \exists u \in L^2(\Omega), u \geq 0, u \leq d_1 + Su, u \leq d_2, (a,u) \leq \alpha\}.$$

We show that E is closed. In fact, if $(\alpha_n, d_{1,n}, d_{2,n}) \to (\alpha, d_1, d_2)$ in $\mathbb{R} \times L^2(\Omega)^2$ for $n \to \infty$, then there are $u_n \in L^2(\Omega)$ such that

$$0 \leq u_n \leq d_{2,n} \quad \forall n = 1, 2, \ldots .$$

Therefore, the sequence $(u_n)_{n=1}^\infty$ is bounded in $L^2(\Omega)$. By weak compactness, we can assume $u_n \rightharpoonup u$ in $L^2(\Omega)$. The weak continuity of S yields $Su_n \rightharpoonup Su$ in $L^2(\Omega)$. Finally, convexity ensures weak closedness, and hence it holds in the limit that

$$0 \leq u \leq d_2, \quad u \leq d_1 + Su, \quad (a,u) \leq \alpha.$$

This means that (α, d_1, d_2) is in E.

(ii) Let $\bar{\alpha}$ denote the minimum for (3.10)–(3.11). Clearly, the primal problem (3.10)–(3.11) has a solution; i.e., its infimum is attained. This follows from weak compactness, since $0 \leq u \leq e$ gives boundedness of the feasible set. By definition, $(\bar{\alpha} - \frac{1}{n}, \bar{c}, e) \notin E$ for all $n \in \mathbb{N}$. Therefore, the element $(\bar{\alpha} - \frac{1}{n}, \bar{c}, e)$ can be separated from the set E by a closed hyperplane in $\mathbb{R} \times L^2(\Omega)^2$: There are $(\beta_n, c_n^*, e_n^*) \in \mathbb{R} \times L^2(\Omega)^2$ such that $(\beta_n, c_n^*, e_n^*) \neq (0,0,0)$, and

$$\beta_n \left(\bar{\alpha} - \frac{1}{n}\right) + (\bar{c}, c_n^*) + (e, e_n^*) < \beta_n \alpha + (d_1, c_n^*) + (d_2, e_n^*) \tag{5.1}$$

holds for all $(\alpha, d_1, d_2) \in E$. Inserting $(\bar{\alpha}, \bar{c}, e)$ in the right-hand side, it follows that $\beta_n > 0$ and w.l.o.g. $\beta_n = 1$. We deduce that $\mu_{2,n} := c_n^*$ and $\mu_{3,n} := e_n^*$ satisfy the constraints

$$\begin{array}{c} \mu_2 + \mu_3 \geq -a + S^*\mu_2, \\ \mu_i \geq 0, \quad i = 2, 3. \end{array} \tag{5.2}$$

To confirm this insert, for arbitrary $u \geq 0$, $d_1 = u - Su$, $d_2 = u$, and $\alpha = (a\,,\, u)$ in the right-hand side of (5.1). Denote its (fixed) left-hand side by γ_n. Then

$$\gamma_n < (a + \mu_{2,n} + \mu_{3,n} - S^*\mu_{2,n}\,,\, u)$$

must hold for all $u \geq 0$. This implies $a + \mu_{2,n} + \mu_{3,n} - S^*\mu_{2,n} \geq 0$, i.e., the upper inequality of (5.2). Inserting $\alpha = 0$, $u = 0$, $d_2 = 0$, and arbitrary $d_1 \geq 0$, we obtain

$$\gamma_n < (d_1\,,\, \mu_{2,n}) \qquad \forall\, d_1 \geq 0,$$

and hence $\mu_{2,n} \geq 0$. In the same way, $\mu_{3,n} \geq 0$ is shown.

Setting $\mu_1 = \mu_2 + \mu_3 + a - S^*\mu_2$, (5.2) becomes equivalent to the constraints of (4.1). Moreover, inserting above $(\alpha, d_1, d_2) = (0, 0, 0) \in E$,

$$-\int_\Omega (\bar{c}(x)\mu_{2,n}(x) + e(x)\mu_{3,n}(x))\,dx > \bar{\alpha} - \frac{1}{n}$$

is obtained for every $n \in \mathbb{N}$. It is true in general that the dual supremum $\bar{\sigma}$ is less than or equal to the primal minimum $\bar{\alpha}$ (cf. (6.13)), and hence

$$\bar{\alpha} \geq \bar{\sigma} > \bar{\alpha} - \frac{1}{n}.$$

Now, $n \to \infty$ yields $\bar{\alpha} = \bar{\sigma}$, the statement of the theorem. □

Remark. The main point in the proof was the closedness of E. It was deduced from the boundedness of the sets $\{u \in L^2(\Omega) |\, 0 \leq u \leq d_{2,n}\}$.

6. Solvability of the dual problem. To prove existence for the dual problem, we rely on the nonnegativity and certain smoothing properties of S and S^*.

LEMMA 6.1. *S and S^* are nonnegative, i.e.,*

$$u \geq 0 \;\Rightarrow\; Su \geq 0 \quad \textit{a.e. in } \Omega.$$

This follows from $G'(\bar{u}) \geq 0$, a consequence of comparison principles for linear elliptic equations: For $u \geq 0$, we consider v^-, the negative part of the solution $v = G'(\bar{u})u$ of (3.1). By using v^- as a test function in (3.1), we see that v^- vanishes. Here, $\beta \geq 0$, $\gamma \geq 0$ are used.

By their definitions, S and S^* are linear and continuous operators in $L^2(\Omega)$. We shall estimate them in different L^p-norms for $1 \leq p \leq \infty$. To match their domain $L^2(\Omega)$, we have to consider the intersection of $L^p(\Omega)$ with $L^2(\Omega)$ in the estimates below. For instance, we know from Theorem 3.1 that

$$\|Su\|_{L^\infty(\Omega)} \leq c_S\, \|u\|_{L^p(\Omega)} \qquad \forall u \in L^p(\Omega) \cap L^2(\Omega), \quad p \geq N/2.$$

LEMMA 6.2. *With constants c_S, $\delta > 0$, the following estimates are fulfilled:*

$$\|S^*v\|_{L^\infty(\Omega)} \leq c_S\, \|v\|_{L^p(\Omega)} \qquad \forall v \in L^p(\Omega) \cap L^2(\Omega), \quad p > N/2. \tag{6.1}$$

$$\|S^*v\|_{L^{p+\delta}(\Omega)} \leq c_S\, \|v\|_{L^p(\Omega)} \qquad \forall v \in L^p(\Omega) \cap L^2(\Omega), \quad 1 \leq p \leq N/2. \tag{6.2}$$

Proof. (i) *First estimate.* S^*v is given by formula (4.3), where q is obtained as the solution of the adjoint elliptic equation (4.2). The mapping $v \mapsto q$ is continuous from $L^p(\Omega)$ to $L^\infty(\Omega)$. This follows from Theorem 3.1 with γ substituted for β. Therefore, the estimate $\|S^*v\|_{L^\infty(\Omega)} = \|\beta\, q\|_{L^\infty(\Omega)} \leq c_S\, \|v\|_{L^p(\Omega)}$ follows immediately.

(ii) *Some continuity properties.* To avoid confusion of operators and their associated spaces, let us first mention that the solution mapping $u \mapsto v$ for (3.1) is linear and continuous from $(H^1(\Omega))^*$ to $H^1(\Omega)$. We denote this operator by $\mathbf{S}$. In this sense, S is the restriction of $\mathbf{S}$ to $L^2(\Omega)$ considered with a range in $L^2(\Omega)$. Define, for $p > N/2$, $S_p : L^p(\Omega) \to L^\infty(\Omega)$ as the restriction of $\mathbf{S}$ to $L^p(\Omega)$. By Theorem 3.1, this operator is continuous. For arbitrary $v \in L^1(\Omega)$, the mapping

$$u \mapsto \int_\Omega v\,(S_p\,u)\,dx$$

defines a functional $\varphi \in L^p(\Omega)^* = L^{p'}(\Omega)$. The mapping $v \mapsto \varphi$ is obviously continuous from $L^1(\Omega)$ to $L^{p'}(\Omega)$. We denote this mapping by $S_p^\otimes$. It is the restriction of S_p^* to $L^1(\Omega)$, and hence we have

$$\|S_p^\otimes v\|_{L^{p'}(\Omega)} \le c_1\,\|v\|_{L^1(\Omega)} \qquad \forall v \in L^1(\Omega).$$

This estimate holds in particular for all $v \in L^2(\Omega)$. For such functions we have $S_p^\otimes v = S^* v$. In fact, we find for all $u \in L^\infty(\Omega)$ and $v \in L^2(\Omega)$,

$$\int_\Omega (S_p^\otimes v)\,u\,dx = \int_\Omega v\,(S_p\,u)\,dx = (v, S_p u) = (v, Su) = (S^* v, u).$$

Since this is true for all $u \in L^\infty(\Omega)$, it must hold that $S_p^\otimes v = S^* v$ for all $v \in L^2(\Omega)$. Therefore, the boundedness of $S_p^\otimes$ extends to S^*:

$$\|S^* v\|_{L^{p'}(\Omega)} \le c_1 \|v\|_{L^1(\Omega)} \qquad \forall v \in L^2(\Omega).$$

This shows that S^* can be extended to a continuous linear operator from $L^1(\Omega)$ to $L^{p'}(\Omega)$. In fact, it is known from Alibert and Raymond [1] and Casas [6] that the solution operator of the adjoint equation (4.2) is linear and continuous from $L^1(\Omega)$ to $W^{1,\sigma}(\Omega)$ for all $\sigma < N/(N-1)$. However, we do not rely on this deep result. We just use the available following information.

We know in particular that S^* is continuous from $L^{p_0}(\Omega)$ to $L^{p_1}(\Omega)$ and from $L^{q_0}(\Omega)$ to $L^{q_1}(\Omega)$ for $p_0 = N$, $p_1 = 2N$, $q_0 = 1$, $q_1 = p' > 1$, where p' is associated with the arbitrary but fixed $p \in (N/2, \infty)$ chosen at the beginning. It is now fairly clear that the second estimate can be obtained by interpolation.

(iii) *Second estimate.* Define, for $\theta \in [0,1]$, $r_j = r_j(\theta)$ by

$$\frac{1}{r_j} = \frac{1-\theta}{p_j} + \frac{\theta}{q_j}, \quad j = 0, 1.$$

Then the classical interpolation theorem in Triebel [16, §1.18.7, Thm. 1] ensures that S^* is continuous from $L^{r_0}(\Omega)$ to $L^{r_1}(\Omega)$. From now on, we need only simple calculations. Inserting the concrete values for p_j, q_j, we find

$$\frac{1}{r_0(\theta)} = \frac{1-\theta}{N} + \theta, \qquad \frac{1}{r_1(\theta)} = \frac{1-\theta}{2N} + \frac{\theta}{p'}.$$

Therefore,

$$\frac{1}{r_0(\theta)} - \frac{1}{r_1(\theta)} = \frac{1-\theta}{N} + \theta - \frac{1-\theta}{2N} - \frac{\theta}{p'} = \frac{1}{2N} + \theta\left(1 - \frac{1}{2N} - \frac{1}{p'}\right).$$

The difference $1/r_0(\theta) - 1/r_1(\theta)$ is a linear function of θ. It admits the values $1/2N$ and $1 - 1/p' > 0$ for $\theta = 0$ and $\theta = 1$, respectively, and thus

$$\frac{1}{r_0(\theta)} - \frac{1}{r_1(\theta)} \geq \delta > 0$$

holds with $\delta = \min(1/2N, 1 - 1/p')$. After multiplication with $r_0 r_1 > 1$, we get $r_1(\theta) - r_0(\theta) \geq \delta r_0 r_1 \geq \delta$ for all $\theta \in [0,1]$, and hence $r_1 \geq r_0 + \delta$. Therefore, the second estimate holds for all $r_0 = p \in [1, N]$, in particular for $p \in [1, N/2]$. □

THEOREM 6.3. *If the regularity condition* (3.5) *is satisfied, then the dual problem* (4.1) *has a solution* $\bar\mu \in (L^\infty(\Omega))^3$.

Proof. (i) *Boundedness in* L^1. We know that the supremum $\bar\alpha$ in (4.1) is finite. Let $(\mu_{i,n})_{n=1}^\infty \in L^2(\Omega)$, $i = 1,2,3$, be sequences such that the constraints of (4.1) are satisfied and

$$-(\bar c\,,\, \mu_{2,n}) - (e\,,\, \mu_{3,n}) \to \bar\alpha$$

for $n \to \infty$. The regularity condition (3.5) gives $\bar c \geq \varepsilon$ and $e \geq \varepsilon$ a.e. on Ω, and hence

$$-(\bar c, \mu_{2,n}) - (e, \mu_{3,n}) \leq -\varepsilon\,(\|\mu_{2,n}\|_{L^1(\Omega)} + \|\mu_{3,n}\|_{L^1(\Omega)}).$$

Therefore, the L^1-norm of $\mu_{2,n}$ and $\mu_{3,n}$ is bounded. The boundedness of $\mu_{1,n}$ in $L^1(\Omega)$ can now be obtained from the constraints of (4.1).

(ii) *Boundedness in* $L^{1+\delta}$. The constraints of (4.1) are equivalent to the constraints (5.2), i.e., $\mu_2 \geq 0$, $\mu_3 \geq 0$, and

$$\mu_2 + \mu_3 \geq -a + S^*\mu_2.$$

In view of this, we have

$$\mu_{2,n} + \mu_{3,n} \geq -a + S^*\mu_{2,n}. \tag{6.3}$$

By Lemma 6.2 and the L^1-boundedness of the sequence $\mu_{2,n}$, the sequence $(S^*\mu_{2,n})$ is bounded in the reflexive Banach space $L^{1+\delta}(\Omega)$, and hence the right-hand side of (6.3) is bounded in $L^{1+\delta}(\Omega)$. Now, we redefine $\mu_{2,n}$, $\mu_{3,n}$ as follows: We take

(6.4)
$$\hat\mu_{2,n}(x) + \hat\mu_{3,n}(x) = \begin{cases} 0, & \text{where } -a(x) + (S^*\mu_{2,n})(x) \leq 0, \\ -a(x) + (S^*\mu_{2,n})(x), & \text{where } -a(x) + (S^*\mu_{2,n}(x)) > 0 \end{cases}$$

and require $0 \leq \hat\mu_{2,n} \leq \mu_{2,n}$ and $0 \leq \hat\mu_{3,n} \leq \mu_{3,n}$. Certainly, this is possible. Then,

$$|\hat\mu_{2,n}(x)| + |\hat\mu_{3,n}(x)| \leq |a| + |(S^*\mu_{2,n})(x)| \text{ a.e. on } \Omega,$$

and hence these new sequences are bounded in $L^{1+\delta}(\Omega)$. Moreover, $\hat\mu_{2,n}, \hat\mu_{3,n} \geq 0$ and $\hat\mu_{2,n} \leq \mu_{2,n}$, together with $S^* \geq 0$, show

$$\hat\mu_{2,n} + \hat\mu_{3,n} \geq -a + S^*\mu_{2,n} \geq -a + S^*\hat\mu_{2,n},$$

and thus the new functions are feasible. They have an objective value greater than or equal to the pair $\mu_{2,n}, \mu_{3,n}$ since $\bar c \geq 0$ and $e \geq 0$. In this way, we have found a sequence with a better objective value that is bounded in the reflexive Banach space $L^{1+\delta}(\Omega)$. By selecting weakly convergent subsequences we can assume that

$$\hat\mu_{2,n} \rightharpoonup \mu_2, \quad \hat\mu_{3,n} \rightharpoonup \mu_3, \quad n \to \infty$$

hold in $L^{1+\delta}(\Omega)$, where $\mu_2, \mu_3 \in L^{1+\delta}(\Omega)$ satisfy all constraints of the dual problem. Moreover, the functions are optimal, and

$$\mu_1 := \mu_2 + \mu_3 + a - S^*\mu_2$$

is the associated optimal μ_1.

(iii) *Solution in* $(L^\infty(\Omega))^3$. We have found an optimal triplet in $(L^{1+\delta}(\Omega))^3$ that satisfies the nonnegativity constraints and

$$\mu_2 + \mu_3 \geq -a + S^*\mu_2. \tag{6.5}$$

By the smoothing property (6.2) of S^*, the function $S^*\mu_2$ belongs to $L^{1+2\delta}(\Omega)$. Applying the technique of (ii), we find $\hat{\mu}_2 \geq 0$, $\hat{\mu}_3 \geq 0$, which satisfy all constraints. Moreover, they belong to $L^{1+2\delta}(\Omega)$ and have the same objective value (clearly, the value cannot be improved further). Applying this bootstrapping technique, we arrive after finitely many steps at the case when $1 + k\delta > N/2$. In the next step, $S^* : L^{1+k\delta}(\Omega) \to L^\infty(\Omega)$ is used to obtain $\hat{\mu}_i \in L^\infty(\Omega)$, $i = 1, 2, 3$. □

THEOREM 6.4. *Suppose that $\bar{u}$ is a local solution of the optimal control problem* (1.1)–(1.3) *that satisfies the regularity condition* (3.5). *Then there exist Lagrange multipliers $\mu_i \in L^\infty(\Omega)$, $i = 1, 2, 3$, satisfying with $\bar{u}$, $\bar{y}$, $\bar{p}$ the optimality system that consists of constraints* (1.2)–(1.3), *the adjoint equation*

$$\begin{array}{rcll} (Ap)(x) + d_y(x, \bar{y}(x))p(x) & = & \psi_y(x, \bar{y}(x), \bar{u}(x)) - \gamma(x)\mu_2(x) & \text{in } \Omega, \\ \partial_\nu p(x) + b_y(x, \bar{y}(x))p(x) & = & \varphi_y(x, \bar{y}(x)) & \text{on } \Gamma, \end{array} \tag{6.6}$$

and, for almost all $x \in \Omega$, the conditions

$$\psi_u(x, \bar{y}(x), \bar{u}(x)) + \beta(x)p(x) + \mu_3(x) + \mu_2(x) - \mu_1(x) = 0, \tag{6.7}$$

$$\mu_i(x) \geq 0, \tag{6.8}$$

$$(\bar{u}(x) - e(x))\mu_3(x) = 0, \tag{6.9}$$

$$(\bar{u}(x) - \gamma(x)\bar{y}(x) - c(x))\mu_2(x) = 0, \tag{6.10}$$

$$\bar{u}(x)\mu_1(x) = 0. \tag{6.11}$$

Proof. We consider the optimal solutions $\mu_i \in L^\infty(\Omega)$ of the dual problem (4.1) and show that they fulfill the required properties. First, we mention

$$\begin{aligned} (a\,,\,\bar{u}) &\geq (a\,,\,\bar{u}) + (\bar{u} - \bar{c} - S\bar{u}\,,\,\mu_2) + (\bar{u} - e\,,\,\mu_3) \\ &= (a + \mu_2 + \mu_3 - S^*\mu_2\,,\,\bar{u}) - (\bar{c}\,,\,\mu_2) - (e\,,\,\mu_3) \end{aligned} \tag{6.12}$$

$$\geq -(\bar{c}\,,\,\mu_2) - (e\,,\,\mu_3), \tag{6.13}$$

since $\bar{u}, \mu_2, \mu_3 \geq 0$ and the inequalities (3.11), (6.5) hold true. From the duality relation of Theorem 5.1,

$$(a\,,\,\bar{u}) = -(\bar{c}\,,\,\mu_2) - (e\,,\,\mu_3),$$

we deduce that the inequalities (6.12), (6.13) must hold as equations, and hence

$$0 = (\bar{u} - \bar{c} - S\bar{u}\,,\,\mu_2) = (\bar{u} - e\,,\,\mu_3) = (a + \mu_2 + \mu_3 - S^*\mu_2\,,\,\bar{u}). \tag{6.14}$$

By definition, $\bar{c} = c + \gamma\,\bar{y} - S\bar{u}$, and hence (6.14) yields

$$\int_\Omega (\bar{u} - c - \gamma\,\bar{y})\mu_2\,dx = \int_\Omega (\bar{u} - e)\mu_3\,dx = 0.$$

Thanks to $\bar u - c - \gamma\bar y \le 0$, $\bar u \le e$, and to the nonnegativity of μ_2, μ_3, this is equivalent to (6.9)–(6.10). Inserting $\mu_1 = a + \mu_2 + \mu_3 - S^*\mu_2$, we obtain from (6.14)

$$\int_\Omega \bar u\, \mu_1\, dx = 0.$$

The nonnegativity of $\bar u$ and μ_1 implies (6.11). By definition, it holds that

$$a + \mu_2 + \mu_3 - \mu_1 - S^*\mu_2 = 0. \tag{6.15}$$

The representation (4.3) for S^* permits us to write

$$S^*\mu_2 = \beta\, q, \tag{6.16}$$

where q solves the adjoint equation (4.2) with right-hand side $\gamma\mu_2$. Moreover,

$$a = \psi_u(\bar y, \bar u) + \beta\, \hat p, \tag{6.17}$$

where $\hat p$ is the solution of (3.9). Obviously, (6.15)–(6.17) are equivalent to (6.6)–(6.7). The complete system of necessary conditions is shown. □

7. An application to pointwise state constraints. Consider the optimal control problem (1.1)–(1.2) subject to the pointwise constraints

$$0 \le u(x), \tag{7.1}$$

$$0 \le c(x) + y(x). \tag{7.2}$$

We know $y \in C(\bar\Omega)$, and hence a Lagrange multiplier μ_2 associated with (7.2) is in general a regular Borel measure, provided that $c \in C(\bar\Omega)$ (cf. Casas [6]). One way to deal with (7.2) numerically by the Lavrentiev-type regularization

$$-\lambda\, u(x) \le c(x) + y(x) \tag{7.3}$$

with $\lambda > 0$ being small [11].

For the results of this section, we assume that the control problem (1.1), (1.2), (7.1), (7.3) has a locally optimal solution $\bar u \in L^\infty(\Omega)$. Notice that this boundedness does not follow from the constraints (7.1), (7.3).

We do not consider here the pass to the limit $\lambda \downarrow 0$. However, we are able to show that, for $\lambda > 0$ fixed, the associated Lagrange multiplier μ_2 can be taken as a function from $L^\infty(\Omega)$ again. For this result, the theory of the preceding sections needs a few modifications. To simplify the notation, we write $\mu := \mu_2$ below. Recall that now $\gamma(x) = 1$, $S = G'(\bar u)$ and therefore $\bar c = c + \bar y - S\,\bar u$.

The state constraints in (7.2) are meaningful only if $c \not\ge 0$, since otherwise $u \ge 0$ would imply $c + y \ge 0$. Therefore, now we cannot assume $\bar c \ge 0$ if we are interested in taking λ small (consider, for instance, the linear case, where $\bar y = S\bar u$, and hence $c = \bar c$).

Our approach is based on linearization and needs the regularity condition (2.4). For the case of the constraints (7.1), (7.3), this condition amounts to the existence of $u_0 \in L^\infty(\Omega)$ and $\varepsilon > 0$ such that

$$\varepsilon \le u_0(x), \quad \varepsilon - \lambda\, u_0(x) \le \bar c(x) + S\, u_0(x) \tag{7.4}$$

for almost all $x \in \Omega$. It is easy to see that this condition is satisfied for the function $u_0(x) \equiv c_0$ if c_0 is taken sufficiently large. Indeed, for $c_0 \to \infty$, the term $-\lambda\, c_0$ in (7.4)

tends to $-\infty$, while $S\,c_0$ remains nonnegative. Moreover, $\bar{c}$ is bounded. Therefore, (7.4) is satisfied for sufficiently large c_0, and we are justified to linearize at any $\bar{u}$.

The associated linear programming problem for a local solution of the problem (1.1), (1.2), (7.2), (7.3) that replaces (3.10)–(3.11) is now

$$\begin{aligned} &\min \quad (a\,,\,u),\\ &-\lambda\,u \le \bar{c} + Su,\\ &\quad u \ge 0,\ u \in L^\infty(\Omega). \end{aligned} \tag{7.5}$$

We know that $\bar{u}$ is a solution of that problem, since $\bar{u}$ is supposed to be a local solution of the optimal control problem in $L^\infty(\Omega)$ and the regularity condition is satisfied. However, this information is related to the space $L^\infty(\Omega)$ that is not suitable for obtaining regular Lagrange multipliers. Therefore, we show that the optimal value of problem (7.5) does not increase if the feasible set is extended from $L^\infty(\Omega)$ to $L^p(\Omega)$ with $p > N/2$.

LEMMA 7.1. *The optimal value of problem* (7.5) *does not change if its feasible set is extended to all* $u \in L^p(\Omega)$, $p > \max\{2, N/2\}$, *satisfying the associated constraints.*

Proof. Let an arbitrary $u \in L^p(\Omega)$ be given such that $u \ge 0$ and

$$-\lambda\,u \le \bar{c} + S\,u \quad \text{a.e. in } \Omega. \tag{7.6}$$

We construct a sequence of functions $u_n \in L^\infty(\Omega)$ such that u_n satisfies (7.6) and $u_n \to u$ in $L^p(\Omega)$ as $n \to \infty$. This implies the statement of the lemma.

By the smoothing property (6.2) of S, we know that $\bar{c} + S\,u \in C(\bar{\Omega})$. Define

$$K = \frac{1}{\lambda}\|\bar{c} + S\,u\|_{L^\infty(\Omega)}$$

and introduce the "cut-off" function

$$v_n(x) = \begin{cases} n, & u(x) > n,\\ u(x), & u(x) \le n. \end{cases}$$

This function is bounded, measurable, and nonnegative. It holds that $v_n \to u$ in $L^p(\Omega)$ as $n \to \infty$. For all $n \ge K$ we have

$$-\lambda\,v_n \le \bar{c} + S\,u \quad \text{a.e. in } \Omega.$$

This inequality is trivial for all x with $u(x) \le n$, since $v_n(x) = u(x)$ is satisfied there. In the remaining points, it holds that

$$-\lambda\,v_n(x) = -\lambda\,n \le -\lambda\,K = -\|\bar{c} + S\,u\|_{L^\infty(\Omega)} \le \bar{c}(x) + (S\,u)(x).$$

In view of $v_n \to u$, the continuity of S from $L^p(\Omega)$ to $C(\bar{\Omega})$ yields

$$-\lambda\,v_n \le \bar{c} + S\,v_n + \varepsilon_n$$

with $\varepsilon_n \downarrow 0$ as $n \to \infty$. For sufficiently large constant $c_0 > 0$, we have

$$-\lambda\,c_0 \le \bar{c} + S\,c_0 - 1.$$

For all $t \in [0,1]$, the convex combination $u_n := (1-t)\,v_n + t\,c_0$ is nonnegative and satisfies

$$-\lambda u_n \le \bar{c} + S\,u_n + (1-t)\,\varepsilon_n - t.$$

We take $t = t_n = \varepsilon_n/(1+\varepsilon_n)$. Then $(1-t_n)\,\varepsilon_n - t_n = \varepsilon_n - t_n\,(1+\varepsilon_n) = 0$, thus

$$-\lambda u_n \le \bar{c} + S\,u_n,$$

and hence u_n fulfills the linearized constraints. Moreover, we have $u_n \to u$, $n \to \infty$. □

Lemma 7.1 ensures that $\bar{u}$ is also a solution of the extended primal problem

$$\begin{aligned} &\min \quad (a\,,\,u),\\ &-\lambda\,u \le \bar{c} + Su,\\ &\quad u \ge 0,\ u \in L^p(\Omega). \end{aligned} \tag{7.7}$$

The dual problem to (7.7) is

$$\begin{aligned} &\max \quad (-\bar{c}\,,\,\mu),\\ &\quad a \ge \lambda\,\mu + S^*\mu,\\ &\quad \mu \ge 0,\ \mu \in L^{p'}(\Omega), \end{aligned} \tag{7.8}$$

where a function $\mu \in L^{p'}(\Omega)$ is to be found. We have already discussed that $S_p^{\otimes}$ maps $L^1(\Omega)$ to $L^{p'}(\Omega)$, and hence S^* can be extended to a continuous operator in $L^{p'}(\Omega)$. Moreover, S^* is nonnegative. We shall see that the constraints above imply $\mu \in L^\infty(\Omega)$, and hence $\mu \in L^2(\Omega)$ so that the notation S^* in (7.8) is justified.

We show that both problems admit the same optimal values and that (7.8) has a solution.

LEMMA 7.2. *The function a is nonnegative a.e. on Ω.*

Proof. As the regularity condition is always satisfied here, Theorem 2.4 shows that $\bar{u}$ solves the linearized problem (7.5). Moreover, Lemma 7.1 ensures that $\bar{u}$ is a solution of the extended problem (7.7), too. Therefore, (7.7) must have at least the optimal solution $\bar{u}$. For all $u \ge 0$, $\bar{u} + u$ satisfies the constraints of (7.5): obviously, we have $\bar{u} + u \ge 0$. Moreover,

$$-\bar{c} \le \lambda\,\bar{u} + S\bar{u} \le \lambda(\bar{u}+u) + S(\bar{u}+u)$$

follows from $\lambda \ge 0$ and $S \ge 0$. Assume that $a(x) < 0$ on $M \subset \Omega$, where M has a positive measure. Take

$$u(x) = \begin{cases} 1 & \text{on } M,\\ 0 & \text{on } \Omega \setminus M. \end{cases}$$

Then $\bar{u} + u$ is feasible, but $(a\,,\,\bar{u}+u) < (a\,,\,\bar{u})$ would hold in contradiction to the optimality of $\bar{u}$. □

THEOREM 7.3. *The dual problem* (7.8) *admits a solution. Moreover, the maximum of the primal problem* (7.5) *is equal to the minimum of* (7.8). *The solution of* (7.8) *is bounded and measurable and is a Lagrange multiplier associated with* (7.2).

Proof. (i) *Solvability of* (7.8). We know from Lemma 7.2 that a is nonnegative. Therefore, $\mu = 0$ is feasible for (7.8). The feasible set is bounded in $L^\infty(\Omega)$, since $\mu \ge 0$ implies $S^*\mu \ge 0$, and hence

$$a \ge \lambda\,\mu + S^*\mu \ge \lambda\mu \ge 0.$$

Now the existence of an optimal solution of (7.8) follows by weak*-compactness.

(ii) *Equality of optimal values.* Rewrite (7.8) in the form of a primal problem, namely,

$$\begin{aligned} &\min \quad (\bar{c}\,,\,\mu),\\ &\lambda\,\mu + S^*\mu \le a,\\ &\mu \ge 0. \end{aligned} \tag{7.9}$$

Its dual admits the form

$$\begin{aligned} &\max \quad (-a\,,\,u),\\ &\lambda\,u + Su \ge -\bar{c},\\ &u \ge 0. \end{aligned} \tag{7.10}$$

(If B is a linear continuous operator, then the general rule for establishing dual problems is as follows: If the primal problem is to minimize $(c\,,\,u)$ subject to $B\,u \le a$ and $u \ge 0$, then its dual is to maximize $(-a\,,\,\mu)$ subject to $B^*\,\mu \ge -c$ and $\mu \ge 0$.)

The optimal values of (7.9) and (7.10) are equal. To show this, we follow the same steps as in the proof of Theorem 5.1. We introduce the cone

$$E := \{(\alpha, d) \in \mathbb{R} \times L^{p'}(\Omega) |\, \exists \mu \ge 0 :\ \lambda\mu + S^*\mu \le d,\ (\bar{c}, \mu) \le \alpha\}.$$

Again, E is closed (boundedness of a sequence $\{\alpha_n, d_n\}$ in $\mathbb{R} \times L^{p'}(\Omega)$ implies that the associated sequence $\{\mu_n\}$ is bounded). Moreover, the element $(\bar{\alpha} - \frac{1}{k}, \bar{c})$ does not belong to E for all $k \in \mathbb{N}$. This element can be separated from E by a closed hyperplane. Continuing as in the proof of Theorem 5.1 (here, the term e does not occur), the equality of the optimal values of (7.9) and (7.10) is obtained. By changing the signs, the same holds for problems (7.8) and (7.7). Lemma 7.1 completes the proof. It follows as in section 6 that the solution $\mu \in L^\infty(\Omega)$ of (7.8) is a Lagrange multiplier associated with the mixed pointwise control-state constraint (7.2). ∎

REFERENCES

[1] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.

[2] E. J. Anderson, *Linear Programming in Infinite-Dimensional Spaces. Theory and Applications*, Wiley, Chichester, UK, 1987.

[3] N. Arada and J.-P. Raymond, *Optimal control problems with mixed control-state constraints*, SIAM J. Control Optim., 39 (2000), pp. 1391–1407.

[4] M. Bergounioux and F. Tröltzsch, *Optimal control of linear bottleneck problems*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 235–250.

[5] M. Bergounioux and F. Tröltzsch, *Optimal control of semilinear parabolic equations with state-constraints of bottleneck type*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 595–608.

[6] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[7] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations*, SIAM J. Control Optim., 38 (2000), pp. 1369–1391.

[8] R. Grinold, *Continuous programming* I, J. Math. Anal. Appl., 28 (1969), pp. 39–51.

[9] W. Krabs, *Lineare Optimierung in halbgeordneten Vektorräumen*, Numer. Math. 11 (1968), pp. 220–231.

[10] N. Levinson, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.

[11] C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl., to appear.

[12] M. K. V. Murthy and G. Stampacchia, *A variational inequality with mixed boundary conditions*, Israel J. Math., 13 (1972), pp. 188–224.
[13] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 167–199.
[14] A. Rösch and F. Tröltzsch, *Sufficient second-order optimality conditions for an elliptic optimal control problem with pointwise control-state constraints*, SIAM J. Optim., to appear.
[15] A. Rösch and F. Tröltzsch, *Sufficient second order optimality conditions for a parabolic optimal control problem with pointwise control-state constraints*, SIAM J. Control Optim., 42 (2003), pp. 138–154.
[16] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., J. A. Barth, Heidelberg, 1995.
[17] F. Tröltzsch, *Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints*, SIAM J. Optim., 15 (2005), pp. 616–634.
[18] F. Tröltzsch, *A minimum principle and a generalized bang-bang-principle for a distributed optimal control problem with constraints on control and state*, Z. Angew. Math. Mech., 59 (1979), pp. 737–739.
[19] F. Tröltzsch, *The generalized bang-bang-principle and the numerical solution of a parabolic boundary-control problem with constraints on the control and the state*, Z. Angew. Math. Mech., 64 (1984), pp. 551–557.
[20] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner-Texte Math. 62, Teubner, Leipzig, 1984.
[21] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen—Theorie, Verfahren und Anwendungen*, Vieweg, Wiesbaden, 2005.
[22] W. F. Tyndall, *A duality theorem for a class of continuous linear programming problems*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 644–666.
[23] J. Zowe and S. Kurcyusz, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 565–579

ON AN OPTIMAL STOPPING PROBLEM OF TIME INHOMOGENEOUS DIFFUSION PROCESSES*

YOICHI OSHIMA†

Abstract. For given quasi-continuous (q.c.) functions g, h with $g \leq h$ and diffusion process $\mathbf{M}$ determined by stochastic differential equations or symmetric Dirichlet forms, characterizations of the value functions $\widetilde{e}_g(s,x) = \sup_\sigma J_{(s,x)}(\sigma)$ and $\bar{w}(s,x) = \inf_\tau \sup_\sigma J_{(s,x)}(\sigma,\tau)$ have been well studied. In this paper, by using the time-dependent Dirichlet forms, we generalize these results to time inhomogeneous diffusion processes. The difficulty of our case arises from the existence of essential semipolar sets [Y. Oshima, *Tohoku Math. J.* (2), 54 (2002), pp. 443–449]. In particular, excessive functions are not necessarily continuous along the sample paths. We get the result by showing such a continuity of the value functions.

Key words. time inhomogeneous diffusion processes, optimal stopping, Dirichlet forms, quasi-variational inequality

AMS subject classifications. 60G40, 49J40, 62L15, 60J60

DOI. 10.1137/040609549

1. Introduction and preliminaries.

Let $\mathbf{M} = (X_t, P_{(s,x)})$ be a not necessarily time homogeneous diffusion process on a locally compact separable metric space X. For given (quasi-) continuous functions g, h on $[0,\infty) \times X$ and stopping times σ and τ, let

$$J_{(s,x)}(\sigma) = E_{(s,x)}\left(e^{-\sigma} g(s+\sigma, X_\sigma)\right), \tag{1.1}$$

$$J_{(s,x)}(\sigma,\tau) = E_{(s,x)}\left(e^{-\sigma\wedge\tau}\left(g(s+\sigma,X_\sigma)I_{\{\sigma\leq\tau\}} + h(s+\tau,X_\tau)I_{\{\tau<\sigma\}}\right)\right). \tag{1.2}$$

The main purpose of this paper is to characterize $\widetilde{e}_g(s,x) = \sup_\sigma J_{(s,x)}(\sigma)$ and $\bar{w}(s,x) = \sup_\sigma \inf_\tau J_{(s,x)}(\sigma,\tau)$.

Usually, such a problem is considered for

$$J^f_{(s,x)}(\sigma,\tau) = E_{(s,x)}\left(\int_0^{\sigma\wedge\tau} e^{-t} f(s+t,X_t)dt\right) + J_{(s,x)}(\sigma,\tau) \tag{1.3}$$

instead of $J_{(s,x)}(\sigma,\tau)$. But we use $J_{(s,x)}(\sigma,\tau)$ because (1.3) is essentially reduced to (1.2) by taking $g + R_1 f$ and $h + R_1 f$ instead of g and h, respectively, in (1.3), where $R_\alpha f$ is the resolvent of $\mathbf{M}$.

There are many works related to our problem. In particular, when $\mathbf{M}$ is a diffusion process determined by a stochastic differential equation with Lipschitz continuous coefficients, the detailed results related to $\widetilde{e}_g$ can be found in [1], [7] and references therein. In the time homogeneous case, Nagai [10], [11] and Zabczyk [19] used (symmetric) Dirichlet form theory to solve the problem. The diffusion process $\mathbf{M}$ corresponding to the generator on R^d determined by

$$A\varphi(x) = \sum_{i,j=1}^d \frac{1}{\rho(x)}\frac{\partial}{\partial x_i}\left(a_{ij}(x)\rho(x)\frac{\partial\varphi}{\partial x_j}\right) \tag{1.4}$$

*Received by the editors June 6, 2004; accepted for publication (in revised form) December 27, 2005; published electronically April 21, 2006.
http://www.siam.org/journals/sicon/45-2/60954.html
†Department of Mathematics, Faculty of Engineering, Kumamoto University, Kumamoto, 860-8555, Japan (oshima@gpo.kumamoto-u.ac.jp).

for uniformly elliptic functions $(a_{ij}(x))_{i,j=1,2,\dots,d}$ and a function $\rho(x) > 0$ belonging to a Sobolev space on R^d is contained in their framework. See also [5] and [8] for related results.

The purpose of this paper is to generalize those results to time inhomogeneous diffusion processes including the case that (a_{ij}) in (1.4) admits to dependence on a time parameter. In this case, the generator for each t is given by

$$A^{(t)}\varphi(x) = \sum_{i,j=1}^{d} \frac{1}{\rho(x)} \frac{\partial}{\partial x_i} \left(a_{ij}(t,x)\rho(x) \frac{\partial \varphi}{\partial x_j} \right) \tag{1.5}$$

and the coresponding Dirichlet form on $L^2(R^d; \rho(x)dx)$ is an extension of

$$E^{(t)}(\varphi, \psi) = \sum_{i,j=1}^{d} \int_{R^d} a_{ij}(t,x) \frac{\partial \varphi}{\partial x_i} \frac{\partial \psi}{\partial x_j} \rho(x) dx. \tag{1.6}$$

In the Lipschitz continuous and time homogeneous cases stated above, the (quasi-) continuity of the value functions $\widetilde{e}_g$ and $\bar{w}$ follows naturally. The essential step in this paper is to prove the fine and cofine continuities of the value functions.

The organization of this paper is as follows. In the rest of this section, the notions of time-dependent Dirichlet forms and the basic properties of the associated time inhomogeneous Markov processes are stated. In section 2, under the separability condition, quasi-variational inequalities and their solutions are given. In section 3, the optimal stopping problem is solved and is divided into three cases: (I) one obstacle cases, (II) two obstacles cases under the separability condition, and (III) general two obstacles cases.

Now we shall start with our settings. Let X be a locally compact separable metric space and m a positive Radon measure on X with full support. We assume that we are given a family $(E^{(t)}, F)_{t \geq 0}$ of Dirichlet forms on $H = L^2(X; m)$ satisfying the following conditions [4]:

(i) For each $t \geq 0$, $(E^{(t)}, F)$ is an m-symmetric Dirichlet form on H.

(ii) For any $\varphi \in F$, $E^{(t)}(\varphi, \varphi)$ is a measurable function of $t \geq 0$ and satisfies

$$\lambda^{-1} \|\psi\|_F^2 \leq E_1^{(t)}(\psi, \psi) \leq \lambda \|\psi\|_F^2 \tag{1.7}$$

for some positive constant λ, where $E_\alpha^{(t)}(\psi, \psi) = E^{(t)}(\psi, \psi) + \alpha(\psi, \psi)_m$ and $\|\psi\|_F^2 = E_1^{(0)}(\psi, \psi)$.

(iii) F is regular; that is, $C_0(X) \cap F$ is uniformly dense in $C_0(X)$ and $\|\cdot\|_F$-dense in F, where $C_0(X)$ is the family of continuous functions on X with compact support.

(iv) For any $t \geq 0$, $(E^{(t)}, F)$ is local; that is, for any $\varphi, \psi \in F \cap C_0(X)$ such that $\varphi \cdot \psi = 0$, $E^{(t)}(\varphi, \psi) = 0$.

For simplicity, we put $E^{(t)} = E^{(0)}$ for $t < 0$. For each t, there exists an operator $A^{(t)}$ from F to F' such that

$$-\langle A^{(t)}\varphi, \psi \rangle = E^{(t)}(\varphi, \psi) \tag{1.8}$$

for any $\varphi, \psi \in F$. To consider an optimal stopping problem related to the time inhomogeneous diffusion process X_t with generator $A^{(t)}$, we shall introduce the space-time process $Z_t = (\tau(t), X_t)$ on $Z = R^1 \times X$ with uniform motion $\tau(t)$. Formally, the resolvent $R_\alpha f$ of Z_t satisfies

$$\left(\alpha - \frac{\partial}{\partial t} - A^{(t)} \right) R_\alpha f(t,x) = f(t,x). \tag{1.9}$$

To define Z_t more rigorously, let us introduce the spaces $\mathcal{H}$, $\mathcal{F}$, and $\mathcal{W}$. Put $\mathcal{H} = \{u(t,x) : u(t,\cdot) \in H, \|u\|_{\mathcal{H}} < \infty\}$, where

$$\|u\|_{\mathcal{H}}^2 = \int_{R^1} \|u(t,\cdot)\|_H^2 dt.$$

The space $\mathcal{F}$ is a family of measurable functions $u \in \mathcal{H}$ such that $u(t,\cdot) \in F$ for all t and $\|u\|_{\mathcal{F}} < \infty$, where

$$\|u\|_{\mathcal{F}}^2 = \int_{R^1} \|u(t,\cdot)\|_F^2 \, dt.$$

The dual space $\mathcal{F}'$ is defined similarly by taking F' instead of F in the definition of $\mathcal{F}$. For any function $f \in \mathcal{F}$, considering f as a function of $t \in R^1$ with value in F', the distribution derivative $\partial f/\partial t$ is defined as a function $g(t,\cdot)$ on R^1 with value in F' such that

$$\int_{R^1} g(t,\cdot)\xi(t)dt = \int f(t,\cdot)\xi'(t)dt$$

for any $\xi \in C_0^\infty(R^1)$. Using this derivative, define the space $(\mathcal{W}, \|\cdot\|_{\mathcal{W}})$ by

$$\mathcal{W} = \left\{ u \in \mathcal{F} : \frac{\partial u}{\partial t} \in \mathcal{F}', \ \|u\|_{\mathcal{W}} < \infty \right\},$$
$$\|u\|_{\mathcal{W}}^2 = \left\| \frac{\partial u}{\partial t} \right\|_{\mathcal{F}'}^2 + \|u\|_{\mathcal{F}}^2 .$$

Further, define the bilinear form $\mathcal{E}$ by

$$\mathcal{E}(u,v) = \begin{cases} -\int_{R^1} \left(\frac{\partial u}{\partial t}, v \right) dt + \int_{R^1} E^{(t)}(u(t,\cdot), v(t,\cdot))dt, & u \in \mathcal{W}, v \in \mathcal{F}, \\ \int_{R^1} \left(\frac{\partial v}{\partial t}, u \right) dt + \int_{R^1} E^{(t)}(u(t,\cdot), v(t,\cdot))dt, & u \in \mathcal{F}, v \in \mathcal{W}. \end{cases}$$

Then for $f \in \mathcal{H}$, $R_\alpha f$ in (1.9) is considered as a version of $G_\alpha f \in \mathcal{W}$ of the solution of

$$\mathcal{E}_\alpha(G_\alpha f, v) = (f, v)_\nu$$

for any $v \in \mathcal{F}$, where $\mathcal{E}_\alpha(\ ,\ ) = \mathcal{E}(\ ,\ ) + \alpha(\ ,\ )_\nu$ and $d\nu(t,x) = dtdm(x)$. This equation is equivalent to

$$-\left(\frac{\partial}{\partial t} G_\alpha f(t,\cdot), \varphi \right) + E_\alpha^{(t)} \left(G_\alpha f(t,\cdot), \varphi \right) = (f(t,\cdot), \varphi)$$

for any $t \geq 0$ and $\varphi \in F$. The dual resolvent $\widehat{G}_\alpha f \in \mathcal{W}$ is defined as a solution of

$$\left(\frac{\partial}{\partial t} \widehat{G}_\alpha f(t,\cdot), \varphi \right) + E_\alpha^{(t)} \left(\varphi, \widehat{G}_\alpha f(t,\cdot) \right) = (f(t,\cdot), \varphi)$$

for any $t \geq 0$ and $\varphi \in F$. Then, for any $f \in \mathcal{F}$ (resp., $f \in \mathcal{H}$), $\|\alpha G_\alpha f\|_{\mathcal{F}} \leq C_1 \|f\|_{\mathcal{F}}$ (resp., $\|\alpha G_\alpha f\| \leq \|f\|_{\mathcal{H}}$) for some constant C_1 and $\lim_{\alpha\to\infty} \alpha G_\alpha f = f$ in $\mathcal{F}$ (resp.,

$\mathcal{H}$) (see [14, Lemma 2.1], [16, I.3, I.4]). Similar results also hold for the dual resolvent $\widehat{G}_\alpha$.

To choose a version $R_\alpha f$ of $G_\alpha f$, we need to define a capacity. A function $u \in \mathcal{F}$ is called *α-excessive* if $\mathcal{E}_\alpha(u, w) \geq 0$ for any nonnegative function $w \in \mathcal{W}$. Then $u \in \mathcal{F}$ is α-excessive if and only if $u \geq 0$ and $\beta G_{\beta+\alpha} u \leq u$ a.e. for all $\beta > 0$ (see [14]). We denote by $\mathcal{P}_\alpha$ the family of all α-excessive functions. In particular, put $\mathcal{P} = \mathcal{P}_1$.

For any function $h \in \mathcal{H}$, let

$$\mathcal{L}_h = \{u \in \mathcal{F} : u \geq h \ \nu \ a.e.\}$$

and $\mathcal{L}_A = \mathcal{L}_{I_A}$. Then the following results hold (see [9] and [15]).

LEMMA 1.1. *For any $\varepsilon > 0$ and $\alpha > 0$, there exists a unique function $h_\varepsilon^\alpha \in \mathcal{W}$ such that*

$$-\left(\frac{\partial h_\varepsilon^\alpha}{\partial t}, \varphi\right) + E_\alpha^{(t)}\left(h_\varepsilon^\alpha(t,\cdot), \varphi\right) = \frac{1}{\varepsilon}\left(\left(h_\varepsilon^\alpha(t,\cdot) - h(t,\cdot)\right)^-, \varphi\right) \tag{1.10}$$

for any $\varphi \in F$.

THEOREM 1.2. *Suppose that $\mathcal{L}_h \cap \mathcal{W} \neq \emptyset$. Then $e_h^\alpha = \lim_{\varepsilon \downarrow 0} h_\varepsilon^\alpha$ converges increasingly, strongly in $\mathcal{H}$ and weakly in $\mathcal{F}$. Furthermore, e_h^α is the minimal function of $\mathcal{P}_\alpha \cap \mathcal{L}_h$ and satisfies*

$$\mathcal{A}_\alpha\left(e_h^\alpha, e_h^\alpha\right) \leq \mathcal{E}_\alpha\left(e_h^\alpha, w\right) \tag{1.11}$$

for any $w \in \mathcal{L}_h \cap \mathcal{W}$.

If $u \in \mathcal{P}_\alpha$, then there exists a positive Radon measure μ_u^α on Z such that

$$\mathcal{E}_\alpha(u, w) = \int_Z w(z) d\mu_u^\alpha(z) \quad \text{for any } w \in C_0(Z) \cap \mathcal{W}. \tag{1.12}$$

We omit the superfix α in e_h^α and μ_u^α if $\alpha = 1$. For any open set A of Z such that $\mathcal{L}_A \cap \mathcal{W} \neq \emptyset$, put $e_A = e_{I_A}$ and $\mu_A = \mu_{e_A}$. Then μ_A is supported by the closure $\bar{A}$ of A. The *capacity* $\mathrm{Cap}(A)$ of A is defined by

$$\mathrm{Cap}(A) = \mu_A(\bar{A}).$$

If there exists $w \in \mathcal{W}$ such that $w = 1$ a.e. on A, then

$$\mathrm{Cap}(A) = \mathcal{E}_\alpha(e_A^\alpha, w). \tag{1.13}$$

The notion of the capacity is extended to any Borel set by the usual manner. A set is called *exceptional* if it is of zero capacity. If a statement holds everywhere except on an exceptional set, then we say the statement holds quasi-everywhere (q.e.).

An increasing sequence of closed sets $\{F_n\}$ is called a *nest* if $\lim_{n\to\infty} \mathrm{Cap}(Z \backslash F_n) = 0$. A function u is called *quasi-continuous* (q.c.) if there exists a nest $\{F_n\}$ of closed sets such that u is continuous on each F_n. The *quasi-lower semicontinuity* is defined similarly. Any function $u \in \mathcal{W}$ has a q.c. modification $\widetilde{u}$. In particular, for any $f \in \mathcal{H}$ and $\alpha > 0$, $G_\alpha f$ and $\widehat{G}_\alpha f$ have q.c. modifications. The relation (1.12) can be extended to $w \in \mathcal{W}$ by taking the q.c. modification. For any α-excessive function $u \in \mathcal{F}$, define its α-excessive modification $\widetilde{u}$ by $\widetilde{u} = \lim_{n\to\infty} nR_{n+\alpha} u$. Since $\widetilde{u}$ is an increasing limit of q.c. functions, $\widetilde{u}$ is quasi-lower semicontinuous. The following theorem and the properties of the associated diffusion process can be found in [12], [14], [16], [17], and [18].

THEOREM 1.3. *There exist diffusion processes* $\mathbf{M} = (Z_t, P_z)$ *and* $\widehat{\mathbf{M}} = (\widehat{Z}_t, \widehat{P}_z)$ *on* Z *satisfying the following conditions:*

(i) *The resolvents* $R_\alpha f$ *and* $\widehat{R}_\alpha f$ *of* $\mathbf{M}$ *and* $\widehat{\mathbf{M}}$ *are q.c. modifications of* $G_\alpha f$ *and* $\widehat{G}_\alpha f$*, respectively.*

(ii) *Let* $Z_t = (\tau(t), X_t)$ *and* $\widehat{Z}_t = (\widehat{\tau}(t), \widehat{X}_t)$ *be the decompositions of* Z_t *and* $\widehat{Z}_t$ *into the processes on* R^1 *and* X*, respectively. Then* $\tau(t) = \tau(0)+t$ *and* $\widehat{\tau}(t) = \widehat{\tau}(0)-t$.

(iii) *For any open set* A *of* Z, $E.(e^{-\alpha\sigma_A})$ *is a quasi-lower semicontinuous modification of* e_A^α*, where* σ_A *is the hitting time of* A.

For later use, we present two lemmas. The proof of Lemma 1.4 can be found in [12, Lemma 3.7].

LEMMA 1.4. *For any* α*-excessive function* $u \in \mathcal{F}$, $\mu_u^{(\alpha)}$ *does not charge any Borel set of zero capacity.*

LEMMA 1.5. *Suppose that a sequence of* 1*-excessive functions* $\{u_n\}$ *converges to zero in* $\mathcal{H}$*. Then there exists a subsequence* $\{u_{n_k}\}$ *such that* $\lim_{k\to\infty} \widetilde{u}_{n_k} = 0$ *quasi-uniformly; that is, there exists a nest* $\{F_n\}$ *such that* $\lim_{k\to\infty} \widetilde{u}_{n_k} = 0$ *uniformly on each* F_n.

Proof. Since $\widetilde{u}_n$ is quasi-lower semicontinuous, there exists an open set N_k such that $\mathrm{Cap}(N_k) < 1/2^k$ and $\widetilde{u}_n$ is lower semicontinuous on $Z \setminus N_k$ for all n. Put $B_k^n = \{z \in Z \setminus N_k : \widetilde{u}_n(z) > 1/2^k\}$ and $D_k^n = B_k^n \cup N_k$. Then D_k^n is open and, noting that $(e_{N_k}, p) = \langle \mu_{N_k}, \widehat{R}_1 p\rangle \le ||p||_\infty \mathrm{Cap}(N_k) < ||p||_\infty/2^k$ for any nonnegative bounded continuous function $p \in \mathcal{H}$, it holds that

$$\begin{aligned}(e_{D_k^n}, p) &\le (e_{B_k^n}, p) + (e_{N_k}, p) \le 2^k (u_n, p) + \|p\|_\infty \mathrm{Cap}(N_k)\\ &\le 2^k \|u_n\|_{\mathcal{H}} \cdot \|p\|_{\mathcal{H}} + \frac{\|p\|_\infty}{2^k}.\end{aligned}$$

For each k, take n_k such that $\|u_n\|_{\mathcal{H}} \le 1/2^{2k}$ for any $n \ge n_k$. Then $F_m = Z \setminus \cup_{k=m}^\infty D_{n_k}^k$ is a closed set. Since $\{Z \setminus F_m\}$ is a decreasing sequence of open sets such that

$$\langle \mu_{Z\setminus F_m}, \widehat{R}_1 p\rangle = \left(e_{Z\setminus F_m}, p\right) \le \frac{1}{2^{m-1}} \left(\|p\|_{\mathcal{H}} + \|p\|_\infty\right)$$

for any p satisfying the stated conditions, $\lim_{m\to\infty} \mathrm{Cap}(Z\setminus F_m) = \lim_{m\to\infty} \mu_{Z\setminus F_m}(Z) = 0$. Furthermore, $\lim_{k\to\infty} u_{n_k} = 0$ uniformly on each F_m. □

2. Quasi-variational inequalities. In this section, we assume that we are given two *obstacles* $g, h \in \mathcal{F}$ which are q.-c. and $g \le h$ q.e. We say that the pair (g, h) satisfies the *separability condition* if there exist $\varphi, \psi \in \mathcal{P}$ such that

$$g \le \widetilde{\varphi} - \widetilde{\psi} \le h \text{ q.e.} \tag{2.1}$$

Define the sequences of 1-excessive functions $\{u_n\}$ and $\{v_n\}$ inductively by

$$u_0 = 0,\ v_n = e_{u_{n-1}-h},\ u_n = e_{v_n+g}.$$

For any $\phi \in \mathcal{P}$, let L_ϕ be a continuous linear functional on $\mathcal{F}$ defined by $L_\phi(w) = \mathcal{A}_1(\phi, w)$. In the proof of Lemma 5.2 in [14], one can see that $2\widehat{G}_1 L_\phi - \phi$ is 1-coexcessive, and thus in particular nonnegative. Hence

$$\phi \le 2\widehat{G}_1 L_\phi \in \mathcal{W}. \tag{2.2}$$

LEMMA 2.1. *Suppose that the separability condition* (2.1) *holds. Then*

(i) u_n, v_n *are well defined.*

(ii) $\lim_{n\to\infty} u_n = \bar{u}$ *and* $\lim_{n\to\infty} v_n = \bar{v}$ *converge increasingly, strongly in* $\mathcal{H}$, *and weakly in* $\mathcal{F}$.

(iii) $\bar{u} \le \varphi$, $\bar{v} \le \psi$, *and* $g \le \bar{u} - \bar{v} \le h$ *a.e.*

Proof. Clearly $u_0 = 0 \le \varphi$. Suppose that u_{n-1} is defined and satisfies $u_{n-1} \le \varphi$. Then by the separability condition, $u_{n-1} - h \le \varphi - h \le \psi$. Thus $u_{n-1} - h \le 2\widehat{G}_1 L_\psi \in \mathcal{W}$ by (2.2). Hence $v_n := e_{u_{n-1}-h}$ is well defined and $v_n \le \psi$ by Theorem 1.2. Again, by (2.2) and the separability condition, $v_n + g = e_{u_{n-1}-h} + g \le \psi + g \le \varphi \le 2\widehat{G}_1 L_\varphi \in \mathcal{W}$. Therefore $u_n := e_{v_n+g}$ is well defined and dominated by φ.

If $u_{n-1} \le u_n$, then $v_n = e_{u_{n-1}-h} \le e_{u_n-h} = v_{n+1}$ and $u_n = e_{v_n+g} \le e_{v_{n+1}+g} = u_{n+1}$. Thus u_n and v_n are well defined and increasing relative to n. By virtue of (2.2),

$$\mathcal{A}_1(u_n.u_n) \le 2\mathcal{E}_1(u_n, \widehat{G}_1 L_\varphi) \le 2\|u_n\|_{\mathcal{F}} \|\widehat{G}_1 L_\varphi\|_{\mathcal{W}}.$$

Hence $\{\|u_n\|_{\mathcal{F}}\}$ is bounded. Similarly, $\{\|v_n\|_{\mathcal{F}}\}$ is bounded and the assertion (ii) holds by Lemma I.2.12 in [6].

Since $u_n \le \varphi$ and $v_n \le \psi$, the first assertion of (iii) holds. Furthermore, from the definition, $u_{n-1} - h \le v_n$ and $v_n + g \le u_n$. This implies the second assertion of (iii). □

THEOREM 2.2. *Under the separability condition,* $\bar{u} = e_{\bar{v}+g}$ *and* $\bar{v} = e_{\bar{u}-h}$. *In particular,*

$$\mathcal{A}_1(\bar{u}, \bar{u}) \le \mathcal{E}_1(\bar{u}, w) \quad \forall w \in \mathcal{L}_{\bar{v}+g} \cap \mathcal{W},$$
$$\mathcal{A}_1(\bar{v}, \bar{v}) \le \mathcal{E}_1(\bar{v}, w) \quad \forall w \in \mathcal{L}_{\bar{u}-h} \cap \mathcal{W}.$$

Moreover, if a pair of 1*-excessive functions* (u, v) *satisfies* $g \le u - v \le h$, *then* $\bar{u} \le u$ *and* $\bar{v} \le v$.

Proof. Since $\bar{u}$ is a 1-excessive function in $\mathcal{L}_{\bar{v}+g}$, clearly $e_{\bar{v}+g} \le \bar{u}$. Conversely, $\bar{u} = \lim_{n\to\infty} u_n = \lim_{n\to\infty} e_{v_n+g} \le e_{\bar{v}+g}$ by Lemma 2.1. Similarly, $\bar{v} = e_{\bar{u}-h}$. The quasi-variational inequalities are already stated in Theorem 1.2. By Lemma 2.1, if g, h satisfy the separability condition with $(u, v) \in \mathcal{P} \times \mathcal{P}$, then $u_n \le u, v_n \le v$ for any n. Since $\lim_{n\to\infty} u_n = \bar{u}, \lim_{n\to\infty} v_n = \bar{v}$, we obtain $\bar{u} \le u, \bar{v} \le v$. □

A similar quasi-variational inequality for $\bar{u} - \bar{v}$ also holds. But it will be given in the next section because we use a probabilistic argument for the proof.

LEMMA 2.3. *For any* $g \in \mathcal{W}$, $\lim_{k\to\infty} e_{g-g^{(k)}} = 0$ *in* $\mathcal{F}$.

Proof. Since $g - g^{(k)} \in \mathcal{L}_{g-g^{(k)}} \cap \mathcal{W}$,

$$\begin{aligned}
\mathcal{A}_1\left(e_{g-g^{(k)}}, e_{g-g^{(k)}}\right) &\le \mathcal{E}_1\left(e_{g-g^{(k)}}, g - g^{(k)}\right) \\
&= -\mathcal{E}_1\left(g - g^{(k)}, e_{g-g^{(k)}}\right) + 2\mathcal{A}_1\left(e_{g-g^{(k)}}, g - g^{(k)}\right) \\
&= -\mathcal{E}_1\left(g, e_{g-g^{(k)}} - k\widehat{G}_k e_{g-g^{(k)}}\right) + 2\mathcal{A}_1\left(e_{g-g^{(k)}}, g - g^{(k)}\right) \qquad (2.3) \\
&\le 2\|g\|_{\mathcal{W}} \|e_{g-g^{(k)}} - k\widehat{G}_k e_{g-g^{(k)}}\|_{\mathcal{F}} + 2\|g - g^{(k)}\|_{\mathcal{F}} \|e_{g-g^{(k)}}\|_{\mathcal{F}} \\
&\le 4\|g\|_{\mathcal{W}}(C_1 + 1)\|e_{g-g^{(k)}}\|_{\mathcal{F}}.
\end{aligned}$$

Hence $\mathcal{A}_1(e_{g-g^{(k)}}, e_{g-g^{(k)}})$ is bounded. By virtue of [16, III.Lemma 2.2], $e_{g-g^{(k)}}$ converges to 0 strongly in $\mathcal{H}$, and hence weakly in $\mathcal{F}$ from [6, Lemma I.2.12]. Since $\|k\widehat{G}_k e_{g-g^{(k)}}\|_{\mathcal{F}} \le C_1 \|e_{g-g^{(k)}}\|_{\mathcal{F}}$, by the same argument, $\lim_{k\to\infty} k\widehat{G}_k e_{g-g^{(k)}} = 0$ strongly in $\mathcal{H}$ and weakly in $\mathcal{F}$. Hence from (2.3), $\lim_{k\to\infty} \mathcal{A}_1(e_{g-g^{(k)}}, e_{g-g^{(k)}}) = 0$. □

3. An optimal stopping problem. Let $\mathbf{M}$ and $\widehat{\mathbf{M}}$ be the diffusion processes given by Theorem 1.3. Denote by R_α and $\widehat{R}_\alpha$ their associated resolvents. For any stopping time σ, define $H_\sigma u$ by $H_\sigma u(z) = E_z\left(e^{-\sigma}u(Z_\sigma)\right)$. In particular, put $H_B = H_{\sigma_B}$ for the hitting time σ_B of the nearly Borel set B.

(I) One obstacle case:

Let g be a q.c. function of $\mathcal{F}$ such that $\mathcal{L}_g \cap \mathcal{W} \neq \emptyset$. As in the previous section, denote by e_g the minimal 1-excessive function of $\mathcal{L}_g$. Then it is the minimal function of $\mathcal{L}_g$ satisfying the quasi-variational inequality (1.11). The following result is a time inhomogeneous version of Nagai's result [10].

THEOREM 3.1. *Suppose that $g \in \mathcal{F}$ is q.c. and $\mathcal{L}_g \cap \mathcal{W} \neq \emptyset$. Then*

$$\widetilde{e}_g(z) = \sup_\sigma J_z(\sigma) = E_z\left(e^{-\sigma_B}g(Z_{\sigma_B})\right) \text{ q.e.}, \tag{3.1}$$

where the supremum is taken over all stopping times σ and $B = \{z : \widetilde{e}_g(z) = g(z)\}$.

Proof. Noting that $\widetilde{e}_g$ is the smallest 1-excessive function dominating g q.e., we have for any stopping time σ,

$$E_z\left(e^{-\sigma}g(Z_\sigma)\right) \le E_z\left(e^{-\sigma}\widetilde{e}_g(Z_\sigma)\right) \le \widetilde{e}_g(z) \text{ q.e.}$$

Hence it is enough to show

$$\widetilde{e}_g(z) = E_z\left(e^{-\sigma_B}g(Z_{\sigma_B})\right). \tag{3.2}$$

This is essentially shown in [14, Lemma 6.2], but we shall give the outline of the proof for completeness. For $\varepsilon_n \downarrow 0$, let g_n be a q.c. version of the solution of $g_n = (1/\varepsilon_n)G_1\left((g_n - g)^-\right)$ determined by Lemma 1.1. Let $B_n = \{z : g_n(z) \le g(z)\}$ and $\dot{\sigma}_n = \dot{\sigma}_{B_n}$, where $\dot{\sigma}_A$ is the first entry time of A defined by $\dot{\sigma}_A = \inf\{t \ge 0 : Z_t \in A\}$. Then

$$\begin{aligned} g_n(z) &= \frac{1}{\varepsilon_n}E_z\left(\int_{\dot{\sigma}_n}^\infty e^{-t}(g_n - g)^-(Z_t)dt\right) \\ &= E_z\left(e^{-\dot{\sigma}_n}g_n(Z_{\dot{\sigma}_n})\right) \le E_z\left(e^{-\dot{\sigma}_n}g(Z_{\dot{\sigma}_n})\right) \\ &\le E_z\left(e^{-\dot{\sigma}_n}\widetilde{e}_g(Z_{\dot{\sigma}_n})\right). \end{aligned}$$

Put $\dot{\sigma} = \lim_{n\to\infty}\dot{\sigma}_n$. Then $\dot{\sigma} \le \dot{\sigma}_B$. By virtue of Theorem 1.2, since $g_n \uparrow e_g$ a.e., we then have, for any nonnegative function $f \in \mathcal{H}$,

$$\begin{aligned} (f, \widetilde{e}_g) &= \lim_{n\to\infty}(f, g_n) = \lim_{n\to\infty} E_{f\cdot\nu}\left(e^{-\dot{\sigma}_n}g_n(Z_{\dot{\sigma}_n})\right) \le \lim_{n\to\infty} E_{f\cdot\nu}\left(e^{-\dot{\sigma}_n}g(Z_{\dot{\sigma}_n})\right) \\ &= E_{f\cdot\nu}\left(e^{-\dot{\sigma}}g(Z_{\dot{\sigma}})\right) \le E_{f\cdot\nu}\left(e^{-\dot{\sigma}}\widetilde{e}_g(Z_{\dot{\sigma}})\right) \le (f, \widetilde{e}_g). \end{aligned}$$

Hence $\widetilde{e}_g(z) = E_z\left(e^{-\dot{\sigma}}g(Z_{\dot{\sigma}})\right)$ for a.e. z. Since $g \le \widetilde{e}_g$ q.e., we also have $g(Z_{\dot{\sigma}}) = \widetilde{e}_g(Z_{\dot{\sigma}})$, and hence $\dot{\sigma}_B \le \dot{\sigma}$ a.s. P_z for a.e. z. Therefore $\dot{\sigma} = \dot{\sigma}_B$ a.s. P_z and $\widetilde{e}_g(z) = E_z\left(e^{-\dot{\sigma}_B}g(Z_{\dot{\sigma}_B})\right)$ for a.e. z. By taking the 1-excessive regularization, we get the result. □

Remark. Since $\dot{\sigma}_B \le \sigma_B$, it holds that

$$\widetilde{e}_g \ge H_{\dot{\sigma}_B}\widetilde{e}_g \ge H_B\widetilde{e}_g \ge H_Bg \quad \text{q.e.}$$

Hence Theorem 3.1 implies that $H_{\dot{\sigma}_B}\widetilde{e}_g = H_B\widetilde{e}_g$ q.e. z. In particular, the set of irregular points of B is exceptional.

(II) Two obstacles case under the separability condition:

We assume that we are given two q.c. functions $g, h \in \mathcal{F}$ such that $g \leq h$ q.e. If the separability condition (2.1) is satisfied, then there exists the minimal pair of finely continuous functions $(\bar{u}, \bar{v})$ given by Theorem 2.2. They are given by $\bar{u} = \lim_{k\to\infty} \bar{u}_k$ and $\bar{v} = \lim_{k\to\infty} \bar{v}_k$ with the solutions $\bar{u}_k, \bar{v}_k \in \mathcal{W}$ of $\bar{u}_k = (1/\varepsilon_k)G_\alpha\left((\bar{u}_k - \bar{v} - g)^-\right)$ and $\bar{v}_k = (1/\varepsilon_k)G_\alpha\left((\bar{v}_k - \bar{u} + h)^-\right)$, respectively. Define the sets C_k and D_k by

$$C_k = \{z : \bar{u}_k(z) \leq (g + \bar{v})(z)\}, \quad D_k = \{z : \bar{v}_k(z) \leq (\bar{u} - h)(z)\}. \tag{3.3}$$

Then $C_k \supset C$ and $D_k \supset D$ for

$$C = \{z : \bar{u}(z) = (g + \bar{v})(z)\}, \quad D = \{z : \bar{v}(z) = (\bar{u} - h)(z)\}. \tag{3.4}$$

As we remarked before (2.2), $\bar{u}$ can be represented as a difference of a function of $\mathcal{W}$ and a coexcessive function. In particular, $\bar{u}$ has a q.e. cofinely continuous modification $\widehat{u}$ given by $\widehat{u} = \lim_{n\to\infty} n\widehat{R}_{n+1}\bar{u}$. Since $\bar{u}$ is q.e. lower semicontinuous, $\widehat{u}(z) = \lim_{t\to 0} \widehat{E}_z\left(\bar{u}(Z_t)\right) \geq \underline{\lim}_{y\to z} \bar{u}(y) \geq \bar{u}(z)$ q.e. Similarly, a q.e. cofinely continuous modification $\widehat{v}$ of $\bar{v}$ exists and satisfies $\widehat{v} \geq \bar{v}$ q.e.

Since $\bar{v}$ is 1-excessive, there exists a positive Radon measure $\mu_{\bar{v}}$ charging no exceptional set such that $\mathcal{E}_1(\bar{v}, w) = \langle \mu_{\bar{v}}, \widetilde{w} \rangle$ for any $w \in \mathcal{W}$. In the following two lemmas, we use the notation $\mu_{\bar{v},F} = \mu_{\bar{v}}|_F, \mu_{\bar{v},F^c} = \mu_{\bar{v}}|_{F^c}, \bar{v}_F = \widetilde{U}_1\mu_{\bar{v},F}$, $\bar{v}_{F^c} = \widetilde{U}_1\mu_{\bar{v},F^c}$, and $\widehat{v}_{F^c} = \lim_{n\to\infty} n\widehat{R}_{n+1}\bar{v}_{F^c}$.

LEMMA 3.2. *Assume that there exists a nonexceptional compact set F such that $F \subset \{z : (\widehat{v} - \bar{v})(z) \geq \delta\}$ for some $\delta > 0$. Then $\widehat{v}_{F^c}(z) = \bar{v}_{F^c}(z)$ for q.e. $z \in F$.*

Proof. For simplicity of the notation, put $\mu_1 = \mu_{\bar{v},F}$, $\mu_2 = \mu_{\bar{v},F^c}$, $\bar{v}_2 = \widetilde{U}_1\mu_2$, and $\widehat{v}_2$ the cofinely continuous modification of $\bar{v}_2$. Assume that there exists a nonexceptional compact subset K of F such that $\bar{v}_2(z) < \widehat{v}_2(z)$ for q.e. $z \in K$. For a decreasing sequence of open sets G_n such that $\bar{G}_{n+1} \subset G_n$ and $\cap_n G_n = K$, since σ_{G_n} increases strictly to σ_K a.s. P_z for q.e. $z \notin K$, the left continuity of $\widehat{v}_2(Z_t)$ implies that

$$\lim_{n\to\infty} H_{G_n}\bar{v}_2(z) = \lim_{n\to\infty} E_z\left(e^{-\sigma_{G_n}}\bar{v}_2(Z_{\sigma_{G_n}})\right) = E_z\left(e^{-\sigma_K}\widehat{v}_2(Z_{\sigma_K})\right) = H_K\widehat{v}_2(z).$$

On the other hand, since $\mu_2(K) = 0$, for any $f \in L^2_+(Z)$, we have from [14, Corollary 5.1] and [2, Theorem I.11.2],

$$\begin{aligned}
\lim_{n\to\infty} (H_{G_n}\bar{v}_2, f) &= \lim_{n\to\infty} \mathcal{E}_1(H_{G_n}\bar{v}_2, \widehat{R}_1 f) = \lim_{n\to\infty} \mathcal{E}_1(U_1\mu_2, \widehat{H}_{G_n}\widehat{R}_1 f) \\
&= \lim_{n\to\infty} \langle \mu_2, \widehat{H}_{G_n}\widehat{R}_1 f\rangle = \langle \mu_2, \widehat{H}_K\widehat{R}_1 f\rangle \\
&= (H_K\bar{v}_2, f).
\end{aligned}$$

Hence $H_K\widehat{v}_2 = H_K\bar{v}_2$ a.e., which contradicts the assumption. □

LEMMA 3.3. *$\widehat{u} = \bar{u}$ and $\widehat{v} = \bar{v}$ q.e.*

Proof. We shall divide the proof into three steps.

Step 1. The sets $\{z : \widehat{v}(z) > \bar{v}(z), \widehat{v}(z) > (\widehat{u} - h)(z)\}$ and $\{z : \widehat{u}(z) > \bar{u}(z), \widehat{u}(z) > (\widehat{v} + g)(z)\}$ are exceptional.

To prove that any compact subset set of $\{z : \widehat{v}(z) > \bar{v}(z), \widehat{v}(z) > (\widehat{u} - h)(z)\}$ is exceptional, assume that there exists a compact nonexceptional subset F of $\{z : \widehat{v}(z) \geq \bar{v}(z) + \delta, \widehat{v}(z) \geq (\widehat{u} - h)(z) + \delta\}$ for some $\delta > 0$. For any cofinely open neighborhood A of F, since $H_A\bar{v}_F = \widetilde{U}_1(\widehat{H}_A\mu_{\bar{v},F}) = \widetilde{U}_1\mu_{\bar{v},F} = \bar{v}_F$, $\bar{v}_F$ takes its supremum on the cofine closure ${}^rF$ of F. Put $\gamma =$ q.e. sup $\bar{v}_F =$ q.e. sup $\widehat{v}_F$ and

$F_\gamma = \{z : \widehat{v}_F(z) = \gamma\} \subset {}^rF$. Then any cofine neighborhood of F_γ has positive capacity. Suppose that q.e. $\sup(\bar{u} - h - \bar{v}_{F^c}) = \gamma$; then q.e. $\sup(\widehat{u} - h - \widehat{v}_{F^c}) = \gamma$. Since $\widehat{u} - h - \widehat{v}_{F^c} \le \widehat{v}_F$, the q.e. supremum of $\widehat{u} - h - \widehat{v}_{F^c}$ is attained on $F_\gamma \subset {}^rF \subset F$. But this is absurd, because $\widehat{u} - h - \widehat{v}_{F^c}$ is a cofinely continuous function dominated by $\widehat{v}_F - \delta$ on F. Therefore $\bar{u} - h - \bar{v}_{F^c} \le \eta$ q.e. for some $\eta < \gamma$. In particular, $\bar{u} - h \le \bar{v}_F \wedge \eta + \bar{v}_{F^c}$. Since $\bar{v}_F \wedge \eta + v_{F^c}$ is an excessive function satisfying

$$g \le \bar{u} - \bar{v} \le \bar{u} - (v_F \wedge \eta + v_{F^c}) \le h,$$

this contradicts the minimality of $\bar{v}$. Now we have shown that the set $\{z : \widehat{v}(z) > \bar{v}(z), \widehat{v}(z) > (\widehat{u} - h)(z)\}$ is exceptional. The exceptionality of $\{z : \widehat{u}(z) > \bar{u}(z), \widehat{v}(u) > (\widehat{v} + g)(z)\}$ follows similarly.

Step 2. $\widehat{u} - \widehat{v} = \bar{u} - \bar{v}$ q.e.

If $(\widehat{u} - \bar{u})(z) < (\widehat{v} - \bar{v})(z)$, then $\widehat{v}(z) > \bar{v}(z)$ and $\widehat{v}(z) > (\widehat{u} - h)(z)$, because $(\widehat{u} - h)(z) < (\bar{u} + \widehat{v} - \bar{v} - h)(z) \le \widehat{v}(z)$. Hence $\{z : (\widehat{u} - \bar{u})(z) < (\widehat{v} - \bar{v})(z)\}$ is exceptional from Step 1. Similarly, $\{z : (\widehat{u} - \bar{u})(z) > (\widehat{v} - \bar{v})(z)\}$ is exceptional.

Step 3. $\widehat{u} = \bar{u}$ and $\widehat{v} = \bar{v}$ q.e.

Note that $\mu_{\bar{u}}$ and $\mu_{\bar{v}}$ are mutually singular. In fact, if we can write $\mu_{\bar{u}} = \mu_{\bar{u}}^{(s)} + f \cdot \mu_{\bar{v}}$ for some nonnegative measure $\mu_{\bar{u}}^{(s)}$ and nonnegative function f such that $\langle \mu_{\bar{v}}, f \rangle > 0$, then

$$\bar{u} - \bar{v} = \widetilde{U}_1 \mu_{\bar{u}} - \widetilde{U}_1 \mu_{\bar{v}} = \widetilde{U}_1 \left(\mu_{\bar{u}}^{(s)} + (f - f \wedge 1) \cdot \mu_{\bar{v}} \right) - \widetilde{U}_1 \left((1 - f \wedge 1) \cdot \mu_{\bar{v}} \right),$$

which contradicts the minimality of $\bar{u}$ and $\bar{v}$ in Theorem 2.2. Hence there exists a Borel set B such that $\mu_{\bar{u}}(\cdot) = \mu_{\bar{u}}(B \cap \cdot)$ and $\mu_{\bar{v}}(\cdot) = \mu_{\bar{v}}(B^c \cap \cdot)$. By virtue of Lemma 3.2, if $\{z : \widehat{v}(z) \ne \bar{v}(z)\}$ is not exceptional, then there exists a compact nonexceptional set $F \subset \{z : \widehat{v}(z) > \bar{v}(z)\}$. Since $\widehat{v}_{F^c} = \bar{v}_{F^c}$ q.e. on F by Lemma 3.2, $\widehat{v}_F > \bar{v}_F$ q.e. on F and, in particular, $\mu_{\bar{v}}(F) = \mu_{\bar{v}}(B^c \cap F) > 0$. We may assume that $F \subset B^c$. Then $\mu_{\bar{u}}(F) = 0$, and hence $\widehat{u} = \bar{u}$ q.e. on F. In fact, if $\widehat{u} > \bar{u}$ on a nonexceptional set $K \subset F$, then $\widehat{u}_{K^c} = \bar{u}_{K^c}$ q.e. on K, and hence $\widehat{u}_K > \bar{u}_K$ q.e. on K. This implies $\mu_{\bar{u}}(K) > 0$, which is impossible because $\mu_{\bar{u}}(F) = 0$. Therefore

$$(\widehat{u} - \widehat{v}) - (\bar{u} - \bar{v}) = -(\widehat{v} - \bar{v}) < 0$$

q.e. on F which contradicts the assertion of Step 2. □

Since $\bar{u}$ is finely continuous, $\bar{u}(Z_t)$ is right continuous a.s. P_z for q.e. z. Similarly, $\widehat{u}(\widehat{Z}_t)$ is right continuous a.s. $\widehat{P}_z$ for q.e. z. Since $\bar{u} = \widehat{u}$, it becomes continuous along the sample paths. In fact, for any $f, g \ge 0$, and $t > 0$,

$$\begin{aligned} &E_{f \cdot \nu} \left(g(Z_t) : \bar{u}(Z_s)_- \ne \bar{u}(Z_s), \exists s \in (0, t) \right) \\ &= \widehat{E}_{g \cdot \nu} \left(f(\widehat{Z}_t) : \bar{u}(\widehat{Z}_s)_+ \ne \bar{u}(\widehat{Z}_s), \exists s \in (0, t) \right) \\ &= \widehat{E}_{g \cdot \nu} \left(f(\widehat{Z}_t) : \widehat{u}(\widehat{Z}_s)_+ \ne \widehat{u}(\widehat{Z}_s), \exists s \in (0, t) \right) = 0. \end{aligned}$$

The similar result also holds for $\bar{v}$. Hence

$$P_z \left(\bar{u}(Z_t) \text{ and } \bar{v}(Z_t) \text{ are continuous for } t > 0 \right) = 1 \tag{3.5}$$

for a.e. z. By operating the transition function p_s and letting $s \to 0$, (3.5) holds for q.e. z.

Let $\mathcal{J} = \{u = u_1 - u_2 + w; u_i \in \mathcal{P}, w \in \mathcal{W}\}$. As in [14, section 5], $\mathcal{E}$ can be extended to $\mathcal{J} \times \mathcal{J}$ by $\mathcal{E}(u,v) = \lim_{\alpha\to\infty} \mathcal{E}(\alpha G_\alpha u, v)$ for $u, v \in \mathcal{J}$.

LEMMA 3.4. *The function $\bar{w} := \bar{u} - \bar{v}$ is the unique function of $\mathcal{J}$ such that $\bar{w} = \widehat{w}$, $g \le \bar{w} \le h$, and, for any $w \in \mathcal{J}$ satisfying $g \le w \le h$,*

$$\mathcal{A}_1(\bar{w}, \bar{w}) \le \mathcal{E}_1(\bar{w}, w). \tag{3.6}$$

Proof. For any $w \in \mathcal{J}$ such that $g \le w$, since

$$\mathcal{E}_1(\bar{u}_n, \bar{u}_n - \bar{v} - w) = \frac{1}{\varepsilon_n}\left((\bar{u}_n - \bar{v} - g)^-, \bar{u}_n - \bar{v} - w\right) \le 0,$$

it holds that

$$\mathcal{A}_1(\bar{u}, \bar{u}) \le \lim_{n\to\infty} \mathcal{E}_1(\bar{u}_n.\bar{u}_n) \le \lim_{n\to\infty} \left(\mathcal{E}_1(\bar{u}_n, \bar{v}) + \mathcal{E}_1(\bar{u}_n, w)\right). \tag{3.7}$$

For any $p \in \mathcal{P}$, since $\mathcal{E}_1(\bar{u}_n, p) = -\mathcal{E}_1(p, \bar{u}_n) + 2\mathcal{A}_1(\bar{u}_n, p)$ and $\alpha G_{\alpha+1}\bar{u}_n$ is increasing relative to α and n, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathcal{E}_1(\bar{u}_n, p) &= -\lim_{n\to\infty}\lim_{\alpha\to\infty} \mathcal{E}_1(p, \alpha G_{\alpha+1}\bar{u}_n) + 2\mathcal{A}_1(\bar{u}, p) \\
&= -\lim_{\alpha\to\infty} \mathcal{E}_1(p, \alpha G_{\alpha+1}\bar{u}) + 2\lim_{\alpha\to\infty} \mathcal{A}_1(\alpha G_{\alpha+1}\bar{u}, p) \\
&= \lim_{\alpha\to\infty} \mathcal{E}_1(\alpha G_{\alpha+1}\bar{u}, p) = \mathcal{E}_1(\bar{u}, p).
\end{aligned}$$

This relation can be extended to all $p \in \mathcal{J}$, and hence by (3.7),

$$\mathcal{A}_1(\bar{u}, \bar{u}) \le \mathcal{E}_1(\bar{u}, \bar{v}) + \mathcal{E}_1(\bar{u}, w).$$

Furthermore, since $\widehat{v} = \bar{v}$ q.e. from Lemma 3.3,

$$\begin{aligned}
\mathcal{E}_1(\bar{u}, \bar{v}) &= \lim_{\alpha\to\infty} \mathcal{E}_1(\alpha G_{\alpha+1}\bar{u}, \bar{v}) = \lim_{\alpha\to\infty} \mathcal{E}_1\left(\bar{u}, \alpha\widehat{G}_{\alpha+1}\bar{v}\right) \\
&= \int \widehat{v} d\mu_{\bar{u}} = \int \bar{v} d\mu_{\bar{u}} \\
&= \lim_{\alpha\to\infty} \mathcal{E}_1(\bar{u}, \alpha G_{\alpha+1}\bar{v}).
\end{aligned} \tag{3.8}$$

Thus we get that

$$\mathcal{A}_1(\bar{u}, \bar{u}) \le \lim_{\alpha\to\infty} \mathcal{E}_1(\bar{u}, \alpha G_{\alpha+1}\bar{v}) + \mathcal{E}_1(\bar{u}, w).$$

Similarly, for any $w \in \mathcal{J}$ such that $w \le h$,

$$\mathcal{A}_1(\bar{v}, \bar{v}) \le \lim_{\alpha\to\infty} \mathcal{E}_1(\alpha G_{\alpha+1}\bar{v}, \bar{u}) - \mathcal{E}_1(\bar{v}, w).$$

Therefore, for any $w \in \mathcal{J}$ such that $g \le w \le h$,

$$\begin{aligned}
\mathcal{E}_1(\bar{u} - \bar{v}, w) &\ge \mathcal{A}_1(\bar{u}, \bar{u}) + \mathcal{A}_1(\bar{v}, \bar{v}) - \lim_{\alpha\to\infty} \{\mathcal{E}_1(\bar{u}, \alpha G_{\alpha+1}\bar{v}) + \mathcal{E}_1(\alpha G_{\alpha+1}\bar{v}, \bar{u})\} \\
&= \mathcal{A}_1(\bar{u}, \bar{u}) + \mathcal{A}_1(\bar{v}, \bar{v}) - 2\lim_{\alpha\to\infty} \mathcal{A}_1(\bar{u}, \alpha G_{\alpha+1}\bar{v}) \\
&= \mathcal{A}_1(\bar{u} - \bar{v}, \bar{u} - \bar{v}),
\end{aligned}$$

that is, (3.6) holds. To prove the uniqueness of the solution, suppose that $w_1, w_2 \in \mathcal{J}$ satisfy the properties of the lemma. Since (3.8) holds for w_1 and w_2 instead of $\bar{u}$ and $\bar{v}$, respectively,

$$\begin{aligned}\mathcal{A}_1(w_1, w_2) + \mathcal{A}_1(w_2.w_1) &= \lim_{\alpha\to\infty} \left(\mathcal{A}_1(w_1, \alpha G_{\alpha+1} w_2) + \mathcal{A}_1(\alpha G_{\alpha+1} w_2, w_1)\right) \\ &= \lim_{\alpha\to\infty} \left(\mathcal{E}_1(w_1, \alpha G_{\alpha+1} w_2) + \mathcal{E}_1(\alpha G_{\alpha+1} w_2, w_1)\right) \\ &= \mathcal{E}_1(w_1, w_2) + \mathcal{E}_1(w_2, w_1).\end{aligned}$$

Hence, from (3.6),

$$\begin{aligned}\mathcal{A}_1(w_1 - w_2, w_1 - w_2) &= \mathcal{A}_1(w_1, w_1) + \mathcal{A}_1(w_2, w_2) - \mathcal{A}_1(w_1, w_2) - \mathcal{A}_1(w_2, w_1) \\ &= \mathcal{A}_1(w_1, w_1) + \mathcal{A}_1(w_2, w_2) - \mathcal{E}_1(w_1, w_2) - \mathcal{E}_1(w_2, w_1) \le 0,\end{aligned}$$

which implies $w_1 = w_2$ a.e. □

Put $\dot{\sigma}_k = \dot{\sigma}_{C_k}$ and $\dot{\tau}_k = \dot{\sigma}_{D_k}$. Since C_k and D_k are decreasing, $\dot{\sigma} = \lim_{k\to\infty} \sigma_k$ and $\dot{\tau} = \lim_{k\to\infty} \tau_k$ exist as increasing limits. Clearly $\dot{\sigma} \le \dot{\sigma}_C$ and $\dot{\tau} \le \dot{\sigma}_D$.

LEMMA 3.5. *For q.e. $z \in Z$, $\dot{\sigma}_C = \dot{\sigma}$ and $\dot{\sigma}_D = \dot{\tau}$ a.s. P_z.*

Proof. We shall prove only the assertion for $\dot{\sigma}_C$. For any $\ell \le k$,

$$\bar{u}_\ell(Z_{\dot{\sigma}_k}) \le \bar{u}_k(Z_{\dot{\sigma}_k}) \le (g + \bar{v})(Z_{\dot{\sigma}_k}).$$

Hence by letting $k \uparrow \infty$ and then $\ell \uparrow \infty$, we get from (3.5) that $\bar{u}(Z_{\dot{\sigma}}) \le (g + \bar{v})(Z_{\dot{\sigma}})$. Hence $\dot{\sigma}_C \le \dot{\sigma}$. □

Since $\bar{u}$ and $\bar{v}$ are 1-excessive,

$$\bar{u}(z) \ge E_z\left(e^{-\sigma}\bar{u}(Z_\sigma)\right), \tag{3.9}$$

$$\bar{v}(z) \ge E_z\left(e^{-\tau}\bar{v}(Z_\tau)\right) \tag{3.10}$$

for any stopping times σ and τ. From the definition, $\bar{u}_k = \frac{1}{\varepsilon_k} R_1\left((\bar{u}_k - g - \bar{v})^-\right)$ q.e. Hence for any stopping time σ such that $\sigma \le \dot{\sigma_k}$,

$$\begin{aligned}\bar{u}_k(z) &= \frac{1}{\varepsilon_k} E_z\left(\int_0^\infty e^{-t}(\bar{u}_k - g - \bar{v})^-(Z_t)dt\right) \\ &= \frac{1}{\varepsilon_k} E_z\left(\int_\sigma^\infty e^{-t}(\bar{u}_k - g - \bar{v})^-(Z_t)dt\right) \\ &= E_z\left(e^{-\sigma}\bar{u}_k(Z_\sigma)\right) \quad \text{q.e. } z. \qquad (3.11)\end{aligned}$$

Similarly, if $\tau \le \dot{\tau}_k$, then

$$\bar{v}_k(z) = E_z\left(e^{-\tau}\bar{v}_k(Z_\tau)\right). \tag{3.12}$$

THEOREM 3.6. *Suppose that g and h are q.c. functions of $\mathcal{F}$ satisfying the separability condition. Then*

$$\bar{u}(z) - \bar{v}(z) = \sup_\sigma \inf_\tau J_z(\sigma, \tau) = \inf_\tau \sup_\sigma J_z(\sigma, \tau) \quad \text{q.e.} \tag{3.13}$$

Furthermore, $(\dot{\sigma}_C, \dot{\sigma}_D)$ is the saddle point of $J_z(\sigma, \tau)$, and $\bar{u} - \bar{v}$ is the unique solution of the quasi-variational inequality (3.6).

Proof. For any stopping time τ, applying (3.10) and (3.11) for $\dot{\sigma}_k \wedge \tau \leq \dot{\sigma}_k$, we have

$$\begin{aligned}\bar{u}_k(z) - \bar{v}(z) &\leq E_z\left(e^{-\dot{\sigma}_k \wedge \tau}(\bar{u}_k - \bar{v})(Z_{\dot{\sigma}_k \wedge \tau})\right) \\ &= E_z\left(e^{-\dot{\sigma}_k}(\bar{u}_k - \bar{v})(Z_{\dot{\sigma}_k}) : \dot{\sigma}_k \leq \tau\right) + E_z\left(e^{-\tau}(\bar{u}_k - \bar{v})(Z_\tau) : \tau < \dot{\sigma}_k\right) \\ &\leq E_z\left(e^{-\dot{\sigma}_k} g(Z_{\dot{\sigma}_k}) : \dot{\sigma}_k \leq \tau\right) + E_z\left(e^{-\tau} h(Z_\tau) : \tau < \dot{\sigma}_k\right).\end{aligned}$$

Then by letting $k \to \infty$, we have

$$\bar{u}(z) - \bar{v}(z) \leq E_z\left(e^{-\dot{\sigma}} g(Z_{\dot{\sigma}}) : \dot{\sigma} \leq \tau\right) + E_z\left(e^{-\tau} h(Z_\tau) : \tau < \dot{\sigma}\right).$$

Similarly, for any stopping time σ, by considering $\sigma \wedge \dot{\tau}_k$, it holds that

$$E_z\left(e^{-\sigma} g(Z_\sigma) : \sigma < \dot{\tau}\right) + E_z\left(e^{-\dot{\tau}} h(Z_{\dot{\tau}}) : \dot{\tau} \leq \sigma\right) \leq \bar{u}(z) - \bar{v}(z).$$

Since $\dot{\sigma} = \dot{\sigma}_C$ and $\dot{\tau} = \dot{\sigma}_D$ from Lemma 3.5, the assertion of the theorem follows. ∎

(III) General two obstacles case:

In this case, we assume that g and h are q.c. functions of $\mathcal{W}$. As in the preceding section, put $g^{(k)} = kR_k g$ and $h^{(k)} = kR_k h$. Then the separability condition holds for the obstacles $(g^{(k)}, h^{(k)})$. For any stopping times σ and τ, put

$$J_z^{(k)}(\sigma, \tau) = E_z\left(e^{-\sigma \wedge \tau}\left(g^{(k)}(Z_\sigma) I_{\{\sigma \leq \tau\}} + h^{(k)}(Z_\tau) I_{\{\tau < \sigma\}}\right)\right). \tag{3.14}$$

Let $(\bar{u}^{(k)}, \bar{v}^{(k)})$ be the 1-excessive modifications of the minimal pair of functions determined by Theorem 2.2 for $(g^{(k)}, h^{(k)})$ and put $\bar{w}^{(k)} = \bar{u}^{(k)} - \bar{v}^{(k)}$. Then $\widehat{w}^{(k)} = \bar{w}^{(k)}$ q.e. and satisfies

$$\bar{w}^{(k)}(z) = \sup_\sigma \inf_\tau J_z^{(k)}(\sigma, \tau) = \inf_\tau \sup_\sigma J_z^{(k)}(\sigma, \tau), \tag{3.15}$$

where $\widehat{w}^{(k)} = \lim_{n \to \infty} n\widehat{R}_n \bar{w}^{(k)}$. By virtue of (3.5), $\bar{w}^{(k)}$ is continuous along the sample paths a.s.

THEOREM 3.7. *Suppose that g, h are q.c. functions of $\mathcal{W}$. Then*

$$\bar{w}(z) := \sup_\sigma \inf_\tau J_z(\sigma, \tau) = \inf_\tau \sup_\sigma J_z(\sigma, \tau) \tag{3.16}$$

belongs to $\mathcal{F}$ and satisfies, for any $w \in \mathcal{W}$ such that $g \leq w \leq h$,

$$\mathcal{A}_1(\bar{w}, \bar{w}) \leq \mathcal{E}_1(\bar{w}, w). \tag{3.17}$$

Moreover, the pair of the entry times $(\dot{\sigma}_C, \dot{\sigma}_D)$ is a saddle point of $J_z(\sigma, \tau)$.

Proof. For any stopping times σ, τ, and q.e. z,

$$E_z\left(e^{-\sigma}|g^{(k)} - g|(Z_\sigma) I_{\{\sigma \leq \tau\}}\right) \leq E_z\left(e^{-\sigma} \widetilde{e}_{|g^{(k)} - g|}(Z_\sigma)\right) \leq \widetilde{e}_{|g^{(k)} - g|}(z) \tag{3.18}$$

and

$$E_z\left(e^{-\tau}|h^{(k)} - h|(Z_\tau) I_{\{\tau < \sigma\}}\right) \leq E_z\left(e^{-\tau} \widetilde{e}_{|h^{(k)} - h|}(Z_\tau)\right) \leq \widetilde{e}_{|h^{(k)} - h|}(z). \tag{3.19}$$

Hence

$$|J_z^{(k)}(\sigma, \tau) - J_z(\sigma, \tau)| \leq \widetilde{e}_{|g^{(k)} - g|}(z) + \widetilde{e}_{|h^{(k)} - h|}(z) \text{ q.e.}$$

By virtue of Lemmas 1.5 and 2.3, there exists an increasing sequence of closed sets $\{F_n\}$ and a subsequence $n_k \uparrow \infty$ such that $\mathrm{Cap}(Z\backslash F_n) \to 0$ and $\lim_{k\to\infty} \widetilde{e}_{|g^{(n_k)}-g|} = 0$, $\lim_{k\to\infty} \widetilde{e}_{|h^{(n_k)}-h|} = 0$ in $\mathcal{F}$ and uniformly on each set F_n. Hence

$$\lim_{k\to\infty} \inf_{\tau} \sup_{\sigma} J_z^{(n_k)}(\sigma,\tau) = \inf_{\tau} \sup_{\sigma} J_z(\sigma,\tau) \quad \text{on } F_n.$$

Equation (3.15) combined with (3.18) and (3.19) implies that

$$|\bar{w}^{(k)} - \bar{w}^{(l)}| \le \widetilde{e}_{|g^{(k)}-g|} + \widetilde{e}_{|g^{(l)}-g|} + \widetilde{e}_{|h^{(k)}-h|} + \widetilde{e}_{|h^{(l)}-h|} \text{ q.e.}$$

Therefore $\lim_{k\to\infty} \bar{w}^{(k)} = \bar{w}^{(\infty)}$ exists in $\mathcal{H}$ and a subsequence converges quasi-uniformly on each F_n and satisfies

$$\bar{w}^{(\infty)} = \sup_{\sigma} \inf_{\tau} J_z(\sigma,\tau) = \inf_{\tau} \sup_{\sigma} J_z(\sigma,\tau) \quad \text{q.e. } z \in F_n.$$

Letting $n \to \infty$, we get that $\bar{w}^{(\infty)} = \bar{w}$ q.e. By virtue of the quasi-variational inequality (3.6) applied for $\bar{w}^{(k)}$,

$$\begin{aligned}\mathcal{A}_1(\bar{w}^{(k)},\bar{w}^{(k)}) &\le \mathcal{E}_1(\bar{w}^{(k)},g^{(k)}) = \mathcal{E}_1(k\widehat{G}_k\bar{w}^{(k)},g)\\ &\le \|g\|_{\mathcal{W}}\|k\widehat{G}_k\bar{w}^{(k)}\|_{\mathcal{F}} \le C_1\|g\|_{\mathcal{W}}\|\bar{w}^{(k)}\|_{\mathcal{F}}.\end{aligned}$$

Hence $\|\bar{w}^{(k)}\|_{\mathcal{F}}$ and $\|k\widehat{G}_k\bar{w}^{(k)}\|_{\mathcal{F}}$ are uniformly bounded relative to k. This combined with $\lim_{k\to\infty} \bar{w}^{(k)} = \lim_{k\to\infty} k\widehat{G}_k\bar{w}^{(k)} = \bar{w}$ in $\mathcal{H}$ implies that $\lim_{k\to\infty} \bar{w}^{(k)} = \lim_{k\to\infty} k\widehat{G}_k\bar{w}^{(k)} = \bar{w}$ weakly in $\mathcal{F}$. Therefore, for any $w \in \mathcal{W}$ such that $g \le w \le h$,

$$\mathcal{A}_1(\bar{w}^{(k)},\bar{w}^{(k)}) \le \mathcal{E}_1(\bar{w}^{(k)}, kG_kw) = \mathcal{E}_1(k\widehat{G}_k\bar{w}^{(k)},w).$$

By letting $k \to \infty$, the quasi-variational inequality (3.17) follows.

The proof of the last part of the theorem is similar to [19]. It suffices to prove the following inequalities for any stopping times σ and τ such that $\tau \le \dot{\sigma}_C$ and $\sigma \le \dot{\sigma}_D$:

$$\bar{w}(z) \le E_z\left(e^{-\tau}\bar{w}(Z_\tau)\right) \quad \text{q.e.} \tag{3.20}$$

$$\bar{w}(z) \ge E_z\left(e^{-\sigma}\bar{w}(Z_\sigma)\right) \quad \text{q.e.} \tag{3.21}$$

In fact, if these hold, by noting $\bar{w}(Z_{\dot{\sigma}_C}) = g(Z_{\dot{\sigma}_C})$, we have for any stopping time τ,

$$\begin{aligned}J_z(\dot{\sigma}_C,\tau) &= E_z\left(e^{-\dot{\sigma}_C}g(Z_{\dot{\sigma}_C})I_{\{\dot{\sigma}_C\le\tau\}} + e^{-\tau}h(Z_\tau)I_{\{\tau<\dot{\sigma}_C\}}\right)\\ &\ge E_z\left(e^{-\dot{\sigma}_C}\bar{w}(Z_{\dot{\sigma}_C})I_{\{\dot{\sigma}_C\le\tau\}} + e^{-\tau}\bar{w}(Z_\tau)I_{\{\tau<\dot{\sigma}_C\}}\right)\\ &= E_z\left(e^{-\dot{\sigma}_C\wedge\tau}\left(\bar{w}(Z_{\dot{\sigma}_C\wedge\tau})\right)\right) \ge \bar{w}(z).\end{aligned}$$

Similarly, $J_z(\sigma,\dot{\sigma}_D) \le \bar{w}$ q.e. for any stopping time σ.

Let $C^{(k)} = \{z : \bar{w}^{(k)}(z) \le g^{(k)}(z)\}$, $D^{(k)} = \{z : h^{(k)}(z) \le \bar{w}^{(k)}(z)\}$, $\dot{\sigma}^{(k)} = \dot{\sigma}_{C^{(k)}}$, and $\dot{\tau}^{(k)} = \dot{\sigma}_{D^{(k)}}$. Further, for each positive number γ, let $\eta_\gamma = \inf\{t \ge 0 : \bar{w}(Z_t)+\gamma \ge h(Z_t)\}$. If $t < \eta_\gamma(\omega)$, then $\bar{w}(Z_t(\omega)) + \gamma < h(Z_t(\omega))$. Noting that $\lim_{k\to\infty} \bar{w}^{(k)} = \bar{w}$ and $\lim_{k\to\infty} h^{(k)} = h$ uniformly on each F_n, we can find k_0 such that $|\bar{w}^{(k)} - \bar{w}| \le \gamma/2$ and $|h^{(k)} - h| < \gamma/2$ on F_n for any $k \ge k_0$. Therefore $t < \dot{\tau}^{(k)}$, and hence $\eta_\gamma \wedge \dot{\sigma}_{F_n} \le \dot{\tau}^{(k)}$. In view of (3.9) and (3.12), since (3.21) holds for $\bar{w}^{(k)}$ and $\dot{\tau}^{(k)}$ instead of $\bar{w}$ and $\dot{\sigma}_D$, respectively, it then holds that

$$\bar{w}^{(k)}(z) \ge E_z\left(e^{-(\eta_\gamma\wedge\tau_{F_n}\wedge\sigma)}w^{(k)}(Z_{\eta_\gamma\wedge\tau_{F_n}\wedge\sigma})\right) \quad \text{q.e.}$$

Letting $k \to \infty$ and then $n \to \infty$, we have

$$\bar{w}(z) \geq E_z\left(e^{-(\eta_\gamma \wedge \sigma)}\bar{w}(Z_{\eta_\gamma \wedge \sigma})\right) \quad \text{q.e.}$$

Since $\bar{w}(Z_{\eta_\gamma}) + \gamma \geq h(Z_{\eta_\gamma})$ and η_γ is increasing as $\gamma \downarrow 0$, $\widehat{w}(Z_{\eta_0}) = \bar{w}(Z_{\eta_0}) \geq h(Z_{\eta_0})$ for $\eta_0 = \lim_{\gamma \to 0} \eta_\gamma$. This yields that $\dot{\sigma}_D \leq \eta_0$, and hence

$$\bar{w}(z) \geq E_z\left(e^{-\sigma}\bar{w}(Z_\sigma)\right)$$

for any stopping time σ such that $\sigma \leq \dot{\sigma}_D$. The proof of (3.20) is similar. □

Example. Let (Z_t, P_z) be the space-time diffusion process corresponding to the generator $A^{(t)}\varphi(x) = \frac{d^2\varphi}{dx^2}$ for $t < 1$ and $A^{(t)}\varphi(x) = \frac{1}{2}\frac{d^2\varphi}{dx^2}$ for $t \geq 1$. Denote by q_t and K_α the transition function and resolvent of a one-dimensional Brownian motion, respectively. Taking a nonnegative, nonzero continuous function ϕ on R^1 with compact support, let $\varphi(x) = K_2\phi(x)$,

$$g(s,x) = \begin{cases} e^{-(3-2s)}q_{2(1-s)}\varphi(x), & s < 1, \\ e^{-s}\varphi(x), & s \geq 1, \end{cases}$$

and

$$h(s,x) = \begin{cases} 2e^{-(3-s)/2}q_{2(1-s)}\varphi(x), & s < 1, \\ 2e^{-s}\varphi(x), & s \geq 1. \end{cases}$$

Furthermore, let $\bar{v} = 0$ and

$$\bar{u}(s,x) = \begin{cases} e^{-(2-s)}q_{2(1-s)}\varphi(x), & s < 1, \\ e^{-s}\varphi(x), & s \geq 1. \end{cases}$$

Then $g \leq \bar{u} - \bar{v} \leq h$. Since $\bar{u}(s,x) = E_{(s,x)}\left(\int_0^\infty e^{-t}\xi \otimes \phi(Z_t)dt\right)$, for $\xi(t) = e^{-t}I_{\{t \geq 1\}}$, $\bar{u}$ is 1-excessive function of Z_t and satisfies

$$\mathcal{A}_1(\bar{u}, \bar{u}) = \frac{1}{2}e^{-2}\int_{-\infty}^{\infty} \varphi(x)\phi(x)dx.$$

Further, for any $w \in \mathcal{J}$ such that $g \leq w \leq h$,

$$\mathcal{E}_1(\bar{u}, w) = \int_1^\infty \int_{-\infty}^\infty e^{-s}\phi(x)w(s,x)dxds \geq \int_1^\infty \int_{-\infty}^\infty e^{-s}\phi(x)g(s,x)dxds = \mathcal{A}_1(\bar{u}, \bar{u}).$$

Therefore $\bar{u} - \bar{v} = \bar{u}$ is a solution of (3.6), and the saddle point $(\dot{\sigma}_C, \dot{\sigma}_D)$ of $J_z(\sigma, \tau)$ in Theorem 3.6 is given by $\dot{\sigma}_C = (1 - \tau(0)) \vee 0$ and $\dot{\sigma}_D = \infty$.

Acknowledgments. The author wishes to thank the anonymous referees for their valuable suggestions by which the proofs are improved and the results are generalized.

REFERENCES

[1] A. Bensoussan and L. L. Lions, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.

[2] R. M. Blumenthal and R. K. Getoor, *Markov Processes and Potential Theory*, Academic Press, New York, 1968.

[3] P. J. Fitzsimmons, *Absolute continuity of symmetric diffusions*, Ann. Probab., 25 (1997), pp. 230–258.

[4] M. Fukushima, Y. Oshima, and M. Takeda, *Dirichlet Forms and Symmetric Markov Processes*, de Gruyter Stud. Math. 19, Walter de Gruyter and Co., Berlin, 1994.

[5] M. Fukushima and M. Taksar, *Dynkin games via Dirichlet forms and singular control of one-dimensional diffusions*, SIAM J. Control Optim., 41 (2002), pp. 682–699.

[6] Z. M. Ma and M. Röckner, *Introduction to the Theory of (Nonsymmetric) Dirichlet Forms*, Universitext, Springer-Verlag, Berlin, 1992.

[7] J. L. Menaldi, *On the optimal impulse control problem for degenerate diffusions*, SIAM J. Control Optim., 18 (1980), pp. 722–739.

[8] K. Menda, *On Dynkin Games, Their Refined Solutions, and a Related One-Dimensional Stefan Problem*, Master's thesis, Osaka University, Osaka, Japan, 2004.

[9] F. Mignot and J. P. Puel, *Inequations d'evolution paraboliques avec convexes dépendant du temps, Applications aux inequations quasi-variationelles d'evolution*, Arch. Rational Mech. Anal., 64 (1977), pp. 59–91.

[10] H. Nagai, *On an optimal stopping problem and a variational inequalities*, J. Math. Soc. Japan, 30 (1978), pp. 303–312.

[11] H. Nagai, *Nonzero-sum stopping games of symmetric Markov processes*, Probab. Theory Related Fields, 75 (1987), pp. 487–497.

[12] Y. Oshima, *On a construction of Markov processes associated with time dependent Dirichlet spaces*, Forum Math., 4 (1992), pp. 395–415.

[13] Y. Oshima, *On the exceptionality of some semipolar sets of time inhomogeneous Markov processes*, Tohoku Math. J. (2), 54 (2002), pp. 443–449.

[14] Y. Oshima, *Time dependent Dirichlet forms and related stochastic calculus*, Infin. Dimens. Anal. Quantum Probab. Relat. Top., 7 (2004), pp. 281–316.

[15] M. Pierre, *Représentant précis d'un potentiel parabolique*, in Séminaire Théorie du Potentiel, Université Paris VI, Lecture Notes in Math. 814, Springer, Berlin, 1980, pp. 186–228.

[16] W. Stannat, *The theory of generalized Dirichlet forms and its applications in analysis and stochastics*, Mem. Amer. Math. Soc., 142 (1999), no. 678.

[17] G. Trutnau, *Stochastic calculus of generalized Dirichlet forms and applications to stochastic differential equations in infinite dimensions*, Osaka J. Math., 37 (2000), pp. 315–343.

[18] G. Trutnau, *On a class of nonsymmetric diffusions containing fully nonsymmetric distorted Brownian motions*, Forum Math., 15 (2003), pp. 409–437.

[19] J. Zabczyk, *Stopping games for symmetric Markov processes*, Probab. Math. Statistics, 4 (1984), pp. 185–196.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 580–598

A VARIATIONAL INEQUALITY SUFFICIENT CONDITION FOR OPTIMAL STOPPING WITH APPLICATION TO AN OPTIMAL STOCK SELLING PROBLEM*

RAYMOND RISHEL† AND KURT HELMES‡

Abstract. We give a variational inequality sufficient condition for optimal stopping problems. This result is illustrated by computing solutions to an optimal stock selling problem. The stock selling problem has a model for the stock price, which initially has a "hot" growth rate and then "tanks." Solution of the conditions allows for computation of both the value function and the optimal stopping times. Optimal stock selling times are computed for two utility functions and a variety of realistic parameter values.

Key words. stock selling times, optimal stopping, variational inequalities

AMS subject classifications. 49J40, 49K15, 49K20, 60G35, 60G40, 62L15, 91B70

DOI. 10.1137/050622699

1. Introduction. Variational inequalities for solving optimal stopping problems were first introduced by Bensoussan and Lions [1], [2]. Many variational inequality sufficient conditions have since been given. Some examples of these are Bensoussan [1, p. 301], Krylov [7, Thm. 7, p. 41], and Oksendal [9, p. 225].

The sufficiency condition we give is an extremely simple extension of that of [1]. However, it points out the apparently previously overlooked requirement of a boundedness condition on the gradient of the value function in determining the correct solution from the possibly many solutions of the variational inequality. The use of this property is illustrated in our solution of the stock selling problem.

A number of different stock selling problems have been considered. For instance, in [9, pp. 219, 227], Oksendal considers a stock selling problem whose stock price is given by a geometric Brownian motion with constant coefficients. In [12], Zhang considers a stock selling problem whose stock price is given by a diffusion process whose coefficients are unobservable finite state jump Markov processes. In [3], Beibel and Lerche consider an optimal stock selling problem for a model which is basically the same as the partially observed model we consider below. We discuss comparison of Zhang's results and Beibel and Lerche's results with ours in section 9.

The problem we consider is motivated by the past behavior of "bubble stocks" such as Enron. The problem is to decide when to sell a stock which has a rapid growth rate and then a rapid rate of decline but, because of randomness in the stock price, it is difficult for the investor to know when the stock price is really in decline. We consider an idealized model of such a situation in which the stock price is given by a geometric Brownian motion, which has an initial positive growth rate, and after a random time, jumps to and remains at a negative growth rate. The investor observes the stock prices but cannot observe the growth rates. Deciding when to sell is based

*Received by the editors January 14, 2005; accepted for publication (in revised form) November 28, 2005; published electronically May 3, 2006.
http://www.siam.org/journals/sicon/45-2/62269.html
†Department of Mathematics, University of Kentucky, Lexington, KY 40506-0027 (rishel@ms.uky.edu).
‡Institute of Operations Research, Humboldt University of Berlin, D-10178 Berlin, Germany (helmes@wiwi.hu-berlin.de).

only on past stock prices. Thus this is a partially observed optimal stopping problem. Converting it to a completely observed optimal stopping problem is done similarly to the methods used by Shiryayev [10, p. 200] in his quickest detection of a disorder problem.

We consider the problem with two different utility functions $U(S) = \ln(S)$ and $U(S) = S$. For the utility function $U(S) = \ln(S)$, there are solutions of the variational inequality with the appropriate boundedness, and optimal stopping times are calculated. For the utility function $U(S) = S$, it appears to be difficult to find solutions of the variational inequality with the appropriate boundedness for the problem in its original form. We use two changes of probability measure to convert the problem into one for which these boundedness conditions are determined and calculations of the optimal stopping time can be carried out.

2. A variational inequality sufficient condition for optimal stopping. We begin by stating an optimal stopping problem and giving variational inequality sufficient conditions for it. Let $g(z)$ and $h(z)$ denote, respectively, an n-dimensional vector valued function and an $(n \times m)$-dimensional matrix valued function of the n-dimensional vector z. Let $W(t)$ denote an m-dimensional Wiener process. Assume $g(z)$ and $h(z)$ are regular enough so that solutions of the stochastic differential equation and initial condition

$$dz(t) = g(z(t))\,dt + h(z(t))\,dW(t), \quad z(0) = z, \tag{2.1}$$

exist and are unique. Let $\mathcal{F}_t$ denote the σ-fields

$$\mathcal{F}_t = \sigma[z(r) : \ 0 \le r \le t] \tag{2.2}$$

generated by the past of $z(t)$. Let $\beta(z)$ be a continuous scalar function and $U(z)$ a twice continuously differentiable utility function. Let $\mathcal{A}$ denote a class of $\mathcal{F}_t$ stopping times. For each stopping time $\tau \in \mathcal{A}$, consider the expected discounted utility

$$E\Big[e^{\int_0^\tau \beta(z(t))\,dt} U(z(\tau))\Big]. \tag{2.3}$$

The optimal stopping problem is as follows: Find τ in $\mathcal{A}$ which achieves the maximum of (2.3).

The following theorem gives variational inequality sufficient conditions for optimality for this problem.

THEOREM 2.1. *Let R be a region in E^n. Assume for each z in R that the solution of* (2.1) *with initial condition z is contained in R. Let $V(z)$ be a scalar valued function defined on R. Let $V(z)$ be regular enough so that Itô's stochastic differential rule holds for $V(z(t))$. Define the differential operator $A[V](z)$ by*

$$A[V](z) = \beta(z)V(z) + V_z(z)g(z) + \frac{1}{2}\,\text{trace}\,(h(z)h(z)'V_{zz}(z)). \tag{2.4}$$

Let $V(z)$ be a solution of the variational inequality

$$\begin{aligned} &A[V](z) \le 0, \quad V(z) \ge U(z), \\ &(V(z) - U(z))\ A[V](z) = 0, \end{aligned} \tag{2.5}$$

and let the condition

$$E\int_0^\tau e^{2\int_0^t \beta(z(r))\,dr} \|V_z(z(t))h(z(t))\|^2\,dt < \infty \tag{2.6}$$

hold for each stopping time τ in $\mathcal{A}$. For $z(t)$ the solution of (2.1) *with initial condition z, let*

$$\tau(z) = \textit{ first time } z(t) \textit{ hits } \{q:\ V(q) = U(q)\}. \tag{2.7}$$

Let

$$\tau(z) \in \mathcal{A} \textit{ for each } z \in R. \tag{2.8}$$

Then

$$V(z) = E\Big[e^{\int_0^{\tau(z)} \beta(z(r))\,dr} U(z(\tau(z)))\Big] = \max_{\tau\in\mathcal{A}} E\Big[e^{\int_0^{\tau} \beta(z(r))\,dr}\, U(z(\tau))\Big]. \tag{2.9}$$

That is, $\tau(z)$ is an optimal stopping time in $\mathcal{A}$ and $V(z)$ is the value function for the optimal stopping problem.

Theorem 2.1 should be known. Its simple proof uses that condition (2.6) implies that the expected value of the stochastic integral term in the Itô differential of $e^{\int_0^t \beta(z(r))\,dr}\, V(z(t))$ is zero. Then standard variational inequality arguments such as in Bensoussan [1, p. 201] give the conclusion. For completeness we give the short proof in an appendix. However, the importance of conditions (2.6) and (2.8) holding along with the variational inequality conditions (2.5) does not appear to have been pointed out. Conditions (2.6) and (2.8) make the correct selection of the possibly many solutions of conditions (2.5).

Remark 2.2. Notice that if

$$\beta(z) \le 0 \text{ and } \|V_z(z)h(z)\|^2 \le K, \tag{2.10}$$

then

$$E\left[\int_0^{\tau} e^{2\int_0^t \beta(z(r))\,dr}\, \|V_z(z(t))h(z(t))\|^2\,dt\right] \le KE(\tau). \tag{2.11}$$

Thus if the boundedness condition (2.10) holds, (2.6) holds for stopping times with finite expectations.

Remark 2.2 motivates looking for solutions of (2.5) for which the boundedness condition (2.10) holds.

3. The stock price model. Consider a stock whose price $S(t)$ satisfies the stochastic differential equation and initial condition

$$dS(t) = S(t)(a(t)dt + \sigma dW(t)), \quad S(0) = S. \tag{3.1}$$

In (3.1) S and σ are positive constants, $W(t)$ is a Wiener process, and $a(t)$ is a random process which jumps from $a > 0$ to $b < 0$ at a random time J, which satisfies

$$Pr[J = 0] = 1 - x, \quad Pr[J > t | J > 0] = e^{-ct}, \tag{3.2}$$

where x and c are constants satisfying $0 \le x \le 1$ and $c > 0$.

Let $U(S)$ be a utility function. Let

$$\mathcal{F}_t = \sigma\{S(r):\ 0 \le r \le t\} \tag{3.3}$$

denote the σ-fields generated by the past of the process $S(\cdot)$ up to times t.

Consider finding, in a class $\mathcal{A}$ of $\mathcal{F}_t$ stopping times τ, a stopping time which maximizes the expected utility

$$E[U(S(\tau))]. \tag{3.4}$$

4. Reduction to a completely observed problem. It is well known that the solution of the stock price equation (3.1) is given by

$$S(t) = Se^{\int_0^t a(s)ds + \sigma W(t) - \frac{1}{2}\sigma^2 t}. \tag{4.1}$$

This can be checked using Itô's differential rule.

Equation (4.1) implies

$$\ln(S(t)) - \ln(S) + \frac{1}{2}\sigma^2 t = \int_0^t a(s)ds + \sigma W(t). \tag{4.2}$$

This implies that observing the process $S(t)$ is equivalent to observing the process $y(t)$, where

$$y(t) = \int_0^t a(s)ds + \sigma W(t), \tag{4.3}$$

and that the σ-fields

$$\sigma\{S(r) :\ 0 \le r \le t\} \quad \text{and} \quad \sigma\{y(r) :\ 0 \le r \le t\} \tag{4.4}$$

are equal.

Our assumptions about $a(t)$ imply it is a jump Markov process with two states a and b and the process jumps once from a to b.

Nonlinear filtering results, for instance, Davis and Marcus [4], Liptser and Shiryayev [8], govern conditional probabilities of states of a jump Markov process given measurements of the type (4.3). In particular, if

$$x(t) = Pr[a(t) = a | y(r),\ 0 \le r \le t], \tag{4.5}$$

then $x(t)$ is a solution of

$$dx(t) = -cx(t)dt + \frac{1}{\sigma}(a-b)(1-x(t))\ x(t)d\nu(t),\ x(0) = x. \tag{4.6}$$

In (4.6) $\nu(t)$ is a Wiener process called the innovations process. It satisfies

$$d\nu(t) = \frac{1}{\sigma}(a(t) - (ax(t) + b(1-x(t))))dt + dW(t). \tag{4.7}$$

The relationship (4.7) between the innovations Wiener process $\nu(t)$ and the original Wiener process $W(t)$ implies

$$a(t)dt + \sigma dW(t) = (ax(t) + b(1-x(t)))dt + \sigma d\nu(t). \tag{4.8}$$

Thus we may rewrite the stock price equation (3.1) in terms of the conditional probabilities $x(t)$ and the innovations Wiener process $\nu(t)$ as

$$dS(t) = S(t)[(ax(t) + b(1-x(t)))dt + \sigma d\nu(t)],\quad S(0) = S. \tag{4.9}$$

Since

$$\mathcal{F}(t) = \sigma\{S(r) :\ 0 \le r \le t\} = \sigma\{y(r) :\ 0 \le r \le t\}, \tag{4.10}$$

and $x(t)$ is defined by (4.5), $x(t)$ is $\mathcal{F}_t$ adapted. Thus the optimization problem given in (3.1)–(3.4) is equivalent to the completely observed optimization problem of choosing the $\mathcal{F}_t$ stopping time τ in our class $\mathcal{A}$ of stopping times to maximize

$$E[U(S(\tau))] \tag{4.11}$$

subject to

$$\begin{aligned} dS(t) &= S(t)[(ax(t) + b(1 - x(t)))dt + \sigma d\nu(t)],\ S(0) = S, \\ dx(t) &= -cx(t)dt + \frac{1}{\sigma}(a - b)(1 - x(t))x(t)d\nu(t),\ x(0) = x. \end{aligned} \tag{4.12}$$

LEMMA 4.1. *For q in $(0, 1)$, let*

$$T(q) = \textit{ first time } x(t) \textit{ hits } [0, q]. \tag{4.13}$$

Then

$$E[T(q)] < \infty. \tag{4.14}$$

We give the proof of Lemma 4.1 in the appendix. We remark that (4.14) is false if the interval is replaced by $[q, 1]$.

5. The logarithmic utility problem. Consider the problem (4.11)–(4.12) with utility function

$$U(S) = \ln(S) \tag{5.1}$$

and class of admissible $\mathcal{F}_t$ stopping times τ given by

$$\mathcal{A} = \{\tau : E(\tau) < \infty\}. \tag{5.2}$$

For this problem, in the notation of Theorem 2.1,

$$z = \begin{pmatrix} S \\ x \end{pmatrix}, \quad \beta(z) = 0, \quad h(z) = \begin{pmatrix} S\sigma \\ \frac{1}{\sigma}(a - b)x(1 - x) \end{pmatrix}. \tag{5.3}$$

The region R is

$$R = \{(S, x) :\ 0 \le S < \infty, \quad 0 \le x \le 1\}. \tag{5.4}$$

The operator $A[V](S, x)$ is

$$\begin{aligned} A[V](S, x) &= S(ax + b(1 - x))V_s(S, x) - cxV_x(S, x) \\ &\quad + \frac{1}{2}S^2\sigma^2 V_{ss}(S, x) + (a - b)(1 - x)xSV_{sx}(S, x) \\ &\quad + \frac{1}{2}\left(\frac{a - b}{\sigma}\right)^2 (1 - x)^2 x^2 V_{xx}(S, x), \end{aligned} \tag{5.5}$$

and

$$V_z(z)h(z) = S\sigma V_s(S, x) + \frac{a - b}{\sigma}(1 - x)xV_x(S, x), \tag{5.6}$$

where z and (S, x) are related by (5.3).

Let us begin the discussion by asking if there are conditions for which it is optimal to always sell the stock immediately. Since $a(t)$ takes on values a and b, and b is negative, $a(t) \leq a$. Thus (4.2) implies

$$E[\ln(S(t))] \leq \ln(S) + \left(a - \frac{1}{2}\sigma^2\right)t. \tag{5.7}$$

If $2a \leq \sigma^2$, then $E[\ln(S(t))] \leq \ln(S)$, so $\tau = 0$ is the optimal selling time.

Let us assume $2a > \sigma^2$ and find optimal stopping times in the class $\mathcal{A}$ under this condition. We shall see that under this condition for an appropriate $f(x)$,

$$V(S,x) = \ln(S) + f(x) \tag{5.8}$$

gives a solution of Theorem 2.1.

Using (5.5),

$$\begin{aligned} A[\ln(S) + f(x)](S,x) \\ = ax + b(1-x) - cxf'(x) - \frac{1}{2}\sigma^2 + \frac{1}{2}\left(\frac{a-b}{\sigma}\right)^2 (1-x)^2x^2f''(x). \end{aligned} \tag{5.9}$$

Calling the righthand side of (5.9) $B[f](x)$, the variational inequality (2.5) reduces to

$$B[f](x) \leq 0, \ f(x) \geq 0, \text{ and } f(x)B[f](x) = 0. \tag{5.10}$$

We also see that

$$S\sigma V_s(S,x) + \frac{a-b}{\sigma}(1-x)xV_x(S,x) = \sigma + \frac{a-b}{\sigma}(1-x)xf'(x). \tag{5.11}$$

Since $0 \leq x \leq 1$, this will be bounded if $f'(x)$ is bounded. Thus (2.10) of Remark 2.2 will be satisfied if $f'(x)$ is bounded.

For Itô's differential rule to hold for $\ln(S(t)) + f(x(t))$, $f(x)$ must be at least once continuously differentiable. This continuous differentiability implies: If $q \in (0,1)$, and if q is a boundary point of an interval on which $f(x) = 0$, then $f'(q) = 0$.

Thus we look for a continuously differentiable solution $f(x)$ of (5.10) for which $f'(x)$ is bounded and $f'(q) = 0$ at boundary points q of intervals on which $f(x) = 0$.

Conditions (5.10) imply that if $f(x) \neq 0$, then $B[f](x) = 0$. From (5.9) this equation is

$$ax + b(1-x) - cxf'(x) - \frac{1}{2}\sigma^2 + \frac{1}{2}\left(\frac{a-b}{\sigma}\right)^2 (1-x)^2x^2f''(x) = 0. \tag{5.12}$$

Since we are interested in conditions on $f'(x)$, we will set $r(x) = f'(x)$ in (5.12) and solve

$$ax + b(1-x) - cxr(x) - \frac{1}{2}\sigma^2 + \frac{1}{2}\left(\frac{a-b}{\sigma}\right)^2 x^2(1-x)^2r'(x) = 0 \tag{5.13}$$

for $r(x)$. The general solution of (5.13) is given by

$$\begin{aligned} r(x) = (1-x)^{-h}x^h e^{\frac{h}{1-x}} \\ \cdot \left[\int \frac{h}{2c}x^{-(2+h)}(1-x)^{(h-2)}e^{\frac{-h}{1-x}}(\sigma^2 - 2b - 2(a-b)x)dx + C\right] \end{aligned} \tag{5.14}$$

in which C is an arbitrary constant and h is given by

$$h = \frac{2c\sigma^2}{(a-b)^2}. \tag{5.15}$$

Since $h > 0$,

$$\lim_{x\to 1}(1-x)^{-h}x^h e^{h/(1-x)} = +\infty. \tag{5.16}$$

Thus for $r(x)$ to be bounded at $x = 1$, we must have

$$\lim_{x\to 1}\left[\int \frac{h}{2c}x^{-(2+h)}(1-x)^{(h-2)}e^{\frac{-h}{1-x}}(\sigma^2 - 2b - 2(a-b)x)\,dx + C\right] = 0. \tag{5.17}$$

Define $g(x)$ by

$$g(x) = \int_x^1 \frac{-h}{2c}y^{-(2+h)}(1-y)^{(h-2)}e^{-h/(1-y)}(\sigma^2 - 2b - 2(a-b)y)\,dy. \tag{5.18}$$

For (5.17) to hold, the quantity in brackets in (5.17) must equal $g(x)$. This follows because both quantities have the same derivative and the same limit at $x = 1$. Thus let

$$r(x) = (1-x)^{-h}x^h e^{h/(1-x)}g(x). \tag{5.19}$$

A calculation using L'Hospital's rule shows that

$$\lim_{x\to 1} r(x) \;=\; \frac{2a-\sigma^2}{2c}. \tag{5.20}$$

LEMMA 5.1. *$r(x)$ has a unique root x^* in $(0,1)$ which satisfies*

$$0 < x^* < \frac{\sigma^2 - 2b}{2(a-b)}, \tag{5.21}$$

and $r(x)$ is positive on $(x^*, 1]$.

Proof. Since on the interval $(0,1)$ the function $r(x)$ is given by positive quantities times $g(x)$, its roots will be the same as those of $g(x)$. The integrand of $g(x)$ is given by positive quantities times the linear term

$$2(a-b)y - \sigma^2 + 2b. \tag{5.22}$$

Since $2a > \sigma^2$, (5.22) is positive near $y = 1$. Since $b < 0$, (5.22) is clearly negative near $y = 0$.

Now (5.22) will be positive on

$$\frac{\sigma^2 - 2b}{2a - 2b} < y \le 1. \tag{5.23}$$

Thus

$$g(x) > 0 \quad \text{if} \quad \frac{\sigma^2 - 2b}{2a-2b} < x < 1. \tag{5.24}$$

The term $y^{-(2+h)}$ with $h > 0$ in the integrand of $g(x)$ and the negativity of (5.22) near $y = 0$ imply

$$\lim_{x\to 0} g(x) = -\infty. \tag{5.25}$$

Thus $g(x)$, and hence $r(x)$ must have a root x^* in $\left(0, \frac{\sigma^2-2b}{2a-2b}\right)$. Since $g(x)$ is monotone increasing on this interval, x^* is unique. This implies $r(x)$ is positive on $(x^*, 1]$. □

THEOREM 5.2. *For $f(x)$ defined by*

$$f(x) = \begin{cases} 0 & \text{if } 0 \le x \le x^*, \\ \displaystyle\int_{x^*}^{x} r(x)\,dx & \text{if } x^* \le x \le 1, \end{cases} \tag{5.26}$$

the function $V(S, x) = \ln(S) + f(x)$ satisfies the conditions of Theorem 2.1*, and*

$$T(x^*) = 1st \text{ time } x(t) \text{ hits } [0, x^*] \tag{5.27}$$

is an optimal stopping time in the class $\mathcal{A}$.

Proof. Since $r(x^*) = 0$, $f(x)$ is continuously differentiable and is twice continuously differentiable, except at x^*. Thus Itô's differential rule holds for $\ln(S) + f(x)$. Lemma 5.1 implies that $f(x) \ge 0$. Using (5.9),

$$B[f](x) = \begin{cases} ax + b(1-x) - \frac{1}{2}\sigma^2 & \text{if } 0 \le x \le x^*, \\ 0 & \text{if } x^* \le x \le 1, \end{cases} \tag{5.28}$$

and since

$$x^* < \frac{\sigma^2 - 2b}{2(a-b)}, \tag{5.29}$$

we have that

$$B[f](x) \le 0. \tag{5.30}$$

Since $f(x) = 0$ if $0 \le x \le x^*$ and $B[f](x) = 0$ if $x^* < x \le 1$, $f(x)B[f](x) = 0$. Thus all the conditions of (5.10) are satisfied. The conditions of (5.10) imply the conditions (2.5) for the function (5.8).

Since $r(x)$ is continuous on $[0, 1)$ and has a finite limit at $x = 1$, it is bounded. Thus $f'(x)$ is bounded, which together with (5.11) and Remark 2.2 implies that condition (2.6) holds for stopping times in $\mathcal{A}$. Condition (2.8) follows from Lemma 4.1. Thus the conditions of Theorem 2.1 are satisfied and $T(x^*)$ given by (5.27) is an optimal stopping time in the class $\mathcal{A}$. □

6. The problem with utility function $U(S) = S$. Consider the problem of maximizing

$$E[S(\tau)], \tag{6.1}$$

subject to (4.12) holding, over a class of stopping times $\mathcal{A}$. As mentioned in the introduction, we shall use changes of probability measure to convert this problem into one for which we can verify the conditions of Theorem 2.1.

The solution of the stock price equation is given by

$$S(t) = Se^{\int_0^t (ax(r)+b(1-x(r)))\,dr+\sigma\nu(t)-\frac{\sigma^2}{2}t}. \tag{6.2}$$

Again this can be checked by using Itô's differential rule. Let $\tilde{P}$ be a probability measure which is locally absolutely continuous with respect to P through the following: For each t and $A \in \mathcal{F}_t$,

$$\tilde{P}(A) = \int_A e^{\sigma\nu(t)-\frac{\sigma^2}{2}t}\,dP. \tag{6.3}$$

Let $\tilde{E}$ denote taking expectation with respect to $\tilde{P}$. It follows from [6, Thm. 3.4, p. 153] that for each finite stopping time τ, $\tilde{P}$ is absolutely continuous with respect to P on $\mathcal{F}_\tau$, and for each $A \in \mathcal{F}_\tau$,

$$\tilde{P}(A) = \int_A e^{\sigma\nu(\tau)-\frac{\sigma^2}{2}\tau}\,dP. \tag{6.4}$$

Thus since (6.2) also holds with t replaced by a finite stopping time τ, we have

$$E[S(\tau)] = \tilde{E}\Big[Se^{\int_0^\tau (ax(r)+b(1-x(r)))\,dr}\Big]. \tag{6.5}$$

Now Girsanov's theorem implies that under $\tilde{P}$, $\tilde{\nu}(t)$ defined by

$$\tilde{\nu}(t) = \nu(t) - \sigma t \tag{6.6}$$

is a Wiener process. Thus $x(t)$ is a solution of

$$\begin{aligned} &dx(t) = (-cx(t) + (a-b)(1-x(t))x(t))\,dt + \frac{a-b}{\sigma}(1-x(t))\,x(t)d\tilde{\nu}(t), \\ &x(0) = x. \end{aligned} \tag{6.7}$$

Thus our original problem is equivalent to the following problem: Find a stopping time τ in $\mathcal{A}$ to maximize

$$\tilde{E}\Big[Se^{\int_0^\tau (ax(r)+b(1-x(r)))\,dr}\Big], \tag{6.8}$$

where $x(t)$ is the solution of (6.7).

We shall convert this problem into yet another equivalent problem. Notice that the solution of (6.7) satisfies the integral equation

$$x(t) = xe^{\int_0^t (-c+(a-b)(1-x(r))-\frac{1}{2}(\frac{a-b}{\sigma})^2(1-x(r))^2)\,dr+\int_0^t \frac{a-b}{\sigma}(1-x(r))\,d\tilde{\nu}(r)}. \tag{6.9}$$

Again this may be checked using Itô's differential rule. Formula (6.9) may be rearranged to give

$$\begin{aligned} &e^{\int_0^t (ax(r)+b(1-x(r)))\,dr} \\ &\quad = \frac{xe^{(a-c)t}}{x(t)}\, e^{\int_0^t \frac{a-b}{\sigma}(1-x(r))\,d\tilde{\nu}(r)-\int_0^t \frac{1}{2}(\frac{a-b}{\sigma})^2(1-x(r))^2\,dr}. \end{aligned} \tag{6.10}$$

This suggests making the change of measures defined through the following: For each t and $A \in \mathcal{F}_t$,

$$\tilde{\tilde{P}}(A) = \int_A e^{\int_0^t \frac{a-b}{\sigma}(1-x(r))\,d\tilde{\nu}(r)-\int_0^t \frac{1}{2}(\frac{a-b}{\sigma})^2(1-x(r))^2\,dr}\,d\tilde{P}. \tag{6.11}$$

Then by setting $y(t) = \frac{1}{x(t)}$, again using Theorem 3.4 of [6], and denoting expectation with respect to $\tilde{\tilde{P}}$ by $\tilde{\tilde{E}}$, (6.10) with t replaced by a finite stopping time τ implies

$$E[S(\tau)] = \tilde{E}\Big[Se^{\int_0^\tau (ax(r)+b(1-x(r)))\,dr}\Big] = \tilde{\tilde{E}}[Sxe^{(a-c)\tau}\, y(\tau)]. \tag{6.12}$$

Girsanov's theorem implies that under $\tilde{\tilde{P}}$,

$$\tilde{\tilde{\nu}}(t) = \tilde{\nu}(t) - \int_0^t \frac{a-b}{\sigma}(1-x(r))\,dr \tag{6.13}$$

is a Wiener process.

A calculation using Itô's rule, (6.7), and (6.13) gives that $y(t)$ is the solution of

$$dy(t) = ((c-a+b)y(t) + a - b)\,dt + \frac{a-b}{\sigma}(y(t)-1)\,d\tilde{\tilde{\nu}}(t),\ y(0) = \frac{1}{x}. \tag{6.14}$$

Thus we arrive at the problem of choosing $\tau \in \mathcal{A}$ to maximize

$$\tilde{\tilde{E}}[Sxe^{(a-c)\tau}y(\tau)], \tag{6.15}$$

where $y(t)$ is the solution of (6.14).

7. Solution of the third equivalent problem. Since S and x are positive constants, we consider the problem of choosing $\tau \in \mathcal{A}$ to maximize

$$\tilde{\tilde{E}}[e^{(a-c)\tau}\, y(\tau)], \tag{7.1}$$

where $y(t)$ is the solution of

$$dy(t) = ((c-a+b)y(t) + a - b)\,dt + \left(\frac{a-b}{\sigma}\right)(y(t)-1)\,d\tilde{\tilde{\nu}}(t),\ y(0) = \frac{1}{x}. \tag{7.2}$$

Define

$$T(q) = \text{ first time } y(t) \in [q, \infty). \tag{7.3}$$

In order to utilize Theorem 2.1, we shall first consider maximizing (7.1) over the set of stopping times $\mathcal{A}$ defined by

$$\mathcal{A} = \{\tau :\ \tau \le T(q) \text{ for some } q\} \tag{7.4}$$

and then show that an optimal stopping time in $\mathcal{A}$ is also optimal in the class $\mathcal{B}$ of all finite stopping times.

Notice that if $a \ge c$ and q_n is a sequence approaching infinity, then

$$\tilde{\tilde{E}}[e^{(a-c)T(q_n)}y(T(q_n))] \ge q_n. \tag{7.5}$$

This leads to the very counterintuitive conclusion that no matter how negative the rate of decline b is, if $a \ge c$, then arbitrarily large expected returns can be obtained by using the stopping times $T(q_n)$.

Let us assume $a < c$ and solve the problem given in terms of (7.1), (7.2), and (7.4). The quantity $A[V](y)$ for this problem is given by

$$\begin{aligned} A[V](y) = (a-c)V(y) + ((c-a+b)y + (a-b))V'(y) \\ + \frac{1}{2}\left(\frac{a-b}{\sigma}\right)^2 (y-1)^2 V''(y). \end{aligned} \tag{7.6}$$

The general solution of $A[V](y) = 0$ is given by

$$V(y) = C_1\, W_{k,m}\left(\frac{h}{y-1}\right) e^{\frac{h}{2(y-1)}}(y-1)^k + C_2\, M_{k,m}\left(\frac{h}{y-1}\right) e^{\frac{h}{2(y-1)}}(y-1)^k, \tag{7.7}$$

where

$$k = 1 + \frac{\sigma^2(a-b-c)}{(a-b)^2}, \quad h = \frac{2c\sigma^2}{(a-b)^2}, \quad m = \left[\left(k-\frac{1}{2}\right)^2 + h\frac{(c-a)}{c}\right]^{\frac{1}{2}}, \tag{7.8}$$

and $W_{k,m}(\cdot)$ and $M_{k,m}(\cdot)$ are the classical Whittaker functions.

LEMMA 7.1. *The condition $c > a$ implies that the following conditions* (a) *and* (b) *hold:*

$$\text{(a)} \quad m + k - \frac{3}{2} > 0, \qquad \text{(b)} \quad \frac{h}{(m-k+\frac{1}{2})(m+k-\frac{1}{2})} > 1.$$

The proof of Lemma 7.1 is given in the appendix.

Let

$$g(y) = W_{k,m}\left(\frac{h}{y-1}\right) e^{\frac{h}{2(y-1)}}(y-1)^k, \tag{7.9}$$

and let us see if we can construct a solution to the conditions of Theorem 2.1 from the function $C_1 g(y)$.

The $W_{k,m}(\cdot)$ function [11, p. 122] has integral representation given by

$$W_{k,m}\left(\frac{h}{y-1}\right) = \left[\int_0^\infty e^{-x} x^{m-k-\frac{1}{2}}\left(1 + \frac{x(y-1)}{h}\right)^{m+k-\frac{1}{2}} dx\right] \cdot \left[\frac{1}{\Gamma(m-k+\frac{1}{2})}\left(\frac{h}{y-1}\right)^k e^{\frac{-h}{2(y-1)}}\right]. \tag{7.10}$$

Thus $g(y)$ given by (7.9) has the representation

$$g(y) = \frac{h^k}{\Gamma(m-k+\frac{1}{2})} \int_0^\infty e^{-x} x^{m-k-\frac{1}{2}}\left(1 + \frac{x(y-1)}{h}\right)^{m+k-\frac{1}{2}} dx. \tag{7.11}$$

Differentiating (7.11) under the integral sign gives

$$g'(y) = \frac{h^{k-1}(m+k-\frac{1}{2})}{\Gamma(m-k+\frac{1}{2})} \int_0^\infty e^{-x} x^{m-k+\frac{1}{2}}\left(1 + \frac{x(y-1)}{h}\right)^{m+k-\frac{3}{2}} dx \tag{7.12}$$

and

$$g''(y) = \left[\int_0^\infty e^{-x} x^{m-k+\frac{3}{2}}\left(1 + \frac{x(y-1)}{h}\right)^{m+k-\frac{5}{2}} dx\right] \cdot \left[\frac{(m+k-\frac{1}{2})(m+k-\frac{3}{2})h^{k-2}}{\Gamma(m-k+\frac{1}{2})}\right]. \tag{7.13}$$

It follows from (a) of Lemma 7.1 that for $1 \le y < \infty$, (7.12) and (7.13) are greater than 0. This implies $g(y)$ is increasing and strictly convex on $1 \le y < \infty$.

In order for

$$V(y) = \begin{cases} C_1 g(y) & \text{if } 1 \le y < y^*, \\ y & \text{if } y^* \le y < \infty \end{cases} \tag{7.14}$$

to satisfy the conditions of Theorem 2.1, it must be at least once continuously differentiable. For this to happen, C_1 and y^* must be a solution of

$$C_1 g(y^*) = y^*, \quad C_1 g'(y^*) = 1. \tag{7.15}$$

If y^* is a solution of

$$\frac{g(y)}{g'(y)} = y, \tag{7.16}$$

(7.15) will hold with $C_1 = \frac{1}{g'(y^*)}$.

LEMMA 7.2. *There is a solution $y^* > 1$ of* (7.16).

Proof.

$$\frac{g(y)}{g'(y)} = \frac{h}{m+k-\frac{1}{2}} \frac{\int_0^\infty e^{-x} x^{m-k-\frac{1}{2}} (1 + \frac{x(y-1)}{h})^{m+k-\frac{1}{2}}\, dx}{\int_0^\infty e^{-x} x^{m-k+\frac{1}{2}} (1 + \frac{x(y-1)}{h})^{m+k-\frac{3}{2}}\, dx}. \tag{7.17}$$

Thus

$$\frac{g(1)}{g'(1)} = \frac{h}{m+k-\frac{1}{2}} \frac{\int_0^\infty e^{-x} x^{m-k-\frac{1}{2}}\, dx}{\int_0^\infty e^{-x} x^{m-k+\frac{1}{2}}\, dx}. \tag{7.18}$$

An integration by parts gives

$$\int_0^\infty e^{-x} x^{m-k+\frac{1}{2}}\, dx = \left(m - k + \frac{1}{2}\right) \int_0^\infty e^{-x} x^{m-k-\frac{1}{2}}\, dx, \tag{7.19}$$

and thus by (7.18) and (b) of Lemma 7.1,

$$\frac{g(1)}{g'(1)} = \frac{h}{(m-k+\frac{1}{2})(m+k-\frac{1}{2})} > 1. \tag{7.20}$$

Dividing both numerator and denominator of (7.17) by $(y-1)^{(m+k-\frac{1}{2})}$ gives

$$\frac{g(y)}{g'(y)} = \frac{h(y-1)}{m+k-\frac{1}{2}} \frac{\int_0^\infty e^{-x} x^{m-k-\frac{1}{2}} (\frac{1}{(y-1)} + \frac{x}{h})^{m+k-\frac{1}{2}}\, dx}{\int_0^\infty e^{-x} x^{m-k+\frac{1}{2}} (\frac{1}{(y-1)} + \frac{x}{h})^{m+k-\frac{3}{2}}\, dx}. \tag{7.21}$$

Now

$$\lim_{y\to\infty} \frac{\int_0^\infty e^{-x} x^{m-k-\frac{1}{2}} (\frac{1}{y-1} + \frac{x}{h})^{m+k-\frac{1}{2}}\, dx}{\int_0^\infty e^{-x} x^{m-k+\frac{1}{2}} (\frac{1}{y-1} + \frac{x}{h})^{m+k-\frac{3}{2}}\, dx} = \frac{1}{h}. \tag{7.22}$$

Thus asymptotically as $y \to \infty$,

$$\frac{g(y)}{g'(y)} \sim \frac{y-1}{m+k-\frac{1}{2}}. \tag{7.23}$$

Since $m + k - \frac{1}{2} > 1$, $\frac{g(y)}{g'(y)}$ asymptotically has slope less than one. Since its value at one is bigger than one, and it has asymptotic slope less than one, $\frac{g(y)}{g'(y)}$ must cross the line given by y at some $y^* > 1$. □

THEOREM 7.3. *For y^* the solution of* (7.16)*, the function*

$$V(y) = \begin{cases} \frac{1}{g'(y^*)} g(y) & \text{if } 1 \le y \le y^*, \\ y & \text{if } y^* \le y < \infty \end{cases} \tag{7.24}$$

satisfies the conditions of Theorem 2.1 *for the problem given by* (7.1)*,* (7.2)*, and* (7.4)*. The stopping time $T(y^*)$ is an optimal stopping time in the class $\mathcal{A}$.*

Proof. Since $g'(y) > 0$, and $g(y)$ is strictly convex, the conditions

$$\frac{1}{g'(y^*)}\, g(y^*) = y^*, \quad \frac{1}{g'(y^*)}\, g'(y^*) = 1 \tag{7.25}$$

imply y is a line of support of $\frac{1}{g'(y^*)}\, g(y)$ at y^*. Thus

$$\frac{1}{g'(y^*)}\, g(y) \ge y, \tag{7.26}$$

which implies

$$V(y) \ge y. \tag{7.27}$$

Since $A[\frac{1}{g'(y^*)}\, g(\cdot)](y) = 0$, for $V(y)$ given by (7.24),

$$A[V](y) = \begin{cases} 0 & \text{if } 1 \le y < y^*, \\ by + (a - b) & \text{if } y^* < y < \infty. \end{cases} \tag{7.28}$$

Evaluating $A[\frac{1}{g'(y^*)}\, g(\cdot)](y) = 0$ at y^*, and using (7.6) and the conditions of (7.25), give

$$(a - c)y^* + (cy^* - (a - b)(y^* - 1)) + \frac{1}{2}\left(\frac{a - b}{\sigma}\right)^2 (y^* - 1)^2\, \frac{g''(y^*)}{g'(y^*)} = 0. \tag{7.29}$$

Since $g(y)$ is increasing and strictly convex, the last term in (7.29) is positive. Thus (7.29) implies

$$by^* + a - b < 0. \tag{7.30}$$

Since b is negative,

$$by + (a - b) < 0 \quad \text{if } \ y^* \le y < \infty, \tag{7.31}$$

or, from (7.28), then $A[V](y) \le 0$ for $V(y)$ given by (7.24). That $(V(y) - y)A[V](y) = 0$ follows immediately from (7.24) and (7.28).

To verify condition (2.6), we see

$$V'(y) = \begin{cases} \frac{1}{g'(y^*)}\, g'(y) & \text{if } 1 \le y \le y^*, \\ 1 & \text{if } y^* < y < \infty. \end{cases} \tag{7.32}$$

From (7.12) and (a) of Lemma 7.1, $g'(1) > 0$. Since $g'(y)$ is increasing, (7.32) implies

$$0 \le V'(y) \le 1. \tag{7.33}$$

The quantity in (2.6) for this problem is

$$\tilde{\tilde{E}}\left[\int_0^\tau e^{2(a-c)t}\left(V'(y(t))^2\left(\frac{a-b}{\sigma}\right)^2(y(t)-1)\right)^2 dt\right]. \tag{7.34}$$

Since $\tau \le T(q)$ for some q and (7.33) holds, (7.34) is bounded by

$$\int_0^\infty e^{2(a-c)t}\left(\frac{a-b}{\sigma}\right)^2(q-1)^2\,dt = \frac{(q-1)^2}{2(c-a)}\left(\frac{a-b}{\sigma}\right)^2. \tag{7.35}$$

Thus condition (2.6) holds. Condition (2.8) follows from Lemma 4.1 and the definition of $T(y^*)$. Thus the conditions of Theorem 2.1 are satisfied and $T(y^*)$ is an optimal stopping time in the class $\mathcal{A}$. □

THEOREM 7.4. *The stopping time $T(y^*)$ is optimal in the class $\mathcal{B}$ of all finite stopping times.*

Proof. If τ is a finite stopping time and q_n is a sequence approaching ∞, satisfying $q_n > 1/x$, the definition of $T(q_n)$ implies

$$y(\tau \wedge T(q_n)) = \begin{cases} y(\tau) & \text{if } \tau < T(q_n), \\ q_n & \text{if } \tau \ge T(q_n). \end{cases} \tag{7.36}$$

It can be shown that $\lim_{n\to\infty} T(q_n) = +\infty$ with probability one.

Hence

$$\lim_{n\to\infty} e^{(a-c)\tau\wedge T(q_n)} y(\tau \wedge T(q_n)) = e^{(a-c)\tau} y(\tau). \tag{7.37}$$

Theorem 7.3 implies

$$\tilde{\tilde{E}}[e^{(a-c)\tau\wedge T(q_n)} y(\tau \wedge T(q_n))] \le \tilde{\tilde{E}}[e^{(a-c)T(y^*)} y(T(y^*))]. \tag{7.38}$$

Fatou's Lemma and (7.37) imply

$$\tilde{\tilde{E}}[e^{(a-c)\tau} y(\tau)] \le \liminf_{n\to\infty} \tilde{\tilde{E}}[e^{(a-c)\tau\wedge T(q_n)} y(\tau \wedge T(q_n))]. \tag{7.39}$$

Thus (7.38) and (7.39) imply

$$\tilde{\tilde{E}}[e^{(a-c)\tau} y(\tau)] \le \tilde{\tilde{E}}[e^{(a-c)T(y^*)} y(T(y^*))], \tag{7.40}$$

which gives the optimality of $T(y^*)$ in the class $\mathcal{B}$ of all finite stopping times. □

8. Computations. In Table 8.1 we give results of computing optimal selling regions for a variety of conditions. To explain the coefficients used in Table 8.1, we note that (4.1) implies the following: If the growth rate $a(t)$ has the constant value a, the expected stock price is

$$E[S(t)] = Se^{at}, \tag{8.1}$$

and similarly, if $a(t)$ has the constant value b, the expected stock price is

$$E[S(t)] = Se^{bt}. \tag{8.2}$$

TABLE 8.1
Optimal selling regions for both criteria and different parameter values.

Problem coefficients				$U(S) = \ln(S)$	$U(S) = S$
Growth	Decline	Mean years	σ	Selling region	Selling region
10%	60%	2	.3	$0 \le x \le .911$	$0 \le x \le .766$
10%	60%	2	.4	$0 \le x \le .978$	$0 \le x \le .831$
10%	60%	2	.5	$0 \le x \le 1.00$	$0 \le x \le .857$
10%	80%	2	.3	$0 \le x \le .927$	$0 \le x \le .760$
10%	80%	2	.4	$0 \le x \le .986$	$0 \le x \le .856$
10%	80%	2	.5	$0 \le x \le 1.00$	$0 \le x \le .891$
20%	60%	2	.3	$0 \le x \le .678$	$0 \le x \le .441$
20%	60%	2	.4	$0 \le x \le .849$	$0 \le x \le .601$
20%	60%	2	.5	$0 \le x \le .939$	$0 \le x \le .685$
20%	80%	2	.3	$0 \le x \le .651$	$0 \le x \le .368$
20%	80%	2	.4	$0 \le x \le .913$	$0 \le x \le .587$
20%	80%	2	.5	$0 \le x \le .953$	$0 \le x \le .709$
20%	80%	6	.4	$0 \le x \le .659$	For the sequence of selling regions $0 \le x \le a_n$ where $a_n \downarrow 0$, the problem is unbounded

TABLE 8.2
Correspondence of parameter values and approximate percents of growth and decline per year.

a	.1	.18	b	−.92	−1.6
% growth	10	20	% decline	60	80

The value $a = .1$ gives $e^{.1} \approx 1.105$, or a growth rate of approximately 10% per year. The value $b = -.92$ gives $e^{-.92} \approx .398$, or a rate of decline of approximately 60% per year. Table 8.1 was computed in terms of values of a, b, c, σ but is displayed in terms of approximate percents of growth and decline to aid intuitive understanding of the results. The relationships between the values of a and b used in the computation of Table 8.1 and the approximate percents of growth and decline per year are listed in Table 8.2.

Except in the last case in Table 8.1, the mean of the jump time was taken to be two years, or $c = \frac{1}{2}$. In literature using actual stock data, values for the stock variance σ near .3 or .4 are common. We use those and also the value $\sigma = .5$ to illustrate a case with large variance.

Table 8.1 shows that the optimal stock selling regions for $U(S) = \ln(S)$ and $U(S) = S$ are quite different. The optimal selling regions for $\ln(S)$ are uniformly larger than those of S. In some cases these differences are large. If the growth rate equals 10% and $\sigma = .5$, the inequality

$$2a = .2 < .25 = \sigma^2$$

holds, and thus selling immediately always occurs for the $ln(S)$ criteria, while for the S criteria selling immediately does not occur if x is large enough. If the growth rate is approximately 20% for an exponentially distributed length of time with mean six years, then

$$a = .18 > \frac{1}{6} = c.$$

This implies that there is a sequence of values of x approaching zero so that stopping

at these values gives unbounded expected return for the criteria S. For this case, the criteria $\ln(S)$ is finite and its optimal selling region is $0 \le x \le .659$.

9. Comparison with Zhang [12] and Beibel and Lerche [3]. In [12] Zhang considers a stock selling problem in which the stock price is given by a geometric Brownian motion whose drift and diffusion coefficients are given by unobservable, finite state, jump Markov processes. This model is more general than that of the current paper, whose a drift coefficient is a one jump Markov process, and where a diffusion coefficient is constant. However, the classes of stopping times optimized over are different. In [12] the stopping times are first times that the stock price leaves a given interval. In the present paper they are stopping times adapted to the past measurements of the stock price. These include the first times the stock price leaves an interval, so there should be a larger optimal expected return in the present case. There is an advantage to using the stopping times of [12] in that they are easier to implement.

The difference between the two problems is especially clear in the case in which the initial probability x of being in the increasing state a is less than one. (Some people assert this is "always" the case by the time a small investor invests in a stock.) The stopping times of [12] do not depend on x. If x is small enough to be in the optimal selling region for the current problem, the investor should sell the stock immediately. If he/she is using stopping times of [12], he/she holds the stock until its price leaves an interval.

In [3] Beibel and Lerche consider an optimal stock selling problem with utility function $U(S) = S$, which is roughly the same as our second equivalent problem. Their methods applied to this problem would be roughly the following. Let $x(t)$ be the solution of (6.7) with initial condition $x(0) = 1$. Let $0 < x < 1$ and

$$T(x) = \text{ first time } x(t) = x. \tag{9.1}$$

Then, using elegant probabilistic arguments, they show

$$E\left[e^{\int_0^{T(x)} a(x(r))+b(1-x(r))\,dr}\right] \tag{9.2}$$

has a maximum attained at some x^* and for the problem with $x(t)$ given by (6.7), with initial condition γ, for which $x^* \le \gamma \le 1$ that

$$T(x^*) = \text{ first time } x(t) = x^* \tag{9.3}$$

maximizes the criteria (6.8).

However, they do not carry out any computations. It appears that computing (9.2) would encounter similar difficulties to those we encountered in solving the variational inequalities for this case.

Appendix. In this appendix we provide the proofs of Theorem 2.1, Lemma 4.1, and Lemma 7.1.

Proof of Theorem 2.1. Itô's differential rule implies that

$$\begin{aligned} de^{\int_0^t \beta(z(r))dr}V(z(t)) &= \beta(z(t))e^{\int_0^t \beta(z(r))dr}V(z(t))dt \\ &+ e^{\int_0^t \beta(z(r))dr}\Big[V_z(z(t))g(z(t)) \\ &\qquad + \frac{1}{2}\operatorname{trace}\left(h(z(t))h'(z(t))V_{zz}(z(t))\right)\Big]dt \\ &+ e^{\int_0^t \beta(z(r))dr}V_z(z(t))h(z(t))dW(t), \end{aligned} \tag{A.1}$$

or, expressed in integrated form and using (2.4), that

$$e^{\int_0^t \beta(z(r))dr}V(z(t)) - V(z) = \int_0^t e^{\int_0^s \beta(z(r))dr}A[V](z(s))ds + \int_0^t e^{\int_0^s \beta(z(r))dr}V_z(z(s))h(z(s))dW(s). \tag{A.2}$$

Since (A.2) holds for each sample function, it also holds with t replaced by a finite stopping time τ.

It follows from Remark 1 on page 29 of Gihman and Skorohod [5] that condition (2.6) implies for each stopping time τ in $\mathcal{A}$ that

$$E\left[\int_0^\tau e^{\int_0^s \beta(z(r))\,dr}\,V_z(z(s))h(z(s))\,dW(s)\right] = 0. \tag{A.3}$$

Thus for each τ in $\mathcal{A}$,

$$E\left[e^{\int_0^\tau \beta(z(r))\,dr}\,V(z(\tau))\right] = V(z) + E\left[\int_0^\tau\ e^{\int_0^s \beta(z(r))\,dr}\,A[V](z(s))\,ds\right]. \tag{A.4}$$

Now from (2.5),

$$V(z) \geq U(z) \quad \text{and} \quad A[V](z) \leq 0,$$

and thus for each τ in $\mathcal{A}$,

$$E\left[e^{\int_0^\tau \beta(z(r))\,dr}U(z(\tau))\right] \leq V(z). \tag{A.5}$$

From (2.5) for

$$\tau(z) = \text{ first time } z(t) \text{ hits } \{y:\ U(y) = V(y)\},$$

it follows that

$$A[V](z(s)) = 0 \quad \text{on } 0 \leq s < \tau(z) \tag{A.6}$$

and

$$U(z(\tau(z))) = V(z(\tau(z))). \tag{A.7}$$

Thus (A.4) implies

$$E\left[e^{\int_0^{\tau(z)}\ \beta(z(r))\,dr}\,U(z(\tau(z)))\right] = V(z). \tag{A.8}$$

Since condition (2.8) requires $\tau(z)$ to belong to $\mathcal{A}$, (A.5) and (A.8) imply $\tau(z)$ is an optimal stopping time in the class $\mathcal{A}$. □

Proof of Lemma 4.1. We may assume $x > q$; otherwise $T(q) = 0$ and $E[T(q)] = 0$. For $x(t)$ given by (4.8), from Theorem 2 on page 149 of [5], appropriately modified, it follows that for $t > 0$,

$$P[x(t) < 1] = 1. \tag{A.9}$$

A solution $K(x)$ of the differential equation

$$-cxK'(x) + \frac{1}{2}r^2(1-x)^2x^2K''(x) + 1 = 0 \tag{A.10}$$

on $[0,1]$ satisfying $K(q) = 0$ and $K'(x)$ bounded on $[q,1]$ is given by

(A.11)

$$K(x) = \int_q^x \left[\frac{2}{r^2} e^{\frac{2c}{r^2(1-z)}} z^{\frac{2c}{r^2}} (1-z)^{-\frac{2c}{r^2}} \int_z^1 (1-y)^{\frac{-2(r^2-c)}{r^2}} y^{\frac{-2(r^2+c)}{r^2}} e^{\frac{-2c}{r^2(1-y)}} dy\right] dz.$$

To see that $K'(x)$ is bounded, a L'Hospital's rule argument implies that

$$\lim_{x\to 1} K'(x) = \frac{r^4}{2c}, \tag{A.12}$$

and from this and the form of $K'(x)$, boundedness follows. Notice that $K(x) \geq 0$ for $x \geq q$.

Set

$$T(q) = \inf\{t : x(t) = q\}, \tag{A.13}$$

and for a fixed time T greater than 0,

$$\tau_T = \min(T, T(q)). \tag{A.14}$$

From (A.9) and (A.13), it follows for $s < \tau_T$ that $x(s)$ is contained in $[q,1]$ over which $K(x)$ is defined. Itô's formula implies that

$$\begin{aligned} &K(x(\tau_T)) - K(x) \\ &\quad = \frac{a-b}{\sigma}\int_0^{\tau_T} (1-x(s))x(s)K'(x(s))\, dW(s) \\ &\qquad + \int_0^{\tau_T} \left[-cx(s)K'(x(s)) + \frac{1}{2}\left(\frac{a-b}{\sigma}\right)^2 (1-x(s))^2x(s)^2K''(x(s))\right] ds. \end{aligned} \tag{A.15}$$

For $K(x)$ the solution of (A.10) with $r = \frac{a-b}{\sigma}$, (A.15) implies

$$K(x(\tau_T)) - K(x) = -\tau_T + \frac{a-b}{\sigma}\int_0^{\tau_T} (1-x(s))x(s)K'(x(s))\, dW(s). \tag{A.16}$$

Since the integrand in the stochastic integral is bounded and τ_T is bounded, the expected value of the stochastic integral is zero, giving

$$E[\tau_T] = K(x) - E[K(x(\tau_T))]. \tag{A.17}$$

Now τ_T increases monotonically to $T(q)$ as T approaches infinity. This and the positivity of $K(x(\tau_T))$ imply

$$E[T(q)] \leq K(x). \qquad \square \tag{A.18}$$

Proof of Lemma 7.1. (a) Since $m \geq 0$, if $k - \frac{3}{2} > 0$ there is nothing to prove. Thus consider $k - \frac{3}{2} \leq 0$. A calculation using (7.8) shows

$$m^2 - \left(\frac{3}{2} - k\right)^2 = \frac{2\sigma^2(-b)}{(a-b)^2} > 0. \tag{A.19}$$

Thus $m^2 > (\frac{3}{2} - k)^2$, and since both m and $\frac{3}{2} - k$ are ≥ 0,

$$m > \frac{3}{2} - k \quad \text{or} \quad m + k - \frac{3}{2} > 0, \tag{A.20}$$

giving (a).

(b) Using (7.8),

$$\begin{aligned} \frac{h}{(m-k+\frac{1}{2})(m+k-\frac{1}{2})} &= \frac{h}{m^2-(k-\frac{1}{2})^2} \\ &= \frac{\frac{2c\sigma^2}{(a-b)^2}}{(k-\frac{1}{2})^2 + \frac{2\sigma^2(c-a)}{(a-b)^2} - (k-\frac{1}{2})^2} \\ &= \frac{c}{c-a} \ > \ 1. \qquad \square \end{aligned} \tag{A.21}$$

REFERENCES

[1] A. Bensoussan, *Stochastic Control by Functional Analysis Methods*, North–Holland, Amsterdam, 1982.

[2] A. Bensoussan and J. L. Lions, *Applications of Variational Inequalities in Stochastic Control*, North–Holland, Amsterdam, 1982.

[3] M. Beibel and H. R. Lerche, *A new look at optimal stopping problems related to mathematical finance*, Statist. Sinica, 7 (1997), pp. 93–108.

[4] M. H. A. Davis and S. I. Marcus, *An introduction to nonlinear filtering,* in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., D. Reidel, Dordrecht, Boston, 1981, pp. 53–75.

[5] I. I. Gihman and A. V. Skorokhod, *Stochastic Differential Equations*, Springer-Verlag, Berlin, New York, 1972.

[6] J. Jacod and A. N. Shiryayev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, New York, 1987.

[7] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, New York, 1980.

[8] R. S. Liptser and A. N. Shiryayev *Statistics of Random Processes* I. *General Theory*, Springer-Verlag, Berlin, New York, 1977.

[9] B. Oksendal, *Stochastic Differential Equations*, 6th ed., Springer-Verlag, Berlin, New York, 2003.

[10] A. N. Shiryayev, *Optimal Stopping Rules*, Springer-Verlag, Berlin, New York, 1978.

[11] B. Spain, *Functions of Mathematical Physics*, Van Nostrand Reinholt, 1970.

[12] Q. Zhang, *Stock trading: An optimal selling rule*, SIAM J. Control Optim., 40 (2001), pp. 64–87.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 599–633

APPROXIMATION OF NONLINEAR FILTERS FOR MARKOV SYSTEMS WITH DELAYED OBSERVATIONS*

ANTONELLA CALZOLARI†, PATRICK FLORCHINGER‡, AND GIOVANNA NAPPO§

Abstract. The aim of this paper is to give some approximation results for a class of nonlinear filtering problems with delay in the observation. First, we point out some general results on the approximation problem for the filter in nonlinear filtering. In particular, we give a general procedure for obtaining some upper bounds for the different approximations we consider. This procedure is then applied in the case of nonlinear filtering problems with delay (X, Y), which can be represented by means of a Markov system $(X, \hat{Y})$, in the sense that $Y_t = \hat{Y}_{a(t)}$. Finally, these upper bounds are computed explicitly in the particular case of Markov jump process with counting observations.

Key words. nonlinear filtering, jump processes, Markov processes, approximation of stochastic processes

AMS subject classifications. 93E11, 60G35, 60J75, 65C50, 62M20

DOI. 10.1137/050623504

1. Introduction.

Consider a partially observed stochastic system $(\boldsymbol{X}, \boldsymbol{Y}) = (X_t, Y_t)_{t\geq 0}$, that is, a *state process* $\boldsymbol{X} = (X_t)_{t\geq 0}$, which cannot be directly observed, and a completely observable process $\boldsymbol{Y} = (Y_t)_{t\geq 0}$, which is referred to as the *observation process*. The aim of stochastic nonlinear filtering is to compute the conditional law π_t of the state process at time t, given the observation process up to time t, i.e., the computation of the so-called filter

$$\pi_t(\varphi) = E\big[\varphi(X_t)/\mathcal{F}_t^Y\big], \tag{1}$$

for all functions φ belonging to a determining class, i.e., the best estimate of $\varphi(X_t)$ given the σ-algebra of the observations up to time t, $\mathcal{F}_t^Y = \sigma\{Y_s,\, s \leq t\}$.

A classical model of partially observed systems arises when both the state and the observation are diffusion processes. In this case it has been shown that the filter solves a stochastic partial differential equation known as the Kushner–Stratonovich equation (see, e.g., Pardoux [20] and the references therein).

Filtering problems involving jump-diffusion processes have been studied by many authors. In particular, the jump system

$$\begin{aligned} x_t = x_0 &+ \int_0^t \int_{D_0(x_{s^-}, y_{s^-})} K_0(x_{s^-}, y_{s^-}; \zeta)\mathcal{N}(ds, d\zeta) \\ &+ \int_0^t \int_{D_1(x_{s^-}, y_{s^-})} K_1(x_{s^-}, y_{s^-}; \zeta)\mathcal{N}(ds, d\zeta), \end{aligned} \tag{2}$$

*Received by the editors January 29, 2005; accepted for publication (in revised form) December 5, 2005; published electronically May 3, 2006. This research was partially supported by MURST project "Metodi stocastici in finanza matematica" Anno 2004 - prot. 2004014572_007.
http://www.siam.org/journals/sicon/45-2/62350.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata," via della Ricerca Scientifica 1, I 00133 Roma, Italy (calzolar@mat.uniroma2.it).

‡Département de Mathématiques, Université de Metz, 23 Allée des Oeillets F 57160 Moulins les Metz, France (florchin@math.univ-metz.fr).

§Dipartimento di Matematica, Università di Roma "La Sapienza," piazzale A. Moro 2, I 00185 Roma, Italy (nappo@mat.uniroma1.it).

$$y_t = \int_0^t \int_{D_1(x_{s^-}, y_{s^-})} \mathcal{N}(ds, d\zeta), \tag{3}$$

where

- $\mathcal{N}(ds, d\zeta)$ is a Poisson measure on $\mathbb{R} \times \Sigma$ with mean measure $ds \times \nu(d\zeta)$,
- the random variable x_0 has values in $\mathbb{R}^k$ and probability distribution μ_0^x,
- the random variable x_0 and the Poisson random measure $\mathcal{N}(ds, d\zeta)$ are independent,
- the sets $D_0(x, y)$ and $D_1(x, y)$ are disjoint,

enters in the more general framework studied by Kliemann, Koch, and Marchetti [15]. Under suitable hypotheses (see [15]), the above system has a unique solution $(x_t, y_t)_{t \geq 0}$, which is a Markov process with formal generator $\boldsymbol{L}$ given by

$$\begin{aligned} \boldsymbol{L} f(x, y) = &\int_{D_0(x,y)} [f(x + K_0(x, y; \zeta), y) - f(x, y)] \, \nu(d\zeta) \\ &+ \int_{D_1(x,y)} [f(x + K_1(x, y; \zeta), y + 1) - f(x, y)] \, \nu(d\zeta). \end{aligned} \tag{4}$$

Then the filter $\pi_t^x(\varphi) = E[\varphi(x_t)/\mathcal{F}_t^y]$ can be obtained via the following normalization procedure:

- Let $\boldsymbol{s} = \{s_i\}_{i \geq 0}$ be an increasing sequence of times such that $s_0 = 0$;
- define

$$B^y \varphi(x) = \int_{D_0(x,y)} [\varphi(x + K_0(x, y; \zeta)) - \varphi(x)] \, \nu(d\zeta) \tag{5}$$

and

$$R^y \varphi(x) = \int_{D_1(x,y)} [\varphi(x + K_1(x, y; \zeta)) - \varphi(x)] \, \nu(d\zeta); \tag{6}$$

- define $\hat{\rho}_t(dx|\boldsymbol{s})$ by the following self-contained procedure:

(7)

$$\hat{\rho}_t(\varphi|\boldsymbol{s}) = E\left[\varphi(X^i_{t-s_i}(\boldsymbol{s})) \exp\left\{-\int_0^{t-s_i} \lambda_1(X^i_u(\boldsymbol{s}), i) du\right\}\right], \qquad s_i \leq t < s_{i+1},$$

where $\lambda_1(x, i) = \nu(D_1(x, i))$, and $\{X_t^i(\boldsymbol{s}); t \geq 0\}$ is a Markov process with generator B^i and initial distribution $\hat{\rho}_{s_i}(dx|\boldsymbol{s})$ defined inductively by

$$\hat{\rho}_{s_0}(\varphi|\boldsymbol{s}) = \mu_0^x(\varphi), \qquad \hat{\rho}_{s_{i+1}}(\varphi|\boldsymbol{s}) = \frac{\hat{\rho}_{s_{i+1}^-}(Q^i \varphi|\boldsymbol{s})}{\hat{\rho}_{s_{i+1}^-}(\lambda_1(\cdot, i)|\boldsymbol{s})} \tag{8}$$

with $Q^i \varphi(x) = R^i \varphi(x) + \lambda_1(x, i)\varphi(x) = \int_{D_1(x,i)} \varphi(x + K_1(x, i; \zeta)) \, \nu(d\zeta)$.

Setting

$$\hat{\Pi}_t(\varphi|\boldsymbol{s}) := \frac{\hat{\rho}_t(\varphi|\boldsymbol{s})}{\hat{\rho}_t(\mathbf{1}|\boldsymbol{s})}, \tag{9}$$

we have

$$\pi_t^x(\varphi) = E[\varphi(x_t)/\mathcal{F}_t^y] = \hat{\Pi}_t(\varphi|\boldsymbol{s})\Big|_{\boldsymbol{s} = \{T_i\}_{i \geq 0}}, \tag{10}$$

where T_i are the observed jump times.

Different approximation procedures and numerical schemes have been studied by many authors in the past years (see, for example, Kushner [17], Di Masi, Pratelli, and Runggaldier [12], Le Gland [18], Elliott and Glowinski [13], Lototsky, Mikulevicius, and Rozovskii [19], Del Moral [11], Calzolari and Nappo [8], or Ceci, Gerardi, and Tardelli [9] and the references therein).

In this paper we are interested in the approximation of the filter when dealing with the following nonlinear filtering problem with delay in the observation: $(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ is a Markov process with generator $\boldsymbol{L}$, and the observation process $\boldsymbol{Y}$ satisfies

$$Y_t = \hat{Y}_{a(t)},$$

where $a(\cdot)$ is a nondecreasing continuous function such that $a(0) = 0$, $a(t) \le t$ for all $t \ge 0$.

For this system the following result has been proved by the authors in [4].

THEOREM 1.1. *The filter π_t associated with the system $(\boldsymbol{X}, \boldsymbol{Y})$ described above can be represented as*

$$\pi_t(\varphi) = E\big[\big(\exp\{\boldsymbol{L}(t - a(t))\}\phi\big)(X_{a(t)}, \hat{Y}_{a(t)})/\mathcal{F}^{\hat{Y}}_{a(t)}\big],$$

where $\phi(x, y) = \varphi(x)$.

Moreover, if the conditional law of X_s given $\mathcal{F}^{\hat{Y}}_s$ is known and denoted by $\hat{\pi}_s$, then

$$\pi_t(\varphi) = \hat{\pi}_{a(t)}\big((\exp\{\boldsymbol{L}(t - a(t))\}\phi)(\cdot, \hat{Y}_{a(t)})\big). \tag{11}$$

Remark 1.2. As an example we can take $(\boldsymbol{X}, \hat{\boldsymbol{Y}}) = (x_t, y_t)_{t \ge 0}$, with x_t and y_t defined as in (2) and (3), respectively. In this case the generator $\boldsymbol{L}$ coincides with the one given in (4).

Remark 1.3. The case when the state process solves the stochastic differential equation

$$X_t = X_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dB_s, \qquad t \ge 0,$$

and the observation is given by

$$\begin{cases} Y_t = 0, & 0 \le t \le \tau, \\ Y_t = \int_\tau^t h(X_{s-\tau})ds + W_t - W_\tau, & t \ge \tau, \end{cases} \tag{12}$$

corresponds to $a(t) = (t - \tau)^+$ and to the system $(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ with the generator

$$\boldsymbol{L}f(x, y) = b(x)\nabla_x f(x, y) + h(x)\nabla_y f(x, y) + \frac{1}{2}\mathrm{tr}\{\sigma(x)\sigma^*(x)\nabla_x^2 f(x, y)\} + \frac{1}{2}\Delta_y f(x, y)$$

in the above theorem. On the other hand, by applying a shift in time of size τ to the system $(\boldsymbol{X}, \boldsymbol{Y})$, we get a stochastic delay system: indeed, setting $X_{(\tau)}(s) := X_{s+\tau}$ and $Y_{(\tau)}(s) := Y_{s+\tau}$, we have

$$dX_{(\tau)}(s) = b\big(X_{(\tau)}(s)\big)ds + \sigma\big(X_{(\tau)}(s)\big)d\tilde{B}_s, \quad s \ge 0, \qquad X_{(\tau)}(s) = \xi(s), \quad -\tau \le s \le 0, \tag{13}$$

$$Y_{(\tau)}(t) = \int_0^t h\big(X_{(\tau)}(s-\tau)\big)ds + \hat{W}_t, \qquad t \geq 0, \tag{14}$$

where $\tilde{B}_s = B_{s+\tau}$, and $\xi(s)$ is the solution of the stochastic differential equation

$$d\xi(s) = b(\xi(s))ds + \sigma(\xi(s))d\tilde{B}_s \quad \text{for } -\tau \leq s \leq 0 \quad \text{with } \xi(-\tau) = X(0).$$

Clearly, in the above system the delay is present only in the observation.

The approximation problem we consider is interesting due to the fact that partially observed systems with delay in the observations appear in stochastic finance. For instance, in [21] Schweizer has given an example of information with delay for a financial model. From [21] one can easily deduce that when the market is incomplete the risk minimization criterion leads to a filtering problem with delayed observation. The risk minimization criterion corresponds to a quadratic loss function, and this is criticized, as it implies the minimization of both earnings and losses. This leads to the use of different loss functions and to the consideration of the hedging problem as an optimal control problem. Then, from a Bayesian point of view, filtering may appear in the case of partial observations, and in this direction diffusion-type models are mainly studied in the literature. More recently, in [14] Kirch and Runggaldier have studied a control problem with partial observation when the stock price evolves as

$$S_t = s_0\, e^{aN_t^+ - bN_t^-},$$

where a, $b > 0$ are constants, and N^+ and N^- are counting processes with random unobservable intensities Λ^+ and Λ^-.

The observation of the stock price is equivalent to the observation of the couple of counting processes N^+ and N^-.

In [14], N^+ and N^- are assumed to be conditionally independent Poisson processes given their random time constant intensities Λ^+ and Λ^-, which are independent and with prior distribution gamma. In this case the filter of (Λ^+, Λ^-) given N^+ and N^- up to time t can be reduced to a couple of filtering problems, each with one counting observation process, and can be computed explicitly.

If the time constant assumption on the intensities were dropped, then the explicit computation of the filter would no longer be feasible, and a filtering approximation problem would naturally arise. Considering also a delay in the information would then lead to an example fitting the framework of section 4. Another related situation fitting the same framework could arise if the interest were focused on a fixed strike lookback call option, that is,

$$\left(\max_{t\in[0,T]} S_t - K\right)^+.$$

In this situation, instead of observing the price, it could be natural to observe the process of the running maximum of the price $\max_{u\in[0,t]} S_u$, and this amounts to observing one counting process.

This paper is divided into four sections and an appendix and is organized as follows. In section 2, we introduce different kinds of approximation for the filter that we will use throughout. These approximation schemes rely on the trajectories actually observed, and upper bounds for these approximations are obtained in Theorem 2.3. A relation between these upper bounds and weak convergence of the sequence of filters is presented in Theorem 2.6. In section 3, we consider the approximation of filters for Markov models with delayed observations for which we obtain upper bounds depending on the trajectory actually observed (see Proposition 3.1). In section 4, the

approximation results obtained in section 2, together with the representation for the filter in Theorem 1.1, will be applied to obtain some explicit upper bounds in the counting observation case, as in Remark 1.2 (see Theorem 4.5).

Some approximation results can be obtained for different systems and by making use of different techniques. As an example we can consider the simple stochastic delay system of Remark 1.3. Taking into account this remark, the computation of the filter is equivalent to the computation of the filter of the stochastic delay system (13), (14), for which the convergence result obtained in [5] can be applied. To our knowledge there are only two other papers dealing with the approximation of nonlinear filtering for delayed diffusion systems: Chang [10] and Twardowska, Marnik, and Pasławaska-Połuniak [22].

2. General considerations on approximation for filters.

2.1. Different kinds of approximation. Suppose that a sequence of stochastic systems $(\boldsymbol{X^n}, \boldsymbol{Y^n})$ with values in $\mathbb{R}^k \times \mathbb{R}^d$ converges to a system $(\boldsymbol{X}, \boldsymbol{Y})$. Then a natural question is whether the corresponding sequence of filters π^n converges to the filter of the limit system π.

Different kinds of convergence can be considered for both the systems and the filters. Convergence of their distributions is the first one, and, moreover, the only one that can be considered when the systems $(\boldsymbol{X^n}, \boldsymbol{Y^n})$ are defined on different probability spaces. The most frequently used is weak convergence of their distributions. Furthermore, in this case one can distinguish between convergence of π_t^n to π_t as random probability measures on $\mathbb{R}^k$, for each t, and convergence of the processes $\boldsymbol{\pi^n} = (\pi_t^n; t \geq 0)$ to $\boldsymbol{\pi} = (\pi_t; t \geq 0)$, as *càdlàg* measure-valued processes. Finally, one can consider different metrics $dist(\nu_1, \nu_2)$ on the space $\mathcal{P} = \mathcal{P}(\mathbb{R}^k)$, which is the space of probability measures on $\mathbb{R}^k$, such as the total variation, i.e.,

$$\|\nu_1 - \nu_2\|_{TV} = \sup\left\{\frac{|\nu_1(\varphi) - \nu_2(\varphi)|}{\|\varphi\|}; \varphi \text{ bounded}\right\},$$

the Kantorovitch metric, i.e.,

$$\boldsymbol{\kappa}(\nu_1, \nu_2) = \sup\left\{\frac{|\nu_1(\varphi) - \nu_2(\varphi)|}{L_\varphi}; \varphi \text{ Lipschitz}\right\},$$

or the bounded-Lipschitz metric, i.e.,

$$d_{BL}(\nu_1, \nu_2) = \sup\left\{\frac{|\nu_1(\varphi) - \nu_2(\varphi)|}{\|\varphi\| \vee L_\varphi}; \varphi \text{ bounded and Lipschitz}\right\},$$

where $\|\varphi\|$ denotes the sup-norm, and L_φ is the Lipschitz constant of φ. We recall that the bounded-Lipschitz metric induces the weak convergence topology on the space of probability measures $\mathcal{P}$.

In the following we will consider $\mathcal{P}$ endowed with one of the above metrics.

When instead all the processes are defined on the same probability space $(\Omega, \mathcal{F}, P)$, then one can also consider other kinds of convergence such as convergence in probability, convergence in $L^p([0,T] \times \Omega, dt \otimes dP)$, and so on.

This situation arises typically when one is interested in the limit system, and the sequence $(\boldsymbol{X^n}, \boldsymbol{Y^n})$ is constructed pathwise starting from the path $(\boldsymbol{X}, \boldsymbol{Y})$ such as, for instance, in the Euler approximations of diffusive systems.

When the filters $\boldsymbol{\pi^n}$ and $\boldsymbol{\pi}$ have a robust version, that is, when for suitable deterministic measure-valued functionals U^n and U, with paths in the Skorohod space $D_{\mathcal{P}}([0,T])$,

$$U^n,\, U:\ [0,T]\times D_{\mathbb{R}^d}([0,T])\mapsto \mathcal{P}$$

such that

$$U^n(t,\boldsymbol{y}) = U^n(t, y(\cdot\wedge t)) \text{ and } U(t,\boldsymbol{y}) = U(t, y(\cdot\wedge t)),$$

$$\pi^n_t(dx) = U^n(t,\boldsymbol{Y^n}; dx) \text{ and } \pi_t(dx) = U(t,\boldsymbol{Y}; dx),$$

then, in order to consider convergence in probability, one has to assume that the processes $\boldsymbol{Y^n}$ and $\boldsymbol{Y}$ are defined on the same probability space. Such deterministic functionals U^n and U satisfying the above properties always exist under very general conditions (see Kurtz and Ocone [16]). Note that the functionals U^n and U depend on the joint distribution of $(\boldsymbol{X^n},\boldsymbol{Y^n})$ and $(\boldsymbol{X},\boldsymbol{Y})$, respectively, and are defined in $D_{\mathbb{R}^d}([0,T])$ almost surely with respect to $P_{\boldsymbol{Y^n}}$, the law of $\boldsymbol{Y^n}$, and with respect to $P_{\boldsymbol{Y}}$, the law of $\boldsymbol{Y}$, respectively.

As an example one can consider approximating and limit models such as the jump model with counting observations, as given by (2) and (3), for which the functionals U and U^n can be computed as in (9), or such as the classical diffusive model, for which the functionals U and U^n are also computable as shown in [16], for example.

A different approach to the problem of the approximation of the filter takes into account that a realistic approximation depends on the trajectory actually observed and uses the robust representation described above. Two different types of situations can arise as follows:

1. The *true* model is $(\boldsymbol{X},\boldsymbol{Y})$, and therefore we observe $\boldsymbol{Y}$, while the models $(\boldsymbol{X^n},\boldsymbol{Y^n})$ are *more manageable* approximations.
2. The *true* model is $(\boldsymbol{X^n},\boldsymbol{Y^n})$, depending on a large parameter n, and therefore we observe $\boldsymbol{Y^n}$, while $(\boldsymbol{X},\boldsymbol{Y})$ is a *more manageable* limit model.

In these situations,

1. the *true* filter is π_t, and it is natural to consider

$$\tilde{\pi}^n_t = U^n(t,\boldsymbol{Y}) \tag{15}$$

   as an approximation of π_t, depending on the trajectory actually observed.
2. the *true* filter is π^n_t, and it is natural to consider

$$\overline{\pi}^n_t = U(t,\boldsymbol{Y^n}) \tag{16}$$

   as an approximation of π^n_t, depending on the trajectory actually observed.

The functional U^n is defined $P_{\boldsymbol{Y^n}}$-almost surely; therefore, in order to define $\tilde{\boldsymbol{\pi}}^{\boldsymbol{n}} = \{\tilde{\pi}^n_t; t\ge 0\}$ almost surely it is natural to assume that $P_{\boldsymbol{Y}}$ is absolutely continuous with respect to $P_{\boldsymbol{Y^n}}$. Indeed, when the functional U^n is defined $P_{\boldsymbol{Y^n}}$-almost surely, there exists a set $D^n\subset D_{\mathbb{R}^d}([0,T])$ with $P_{\boldsymbol{Y^n}}(D^n)=1$ such that the filter π^n_t is almost surely equal to $\mathbf{1}_{D^n}(\boldsymbol{Y^n})U^n(t,\boldsymbol{Y^n}) + (1-\mathbf{1}_{D^n}(\boldsymbol{Y^n}))\delta_{\{x_0\}}$ for any fixed point x_0, and one could always define

$$\tilde{\pi}^n_t = \mathbf{1}_{D^n}(\boldsymbol{Y})U^n(t,\boldsymbol{Y}) + (1-\mathbf{1}_{D^n}(\boldsymbol{Y}))\delta_{\{x_0\}}.$$

Furthermore if $P_{\boldsymbol{Y}}(D^n) = 0$, as an extreme example, then the chosen approximation would be clearly useless: indeed, in this case $\tilde{\pi}_t^n = \delta_{\{x_0\}}$ almost surely, and this is clearly not a good approximation, while if $P_{\boldsymbol{Y}}(D^n) = 1$, then the approximation coincides almost surely with $U^n(t, \boldsymbol{Y})$.

In analogy, the functional U is defined $P_{\boldsymbol{Y}}$-almost surely, and to define $\overline{\boldsymbol{\pi}}^{\boldsymbol{n}} = \{\overline{\pi}_t^n; t \geq 0\}$ almost surely, it is natural to assume that $P_{\boldsymbol{Y}^{\boldsymbol{n}}}$ is absolutely continuous with respect to $P_{\boldsymbol{Y}}$.

When using these kinds of approximations one is interested in evaluating

$$dist(\pi_t, \tilde{\pi}_t^n) = dist(U(t, \boldsymbol{Y}), U^n(t, \boldsymbol{Y})), \tag{17}$$

when we observe $\boldsymbol{Y}$ (situation 1), or

$$dist(\pi_t^n, \overline{\pi}_t^n) = dist(U^n(t, \boldsymbol{Y}^{\boldsymbol{n}}), U(t, \boldsymbol{Y}^{\boldsymbol{n}})), \tag{18}$$

when we observe $\boldsymbol{Y}^{\boldsymbol{n}}$ (situation 2). Here the $dist(\nu_1, \nu_2)$ can be, for example, the total variation, the Kantorovitch metric, or the bounded-Lipschitz metric.

Note that, in contrast to the approximations previously considered, neither $\tilde{\pi}_t^n$ nor $\overline{\pi}_t^n$ are conditional laws. Moreover, with this kind of approximation it is not even necessary that the sequence of processes $\{\boldsymbol{Y}^{\boldsymbol{n}}\}$ and $\boldsymbol{Y}$ be defined on the same probability space, and both the almost sure convergence and the convergence in probability can be considered.

For instance, in situation 1, the convergence in probability is implied by the convergence to zero of

$$\begin{aligned} E[dist(\pi_t, \tilde{\pi}_t^n)] &= E[dist(U(t, \boldsymbol{Y}), U^n(t, \boldsymbol{Y}))] \\ &= \int_{D_{\mathbb{R}^d}([0,T])} dist(U(t, \boldsymbol{y}), U^n(t, \boldsymbol{y})) dP_{\boldsymbol{Y}}(\boldsymbol{y}), \end{aligned} \tag{19}$$

while, in situation 2, the convergence in probability is implied by the convergence to zero of

$$\begin{aligned} E[dist(\pi_t^n, \bar{\pi}_t^n)] &= E[dist(U^n(t, \boldsymbol{Y}^{\boldsymbol{n}}), U(t, \boldsymbol{Y}^{\boldsymbol{n}}))] \\ &= \int_{D_{\mathbb{R}^d}([0,T])} dist(U^n(t, \boldsymbol{y}), U(t, \boldsymbol{y})) dP_{\boldsymbol{Y}^{\boldsymbol{n}}}(\boldsymbol{y}). \end{aligned} \tag{20}$$

Finally, note that when the distance between π_t and $\tilde{\pi}_t^n$ is given in terms of one of the integral metrics quoted at the beginning of this section, then

$$E[dist(\pi_t, \tilde{\pi}_t^n)] = E\left[\sup_{\varphi \in \mathcal{K}} |\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)|\right] \tag{21}$$

and

$$E[dist(\pi_t^n, \overline{\pi}_t^n)] = E\left[\sup_{\varphi \in \mathcal{K}} |\pi_t^n(\varphi) - \overline{\pi}_t^n(\varphi)|\right], \tag{22}$$

where $\mathcal{K}$ is a suitable class of functions: the class $\mathcal{K}_{TV}$ of functions φ with $\|\varphi\| \leq 1$ for the total variation, the class $\mathcal{K}_{BL}$ of functions φ with $\|\varphi\| \vee L_\varphi \leq 1$ for the bounded-Lipschitz metric, and the class $\mathcal{K}_K$ of functions φ with $L_\varphi \leq 1$ for the Kantorovitch metric.

2.2. General upper bounds for the approximations of the filters. In the following, we study the approximation problem for the filter by using a general procedure leading to a partial result, which is summarized in Theorem 2.3. Moreover, in the following we also consider the connections with weak convergence of the filters (cf. Theorem 2.6 and Remark 2.7). However, as will become clear in section 3, since these results are not sufficient for application to the models with delayed observations, we conclude this section with two results (cf. Theorem 2.9 and Proposition 2.11). These results will be applied to the Markov jump models with delayed counting observations in section 4.

Assume that it is possible to construct copies $(\tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}}^{\boldsymbol{n}})$ and $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ of the pairs $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y}^{\boldsymbol{n}})$ and $(\boldsymbol{X}, \boldsymbol{Y})$ on the same measurable space $(\Omega, \mathcal{F})$, equipped with different probability measures $\mathbb{P}$ and $\mathbb{P}^n$ and with the property that $\tilde{\boldsymbol{Y}}^{\boldsymbol{n}} = \tilde{\boldsymbol{Y}}$, i.e.,

(a) on $(\Omega, \mathcal{F}, \mathbb{P})$ the model $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ has the same law as the model $(\boldsymbol{X}, \boldsymbol{Y})$,

($\mathrm{a_n}$) on $(\Omega, \mathcal{F}, \mathbb{P}^n)$ the model $(\tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}}^{\boldsymbol{n}}) = (\tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}})$ has the same law as the model $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y}^{\boldsymbol{n}})$.

Remark 2.1. In particular, when the systems $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y}^{\boldsymbol{n}})$ are Markovian, condition (a) means that under $\mathbb{P}$ the pair $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ has the generator $\boldsymbol{L}$ of $(\boldsymbol{X}, \boldsymbol{Y})$, and the initial condition $\mu = \mu_0^{X,Y}$, while condition ($\mathrm{a_n}$) means that under $\mathbb{P}^n$ the pair $(\tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}})$ has the generator $\boldsymbol{L}^n$ of $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y}^{\boldsymbol{n}})$, and the initial condition $\mu^n = \mu_0^{X^n,Y^n}$. Moreover, if the systems are Markov with generators $\boldsymbol{L}$ and $\boldsymbol{L}^{\boldsymbol{n}}$ and initial conditions $\mu_0^{X,Y} = \delta_{\{x,y\}}$ and $\mu_0^{X^n,Y^n} = \delta_{\{x,y\}}$, then the semigroups can be represented, respectively, as

$$\exp\{\boldsymbol{L}t\}\phi(x,y) = \mathbb{E}\big[\phi(\tilde{X}_t, \tilde{Y}_t)\big] \quad \text{and} \quad \exp\{\boldsymbol{L}^{\boldsymbol{n}}t\}\phi(x,y) = \mathbb{E}^n\big[\phi(\tilde{X}_t^n, \tilde{Y}_t)\big].$$

As will become clear in the application to the jump models in section 4, it is natural to construct the probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega, \mathcal{F}, \mathbb{P}^n)$ starting from a given probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ in such a way that

(b1) $\mathbb{P}$ and $\mathbb{P}^n$ are absolutely continuous with respect to $\mathbb{Q}$ on $\tilde{\mathcal{F}}_t = \mathcal{F}_t^{\tilde{X},\tilde{X}^n,\tilde{Y}}$ for all $t \geq 0$.

The above condition implies that $\mathbb{P}$ and $\mathbb{P}^n$ are absolutely continuous with respect to $\mathbb{Q}$ on $\mathcal{F}_t^{\tilde{Y}}$. For technical reasons, it is more convenient to assume the further condition that

(b2) the probability measures $\mathbb{Q}$, $\mathbb{P}$, and $\mathbb{P}^n$ are equivalent on $\mathcal{F}_t^{\tilde{Y}}$ for all $t \geq 0$.

Indeed, if conditions (a), ($\mathrm{a_n}$), and (b2) hold, then

$$U(t, \tilde{\boldsymbol{Y}}; \varphi) = \mathbb{E}[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}],\ \mathbb{P}\text{-}a.s. \quad \text{(and therefore also } \mathbb{Q}\text{-a.s. and } \mathbb{P}^n\text{-a.s.)},$$

and analogously,

$$U^n(t, \tilde{\boldsymbol{Y}}; \varphi) = \mathbb{E}^n[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}],\ \mathbb{P}^n\text{-}a.s. \quad \text{(and therefore also } \mathbb{Q}\text{-a.s. and } \mathbb{P}\text{-a.s.)}.$$

Therefore,

$$E[dist(\pi_t, \tilde{\pi}_t^n)] = E\left[\sup_{\varphi\in\mathcal{K}} |\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)|\right] \tag{23}$$

$$= E\left[\sup_{\varphi\in\mathcal{K}} |U(t, \boldsymbol{Y}; \varphi) - U^n(t, \boldsymbol{Y}; \varphi)|\right] \tag{24}$$

$$= \mathbb{E}\left[\sup_{\varphi\in\mathcal{K}} |U(t, \tilde{\boldsymbol{Y}}; \varphi) - U^n(t, \tilde{\boldsymbol{Y}}; \varphi)|\right]$$

$$= \mathbb{E}\left[\sup_{\varphi\in\mathcal{K}} |\mathbb{E}[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}] - \mathbb{E}^n[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}]|\right],$$

where we stress that, under $\mathbb{P}$, the law of $U^n(t,\tilde{\boldsymbol{Y}};\varphi)=\mathbb{E}^n[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}]$ is the same as the law of $\tilde{\pi}_t^n(\varphi)$, and therefore not the law of $\pi_t^n(\varphi)$, the filter of the approximating system, evaluated at φ. Analogously,

$$
\begin{aligned}
(25)\qquad E[dist(\pi_t^n,\overline{\pi}_t^n) &= E\left[\sup_{\varphi\in\mathcal{K}}|\pi_t^n(\varphi)-\overline{\pi}_t^n(\varphi)|\right]\\
&= E\left[\sup_{\varphi\in\mathcal{K}}|U^n(t,\boldsymbol{Y^n};\varphi)-U(t,\boldsymbol{Y^n};\varphi)|\right]\\
&= \mathbb{E}^n\left[\sup_{\varphi\in\mathcal{K}}|U^n(t,\tilde{\boldsymbol{Y}};\varphi)-U(t,\tilde{\boldsymbol{Y}};\varphi)|\right]\\
&= \mathbb{E}^n\left[\sup_{\varphi\in\mathcal{K}}|\mathbb{E}^n[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}]-\mathbb{E}[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}]|\right].
\end{aligned}
$$

Thus our aim is equivalent to proving the convergence to zero of either

$$\mathbb{E}[dist(U^n(t,\tilde{\boldsymbol{Y}}),U(t,\tilde{\boldsymbol{Y}}))]$$

or

$$\mathbb{E}^n[dist(U^n(t,\tilde{\boldsymbol{Y}}),U(t,\tilde{\boldsymbol{Y}}))],$$

depending on which metric and situation we are interested in. In any case it is natural to start by looking for $\mathbb{Q}$-a.s. upper bounds of

$$\sup_{\varphi\in\mathcal{K}}\left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right]-\mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right|,$$

with $\mathcal{K}$ depending on the metric, and then to take the expectation with respect to either $\mathbb{E}$ or $\mathbb{E}^n$, depending on the situation we are interested in.

Furthermore, if condition (b1) holds, it can be shown (e.g., by Lemma 4.1 of [8]) that, for any bounded function φ,

$$
(26)\qquad \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right]-\mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right|
$$

is bounded above either by

$$
2\|\varphi\|\,\frac{\mathbb{E}_{\mathbb{Q}}\left[|(d\mathbb{P}^n/d\mathbb{Q})|_{\widetilde{\mathcal{F}}_t}-(d\mathbb{P}/d\mathbb{Q})|_{\widetilde{\mathcal{F}}_t}|/\mathcal{F}_t^{\tilde{Y}}\right]}{\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}^n/d\mathbb{Q})|_{\widetilde{\mathcal{F}}_t}/\mathcal{F}_t^{\tilde{Y}}\right]}+\mathbb{E}^n\left[|\varphi(\tilde{X}_t)-\varphi(\tilde{X}_t^n)|/\mathcal{F}_t^{\tilde{Y}}\right],
$$

$\mathbb{Q}$-a.s., and therefore also $\mathbb{P}$-a.s. and $\mathbb{P}^n$-a.s., or by

$$
2\|\varphi\|\,\frac{\mathbb{E}_{\mathbb{Q}}\left[|(d\mathbb{P}/d\mathbb{Q})|_{\widetilde{\mathcal{F}}_t}-(d\mathbb{P}^n/d\mathbb{Q})|_{\widetilde{\mathcal{F}}_t}|/\mathcal{F}_t^{\tilde{Y}}\right]}{\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}/d\mathbb{Q})|_{\widetilde{\mathcal{F}}_t}/\mathcal{F}_t^{\tilde{Y}}\right]}+\mathbb{E}\left[|\varphi(\tilde{X}_t)-\varphi(\tilde{X}_t^n)|/\mathcal{F}_t^{\tilde{Y}}\right],
$$

$\mathbb{Q}$-a.s., and therefore also $\mathbb{P}$-a.s. and $\mathbb{P}^n$-a.s. Again we use the first or the second upper bound, depending on which situation we are interested in.

Taking into account that

$$|\varphi(\tilde{X}_t)-\varphi(\tilde{X}_t^n)|\le 2\|\varphi\|\,\mathbb{I}_{\{\tilde{X}_t\neq\tilde{X}_t^n\}},$$

or that, if φ is also a Lipschitz function,

$$|\varphi(\tilde{X}_t)-\varphi(\tilde{X}_t^n)|\le L_\varphi\left|\tilde{X}_t-\tilde{X}_t^n\right|,$$

the previous observations can be used to get upper bounds for $E[dist(\pi_t, \tilde{\pi}_t^n)]$ and for $E[dist(\pi_t^n, \overline{\pi}_t^n)]$ when using the total variation or the bounded-Lipschitz metric. Indeed, if one is interested in total variation, one has the following upper bounds, depending on the trajectories actually observed: either

$$\sup_{\varphi\in\mathcal{K}_{TV}} \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right] - \mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right| \le 2\,\mathcal{Z}_t^n + 2\,\mathbb{P}^n\left(\{\tilde{X}_t \neq \tilde{X}_t^n\}/\mathcal{F}_t^{\tilde{Y}}\right)$$

or

$$\sup_{\varphi\in\mathcal{K}_{TV}} \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right] - \mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right| \le 2\,\mathcal{Z}_t + 2\,\mathbb{P}\left(\{\tilde{X}_t \neq \tilde{X}_t^n\}/\mathcal{F}_t^{\tilde{Y}}\right),$$

where, as in the rest of this paper, the following notation is used:

$$\mathcal{Z}_t^n = \frac{\mathbb{E}_{\mathbb{Q}}\left[\left|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\right|/\mathcal{F}_t^{\tilde{Y}}\right]}{\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}/\mathcal{F}_t^{\tilde{Y}}\right]} \tag{27}$$

and

$$\mathcal{Z}_t = \frac{\mathbb{E}_{\mathbb{Q}}\left[\left|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\right|/\mathcal{F}_t^{\tilde{Y}}\right]}{\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}/\mathcal{F}_t^{\tilde{Y}}\right]}. \tag{28}$$

If one is interested in the bounded-Lipschitz metric, one has either

$$\sup_{\varphi\in\mathcal{K}_{BL}} \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right] - \mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right| \le 2\,\mathcal{Z}_t^n + \mathbb{E}^n\left[\left|\tilde{X}_t - \tilde{X}_t^n\right|/\mathcal{F}_t^{\tilde{Y}}\right]$$

or

$$\sup_{\varphi\in\mathcal{K}_{BL}} \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right] - \mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right| \le 2\,\mathcal{Z}_t + \mathbb{E}\left[\left|\tilde{X}_t - \tilde{X}_t^n\right|/\mathcal{F}_t^{\tilde{Y}}\right].$$

Remark 2.2. Considering that

$$\begin{aligned}|\varphi(\tilde{X}_t) - \varphi(\tilde{X}_t^n)| &\le \max(2\|\varphi\|, L_\varphi)\left(\left|\tilde{X}_t - \tilde{X}_t^n\right| \wedge 1\right)\\ &\le 2\max(\|\varphi\|, L_\varphi)\left(\left|\tilde{X}_t - \tilde{X}_t^n\right| \wedge 1\right),\end{aligned}$$

one can also have the bounds with either

$$\sup_{\varphi\in\mathcal{K}_{BL}} \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right] - \mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right| \le 2\,\mathcal{Z}_t^n + 2\,\mathbb{E}^n\left[\left|\tilde{X}_t - \tilde{X}_t^n\right| \wedge 1/\mathcal{F}_t^{\tilde{Y}}\right]$$

or

$$\sup_{\varphi\in\mathcal{K}_{BL}} \left|\mathbb{E}\left[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\right] - \mathbb{E}^n\left[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\right]\right| \le 2\,\mathcal{Z}_t + 2\,\mathbb{E}\left[\left|\tilde{X}_t - \tilde{X}_t^n\right| \wedge 1/\mathcal{F}_t^{\tilde{Y}}\right].$$

Moreover, an easy computation gives

$$\mathbb{E}^n\left[\mathcal{Z}_t^n\right] = \mathbb{E}\left[\mathcal{Z}_t\right] = \mathbb{E}_{\mathbb{Q}}\left[\left|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\right|\right].$$

Indeed,

$$\begin{aligned}\mathbb{E}^n\left[\mathcal{Z}_t^n\right] &= \mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} \frac{\mathbb{E}_{\mathbb{Q}}\left[\left|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\right|/\mathcal{F}_t^{\tilde{Y}}\right]}{\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}/\mathcal{F}_t^{\tilde{Y}}\right]}\right]\\ &= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{I}_{\{(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\neq 0\}}\mathbb{E}_{\mathbb{Q}}\left[\left|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\right|/\mathcal{F}_t^{\tilde{Y}}\right]\right]\\ &= \mathbb{E}_{\mathbb{Q}}\left[\left|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\right|\right], \qquad (29)\end{aligned}$$

where in the last equality we have used condition (b2), and analogously we get

$$\mathbb{E}\big[\mathcal{Z}_t\big] = \mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\big].$$

As a consequence of the above analysis, by taking (23) and (25) into account, we get the following upper bounds.

THEOREM 2.3. *Under conditions* (a), (a_n), (b1), *and* (b2),

$$E\big[\|\pi_t - \tilde{\pi}_t^n\|_{TV}\big] \le 2\mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\big] + 2\mathbb{P}\big(\{\tilde{X}_t \neq \tilde{X}_t^n\}\big) \tag{30}$$

or

$$E\big[\|\pi_t^n - \overline{\pi}_t^n\|_{TV}\big] \le 2\mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\big] + 2\mathbb{P}^n\big(\{\tilde{X}_t \neq \tilde{X}_t^n\}\big), \tag{31}$$

and

$$E\big[d_{BL}(\pi_t, \tilde{\pi}_t^n)\big] \le 2\mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\big] + \mathbb{E}\big[\big|\tilde{X}_t - \tilde{X}_t^n\big|\big] \tag{32}$$

or

$$E\big[d_{BL}(\pi_t^n, \overline{\pi}_t^n)\big] \le 2\mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\big] + \mathbb{E}^n\big[\big|\tilde{X}_t - \tilde{X}_t^n\big|\big]. \tag{33}$$

Remark 2.4. It is clear from the proof that we can rewrite the above inequalities (30)–(33) by substituting $\mathbb{E}^n[\mathcal{Z}_t^n]$ or $\mathbb{E}[\mathcal{Z}_t]$ in the first addend. Moreover, when in addition $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{P}^n$ on $\tilde{\mathcal{F}}_t$, we can substitute

$$\mathbb{E}\left[\left|\frac{d\mathbb{P}^n}{d\mathbb{P}}\big|_{\tilde{\mathcal{F}}_t} - 1\right|\right]$$

in the first addend, since

$$\mathbb{E}^n\left[\left|1 - \frac{d\mathbb{P}}{d\mathbb{P}^n}\big|_{\tilde{\mathcal{F}}_t}\right|\right] = \mathbb{E}_{\mathbb{Q}}\left[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\right].$$

An analogous substitution can be done when, in addition, $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{P}^n$ on $\tilde{\mathcal{F}}_t$, since

$$\mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\big|\big] = \mathbb{E}\left[\left|\frac{d\mathbb{P}^n}{d\mathbb{P}}\big|_{\tilde{\mathcal{F}}_t} - 1\right|\right].$$

Remark 2.5. By taking into account the estimates obtained in Remark 2.2 one can substitute in (32) and (33) the expectations of $\big|\tilde{X}_t - \tilde{X}_t^n\big|$ with the corresponding expectations of $2\big|\tilde{X}_t - \tilde{X}_t^n\big| \wedge 1$.

In the following theorem and remark we highlight the connections of the above considerations with the weak convergence of the sequence of filters π_t^n to π_t (see [1] and [2] for similar considerations).

THEOREM 2.6. *Assume that*

1. (a), (a_n), *and* (b2) *are satisfied;*
2. *the sequence* $U^n(t, \tilde{\boldsymbol{Y}})$ *converges in* $\mathbb{Q}$*-probability to* $U(t, \tilde{\boldsymbol{Y}})$ *as random variables with values in* $\mathcal{P}$, *the space of probability measures endowed with the metric dist, that is, for all* $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{Q}\big(A_\varepsilon^n\big) = 0, \quad \textit{where } A_\varepsilon^n := \big\{dist\big(U^n(t, \tilde{\boldsymbol{Y}}), U(t, \tilde{\boldsymbol{Y}})\big) \ge \varepsilon\big\};$$

3. $\lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} - (d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\big|\big] = 0.$

Then, the filters converge in distribution for each t as random variables with values in $\mathcal{P}$, i.e.,

$$\pi_t^n \Rightarrow \pi_t.$$

Remark 2.7. (I) It is important to stress the connection of the above results to the convergence in probability of (17) and (18). Indeed, conditions 2 and (b2) imply that $\mathbb{P}\big(A_\varepsilon^n\big)$ converge to zero, and thus by condition (a),

$$P\big(dist\big(\tilde{\pi}_t^n;\pi_t\big)\geq\varepsilon\big) = P\big(dist\big(U^n(t,\boldsymbol{Y}),U(t,\boldsymbol{Y})\big)\geq\varepsilon\big) = \mathbb{P}\big(A_\varepsilon^n\big)\to 0.$$

Similarly, by condition ($\mathrm{a_n}$),

$$P\big(dist\big(\pi_t^n;\overline{\pi}_t^n\big)\geq\varepsilon\big) = P\big(dist\big(U^n(t,\boldsymbol{Y^n}),U(t,\boldsymbol{Y^n})\big)\geq\varepsilon\big) = \mathbb{P}^n\big(A_\varepsilon^n\big),$$

and, by conditions 2, 3, and (b2), $\mathbb{P}^n(A_\varepsilon^n)$ converge to zero: the set

$$A_\varepsilon^n = \big\{dist\big(U^n(t,\tilde{\boldsymbol{Y}}),U(t,\tilde{\boldsymbol{Y}})\big)\geq\varepsilon\big\}$$

is $\mathcal{F}_t^{\tilde{Y}}$-measurable, and by condition (b2),

$$\begin{aligned}\mathbb{P}^n(A_\varepsilon^n) &= \mathbb{E}_{\mathbb{Q}}\big[\ (d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\,\mathbb{I}_{A_\varepsilon^n}\big]\\ &= \mathbb{E}_{\mathbb{Q}}\big[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\mathbb{I}_{A_\varepsilon^n}\big] + \mathbb{E}_{\mathbb{Q}}\big[\big[(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} - (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\big]\mathbb{I}_{A_\varepsilon^n}\big]\\ &\leq \mathbb{E}_{\mathbb{Q}}\big[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\mathbb{I}_{A_\varepsilon^n}\big] + \mathbb{E}_{\mathbb{Q}}\big[\big|(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} - (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\big|\big].\end{aligned}$$

The first addend in the last line, i.e., $\mathbb{E}_{\mathbb{Q}}[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}\mathbb{I}_{A_\varepsilon^n}] = \mathbb{P}(A_\varepsilon^n)$, converges to zero by conditions 2 and (b2), while condition 3 asserts that the second addend converges to zero.

(II) Note that if condition (b1) also holds, then condition 3 is implied by the stronger condition

$$3'\ \lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}}\big[|\ (d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - \ (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}\ |\big] = 0,$$

that is, the convergence to zero of the first addend in all the upper bounds (30)–(33).

Therefore, if we can prove that the second addend in upper bound (30) or (32), depending on the chosen metric, also converges to zero, we get the convergence to zero of $\mathbb{P}(A_\varepsilon^n)$. As a consequence condition 2 is automatically satisfied and, provided condition 1 is enforced, we get the weak convergence of the filters. The convergence to zero in condition $3'$ can also be used to get the upper bounds for the second addends, as in section 4 for the counting observation case.

An analogous result holds for the convergence to zero of $\mathbb{P}^n(A_\varepsilon^n)$, when assuming that the second addend in the upper bound (31) or (33), depending on the chosen metric, also converges to zero, since one can prove that $\mathbb{P}(A_\varepsilon^n)$ converge to zero.

Proof of Theorem 2.6. For any bounded function $G:\ \mathcal{P}\to\mathbb{R}$, continuous with respect to *dist*,

$$\begin{aligned}E[G(\pi_t^n)] \ &= \ \mathbb{E}^n\big[G\big(U^n(t,\tilde{Y})\big)\big] = \mathbb{E}_{\mathbb{Q}}\big[(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}G\big(U^n(t,\tilde{Y})\big)\big]\\ &\to \mathbb{E}_{\mathbb{Q}}\big[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}G\big(U(t,\tilde{Y})\big)\big] = \mathbb{E}\big[G\big(U(t,\tilde{Y})\big)\big] = E[G(\pi_t)].\end{aligned}$$

All the equalities in the above relations are obvious; we need only prove the convergence result.

We can restrict to G uniformly continuous, with modulus of continuity $\boldsymbol{w}_G(\varepsilon)$:

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} G\big(U^n(t,\tilde{Y})\big)\right] - \mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} G\big(U(t,\tilde{Y})\big)\right]\right| \\
&\quad \le \mathbb{E}_{\mathbb{Q}}\left[|(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} - (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}|\,|G\big(U^n(t,\tilde{Y})\big)|\right] \\
&\qquad + \mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} |G\big(U^n(t,\tilde{Y})\big) - G\big(U(t,\tilde{Y})\big)|\right] \\
&\quad \le \|G\|\,\mathbb{E}_{\mathbb{Q}}\left[|(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} - (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}}|\right] \\
&\quad + \mathbb{E}_{\mathbb{Q}}\Big[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{Y}}} \Big(2\|G\|\,\mathbf{1}_{\{dist(U^n(t,\tilde{Y}),U(t,\tilde{Y}))\ge\varepsilon\}} + \boldsymbol{w}_G(\varepsilon)\Big)\Big].
\end{aligned}$$

The first addend converges to zero by condition 3. The second addend is equal to

$$2\|G\|\mathbb{P}(A_\varepsilon^n) + \boldsymbol{w}_G(\varepsilon)$$

and converges to zero by conditions 2 and (b2) and the arbitrariness of ε. □

This result has been obtained by using the same kind of techniques as in [1]. In the latter paper the context is more general in the sense that the state space $\mathcal{S}$ is a complete separable metric space, and the filters converge weakly as measure-valued processes in $D_{\mathcal{P}(\mathcal{S})}([0,T])$. However, in [1] the space $\mathcal{P}(\mathcal{S})$ is always endowed with the Prokhorov topology, while in this paper we consider different kinds of metrics.

Remark 2.8. If one is interested in weak convergence of the filters $\boldsymbol{\pi^n}$ to $\boldsymbol{\pi}$ as random processes in $\mathcal{D}_{\mathcal{P}}[0,T]$, then condition 2 can be replaced by

$2'$ $\lim_{n\to\infty} \mathbb{Q}(\sup_{t\in[0,T]} dist(U^n(t,\tilde{\boldsymbol{Y}}),U(t,\tilde{\boldsymbol{Y}})) \ge \varepsilon) = 0$,

and condition 3 must be satisfied for $t = T$:

$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{Q}}\left[|\ (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} - (d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}}\ |\right] = 0.$$

Indeed, for any bounded function $G : \mathcal{D}_{\mathcal{P}}[0,T] \to \mathbb{R}$, continuous with respect to ***Dist***, the Skorokhod metric, we have

$$\begin{aligned}
E[G(\boldsymbol{\pi^n})] &= \mathbb{E}^n\left[G\big(U^n(\cdot,\tilde{Y})\big)\right] = \mathbb{E}_{\mathbb{Q}}\left[\ (d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}}\, G\big(U^n(\cdot,\tilde{Y})\big)\right] \\
&\to \mathbb{E}_{\mathbb{Q}}\left[\ (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}}\, G\big(U(\cdot,\tilde{Y})\big)\right] = \mathbb{E}\left[G\big(U(\cdot,\tilde{Y})\big)\right] = E[G(\boldsymbol{\pi})].
\end{aligned}$$

All the equalities in the above relations are obvious; we need only prove the convergence result.

We can restrict to G uniformly continuous, with modulus of continuity $\boldsymbol{w}_G^{\mathcal{S}}(\varepsilon)$ with respect to the Skorokhod metric:

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} G\big(U^n(\cdot,\tilde{Y})\big)\right] - \mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} G\big(U(\cdot,\tilde{Y})\big)\right]\right| \\
&\quad \le \mathbb{E}_{\mathbb{Q}}\left[|(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} - (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}}|\,|G\big(U^n(\cdot,\tilde{Y})\big)|\right] \\
&\qquad + \mathbb{E}_{\mathbb{Q}}\left[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} |G\big(U^n(\cdot,\tilde{Y})\big) - G\big(U(\cdot,\tilde{Y})\big)|\right] \\
&\quad \le \|G\|\,\mathbb{E}_{\mathbb{Q}}\left[|(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} - (d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}}|\right] \\
&\qquad + \mathbb{E}_{\mathbb{Q}}\Big[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_T^{\tilde{Y}}} \Big(2\|G\|\,\mathbf{1}_{\{\boldsymbol{Dist}(U^n(\cdot,\tilde{Y}),U(\cdot,\tilde{Y}))\ge\varepsilon\}} + \boldsymbol{w}_G^{\mathcal{S}}(\varepsilon)\Big)\Big].
\end{aligned}$$

The first addend converges to zero by condition 3 for $t = T$. The second addend is

equal to

$$2\|G\|\mathbb{P}(\boldsymbol{Dist}\big(U^n(\cdot,\tilde{Y}),U(\cdot,\tilde{Y})\big)\geq\varepsilon)+\boldsymbol{w}_G^{\mathcal{S}}(\varepsilon)$$
$$\leq 2\|G\|\mathbb{P}\bigg(\sup_{[0,T]} dist\big(U^n(t,\tilde{Y}),U(t,\tilde{Y})\big)\geq\varepsilon\bigg)+\boldsymbol{w}_G^{\mathcal{S}}(\varepsilon)$$

and converges to zero by conditions $2'$ and (b2) and the arbitrariness of ε.

2.3. Further upper bounds. In the next section we will use a slight modification of the above procedure, taking into account the delay structure of the model we are dealing with in order to prove our approximation results. Indeed, for technical reasons we will need the obvious observation

$$\sup_{\varphi\in\mathcal{K}}\big|\mathbb{E}\big[\varphi(\tilde{X}_t)/\mathcal{F}_t^{\tilde{Y}}\big]-\mathbb{E}^n\big[\varphi(\tilde{X}_t^n)/\mathcal{F}_t^{\tilde{Y}}\big]\big|\leq\sup_{\phi\in\mathcal{K}'}\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)/\mathcal{F}_t^{\tilde{Y}}\big]-\mathbb{E}^n\big[\phi(\tilde{X}_t^n,\tilde{Y}_t)/\mathcal{F}_t^{\tilde{Y}}\big]\big|$$

whenever $\mathcal{K}'$ contains all functions $\phi(x,y)$ such that $\phi(x,y)=\varphi(x)$, with $\varphi\in\mathcal{K}$. The r.h.s. of the last inequality is equal to

$$\sup_{\phi\in\mathcal{K}'}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|,$$

which coincides with

$$\sup_{\phi\in\mathcal{K}'}|\pi_t(\phi(\cdot,\tilde{Y}_t))-\tilde{\pi}_t^n(\phi(\cdot,\tilde{Y}_t))|,\qquad\mathbb{P}\text{-a.s.},$$

and with

$$\sup_{\phi\in\mathcal{K}'}|\bar{\pi}_t^n(\phi(\cdot,\tilde{Y}_t))-\pi_t^n(\phi(\cdot,\tilde{Y}_t))|,\qquad\mathbb{P}^n\text{-a.s.},$$

and hence, our aim will be accomplished once we prove the convergence to zero of the expectation of the latter quantity for a suitable choice of $\mathcal{K}'$, with the expectation taken with respect to $\mathbb{P}$ or $\mathbb{P}^n$, depending on which situation we are interested in. With this aim, denote by $\mathcal{K}_{TV}(\alpha)$ the class of measurable functions $\varphi(x)$ bounded above by α, and denote by $\mathcal{K}_{BL}(\alpha,\Lambda)$ the class of measurable functions $\varphi(x)$ bounded above by α, such that for all x, x',

$$|\varphi(x)-\varphi(x')|\leq\Lambda|x-x'|,$$

and thus $\mathcal{K}_{TV}=\mathcal{K}_{TV}(1)$ and $\mathcal{K}_{BL}=\mathcal{K}_{BL}(1,1)$. $\mathcal{K}'_{TV}(\alpha)$ is the class of measurable functions $\phi(x,y)$ bounded above by α, and $\mathcal{K}'_{BL}(\alpha,\Lambda)$ is the class of measurable functions $\phi(x,y)$ bounded above by α, such that for all x, x', and y,

$$|\phi(x,y)-\phi(x',y)|\leq\Lambda|x-x'|.$$

Note that $\mathcal{K}'_{TV}(\alpha)$ contains all functions $\phi(x,y)=\varphi(x)$, with $\varphi\in\mathcal{K}_{TV}(\alpha)$, and that, analogously, $\mathcal{K}'_{BL}(\alpha,\Lambda)$ contains all functions $\phi(x,y)=\varphi(x)$, with $\varphi\in\mathcal{K}_{BL}(\alpha,\Lambda)$.

THEOREM 2.9. *Assume that conditions* (a), (a_n), (b1), *and* (b2) *are satisfied; then the following inequalities hold:*

$$\mathbb{E}^n\bigg[\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\bigg]\leq 2\alpha\,\mathbb{E}^n\big[\mathcal{Z}_t^n\big]+2\alpha\,\mathbb{P}^n\big(\{\tilde{X}_t\neq\tilde{X}_t^n\}\big),$$

$$\mathbb{E}\Big[\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]\leq 2\alpha\,\mathbb{E}\big[\mathcal{Z}_t\big]+2\alpha\,\mathbb{P}\big(\{\tilde{X}_t\neq\tilde{X}_t^n\}\big),$$

$$\mathbb{E}^n\Big[\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]\leq 2\alpha\,\mathbb{E}^n\big[\mathcal{Z}_t^n\big]+\Lambda\,\mathbb{E}^n\big[\big|\tilde{X}_t-\tilde{X}_t^n\big|\big],$$

$$\mathbb{E}\Big[\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]\leq 2\alpha\,\mathbb{E}\big[\mathcal{Z}_t\big]+\Lambda\,\mathbb{E}\big[\big|\tilde{X}_t-\tilde{X}_t^n\big|\big],$$

where $\mathcal{Z}_t^n$ and $\mathcal{Z}_t$ are defined as in (27) *and* (28)*, respectively.*

Proof. It can be shown (e.g., by Lemma 4.1 of [8]) that, for any bounded function ϕ, the following inequalities hold $\mathbb{Q}$-a.s., and therefore also $\mathbb{P}$-a.s. and $\mathbb{P}^n$-a.s.:

$$\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)/\mathcal{F}_t^{\tilde{Y}}\big]-\mathbb{E}^n\big[\phi(\tilde{X}_t^n,\tilde{Y}_t)/\mathcal{F}_t^{\tilde{Y}}\big]\big|\leq 2\|\phi\|\,\mathcal{Z}_t^n+\mathbb{E}^n\big[|\phi(\tilde{X}_t,\tilde{Y}_t)-\phi(\tilde{X}_t^n,\tilde{Y}_t)|/\mathcal{F}_t^{\tilde{Y}}\big]$$

and

$$\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)/\mathcal{F}_t^{\tilde{Y}}\big]-\mathbb{E}^n\big[\phi(\tilde{X}_t^n,\tilde{Y}_t)/\mathcal{F}_t^{\tilde{Y}}\big]\big|\leq 2\|\phi\|\,\mathcal{Z}_t+\mathbb{E}\big[|\phi(\tilde{X}_t,\tilde{Y}_t)-\phi(\tilde{X}_t^n,\tilde{Y}_t)|/\mathcal{F}_t^{\tilde{Y}}\big].$$

Taking into account that, if $\phi\in\mathcal{K}'_{TV}(\alpha)$, we have

$$|\phi(\tilde{X}_t,\tilde{Y}_t)-\phi(\tilde{X}_t^n,\tilde{Y}_t)|\leq 2\alpha\,\mathbb{I}_{\{\tilde{X}_t\neq\tilde{X}_t^n\}},$$

or that, if $\phi\in\mathcal{K}'_{BL}(\alpha;\Lambda)$, we have

$$|\phi(\tilde{X}_t,\tilde{Y}_t)-\phi(\tilde{X}_t^n,\tilde{Y}_t)|\leq\Lambda\,\big|\tilde{X}_t-\tilde{X}_t^n\big|,$$

the previous observations can be used to get upper bounds for $E\big[dist(\pi_t,\tilde{\pi}_t^n)\big]$ and $E\big[dist(\pi_t^n,\overline{\pi}_t^n)\big]$ when using the total variation or the bounded-Lipschitz metric. Indeed, if we are interested in total variation, one has either

$$\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\leq 2\alpha\,\mathcal{Z}_t^n+2\alpha\,\mathbb{P}^n\big(\{\tilde{X}_t\neq\tilde{X}_t^n\}/\mathcal{F}_t^{\tilde{Y}}\big)$$

or

$$\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\leq 2\alpha\,\mathcal{Z}_t+2\alpha\,\mathbb{P}\big(\{\tilde{X}_t\neq\tilde{X}_t^n\}/\mathcal{F}_t^{\tilde{Y}}\big),$$

and, if we are interested in the bounded-Lipschitz metric, one has either

$$\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\leq 2\alpha\,\mathcal{Z}_t^n+\Lambda\,\mathbb{E}^n\big[\big|\tilde{X}_t-\tilde{X}_t^n\big|/\mathcal{F}_t^{\tilde{Y}}\big]$$

or

$$\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\leq 2\alpha\,\mathcal{Z}_t+\Lambda\,\mathbb{E}\big[\big|\tilde{X}_t-\tilde{X}_t^n\big|/\mathcal{F}_t^{\tilde{Y}}\big].$$

Taking the expectation we get the stated inequalities. □

Remark 2.10. Noting that $\mathcal{K}'_{BL}(\alpha,\Lambda)$ is contained in $\mathcal{K}'_{TV}(\alpha)$ from the upper bounds in the last two inequalities in the above Theorem 2.9 we get

$$\mathbb{E}^n\Big[\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]$$
$$\leq 2\alpha\,\mathbb{E}^n\big[\mathcal{Z}^n_t\big]+\min\Big(2\alpha\,\mathbb{P}^n\big(\{\tilde{X}_t\neq\tilde{X}^n_t\}\big);\,\Lambda\,\mathbb{E}^n\big[\big|\tilde{X}_t-\tilde{X}^n_t\big|\big]\Big),$$

$$\mathbb{E}\Big[\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]$$
$$\leq 2\alpha\,\mathbb{E}\big[\mathcal{Z}_t\big]+\min\Big(2\alpha\,\mathbb{P}\big(\{\tilde{X}_t\neq\tilde{X}^n_t\}\big);\,\Lambda\,\mathbb{E}\big[\big|\tilde{X}_t-\tilde{X}^n_t\big|\big]\Big).$$

Furthermore, observing that for any $\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)$,

$$|\phi(x,y)-\phi(x',y)|\leq\max(2\alpha;\,\Lambda)\big(|x-x'|\wedge 1\big),$$

by arguments similar to the proof of Theorem 2.9, we get

$$\mathbb{E}^n\Big[\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]$$
$$\leq 2\alpha\,\mathbb{E}^n\big[\mathcal{Z}^n_t\big]+\max(2\alpha;\,\Lambda)\,\mathbb{E}^n\big[\big|\tilde{X}_t-\tilde{X}^n_t\big|\wedge 1\big],$$

$$\mathbb{E}\Big[\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}|U(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))-U^n(t,\tilde{\boldsymbol{Y}};\phi(\cdot,\tilde{Y}_t))|\Big]$$
$$\leq 2\alpha\,\mathbb{E}\big[\mathcal{Z}_t\big]+\max(2\alpha;\,\Lambda)\,\mathbb{E}\big[\big|\tilde{X}_t-\tilde{X}^n_t\big|\wedge 1\big].$$

Moreover, in the next section we will also use the following result.

PROPOSITION 2.11. *Assume that conditions* (a), ($\mathrm{a_n}$), (b1), *and* (b2) *are satisfied; then the following inequalities hold:*

$$\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)\big]-\mathbb{E}^n\big[\phi(\tilde{X}^n_t,\tilde{Y}_t)\big]\big|\leq\alpha\,\mathbb{E}^n\big[\mathcal{Z}^n_t\big]+2\alpha\,\mathbb{P}^n\big(\{\tilde{X}_t\neq\tilde{X}^n_t\}\big),$$

$$\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)\big]-\mathbb{E}^n\big[\phi(\tilde{X}^n_t,\tilde{Y}_t)\big]\big|\leq\alpha\,\mathbb{E}\big[\mathcal{Z}_t\big]+2\alpha\,\mathbb{P}\big(\{\tilde{X}_t\neq\tilde{X}^n_t\}\big),$$

$$\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)\big]-\mathbb{E}^n\big[\phi(\tilde{X}^n_t,\tilde{Y}_t)\big]\big|\leq\alpha\,\mathbb{E}^n\big[\mathcal{Z}^n_t\big]+\Lambda\,\mathbb{E}^n\big[\big|\tilde{X}_t-\tilde{X}^n_t\big|\big],$$

$$\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}\big|\mathbb{E}\big[\phi(\tilde{X}_t,\tilde{Y}_t)\big]-\mathbb{E}^n\big[\phi(\tilde{X}^n_t,\tilde{Y}_t)\big]\big|\leq\alpha\,\mathbb{E}\big[\mathcal{Z}_t\big]+\Lambda\,\mathbb{E}\big[\big|\tilde{X}_t-\tilde{X}^n_t\big|\big],$$

where $\mathcal{Z}^n_t$ *and* $\mathcal{Z}_t$ *are defined as in* (27) *and* (28), *respectively.*

Proof. For any $\phi\in\mathcal{K}'_{TV}(\alpha)$,

$$\begin{aligned}&|\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)]-\mathbb{E}^n[\phi(\tilde{X}^n_t,\tilde{Y}_t)]|\\&\leq|\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)]-\mathbb{E}[\phi(\tilde{X}^n_t,\tilde{Y}_t)]|\\&\qquad+|\mathbb{E}[\phi(\tilde{X}^n_t,\tilde{Y}_t)]-\mathbb{E}_{\mathbb{Q}}[(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}^{\tilde{X}^n,\tilde{Y}}_t}\phi(\tilde{X}^n_t,\tilde{Y}_t)]|\\&\leq|\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)]-\mathbb{E}[\phi(\tilde{X}^n_t,\tilde{Y}_t)]|+\alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}^{\tilde{X}^n,\tilde{Y}}_t}-(d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}^{\tilde{X}^n,\tilde{Y}}_t}|].\end{aligned}$$

Since, by the Jensen inequality, $|E(Z/\mathcal{G})| = |E(E(Z/\mathcal{H})/\mathcal{G})| \le E(|E(Z/\mathcal{H})|/\mathcal{G})$ for any integrable random variable Z and σ-algebras $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}$, we have

$$\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{X}^n,\tilde{Y}}} - (d\mathbb{P}^n/d\mathbb{Q})|_{\mathcal{F}_t^{\tilde{X}^n,\tilde{Y}}}|] \le \mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|],$$

and therefore

$$\begin{aligned}&|\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}^n[\phi(\tilde{X}_t^n,\tilde{Y}_t)]|\\ &\quad\le |\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}[\phi(\tilde{X}_t^n,\tilde{Y}_t)]| + \alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|].\end{aligned}$$

As a consequence,

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)} |\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}^n[\phi(\tilde{X}_t^n,\tilde{Y}_t)]|\\ &\quad\le \alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|] + 2\,\alpha\,\mathbb{P}(\{\tilde{X}_t \ne \tilde{X}_t^n\}).\end{aligned}$$

Interchanging the role of $\mathbb{P}$ and $\mathbb{P}^n$, we get

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)} |\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}^n[\phi(\tilde{X}_t^n,\tilde{Y}_t)]|\\ &\quad\le \alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|] + 2\,\alpha\,\mathbb{P}^n(\{\tilde{X}_t \ne \tilde{X}_t^n\}).\end{aligned}$$

By using similar arguments, it is easy to get

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} |\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}^n[\phi(\tilde{X}_t^n,\tilde{Y}_t)]|\\ &\quad\le \alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|] + \Lambda\,\mathbb{E}[|\tilde{X}_t - \tilde{X}_t^n|]\end{aligned}$$

and

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} |\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}^n[\phi(\tilde{X}_t^n,\tilde{Y}_t)]|\\ &\quad\le \alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|] + \Lambda\,\mathbb{E}^n[|\tilde{X}_t - \tilde{X}_t^n|]. \qquad \square\end{aligned}$$

Remark 2.12. As in Remark 2.10, by arguments similar to the proof of Theorem 2.9, we also get

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} |\mathbb{E}[\phi(\tilde{X}_t,\tilde{Y}_t)] - \mathbb{E}^n[\phi(\tilde{X}_t^n,\tilde{Y}_t)]|\\ &\quad\le \alpha\mathbb{E}_{\mathbb{Q}}[|(d\mathbb{P}/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t} - (d\mathbb{P}^n/d\mathbb{Q})|_{\tilde{\mathcal{F}}_t}|]\\ &\qquad\qquad + \max(2\alpha;\Lambda)\,\min(\mathbb{E}^n[|\tilde{X}_t - \tilde{X}_t^n| \wedge 1];\ \mathbb{E}[|\tilde{X}_t - \tilde{X}_t^n| \wedge 1]).\end{aligned}$$

3. Approximation in filtering for Markov models with delayed observations. In this section we show how to use the approximations $\tilde{\pi}_t^n$ and $\overline{\pi}_t^n$ introduced in section 2 for the filter in the case of Markov models with delayed observations, in order to obtain upper bounds for the expectations in (19) and (20) when dealing with the bounded-Lipschitz distance or the total variation distance. For this purpose, we give some general upper bounds, depending on the trajectories actually observed (see (35) and (38) below) which lead to the main results of this section stated in Propositions 3.1 and 3.2. In the next section we use these bounds, and we exploit the results

about the sensitivity of the filter for Markov jump processes with counting observations given in [6], [7], and [8]. As we have already noticed at the end of section 1, some approximation results hold also for different systems and can be obtained with different techniques.

Consider the systems $(\boldsymbol{X^n}, \boldsymbol{Y^n})$ and $(\boldsymbol{X}, \boldsymbol{Y})$ such that

$$(\boldsymbol{X^n}, \hat{\boldsymbol{Y}}^{\boldsymbol{n}}) \text{ is a Markov system with generator } \boldsymbol{L^n}$$

and

$$Y_t^n = \hat{Y}_{a(t)}^n;$$

analogously,

$$(\boldsymbol{X}, \hat{\boldsymbol{Y}}) \text{ is a Markov system with generator } \boldsymbol{L}$$

and

$$Y_t = \hat{Y}_{a(t)},$$

where $a(\cdot)$ has the same properties as in Theorem 1.1.

As explained in Theorem 1.1 the filters π_t^n and π_t can be expressed in terms of the filters $\hat{\pi}_t^n$ and $\hat{\pi}_t$ and the semigroups $\exp\{\boldsymbol{L^n}s\}$ and $\exp\{\boldsymbol{L}s\}$ of the corresponding Markovian systems. More precisely (see (11))

$$\pi_t^n(\varphi) = \hat{\pi}_{a(t)}^n\big(\exp\{\boldsymbol{L^n}(t-a(t))\}\phi(\cdot, \hat{Y}_{a(t)}^n)\big)$$

and

$$\pi_t(\varphi) = \hat{\pi}_{a(t)}\big(\exp\{\boldsymbol{L}(t-a(t))\}\phi(\cdot, \hat{Y}_{a(t)})\big),$$

where $\phi(x,y) = \varphi(x)$, as in Theorem 1.1 and section 2.3.

Therefore, on the one hand, the convergence of the filters may be investigated in terms of the convergence of $\hat{\pi}_t^n$ to $\hat{\pi}_t$ and of $\exp\{\boldsymbol{L^n}s\}$ to $\exp\{\boldsymbol{L}s\}$. On the other hand, we are interested in the approximations (15) and (16) of the filter, and therefore we need a representation formula for the functionals U^n and U. These functionals can be expressed in terms of the corresponding functionals $\hat{U}^n$ and $\hat{U}$ of the underlying Markov systems, i.e., the functionals such that

$$\hat{\pi}_t^n = \hat{U}^n\big(t, \hat{\boldsymbol{Y}}^{\boldsymbol{n}}\big) \text{ and } \hat{\pi}_t = \hat{U}\big(t, \hat{\boldsymbol{Y}}\big),$$

where the functionals $\hat{U}^n$ and $\hat{U}$ depend on the initial distributions and the generators of the corresponding Markov systems.

As a consequence,

$$\pi_t^n(\varphi) = \int_{\mathbb{R}^k} \hat{U}^n\big(a(t), \boldsymbol{Y^n} \circ \boldsymbol{\mathcal{A}^{-1}}; dx\big) \exp\{\boldsymbol{L^n}(t-a(t))\}\phi(x, Y_t^n)$$

and

$$\pi_t(\varphi) = \int_{\mathbb{R}^k} \hat{U}\big(a(t), \boldsymbol{Y} \circ \boldsymbol{\mathcal{A}^{-1}}; dx\big) \exp\{\boldsymbol{L}(t-a(t))\}\phi(x, Y_t),$$

where again $\phi(x,y) = \varphi(x)$ and hence, the r.h.s.'s of the previous equalities define implicitly the functionals U^n and U, respectively.

Recalling that

$$(Y \circ \mathcal{A}^{-1})_s = Y_{a^{-1}(s)} = \hat{Y}_{a(a^{-1}(s))} = \hat{Y}_s, \tag{34}$$

since $a(a^{-1}(s)) = s$, $Y_t = \hat{Y}_{a(t)}$, and $Y_t^n = \hat{Y}_{a(t)}^n$, we can rewrite the above formulas as

$$\begin{aligned}\pi_t^n(\varphi) &= \int_{\mathbb{R}^k} \hat{U}^n\big(a(t), \hat{\boldsymbol{Y}}^{\boldsymbol{n}}; dx\big) \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t - a(t))\}\phi(x, \hat{Y}_{a(t)}^n) \\ &= \hat{U}^n\big(r, \boldsymbol{y}; \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(\cdot, y_r)\big)\big|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}^{\boldsymbol{n}}}\end{aligned}$$

and

$$\begin{aligned}\pi_t(\varphi) &= \int_{\mathbb{R}^k} \hat{U}\big(a(t), \hat{\boldsymbol{Y}}; dx\big) \exp\{\boldsymbol{L}(t - a(t))\}\phi(x, \hat{Y}_{a(t)}) \\ &= \hat{U}\big(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r)\big)\big|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}}.\end{aligned}$$

Therefore, when we observe $\boldsymbol{Y}$ (situation 1 of section 2), either the total variation distance or the bounded-Lipschitz distance between π_t and $\tilde{\pi}_t^n$ can be evaluated in terms of the supremum of

$$\begin{aligned}|\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)| &= \big|U\big(t, \boldsymbol{Y}; \varphi\big) - U^n\big(t, \boldsymbol{Y}; \varphi\big)\big| \\ &= \big|\hat{U}\big(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r)\big) - \hat{U}^n\big(r, \boldsymbol{y}; \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(\cdot, y_r)\big)\big|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}}\end{aligned}$$

over φ in the class of functions $\mathcal{K}_{TV}$ or $\mathcal{K}_{BL}$, respectively. When instead we observe $\boldsymbol{Y}^{\boldsymbol{n}}$ (situation 2 of section 2), the distance between π_t^n and $\bar{\pi}_t^n$ can be evaluated in terms of the supremum of

$$\begin{aligned}|\pi_t^n(\varphi) - \bar{\pi}_t^n(\varphi)| &= \big|U^n\big(t, \boldsymbol{Y}^{\boldsymbol{n}}; \varphi\big) - U\big(t, \boldsymbol{Y}^{\boldsymbol{n}}; \varphi\big)\big| \\ &= \big|\hat{U}^n\big(r, \boldsymbol{y}; \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(\cdot, y_r)\big) - \hat{U}\big(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r)\big)\big|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}^{\boldsymbol{n}}}\end{aligned}$$

over φ in the class of functions $\mathcal{K}_{TV}$ or $\mathcal{K}_{BL}$, respectively.

In situation 1 we have

$$\begin{aligned}&|\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)| \qquad (35)\\ &\le |\hat{U}(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r)) - \hat{U}^n(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r))|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}} \qquad (36)\\ &+ |\hat{U}^n(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r)) - \hat{U}^n(r, \boldsymbol{y}; \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(\cdot, y_r))|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}}, \qquad (37)\end{aligned}$$

and in situation 2,

$$\begin{aligned}&|\pi_t^n(\varphi) - \bar{\pi}_t^n(\varphi)| \qquad (38)\\ &\le |\hat{U}^n(r, \boldsymbol{y}; \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(\cdot, y_r)) - \hat{U}^n(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r))|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}^{\boldsymbol{n}}} \qquad (39)\\ &+ |\hat{U}^n(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r)) - \hat{U}(r, \boldsymbol{y}; \exp\{\boldsymbol{L}(t-r)\}\phi(\cdot, y_r))|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}^{\boldsymbol{n}}}. \qquad (40)\end{aligned}$$

Note that the two previous upper bounds differ only by the fact that they are evaluated on $\boldsymbol{y} = \hat{\boldsymbol{Y}}$ and $\boldsymbol{y} = \hat{\boldsymbol{Y}}^{\boldsymbol{n}}$. Moreover, natural upper bounds for (37) are

$$\begin{aligned}&\sup_x |\exp\{\boldsymbol{L}(t-r)\}\phi(x, y_r) - \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(x, y_r)|_{r=a(t), \boldsymbol{y}=\hat{\boldsymbol{Y}}} \qquad (41)\\ &\quad \le \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x, y) - \exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(x, y)|_{r=a(t)}. \qquad (42)\end{aligned}$$

The same bounds are valid for (39) when substituting $\boldsymbol{y} = \hat{\boldsymbol{Y}}$ with $\boldsymbol{y} = \hat{\boldsymbol{Y}}^{\boldsymbol{n}}$.

Then in both situations the distance between the filter and the approximation can be expressed by means of the distance between $\hat{U}^n$ and $\hat{U}$ and the distance between the semigroups $\exp\{\boldsymbol{L^n}s\}$ and $\exp\{\boldsymbol{L}s\}$, as can be deduced by the following results.

PROPOSITION 3.1. *For the system with delayed observations described at the beginning of this section we have*

(43)
$$\sup_{\varphi\in\mathcal{K}_{TV}(\alpha)} |\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)| \le \sup_{\psi\in\mathcal{K}'_{TV}(\alpha)} |\hat{U}(r,\boldsymbol{y};\psi(\cdot,y_r)) - \hat{U}^n(r,\boldsymbol{y};\psi(\cdot,y_r))|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}}$$

$$+ \sup_{\phi\in\mathcal{K}'_{TV}(\alpha)} \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x,y) - \exp\{\boldsymbol{L^n}(t-r)\}\phi(x,y)|_{r=a(t)}, \tag{44}$$

where $\mathcal{K}_{TV}(\alpha)$ and $\mathcal{K}'_{TV}(\alpha)$ are defined as in section 2.3.

Furthermore, if for all $u \ge 0$ there exists a constant $\Lambda'(u,\Lambda)$ such that

$$\exp\{\boldsymbol{L}u\}\big(\mathcal{K}'_{BL}(\alpha,\Lambda)\big) \subseteq \mathcal{K}'_{BL}(\alpha,\Lambda'(u,\Lambda)), \tag{45}$$

then we have

$$\sup_{\varphi\in\mathcal{K}_{BL}(\alpha,\Lambda)} |\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)|$$

$$\le \sup_{\psi\in\mathcal{K}'_{BL}(\alpha,\Lambda'(t-r,\Lambda))} |\hat{U}(r,\boldsymbol{y};\psi(\cdot,y_r)) - \hat{U}^n(r,\boldsymbol{y};\psi(\cdot,y_r))|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}} \tag{46}$$

$$+ \sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x,y) - \exp\{\boldsymbol{L^n}(t-r)\}\phi(x,y)|_{r=a(t)}, \tag{47}$$

where $\mathcal{K}_{BL}(\alpha,\Lambda)$ and $\mathcal{K}'_{BL}(\alpha,\Lambda)$ are defined as in section 2.3.

Proof. Taking into account (35)–(37) we get

$$\begin{aligned}
\sup_{\varphi\in\mathcal{K}_{TV}(\alpha)} |\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)| &\le \sup_{\phi\in\mathcal{K}'_{TV}(\alpha)} |\hat{U}\big(r,\boldsymbol{y};\exp\{\boldsymbol{L}(t-r)\}\phi(\cdot,y_r)\big) \\
&\quad - \hat{U}^n\big(r,\boldsymbol{y};\exp\{\boldsymbol{L}(t-r)\}\phi(\cdot,y_r)\big)|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}} \\
&\quad + \sup_{\phi\in\mathcal{K}'_{TV}(\alpha)} \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x,y) - \exp\{\boldsymbol{L^n}(t-r)\}\phi(x,y)|_{r=a(t)} \\
&\le \sup_{\psi\in\mathcal{K}'_{TV}(\alpha)} |\hat{U}\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big) - \hat{U}^n\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}} \\
&\quad + \sup_{\phi\in\mathcal{K}'_{TV}(\alpha)} \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x,y) - \exp\{\boldsymbol{L^n}(t-r)\}\phi(x,y)|_{r=a(t)}.
\end{aligned}$$

Moreover, we also have

$$\begin{aligned}
\sup_{\varphi\in\mathcal{K}_{BL}(\alpha,\Lambda)} |\pi_t(\varphi) - \tilde{\pi}_t^n(\varphi)| &\le \sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} |\hat{U}\big(r,\boldsymbol{y};\exp\{\boldsymbol{L}(t-r)\}\phi(\cdot,y_r)\big) \\
&\quad - \hat{U}^n\big(r,\boldsymbol{y};\exp\{\boldsymbol{L}(t-r)\}\phi(\cdot,y_r)\big)|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}} \\
&\quad + \sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x,y) - \exp\{\boldsymbol{L^n}(t-r)\}\phi(x,y)|_{r=a(t)} \\
&\le \sup_{\psi\in\exp\{\boldsymbol{L}(t-r)\}\mathcal{K}'_{BL}(\alpha,\Lambda)} |\hat{U}\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big) - \hat{U}^n\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}} \\
&\quad + \sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)} \sup_{x,y} |\exp\{\boldsymbol{L}(t-r)\}\phi(x,y) - \exp\{\boldsymbol{L^n}(t-r)\}\phi(x,y)|_{r=a(t)},
\end{aligned}$$

and invoking (45) we get (46) by noticing that the supremum over

$$\exp\{\boldsymbol{L}(t-r)\}\big(\mathcal{K}'_{BL}(\alpha,\Lambda)\big)$$

in the bound given above can be replaced by the supremum over $\mathcal{K}'_{BL}(\alpha,\Lambda'(t-r,\Lambda))$. □

Taking into account (38)–(40) and changing $\hat{\boldsymbol{Y}}$ into $\hat{\boldsymbol{Y}}^{\boldsymbol{n}}$ we have the following result.

PROPOSITION 3.2. *For the system with delayed observations described at the beginning of this section we have*

$$\sup_{\varphi\in\mathcal{K}_{TV}(\alpha)}|\pi_t^n(\varphi)-\overline{\pi}_t^n(\varphi)|$$

$$\leq \sup_{\psi\in\mathcal{K}'_{TV}(\alpha)}\big|\hat{U}\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)-\hat{U}^n\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)\big|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}^{\boldsymbol{n}}} \tag{48}$$

$$+\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}\sup_{x,y}|\exp\{\boldsymbol{L}(t-r)\}\phi(x,y)-\exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(x,y)|_{r=a(t)}\,. \tag{49}$$

Furthermore, if condition (45) *holds, then we have*

$$\sup_{\varphi\in\mathcal{K}_{BL}(\alpha,\Lambda)}|\pi_t^n(\varphi)-\overline{\pi}_t^n(\varphi)|$$

$$\leq \sup_{\psi\in\mathcal{K}'_{BL}(\alpha,\Lambda'(t-r,\Lambda))}|\hat{U}(r,\boldsymbol{y};\psi(\cdot,y_r))-\hat{U}^n(r,\boldsymbol{y};\psi(\cdot,y_r))|_{r=a(t),\boldsymbol{y}=\hat{\boldsymbol{Y}}^{\boldsymbol{n}}} \tag{50}$$

$$+\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}\sup_{x,y}|\exp\{\boldsymbol{L}(t-r)\}\phi(x,y)-\exp\{\boldsymbol{L}^{\boldsymbol{n}}(t-r)\}\phi(x,y)|_{r=a(t)}. \tag{51}$$

The above upper bounds will be evaluated when studying filtering systems with counting observations in the following section. In particular, upper bounds for the expectations of the r.h.s. of (43) and (46) will be obtained by a slight modification of the arguments used in the previous section (see Theorem 2.9). The same kind of consideration holds for the expectations of (44) and (47) (see Proposition 2.11).

4. Counting processes observations.

4.1. Main results. In this section, we accomplish the above-mentioned general program in the case of a Markov jump process with counting delayed observations. More precisely we assume that the pairs $(\boldsymbol{X},\hat{\boldsymbol{Y}})$ and $(\boldsymbol{X}^{\boldsymbol{n}},\hat{\boldsymbol{Y}}^{\boldsymbol{n}})$ are Markov systems in $\mathbb{R}\times\mathbb{N}$, with respective initial conditions $\mu_0^X\otimes\delta_{\{y\}}$ and $\mu_0^{X^n}\otimes\delta_{\{y\}}$, and with respective generators $\boldsymbol{L}$ and $\boldsymbol{L}^{\boldsymbol{n}}$, where

$$\begin{aligned}\boldsymbol{L}\phi(x,y)&=\lambda_0(x,y)\int\big(\phi(x',y)-\phi(x,y)\big)\mu_0(x,y;dx')\\&\quad+\lambda_1(x,y)\int\big(\phi(x',y+1)-\phi(x,y)\big)\mu_1(x,y;dx')\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{L}^{\boldsymbol{n}}\phi(x,y)&=\lambda_0^n(x,y)\int\big(\phi(x',y)-\phi(x,y)\big)\mu_0^n(x,y;dx')\\&\quad+\lambda_1^n(x,y)\int\big(\phi(x',y+1)-\phi(x,y)\big)\mu_1^n(x,y;dx'),\end{aligned}$$

and thus the predictable intensities of $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}^{\boldsymbol{n}}$ are $\lambda_1(X_{t^-}, \hat{Y}_{t^-})$ and $\lambda_1^n(X^n_{t^-}, \hat{Y}^n_{t^-})$, respectively. Furthermore we assume such that the following hypotheses are satisfied:
(H$_0$) $0 \le \underline{\lambda}_i \le \lambda_i(x,y), \lambda_i^n(x,y) \le \overline{\lambda}_i \quad$ for $i = 0, 1$;
(H$_1$) $0 < \lambda_1(x,y), \lambda_1^n(x,y)$;
(H$_2$) $\Delta_i^n := \sup_{x,y} \boldsymbol{\kappa}(\mu_i(x,y;\cdot), \mu_i^n(x,y;\cdot)) < \infty \quad$ for $i = 0, 1$.

Note that the operator $\boldsymbol{L}$ given above can be expressed in the same form as the one defined by (4) and vice versa.

From now on we will say that the operator $\boldsymbol{L}$ is *bL-regular* when it satisfies the following conditions:
(R$_0$) For every z the function $x \mapsto \lambda_i(x,z)$, $i = 0, 1$, is bounded-Lipschitz continuous and the Lipschitz constant is bounded from above by L_{λ_i}.
(R$_1$) $\sup_y \boldsymbol{\kappa}(\mu_i(x,y;\cdot), \mu_i(x',y;\cdot)) \le \Gamma_{\mu_i}|x - x'|$ for $i = 0, 1$.
(R$_2$) $\sup_{x,y} \int |z - x| \mu_0(x,y;dz) \le b_0, \qquad \int |z - x| \mu_1(x,y;dz) \le a_1(|x| + y) + b_1$.

It is important to stress that, if $\boldsymbol{L}$ is bL-regular, then $\exp\{\boldsymbol{L}u\}(\mathcal{K}'_{BL}(\alpha, \Lambda)) \subseteq \mathcal{K}'_{BL}(\alpha, \Lambda e^{Mu})$ (see Corollary 4.9) for a suitable constant M (given in (67)), and thus condition (45) is satisfied with

$$\Lambda'(u, \Lambda) = \Lambda e^{Mu}. \tag{52}$$

Finally, we will denote by $\|\cdot\|_\infty$ the norm defined by $\|\mu\|_\infty = \sup_{x,y} \|\mu(x,y,\cdot)\|_{TV}$ for any regular nucleus $\mu(x,y;dz)$.

The central tool for the proofs of our bounds (see Theorem 4.5 below) is a particular construction of the pairs $(\boldsymbol{X}^{\boldsymbol{n}}, \hat{\boldsymbol{Y}}^{\boldsymbol{n}})$ and $(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ on the same measurable space $(\Omega, \mathcal{F})$, equipped with three different probability measures $\mathbb{Q}$, $\mathbb{P}$, and $\mathbb{P}^n$, in a similar way to that used in section 2, and such that
(â) under $\mathbb{P}$ the pair $(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ has generator $\boldsymbol{L}$ and initial condition $\mu = \mu_0^X \otimes \delta_{\{y\}}$,
(â$_n$) under $\mathbb{P}^n$ the pair $(\boldsymbol{X}^{\boldsymbol{n}}, \hat{\boldsymbol{Y}})$ has generator $\boldsymbol{L}^n$ and initial condition $\mu^n = \mu_0^{X^n} \otimes \delta_{\{y\}}$,
(b̂) the probability measures $\mathbb{Q}$, $\mathbb{P}$, and $\mathbb{P}^n$ are equivalent on $= \mathcal{F}_t^{X, X^n, \hat{Y}}$.

Note that the use of $(\boldsymbol{X}^{\boldsymbol{n}}, \hat{\boldsymbol{Y}}^{\boldsymbol{n}})$ and $(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ is a slight abuse of notation: as at the end of the previous section we should write instead $(\tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\hat{\boldsymbol{Y}}}^{\boldsymbol{n}})$ and $(\tilde{\boldsymbol{X}}, \tilde{\hat{\boldsymbol{Y}}})$. Nevertheless we prefer the slight abuse of notation, in order to avoid the notation $\tilde{\hat{Y}}$.

Therefore, when dealing with the triplet $(\boldsymbol{X}, \boldsymbol{X}^{\boldsymbol{n}}, \hat{\boldsymbol{Y}})$ instead of $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}})$, conditions (a), (a$_n$), (b1), and (b2) are satisfied. Then, in this framework, Theorem 2.9 applies with $(\boldsymbol{X}, \boldsymbol{X}^{\boldsymbol{n}}, \hat{\boldsymbol{Y}})$ instead of $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}})$ and with $\hat{U}$ and $\hat{U}^n$ instead of U and U^n, respectively; the same considerations are valid for Proposition 2.11. The explicit expression of $\hat{U}$ is given by

$$\hat{U}(t, \boldsymbol{y}; \varphi) = \begin{cases} \hat{\Pi}_t(\varphi | \boldsymbol{s}) & \text{for } y(t) = \sum_{i=1}^\infty \mathbb{I}_{[0,s_i]}(t), \\ \varphi(x_0) & \text{otherwise} \end{cases}$$

(see (9) and (10)). Obviously the definition of the probability measure $\hat{U}(t, \boldsymbol{y}; dx)$ on the complement of the set $\{\boldsymbol{y}$ such that $y(t) = \sum_{i=1}^\infty \mathbb{I}_{[0,s_i]}(t)\}$ is not unique. An analogous representation holds for $\hat{U}^n$.

Basically the above conditions coincide with Condition A of [8], and the construction we use is the same as the one in section 3 of [8]; the only difference is that in [8] we assume $\hat{Y}_0 = 0$. By this construction the model $(\boldsymbol{X}, \boldsymbol{Y})$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking $Y_t = \hat{Y}_{a(t)}$, and the model $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y}^{\boldsymbol{n}})$ is defined on $(\Omega, \mathcal{F}, \mathbb{P}^n)$ taking $\hat{\boldsymbol{Y}}^{\boldsymbol{n}} = \hat{\boldsymbol{Y}}$ and $Y_t^n = \hat{Y}^n_{a(t)}$ and thus $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y}^{\boldsymbol{n}})$ coincides with $(\boldsymbol{X}^{\boldsymbol{n}}, \boldsymbol{Y})$ on $(\Omega, \mathcal{F}, \mathbb{P}^n)$. Moreover,

if $\mu_0^X = \delta_{\{x\}}$ and $\mu_0^{X^n} = \delta_{\{x\}}$, then the semigroups can be represented, respectively, as

$$\exp\{\boldsymbol{L}t\}\phi(x,y) = \mathbb{E}\big[\phi(X_t, \hat{Y}_t)\big] \tag{53}$$

and

$$\exp\{\boldsymbol{L^n}t\}\phi(x,y) = \mathbb{E}^n\big[\phi(X_t^n, \hat{Y}_t)\big]. \tag{54}$$

The construction of the triplet $(\boldsymbol{X}, \boldsymbol{X^n}, \hat{\boldsymbol{Y}})$ is possible under the assumptions (H_0) and (H_1).

The space $(\Omega, \mathcal{F}, \mathbb{Q})$ is a probability space on which two independent Poisson random measures are defined: $\mathcal{N}_0(ds, d\zeta)$ on $[0,T] \times [0, \bar{\lambda}_0]$, with intensity measure $ds\, d\zeta$, and $\mathcal{N}_1(ds, d\zeta)$ on $[0,T] \times [0,1]$, with intensity measure $ds\, d\zeta$. Then for suitable functions K_0, K_1, K_0^n, K_1^n, the triplet $(\boldsymbol{X}, \boldsymbol{X^n}, \hat{\boldsymbol{Y}})$ is

$$\begin{aligned} X_t = X_0 &+ \int_0^t \int_0^{\bar{\lambda}_0} K_0(X_{s^-}, X^n_{s^-}, \hat{Y}_{s^-}; \zeta)\mathcal{N}_0(ds, d\zeta) \\ &+ \int_0^t \int_0^1 K_1(X_{s^-}, X^n_{s^-}, \hat{Y}_{s^-}; \zeta)\mathcal{N}_1(ds, d\zeta), \end{aligned} \tag{55}$$

$$\begin{aligned} X^n_t = X^n_0 &+ \int_0^t \int_0^{\bar{\lambda}_0} K^n_0(X_{s^-}, X^n_{s^-}, \hat{Y}_{s^-}; \zeta)\mathcal{N}_0(ds, d\zeta) \\ &+ \int_0^t \int_0^1 K^n_1(X_{s^-}, X^n_{s^-}, \hat{Y}_{s^-}; \zeta)\mathcal{N}_1(ds, d\zeta), \end{aligned} \tag{56}$$

$$\hat{Y}_t = y + \int_0^t \int_0^1 \mathcal{N}_1(ds, d\zeta) = y + \mathcal{N}_1((0,t] \times [0,1]), \tag{57}$$

with initial conditions (X_0, X_0^n) independent of $\mathcal{N}_0$ and $\mathcal{N}_1$, and

$$\left.\frac{d\mathbb{P}}{d\mathbb{Q}}\right|_{\mathcal{F}_t^{X, X^n, \hat{Y}}} = \exp\left\{\int_0^t \log\big(\lambda_1(X_{s^-}, \hat{Y}_{s^-})\big) d\hat{Y}_s - \int_0^t \big(\lambda_1(X_{s^-}, \hat{Y}_{s^-}) - 1\big) ds\right\},$$

$$\left.\frac{d\mathbb{P}^n}{d\mathbb{Q}}\right|_{\mathcal{F}_t^{X, X^n, \hat{Y}}} = \exp\left\{\int_0^t \log\big(\lambda_1^n(X^n_{s^-}, \hat{Y}_{s^-})\big) d\hat{Y}_s - \int_0^t \big(\lambda_1^n(X^n_{s^-}, \hat{Y}_{s^-}) - 1\big) ds\right\}.$$

Remark 4.1. By using Lemma 3.7 in [8], explicit expressions for the functions K_0, K_1, K_0^n, K_1^n can be given. In particular, $K_0(x, \tilde{x}, y)$ and $K_0^n(x, \tilde{x}, y)$ are constructed using $\lambda(x) = \lambda_0(x,y)$, $\tilde{\lambda}(\tilde{x}) = \lambda_0^n(\tilde{x}, y)$, $\bar{\lambda} = \bar{\lambda}_0$, $\mu(x; dz) = \mu_0(x, y; dz)$, and $\tilde{\mu}(\tilde{x}; dz) = \mu_0^n(\tilde{x}, y; dz)$, while $K_1(x, \tilde{x}, y)$ and $K_1^n(x, \tilde{x}, y)$ are constructed using $\lambda(x) = \tilde{\lambda}(\tilde{x}) = \bar{\lambda} = 1$, $\mu(x; dz) = \mu_1(x, y; dz)$, and $\tilde{\mu}(\tilde{x}; dz) = \mu_1^n(\tilde{x}, y; dz)$.

Note that this construction relies on the fact that the state space is one-dimensional. Similar expressions could be obtained in higher dimensions.

First, we deduce from the above Radon–Nikodym derivatives some upper bounds which will play a key role in what follows. By using Lemma 4.7 in [8], we have

$$\mathbb{E}^n\left[\left|\left.\frac{d\mathbb{P}}{d\mathbb{P}^n}\right|_{\mathcal{F}_t^{X,X^n,\hat{Y}}}-1\right|\right]=\mathbb{E}\left[\left|\left.\frac{d\mathbb{P}^n}{d\mathbb{P}}\right|_{\mathcal{F}_t^{X,X^n,\hat{Y}}}-1\right|\right]$$

$$\leq 2\mathbb{E}\wedge\mathbb{E}^n\left[\int_0^t\left|\lambda_1(X_s,\hat{Y}_s)-\lambda_1^n(X_s^n,\hat{Y}_s)\right|ds\right] \tag{58}$$

$$\leq 2\int_0^t\mathbb{E}\vee\mathbb{E}^n\left[\left|\lambda_1(X_s,\hat{Y}_s)-\lambda_1^n(X_s^n,\hat{Y}_s)\right|\right]ds, \tag{59}$$

where $\mathbb{E}\wedge\mathbb{E}^n[Z]=\min\{\mathbb{E}[Z],\mathbb{E}^n[Z]\}$, and $\mathbb{E}\vee\mathbb{E}^n[Z]=\max\{\mathbb{E}[Z],\mathbb{E}^n[Z]\}$, for any random variable Z.

Second, note that since in our construction $\hat{\boldsymbol{Y}}^{\boldsymbol{n}}=\hat{\boldsymbol{Y}}$, the r.h.s. of (43) and (48) coincide, and thus the first upper bounds in Propositions 3.1 and 3.2 also coincide. The same holds for (46) and (50) and consequently for the last upper bounds. Then, taking $r=a(t)$, the following result can be used to obtain upper bounds for the expectations of the r.h.s. of (43) and (46).

PROPOSITION 4.2. *Under the hypotheses* (H_0) *and* (H_1),

$$\mathbb{E}\vee\mathbb{E}^n\left[\sup_{\psi\in\mathcal{K}'_{TV}(\alpha)}\left|\hat{U}\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)-\hat{U}^n\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)\right|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}}\right] \tag{60}$$

$$\leq\alpha\Big[4\|\lambda_1-\lambda_1^n\|r$$

$$+\exp\{(\bar{\lambda}_1-\underline{\lambda}_1)\,r\}\left(\|\mu_0^X-\mu_0^{X^n}\|_{TV}\left(\frac{2\bar{\lambda}_1}{\bar{\lambda}_1-\underline{\lambda}_1}+1\right)+(\bar{\lambda}_0+\bar{\lambda}_1)J^n\,A(r)\right)\Big],$$

where

$$A(r)=2r+4\bar{\lambda}_1\frac{\exp\{-(\bar{\lambda}_1-\underline{\lambda}_1)\,r\}+(\bar{\lambda}_1-\underline{\lambda}_1)r-1}{(\bar{\lambda}_1-\underline{\lambda}_1)^2} \tag{61}$$

and

$$J^n=\max(\|p_0-p_0^n\|_\infty,\|\mu_1-\mu_1^n\|_\infty)/2, \tag{62}$$

with

$$p_0(x,y;dx')=\big(1-(\lambda_0(x,y)/\bar{\lambda}_0)\big)\delta_x(dx')+\big(\lambda_0(x,y)/\bar{\lambda}_0\big)\;\mu_0(x,y;dx'), \tag{63}$$

$$p_0^n(x,y;dx')=\big(1-(\lambda_0^n(x,y)/\bar{\lambda}_0)\big)\delta_x(dx')+\big(\lambda_0^n(x,y)/\bar{\lambda}_0\big)\;\mu_0^n(x,y;dx'). \tag{64}$$

If furthermore the operator $\boldsymbol{L}$ *is bL-regular, assumption* (H_2) *holds, and the initial distributions* μ_0^X *and* $\mu_0^{X^n}$ *have finite first moments, then*

$$\mathbb{E}\vee\mathbb{E}^n\left[\sup_{\psi\in\mathcal{K}'_{BL}(\alpha,\Lambda')}\left|\hat{U}\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)-\hat{U}^n\big(r,\boldsymbol{y};\psi(\cdot,y_r)\big)\right|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}}\right] \tag{65}$$

$$\leq 4\alpha\big[\|\lambda_1-\lambda_1^n\|r+L_{\lambda_1}H^n\,\boldsymbol{\mathcal{E}}_2^M(r)\big]+\Lambda' H^n\,\boldsymbol{\mathcal{E}}_1^M(r)$$

$$+\boldsymbol{\kappa}\big(\mu_0^X,\mu_0^{X^n}\big)\big(4\alpha L_{\lambda_1}\boldsymbol{\mathcal{E}}_1^M(r)+\Lambda'\boldsymbol{\mathcal{E}}_0^M(r)\big),$$

where

$$H^n = \bar{\lambda}_0 \Delta_0^n + b_0 \, \|\lambda_0 - \lambda_0^n\| + \bar{\lambda}_1 \Delta_1^n, \tag{66}$$

$$M = \bar{\lambda}_0 (\Gamma_{\mu_0} - 1)^+ - \underline{\lambda}_0 (1 - \Gamma_{\mu_0})^+ + b_0 L_{\lambda_0} + \bar{\lambda}_1 (\Gamma_{\mu_1} - 1)^+ - \underline{\lambda}_1 (1 - \Gamma_{\mu_1})^+, \tag{67}$$

and

$$\boldsymbol{\mathcal{E}}_k^M(t) := \frac{1}{M^k} \sum_{h=k}^{\infty} \frac{(tM)^h}{h!} = \frac{1}{M^k} \left(\exp\{tM\} - \sum_{h=0}^{k-1} \frac{(tM)^h}{h!} \right). \tag{68}$$

The proof of the above result is given in the next subsection.

Remark 4.3. As a consequence, taking $\alpha = 1$, we can obtain

(69)

$$\begin{aligned} &\mathbb{E} \vee \mathbb{E}^n \left[\|\hat{U}(r, \boldsymbol{y}) - \hat{U}^n(r, \boldsymbol{y})\|_{TV} \big|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}} \right] \\ &\quad \le \Bigg[4 \|\lambda_1 - \lambda_1^n\| r \\ &\qquad + \exp\{(\bar{\lambda}_1 - \underline{\lambda}_1)\, r\} \left(\|\mu_0^X - \mu_0^{X^n}\|_{TV} \left(\frac{2\bar{\lambda}_1}{\bar{\lambda}_1 - \underline{\lambda}_1} + 1 \right) + (\bar{\lambda}_0 + \bar{\lambda}_1) J^n \, A(r) \right) \Bigg], \end{aligned}$$

and similarly, taking $\Lambda' = 1$,

(70)

$$\begin{aligned} &\mathbb{E} \vee \mathbb{E}^n \left[d_{BL}\big(\hat{U}(r, \boldsymbol{y}), \hat{U}^n(r, \boldsymbol{y})\big) \big|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}} \right] \\ &\quad \le 4 \Big[\|\lambda_1 - \lambda_1^n\| r + L_{\lambda_1} H^n \, \boldsymbol{\mathcal{E}}_2^M(r) \Big] + H^n \, \boldsymbol{\mathcal{E}}_1^M(r) \\ &\qquad + \boldsymbol{\kappa}\big(\mu_0^X, \mu_0^{X^n}\big) \big(4 L_{\lambda_1} \boldsymbol{\mathcal{E}}_1^M(r) + \boldsymbol{\mathcal{E}}_0^M(r) \big). \end{aligned}$$

The above upper bounds are interesting by themselves, since they correspond to upper bounds for the approximations of the system without delay. The last bound was essentially obtained in [8, Remark 2.7]. However, in that bound there was an imprecision, and (70) corrects it.

Note that the upper bound for (69) grows exponentially fast in r, and therefore is valid only in a finite interval of time, namely, when the l.h.s. is less than or equal to 2. A similar consideration holds for the upper bound for (60). Moreover, note that we have used implicitly $\bar{\lambda}_1 - \underline{\lambda}_1 > 0$. The upper bound (69) grows exponentially fast also when $\boldsymbol{L} = \boldsymbol{L^n}$, i.e., when the approximating system and the limit system differ only in the distribution of the initial condition, but it simplifies to

$$\begin{aligned} &\mathbb{E} \vee \mathbb{E}^n \left[\|\hat{U}(r, \boldsymbol{y}) - \hat{U}^n(r, \boldsymbol{y})\|_{TV} \big|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}} \right] \\ &\quad \le \exp\{(\bar{\lambda}_1 - \underline{\lambda}_1)\, r\} \left(\frac{2\bar{\lambda}_1}{\bar{\lambda}_1 - \underline{\lambda}_1} + 1 \right) \|\mu_0^X - \mu_0^{X^n}\|_{TV}, \end{aligned} \tag{71}$$

since in this case $\|\lambda_1 - \lambda_1^n\| = 0$ and $J^n = 0$.

As far as the bounds (70) and (65) are concerned, note that if $\boldsymbol{L} = \boldsymbol{L^n}$, then the bound (70) simplifies to

$$\mathbb{E} \vee \mathbb{E}^n \left[d_{BL}\big(\hat{U}(r, \boldsymbol{y}), \hat{U}^n(r, \boldsymbol{y})\big) \big|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}} \right] \le \boldsymbol{\kappa}\big(\mu_0^X, \mu_0^{X^n}\big) \big(4 L_{\lambda_1} \boldsymbol{\mathcal{E}}_1^M(r) + \boldsymbol{\mathcal{E}}_0^M(r) \big), \tag{72}$$

since in this case $\|\lambda_1 - \lambda_1^n\| = 0$ and $H^n = 0$.

Moreover, note that the constants H^n and M correspond to the constants defined in (2.9) and (2.10) of Theorem 2.1 in [8], respectively, and that the constant M can be either positive or negative; for instance, if $\Gamma_{\mu_1} < 1$, then for any $\lambda_0(x,y)$ and $\mu_0(x,y;dx')$, $M = \bar{\lambda}_0(\Gamma_{\mu_0} - 1)^+ - \underline{\lambda}_0(1 - \Gamma_{\mu_0})^+ + b_0 L_{\lambda_0} - \underline{\lambda}_1(1 - \Gamma_{\mu_1})$, and thus for $\underline{\lambda}_1$ sufficiently large, the constant M is negative. The condition that the value of $\underline{\lambda}_1$ is sufficiently large can be interpreted as the condition that the times of observation are sufficiently frequent. In this case the error evaluated in the bounded-Lipschitz metric is of the same order as the error at time zero, evaluated in the Kantorovitch metric. If instead of $\Gamma_{\mu_1} < 1$ we have $\Gamma_{\mu_1} \geq 1$ and $\Gamma_{\mu_0} < 1$, then $M = -\underline{\lambda}_0(1 - \Gamma_{\mu_0})^+ + b_0 L_{\lambda_0} + \bar{\lambda}_1(\Gamma_{\mu_1} - 1)$, and thus for $\underline{\lambda}_0$ sufficiently large, M is again negative.

Furthermore it is interesting to note that in the case of the observation of a fixed parameter, i.e., $X_t = X_0$ and $X_t^n = X_0^n$, observed at times which form a standard Poisson process independent of $\boldsymbol{X}$, we have $M = 0$ since in this case $\lambda_0(x) = \lambda_0^n(x) = 0$, $L_{\lambda_0} = 0$, $\lambda_1(x) = \lambda_1^n(x) = 1$, and $\Gamma_{\mu_1} = 1$.

Finally, note that for any $k \geq 0$ and for positive M, $\boldsymbol{\mathcal{E}}_k^M(t)$ increases to infinity exponentially fast, with $\boldsymbol{\mathcal{E}}_k^M(t) \leq \frac{t^k \exp\{tM\}}{k!}$. Hence the bounds (65) and (70) are meaningful only in a finite interval of time. When M is negative, we have that

$$\lim_{t\to\infty} \boldsymbol{\mathcal{E}}_k^M(t) = \infty \qquad \text{for } k \geq 2,$$

but the rate of convergence is polynomial (indeed, $\lim_{t\to\infty} \boldsymbol{\mathcal{E}}_k^M(t)/t^{k-1} = (k-1)!/|M|$), while

$$\lim_{t\to\infty} \boldsymbol{\mathcal{E}}_0^M(t) = \lim_{t\to\infty} e^{Mt} = 0$$

and

$$\lim_{t\to\infty} \boldsymbol{\mathcal{E}}_1^M(t) = \lim_{t\to\infty} \frac{e^{Mt} - 1}{M} = \frac{1}{|M|}.$$

Hence for negative M, though the general bounds (65) and (70) are meaningful only in a finite interval of time, when the systems differ only for the initial conditions, i.e., when $\boldsymbol{L} = \boldsymbol{L^n}$, we can use (72) and get that

$$\lim_{r\to\infty} \mathbb{E} \vee \mathbb{E}^n \Big[d_{BL}\big(\hat{U}(r,\boldsymbol{y}), \hat{U}^n(r,\boldsymbol{y})\big)\big|_{\boldsymbol{y}=\hat{\boldsymbol{Y}}} \Big] \leq \boldsymbol{\kappa}\big(\mu_0^X, \mu_0^{X^n}\big) \frac{4L_{\lambda_1}}{|M|}. \tag{73}$$

The latter result can be interpreted as a kind of stability property for the filter with respect to its initial condition. Asymptotic stability properties with respect to the Hilbert projective metric have been studied by Budhiraja and Kushner [3] in the case when $\boldsymbol{X}$ is an ergodic Markov process with generator $\boldsymbol{A}$, and the system $(\boldsymbol{X}, \boldsymbol{Y})$ is described by $\boldsymbol{L}\phi(x,y) = (\boldsymbol{A}\phi(\cdot,y))(x) + \lambda_1(x,y)[\phi(x,y+1) - \phi(x,y)]$. Their result does not require $\boldsymbol{X}$ to be a jump process; however, when this is the case simultaneous jump times of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not allowed. In order to compare their result with (73), note that on the one hand we need $\boldsymbol{X}$ to be a jump process, while on the other hand simultaneous jump times are not prohibited and we do not need $\boldsymbol{X}$ to be Markov.

The same kind of considerations hold for the next upper bounds and can be used to give upper bounds for (44) and (47).

PROPOSITION 4.4. *Under the assumptions* $(H_0), (H_1)$ *and the notation of Proposition* 4.2,

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{TV}(\alpha)}\sup_{x,y}|\exp\{\boldsymbol{L}t\}\phi(x,y)-\exp\{\boldsymbol{L^n}t\}\phi(x,y)|\\&\le\alpha\big(2\|\lambda_1-\lambda_1^n\|\,t+\exp\{(\bar\lambda_1-\underline{\lambda}_1)t\}(\bar\lambda_0+\bar\lambda_1)\,J^n\,B(t)\big),\end{aligned}\tag{74}$$

where

$$B(t)=2t+2\bar\lambda_1\frac{\exp\{-(\bar\lambda_1-\underline{\lambda}_1)\,t\}+(\bar\lambda_1-\underline{\lambda}_1)t-1}{(\bar\lambda_1-\underline{\lambda}_1)^2}=t+\frac{A(t)}{2}.\tag{75}$$

If furthermore the operator $\boldsymbol{L}$ *is bL-regular, assumption* (H_2) *holds, and the initial distributions* μ_0^X *and* $\mu_0^{X^n}$ *have finite first moments, then*

$$\begin{aligned}&\sup_{\phi\in\mathcal{K}'_{BL}(\alpha,\Lambda)}\sup_{x,y}|\exp\{\boldsymbol{L}t\}\phi(x,y)-\exp\{\boldsymbol{L^n}t\}\phi(x,y)|\\&\le2\alpha\big[\|\lambda_1-\lambda_1^n\|t+L_{\lambda_1}H^n\,\boldsymbol{\mathcal{E}}_2^M(t)\big]+\Lambda H^n\,\boldsymbol{\mathcal{E}}_1^M(t).\end{aligned}\tag{76}$$

The proof of the above result is given in the next subsection.

Using the results and the notation of Propositions 3.1 and 3.2, with $\alpha=1$, $\Lambda=1$, and $\Lambda'(t-r,1)=e^{M(t-r)}$ (see (52)), then using Proposition 4.2 with $\Lambda'=\Lambda'(t-r,1)$ and Proposition 4.4, we get immediately the main result of this section.

THEOREM 4.5. *Under the assumptions* (H_0) *and* (H_1)*, and the notation introduced in Propositions* 4.2 *and* 4.4,

$$\begin{aligned}&E\big[\|\pi_t-\tilde\pi_t^n\|_{TV}\big]\vee E\big[\|\pi_t^n-\overline{\pi}_t^n\|_{TV}\big]\\&\le\Big[2\|\lambda_1-\lambda_1^n\|(t+r)+\|\mu_0^X-\mu_0^{X^n}\|_{TV}\exp\{(\bar\lambda_1-\underline{\lambda}_1)\,r\}\Big(\frac{2\bar\lambda_1}{\bar\lambda_1-\underline{\lambda}_1}+1\Big)\\&\quad+(\bar\lambda_0+\bar\lambda_1)J^n\big(\exp\{(\bar\lambda_1-\underline{\lambda}_1)\,r\}A(r)+\exp\{(\bar\lambda_1-\underline{\lambda}_1)\,(t-r)\}B(t-r)\big)\Big]_{r=a(t)}.\end{aligned}\tag{77}$$

Furthermore assume (H_2)*, the bL-regularity of* $\boldsymbol{L}$*, and that the initial distributions* μ_0^X *and* $\mu_0^{X^n}$ *have finite first moments. Then*

$$\begin{aligned}&E\big[d_{BL}(\pi_t,\tilde\pi_t^n)\big]\vee E\big[d_{BL}(\pi_t^n,\overline{\pi}_t^n)\big]\\&\le4\big[\|\lambda_1-\lambda_1^n\|r+L_{\lambda_1}H^n\,\boldsymbol{\mathcal{E}}_2^M(r)\big]+\exp\{M(t-r)\}H^n\,\boldsymbol{\mathcal{E}}_1^M(r)\\&\qquad+\boldsymbol{\kappa}\big(\mu_0^X,\mu_0^{X^n}\big)\big(4L_{\lambda_1}\boldsymbol{\mathcal{E}}_1^M(r)+\boldsymbol{\mathcal{E}}_0^M(t)\big)\\&\qquad+2\big(\|\lambda_1-\lambda_1^n\|(t-r)+L_{\lambda_1}H^n\,\boldsymbol{\mathcal{E}}_2^M(t-r)\big)+H^n\,\boldsymbol{\mathcal{E}}_1^M(t-r)\big|_{r=a(t)}.\end{aligned}\tag{78}$$

To conclude this section we work out an example highlighting the bounds in (78) by using a particular case of the model considered in Remark 2.5 in [8].

Example. Assume that the measures $\mu_i(x,y,\cdot)$ are Gaussian with mean $m_i(x,y)$ and variance $\sigma_i^2(x,y)$, and that the functions $\lambda_i(z,y)$, $m_i(z,y)$, and $\sigma_i(z,y)$ are Lipschitz in z, uniformly in y, with Lipschitz constant L_{λ_i}, L_{m_i} and L_{σ_i}, respectively. We assume also that the functions $\lambda_i(z,y)$, $\sigma_0(z,y)$, and $\mathsf{m}_0(z,y):=m_0(z,y)-z$ are uniformly bounded by $\overline{\lambda}_i$, $\overline{\sigma}_0$, and $\overline{\mathsf{m}}_0$, respectively, and that $\sigma_1(0,y)\le\mathsf{K}_1^\sigma(y+1)$ and $|m_1(0,y)|\le\mathsf{K}_1^m(y+1)$.

Then conditions (R_0), (R_1), and (R_2) are satisfied with

$$\Gamma_{\mu_i}=\sqrt{\tfrac{2}{\pi}}\,L_{\sigma_i}+L_{m_i},$$

$$b_0 = \overline{\sigma}_0 + \overline{\mathsf{m}}_0,\ b_1 = \mathsf{K}_1^{\sigma} + \mathsf{K}_1^{m},\ \text{and } a_1 = (L_{\sigma_1} + L_{m_1}) \vee b_1,$$

as can be easily deduced by the assumptions, the upper bound

$$\begin{aligned}\boldsymbol{\kappa}\big(\mu_i(x,y;\cdot),\mu_i(x',y;\cdot)\big) &\le \sqrt{\tfrac{2}{\pi}}\,|\sigma_i(x,y)-\sigma_i(x',y)| + |m_i(x,y)-m_i(x',y)| \\ &\le \left(\sqrt{\tfrac{2}{\pi}}\,L_{\sigma_i} + L_{m_i}\right)|x-x'|,\end{aligned}$$

and the bound

$$\begin{aligned}\int_{\mathbb{R}} |z-x|\,\mu_i(x,y;dz) &\le \int_{\mathbb{R}} |z-m_i(x,y)|\,\mu_i(x,y;dz) + \int_{\mathbb{R}} |m_i(x,y)-x|\,\mu_i(x,y;dz) \\ &\le \sigma_i(x,y) + |m_i(x,y)-x|.\end{aligned}$$

As far as the approximating process is concerned, we define

$$\lambda_i^n(x,y) = \lambda\left(\tfrac{\lfloor nx\rfloor}{n}, y\right), \quad \mu_i^n(x,y;A) := \mathcal{T}^n\mu_i\left(\tfrac{\lfloor nx\rfloor}{n}, y; A\right), \quad \text{and} \quad \pi_0^n(A) = \mathcal{T}^n\pi_0(A),$$

where, for any probability measure ν,

$$\mathcal{T}^n\nu(A) = \sum_k \nu\left(\left[\tfrac{k}{n}, \tfrac{k+1}{n}\right)\right)\,\delta_{\{\frac{k}{n}\}}(A).$$

Then condition (H_0) is obviously satisfied, condition (H_1) is satisfied if we assume furthermore that $\lambda_1(x,y)$ is strictly positive, and finally, condition (H_2) is satisfied with

$$\Delta_i^n = \left(\sqrt{\tfrac{2}{\pi}}\,L_{\sigma_i} + L_{m_i} + 1\right)\tfrac{1}{n}$$

since $\boldsymbol{\kappa}(\nu, \mathcal{T}^n\nu) \le \frac{1}{n}$, and thus

$$\begin{aligned}\boldsymbol{\kappa}(\mu_i(x,y;\cdot),\mu_i^n(x,y;\cdot)) &\le \boldsymbol{\kappa}\left(\mu_i(x,y;\cdot),\mu_i\left(\tfrac{\lfloor nx\rfloor}{n},y;\cdot\right)\right) \\ &\quad + \boldsymbol{\kappa}\left(\mu_i\left(\tfrac{\lfloor nx\rfloor}{n},y;\cdot\right),\mathcal{T}^n\mu_i\left(\tfrac{\lfloor nx\rfloor}{n},y;\cdot\right)\right) \\ &\le \left(\sqrt{\tfrac{2}{\pi}}\,L_{\sigma_i} + L_{m_i} + 1\right)\tfrac{1}{n}.\end{aligned}$$

Then $E\big[d_{BL}(\pi_t,\tilde{\pi}_t^n)\big]$ and $E\big[d_{BL}(\pi_t^n,\overline{\pi}_t^n)\big]$ are of order $\frac{1}{n}$, since the quantities involved in the upper bound (78) are of order $\frac{1}{n}$. Indeed,

$$\|\lambda_i - \lambda_i^n\| \le L_{\lambda_i}\,\tfrac{1}{n}, \quad \boldsymbol{\kappa}(\pi_0,\pi_0^n) \le \tfrac{1}{n},$$

and finally, recalling (66) and the previously obtained upper bounds,

$$H^n \le \left[\bar{\lambda}_0\left(\sqrt{\tfrac{2}{\pi}}\,L_{\sigma_0} + L_{m_0} + 1\right) + (\overline{\sigma}_0 + \overline{\mathsf{m}}_0)\,L_{\lambda_0} + \bar{\lambda}_1\left(\sqrt{\tfrac{2}{\pi}}\,L_{\sigma_1} + L_{m_1} + 1\right)\right]\,\tfrac{1}{n}.$$

4.2. Proofs of Propositions 4.2 and 4.4. Using the construction of the triplet $(\boldsymbol{X}, \boldsymbol{X^n}, \hat{\boldsymbol{Y}})$, it is easy to prove the following auxiliary results. In the first two results we need the processes

$$N_0(t) = \mathcal{N}_0\big((0,t] \times [0, \bar{\lambda}_0]\big), \quad N_1(t) = \mathcal{N}_1\big((0,t] \times [0,1]\big),$$

which are independent Poisson processes under $\mathbb{Q}$, with parameter $\bar{\lambda}_0$ and 1, respectively.

PROPOSITION 4.6. *If the assumptions* $(\mathrm{H}_0), (\mathrm{H}_1)$ *hold and if the initial conditions* (X_0, X_0^n) *in* (55) *and* (56) *are coupled in such a way that* $\mathbb{Q}(X_0 \neq X_0^n) = \|\mu_0^X - \mu_0^{X^n}\|_{TV}/2$ *(see, e.g.,* [8]*), then*

$$\mathbb{P} \vee \mathbb{P}^n\big(X_t \neq X_t^n\big) \leq \exp\{(\bar{\lambda}_1 - \underline{\lambda}_1)\, t\}\big(\|\mu_0^X - \mu_0^{X^n}\|_{TV}/2 + (\bar{\lambda}_0 + \bar{\lambda}_1)\, t J^n\big),$$

where the constant J^n *has been introduced in* (62) *of Proposition* 4.2.

Proof. To get a bound for $\mathbb{P} \vee \mathbb{P}^n(X_t \neq X_t^n)$, observe that

$$\mathbb{P}\big(X_t \neq X_t^n\big) = \mathbb{E}^{\mathbb{Q}}\big[(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{X, X^n, \hat{Y}}}\, \mathbb{I}_{\{X_t \neq X_t^n\}}\big]$$

and that

$$\begin{aligned}
(d\mathbb{P}/d\mathbb{Q})|_{\mathcal{F}_t^{X,X^n,\hat{Y}}} &= \exp\left\{\int_0^t \log\big(\lambda_1(X_{s^-}, \hat{Y}_{s^-})\big) d\hat{Y}_s - \int_0^t \big(\lambda_1(X_{s^-}, \hat{Y}_{s^-}) - 1\big) ds\right\} \\
&= \exp\left\{\int_0^t \log\big(\lambda_1(X_{s^-}, \hat{Y}_{s^-})\big) dN_1(s) - \int_0^t \big(\lambda_1(X_{s^-}, \hat{Y}_{s^-}) - 1\big) ds\right\} \\
&\leq \big(\bar{\lambda}_1\big)^{N_1(t)} \exp\{-(\underline{\lambda}_1 - 1)t\}.
\end{aligned}$$

As a consequence,

$$\mathbb{P}\big(X_t \neq X_t^n\big) \leq \mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar{\lambda}_1\big)^{N_1(t)}\, \mathbb{I}_{\{X_t \neq X_t^n\}}\big] \exp\{-(\underline{\lambda}_1 - 1)t\}. \tag{79}$$

In a similar way we prove that

$$\mathbb{P}^n\big(X_t \neq X_t^n\big) \leq \mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar{\lambda}_1\big)^{N_1(t)}\, \mathbb{I}_{\{X_t \neq X_t^n\}}\big] \exp\{-(\underline{\lambda}_1 - 1)t\}. \tag{80}$$

The result then follows immediately from the next lemma, which is proved in the appendix. □

LEMMA 4.7. *Under the hypotheses and notation of Proposition* 4.6*,*

$$\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar{\lambda}_1\big)^{N_1(t)}\, \mathbb{I}_{\{X_t \neq X_t^n\}}\big] \leq \exp\{(\bar{\lambda}_1 - 1)\, t\}\big(\|\mu_0^X - \mu_0^{X^n}\|_{TV}/2 + (\bar{\lambda}_0 + \bar{\lambda}_1)\, t J^n\big). \tag{81}$$

The following result plays a key role in our analysis in order to get the upper bounds when dealing with the bounded-Lipschitz metric.

PROPOSITION 4.8. *If the operator* $\boldsymbol{L}$ *is bL-regular, the hypotheses* (H_0)–(H_2) *hold, and if the initial conditions* (X_0, X_0^n) *in* (55) *and* (56) *have finite first moments and are coupled in such a way that* $\mathbb{E}_Q\big[\big|X_0 - X_0^n\big|\big] = \boldsymbol{\kappa}(\mu_0^X, \mu_0^{X^n})$ *(see, e.g.,* [8]*), then*

$$\mathbb{E} \vee \mathbb{E}^n\big[\,|X_t - X_t^n|\,\big] \leq \boldsymbol{\kappa}\big(\mu_0^X, \mu_0^{X^n}\big)\boldsymbol{\mathcal{E}}_0^M(t) + H^n \boldsymbol{\mathcal{E}}_1^M(t)$$

where the constants H^n*,* M *and the functions* $\boldsymbol{\mathcal{E}}_k^M$ *are as introduced in Proposition* 4.2.

Proof. The proof of this result is the same as the proof of Lemma 5.1 in [8], with obvious changes in the notation; in particular we have to substitute $\tilde{X}$ with X^n. Indeed, as in that proof, one can show

$$\mathbb{E}\big[|X_t - X_t^n|\big] \le \mathbb{E}\big[|X_0 - X_0^n|\big] + \int_0^t \mathbb{E}\big[H + M|X_s - X_s^n|\big]ds$$

and a similar inequality with respect to $\mathbb{E}^n$. Then the statement follows by the Gronwall inequality. We note also that there was a mistake in the statement of Lemma 5.1 of [8], and this result corrects it. □

As a consequence of Proposition 4.8 we get immediately that the class of bounded functions ψ, which are Lipschitz in x for all y, uniformly in y, is invariant under the action of the semigroup $\exp\{\boldsymbol{L}t\}$.

COROLLARY 4.9. *Let $\boldsymbol{L}$ be a bL-regular operator, with λ_i, $i = 0, 1$, satisfying hypotheses* (H_0) *and* (H_1)*, and let ψ be a measurable bounded function such that, for all x, x', y, $|\psi(x,y) - \psi(x',y)| \le L_\psi|x - x'|$.*

Then

$$|\exp\{\boldsymbol{L}t\}\psi(x,y) - \exp\{\boldsymbol{L}t\}\psi(x',y)| \le e^{Mt}L_\psi|x - x'| \qquad \textit{for all } x, x', y,$$

and

$$\exp\{\boldsymbol{L}t\}\big(K'(\alpha,\Lambda)\big) \subseteq K'(\alpha, e^{Mt}\Lambda).$$

Proof. Assume that $\boldsymbol{L^n} = \boldsymbol{L}$, and thus $H^n = 0$, and assume that $\mu_0^X = \delta_{\{x\}}$, $\mu_0^{X^n} = \delta_{\{x'\}}$, and thus $\boldsymbol{\kappa}(\mu_0^X, \mu_0^{X^n}) = |x - x'|$. Then

$$\begin{aligned}|\exp\{\boldsymbol{L}t\}\psi(x,y) - \exp\{\boldsymbol{L}t\}\psi(x',y)| &= \Big|\mathbb{E}\big[\psi(X_t, \hat{Y}_t)\big] - \mathbb{E}\big[\psi(X_t^n, \hat{Y}_t)\big]\Big| \\ &\le L_\psi\mathbb{E}[|X_t - X_t^n|],\end{aligned}$$

and, observing that (H_2) is trivially satisfied and that $\boldsymbol{\mathcal{E}}_0^M(t) = e^{Mt}$, Proposition 4.8 implies the result.

The second statement is an immediate consequence of the first one. □

Remark 4.10. We note that the boundedness condition on the function ψ is not necessary and that we could ask for a bound of the kind $|\psi(x,y)| \le A(|x| + y) + B$; indeed, under the conditions of Corollary 4.9, $\exp(\boldsymbol{L}t)\psi(x,y)$ is finite for any ψ Lipschitz in x uniformly in y. In Proposition 4.8 the condition on the first moments of (X_0, X_0^n) guarantees the integrability of (X_t, X_t^n) for all $t \ge 0$, and when $(X_0, X_0^n) = (x, x')$ this condition is obviously satisfied and implies that $\exp(\boldsymbol{L}t)\psi(x,y) = \mathbb{E}\big[\psi(X_t, \hat{Y}_t)\big]$ is finite for all such functions which are Lipschitz in x uniformly in y.

Proof of Proposition 4.2. In order to get the first estimate, we can use the first part of Theorem 2.9, with $(\boldsymbol{X}, \boldsymbol{X^n}, \hat{\boldsymbol{Y}})$ and r instead of $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}^{\boldsymbol{n}}, \tilde{\boldsymbol{Y}})$ and t, and get that (60) is bounded above by

$$2\alpha \max\big(\mathbb{E}[\hat{\mathcal{Z}}_r] + \mathbb{P}(X_r \ne X_r^n)\,;\, \mathbb{E}^n[\hat{\mathcal{Z}}_r^n] + \mathbb{P}^n(X_r \ne X_r^n)\big),$$

where $\hat{\mathcal{Z}}_r$ and $\hat{\mathcal{Z}}_r^n$ are defined as $\mathcal{Z}_r$ and $\mathcal{Z}_r^n$, with obvious changes. Using Remark 2.4 we get that $\mathbb{E}[\hat{\mathcal{Z}}_r]$ and $\mathbb{E}^n[\hat{\mathcal{Z}}_r^n]$ are equal to

$$\mathbb{E}^n\left[\left|\frac{d\mathbb{P}}{d\mathbb{P}^n}\Big|_{\mathcal{F}_t^{X,X^n,\hat{Y}}} - 1\right|\right] = \mathbb{E}\left[\left|\frac{d\mathbb{P}^n}{d\mathbb{P}}\Big|_{\mathcal{F}_t^{X,X^n,\hat{Y}}} - 1\right|\right].$$

Therefore, by (59), (60) is bounded above by

$$2\alpha\Big(2\int_0^r \mathbb{E}\vee\mathbb{E}^n\left[\left|\lambda_1(X_s,\hat{Y}_s)-\lambda_1^n(X_s^n,\hat{Y}_s)\right|\right]ds+\mathbb{P}\vee\mathbb{P}^n(X_r\neq X_r^n)\Big)$$
$$\leq 4\alpha\int_0^r\left(\|\lambda_1-\lambda_1^n\|+\bar{\lambda}_1\mathbb{P}\vee\mathbb{P}^n\left(X_s\neq X_s^n\right)\right)ds+2\alpha\,\mathbb{P}\vee\mathbb{P}^n\left(X_r\neq X_r^n\right).$$

The thesis is achieved by applying Proposition 4.6.

To obtain the second estimate, we can use the second part of Theorem 2.9, again with $(\boldsymbol{X},\boldsymbol{X^n},\hat{\boldsymbol{Y}})$ and r instead of $(\tilde{\boldsymbol{X}},\tilde{\boldsymbol{X}}^{\boldsymbol{n}},\tilde{\boldsymbol{Y}})$ and t, and get that (65) is bounded above by

$$\max\left(2\alpha\,\mathbb{E}[\hat{\mathcal{Z}}_r]+\Lambda\mathbb{E}(|X_r-X_r^n|)\,;\,2\alpha\,\mathbb{E}^n[\hat{\mathcal{Z}}_r^n]+\Lambda\,\mathbb{E}^n(|X_r-X_r^n|)\right).$$

Therefore, by (59), (65) is bounded above by

$$4\alpha\int_0^r \mathbb{E}\vee\mathbb{E}^n\left[\left|\lambda_1(X_s,\hat{Y}_s)-\lambda_1^n(X_s^n,\hat{Y}_s)\right|\right]ds+\Lambda\,\mathbb{E}\vee\mathbb{E}^n[|X_r-X_r^n|]$$
$$\leq 4\alpha\int_0^r\left(\|\lambda_1-\lambda_1^n\|+L_{\lambda_1}\mathbb{E}\vee\mathbb{E}^n\left[|X_s-X_s^n|\right]\right)ds+\Lambda\,\mathbb{E}\vee\mathbb{E}^n\left[|X_r-X_r^n|\right].$$

The thesis is achieved by applying Proposition 4.8 and using the fact that $\boldsymbol{\mathcal{E}}_{k+1}^M(r)=\int_0^r\boldsymbol{\mathcal{E}}_k^M(s)\,ds$. □

The upper bounds in Proposition 4.4 are similar to the upper bounds in Proposition 4.2, and the last fact is due to the similarity in the proof: the main changes from the proof of Proposition 4.2 consist in using Proposition 2.11 instead of Theorem 2.9 and in taking initial distributions $\mu_0^X=\mu_0^{X^n}=\delta_x$.

Remark 4.11. When the first moments are not finite the result of Proposition 4.8 cannot be used. However, we can prove that

$$\mathbb{E}\vee\mathbb{E}^n\left[\,|X_t-X_t^n|\wedge 1\right]\leq d_{BL}\big(\mu_0^X,\mu_0^{X^n}\big)\boldsymbol{\mathcal{E}}_0^{\bar{M}}(t)+\bar{H}^n\boldsymbol{\mathcal{E}}_1^{\bar{M}}(t), \tag{82}$$

where $\bar{M}=(M_0)^+ + \bar{\lambda}_1(\Gamma_{\mu_1}-1)^+$ and $\bar{H}^n=H_0^n\wedge\bar{\lambda}_0+\bar{\lambda}_1(\Delta_1^n\wedge 1)$, with $H_0^n=\bar{\lambda}_0\Delta_0^n+\|\lambda_0-\lambda_0^n\|b_0$ and $M_0=\bar{\lambda}_0(\Gamma_{\mu_0}-1)^+-\underline{\lambda}_0(1-\Gamma_{\mu_0})^+ + b_0L_{\lambda_0}$. Indeed, using the same techniques as in the proofs of Lemma 5.1 in [8] and Lemma 4.3 in [7] when setting $g_1(x,x')=|x-x'|\wedge 1$ we get

$$\begin{aligned}\mathbb{L}g_1(x,x')\leq{}&\big(H_0^n+M_0|x-x'|+\bar{\lambda}_0|x-x'|\big)\wedge\bar{\lambda}_0-\bar{\lambda}_0(|x-x'|\wedge 1)\\&+\lambda_1(x,y)\left[(\Delta_1^n+\Gamma_{\mu_1}|x-x'|)\wedge 1-|x-x'|\wedge 1\right].\end{aligned}$$

Then taking into account that $M_0+\bar{\lambda}_0\geq 0$, and that if $a,b,\lambda\geq 0$, then $(a+b)\wedge\lambda\leq a\wedge\lambda+b\wedge\lambda$ and $(ab)\wedge\lambda\leq(a\vee\lambda)(b\wedge 1)$, we have

$$\begin{aligned}\mathbb{L}g_1(x,x')&\leq H_0^n\wedge\bar{\lambda}_0+\big((M_0+\bar{\lambda}_0)\vee\bar{\lambda}_0\big)(|x-x'|\wedge 1)-\bar{\lambda}_0(|x-x'|\wedge 1)\\&\quad+\lambda_1(x,y)\big[\Delta_1^n\wedge 1+(\Gamma_{\mu_1}\vee 1)(|x-x'|\wedge 1)-(|x-x'|\wedge 1)\big]\\&=H_0^n\wedge\bar{\lambda}_0+\lambda_1(x,y)(\Delta_1^n\wedge 1)+\big((M_0)^+ +\lambda_1(x,y)(\Gamma_{\mu_1}-1)^+\big)(|x-x'|\wedge 1)\\&\leq\bar{H}^n+\bar{M}(|x-x'|\wedge 1).\end{aligned}$$

Therefore, when the hypothesis on the first moments does not hold, taking into account (82) and the estimates in Remarks 2.5, 2.10, and 2.12, the upper bound (78) of Theorem 4.5 remains valid with minor changes in the constants and by substituting M and H^n with $\bar{M}$ and $\bar{H}^n$, respectively.

Appendix: Proof of Lemma 4.7. In order to prove (81) it is sufficient to prove that

$$\mathbb{E}^{\mathbb{Q}}\big[\,(\bar\lambda_1)^{N_1(t)}\,\mathbb{I}_{\{X_t\neq X_t^n\}}\big] \le \mathbb{E}^{\mathbb{Q}}\big[\,(\bar\lambda_1)^{N_1(t)}\,\big]\|\mu_0^X-\mu_0^{X^n}\|_{TV}/2 + \mathbb{E}^{\mathbb{Q}}\big[\big(N_0(t)+N_1(t)\big)\,(\bar\lambda_1)^{N_1(t)}\,\big]\,J^n. \tag{83}$$

Indeed the r.h.s. of (83) is equal to the r.h.s. of (81), since $\boldsymbol{N_0}$ and $\boldsymbol{N_1}$ are independent under $\mathbb{Q}$ and for any Poisson process $M(t)$ of parameter ρ, and for any $a>0$, it holds that $\mathbb{E}[M(t)]=\rho\,t$, $\mathbb{E}\big[a^{M(t)}\big]=\exp\{\rho(a-1)t\}$, $\mathbb{E}\big[M(t)\,a^{M(t)}\big]=\rho a\,t\exp\{\rho(a-1)t\}$.

Let T_k denote the jump times of the Poisson process $N(t)=N_0(t)+N_1(t)$, where we recall that $N_0(t)=\mathcal{N}_0\big([0,\bar\lambda_0]\times(0,t]\big)$ and $N_1(t)=\mathcal{N}_1\big([0,1]\times(0,t]\big)$. Then the processes defined by (55), (56), and (57), i.e.,

$$\big(X_t,X_t^n,\hat Y_t\big)=\big(X_t,X_t^n,y+N_1(t)\big),$$

are obtained as a Markov chain $(\eta_k,\eta_k^n,\hat y_k)$ evaluated at time $k=N(t)$, and thus $(\eta_k,\eta_k^n,\hat y_k)=(X_{T_k},X^n_{T_k},\hat Y_{T_k})$. Moreover, the Markov chain and the Poisson process $\{N(t),t>0\}$ are independent.

Let $S\in\{T_h,\,h\ge0\}$ be the first time when $\{X_t\neq X_t^n\}$; then

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[(\bar\lambda_1)^{N_1(t)}\mathbb{I}_{\{X_t\neq X_t^n\}}\big] \le \mathbb{E}^{\mathbb{Q}}\big[(\bar\lambda_1)^{N_1(t)}\mathbb{I}_{\{S\le t\}}\big]\\
&\quad=\mathbb{E}^{\mathbb{Q}}\big[(\bar\lambda_1)^{N_1(t)}\mathbb{I}_{\{X_0\neq X_0^n\}}\big]+\sum_{k=1}^{\infty}\mathbb{E}^{\mathbb{Q}}\left[\mathbb{I}_{\{N(t)=k\}}\sum_{h=1}^{k}(\bar\lambda_1)^{\hat Y_t-y}\mathbb{I}_{\{S=T_h\}}\right]\\
&\quad=\mathbb{E}^{\mathbb{Q}}\big[(\bar\lambda_1)^{N_1(t)}\big]\mathbb{Q}(X_0\neq X_0^n)+\sum_{k=1}^{\infty}\sum_{h=1}^{k}\mathbb{E}^{\mathbb{Q}}\big[\mathbb{I}_{\{N(t)=k\}}(\bar\lambda_1)^{\hat y_k-y}\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\big].
\end{aligned} \tag{84}$$

The first addend is equal to the first addend in the r.h.s. of inequality (83), since $\mathbb{Q}(X_0\neq X_0^n)=\|\mu_0^X-\mu_0^{X^n}\|_{TV}/2$. Moreover, for any k and $1\le h\le k$, we have

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[\mathbb{I}_{\{N(t)=k\}}\big(\bar\lambda_1\big)^{\hat y_k-y}\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\big]\\
&\quad=\mathbb{Q}(N(t)=k)\,\mathbb{E}^{\mathbb{Q}}\big[\big(\bar\lambda_1\big)^{\hat y_k-y}\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\big],
\end{aligned}$$

since under $\mathbb{Q}$ the Poisson process $N(t)$ and the Markov chain $(\eta_k,\eta_k^n,\hat y_k)$ are independent.

Therefore, according to Lemma A.1 below, the second addend in the last line of (84) is bounded above by

$$\begin{aligned}
&\sum_{k=1}^{\infty}\mathbb{Q}(N(t)=k)\,k\,\mathbb{E}^{\mathbb{Q}}\big[\big(\bar\lambda_1\big)^{\hat y_k-y}\big]\;J^n\\
&=\sum_{k=1}^{\infty}\mathbb{E}^{\mathbb{Q}}\big[\mathbb{I}_{\{N(t)=k\}}\,k\,\big(\bar\lambda_1\big)^{\hat y_k-y}\,\big]\;J^n=\mathbb{E}^{\mathbb{Q}}\big[N(t)\,\big(\bar\lambda_1\big)^{N_1(t)}\,\big]\;J^n,
\end{aligned}$$

and this ends the proof of inequality (83).

LEMMA A.1. *For any $k\ge1$ and $1\le h\le k$,*

$$\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_k-y}\,\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\,\big]\le\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_k-y}\,\big]\;J^n. \tag{85}$$

To prove this inequality we need the following technical result.

LEMMA A.2. *Under the hypotheses and notation of Proposition* 4.6, *the following upper bound holds for any* $a > 0$*:*

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[a^{\hat{y}_{k+1}-\hat{y}_k}\mathbb{I}_{\{\eta_{k+1}\neq\eta^n_{k+1}\}}|\eta_k = z, \eta^n_k = z, \hat{y}_k = w\big] \\
&\quad = a^1\,\frac{1}{1+\bar{\lambda}_0}\|\mu_1(z,w,\cdot) - \mu^n_1(z,w,\cdot)\|_{TV}/2 \qquad (86)\\
&\qquad\quad + a^0\,\frac{\bar{\lambda}_0}{1+\bar{\lambda}_0}\|p_0(z,w,\cdot) - p^n_0(z,w,\cdot)\|_{TV}/2 \\
&\quad \le \mathbb{E}^{\mathbb{Q}}\big[a^{\hat{y}_{k+1}-\hat{y}_k}\big]\max(\|p_0 - p^n_0\|_\infty, \|\mu_1 - \mu^n_1\|_\infty)/2. \qquad (87)
\end{aligned}$$

Proof. Without loss of generality, we can assume that $\mathcal{N}_0 = \{(T^0_k, \zeta^0_k), k \ge 1\}$ is the restriction to $(0,\infty)\times(0,\bar{\lambda}_0]$ of a Poisson measure $\mathcal{N} = \{(T_k,\zeta_k), k \ge 1\}$ on $(0,\infty)\times(-1,\bar{\lambda}_0]$. Moreover, if $\overline{\mathcal{N}}_1 = \{(\bar{T}^1_k, \bar{\zeta}^1_k), k \ge 1\}$ is the restriction of $\mathcal{N}$ to $(0,\infty)\times(-1,0]$, we can assume that $\mathcal{N}_1 = \{(T^1_k,\zeta^1_k), k \ge 1\}$ is obtained by taking $(T^1_k,\zeta^1_k) = (\bar{T}^1_k, -\bar{\zeta}^1_k)$.

We recall that under $\mathbb{Q}$ the processes

$$\big(X_t, X^n_t, \hat{Y}_t\big) = \big(X_t, X^n_t, y + N_1(t)\big)$$

are constructed by defining K_0, K^n_0, K_1, and K^n_1 as explained in Definition 3.8 of [8]; that is, K_0 and K^n_0 are defined by means of Lemma 3.7 of [8] using the measures $p_0(x,y;dx')$ and $p^n_0(x,y;dx')$ defined by (63) and (64), respectively, while K_1 and K^n_1 are defined using the measures $\mu_1(x,y;dz)$ and $\mu^n_1(x,y;dz)$.

Equality (86) can be derived as follows:

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[a^{\hat{y}_{k+1}-\hat{y}_k}\mathbb{I}_{\{\eta_{k+1}\neq\eta^n_{k+1}\}}|\ \eta_k = z, \eta^n_k = z, \hat{y}_k = w\big] \\
&\quad = \mathbb{E}^{\mathbb{Q}}\big[a^1\,\mathbb{I}_{\{\zeta_{k+1}\in(-1,0]\}}\,\mathbb{I}_{\{K_1(z,z,w;-\zeta_{k+1})\neq K^n_1(z,z,w;-\zeta_{k+1})\}}\big] \\
&\qquad + \mathbb{E}^{\mathbb{Q}}\big[a^0\,\mathbb{I}_{\{\zeta_{k+1}\in(0,\bar{\lambda}_0]\}}\,\mathbb{I}_{\{K_0(z,z,w;\zeta_{k+1})\neq K^n_0(z,z,w;\zeta_{k+1})\}}\big] \\
&\quad = a^1\mathbb{Q}(\zeta_{k+1}\in(-1,0], K_1(z,z,w;-\zeta_{k+1}) \neq K^n_1(z,z,w;-\zeta_{k+1})) \\
&\qquad + a^0\mathbb{Q}(\zeta_{k+1}\in(0,\bar{\lambda}_0], K_0(z,z,w;\zeta_{k+1}) \neq K^n_0(z,z,w;\zeta_{k+1})) \\
&\quad = a^1\,\frac{1}{1+\bar{\lambda}_0}\int_0^1 \mathbb{I}_{\{u:\,K_1(z,z,w;u)\neq K^n_1(z,z,w;u)\}}(v)\,dv \\
&\qquad + a^0\,\frac{\bar{\lambda}_0}{1+\bar{\lambda}_0}\int_0^{\bar{\lambda}_0} \mathbb{I}_{\{u:\,K_0(z,z,w;u)\neq K^n_0(z,z,w;u)\}}(v)\,\frac{dv}{\bar{\lambda}_0} \\
&\quad = a^1\,\frac{1}{1+\bar{\lambda}_0}\|\mu_1(z,w,\cdot) - \mu^n_1(z,w,\cdot)\|_{TV}/2 \\
&\qquad + a^0\,\frac{\bar{\lambda}_0}{1+\bar{\lambda}_0}\|p_0(z,w,\cdot) - p^n_0(z,w,\cdot)\|_{TV}/2,
\end{aligned}$$

where the last equality is due to the specific choice of K_0, K^n_0, K_1, and K^n_1. Observing that

$$a^1\,\frac{1}{1+\bar{\lambda}_0} + a^0\,\frac{\bar{\lambda}_0}{1+\bar{\lambda}_0} = \mathbb{E}^{\mathbb{Q}}\big[a^{\hat{y}_{k+1}-\hat{y}_k}\big],$$

we see that inequality (87) is an immediate consequence of (86). ∎

Proof of Lemma A.1. We observe that

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_k - y}\,\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\big]\\
&=\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h - y}\;\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\;\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_k-\hat y_h}\;\big|\mathcal{F}_h^{\eta,\eta^n,\hat y}\big]\big]\\
&=\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h - y}\;\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_l=\eta_l^n,\,l\le h-1\}}\;\big]\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_k-\hat y_h}\,\big]\\
&\le\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h - y}\;\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_{h-1}=\eta_{h-1}^n\}}\;\big]\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_k-\hat y_h}\,\big].
\end{aligned}$$

Then (85) follows, taking into account that

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h - y}\;\mathbb{I}_{\{\eta_h\neq\eta_h^n,\,\eta_{h-1}=\eta_{h-1}^n\}}\,\big]\\
&=\mathbb{E}^{\mathbb{Q}}\big[\mathbb{I}_{\{\eta_{h-1}=\eta_{h-1}^n\}}\;\big(\bar\lambda_1\big)^{\hat y_{h-1} - y}\,\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h-\hat y_{h-1}}\;\mathbb{I}_{\{\eta_h\neq\eta_h^n\}}\;\big|\mathcal{F}_{h-1}^{\eta,\eta^n,\hat y}\big]\big]\\
&=\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_{h-1} - y}\,\mathbb{I}_{\{\eta_{h-1}=\eta_{h-1}^n\}}\;\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h-\hat y_{h-1}}\;\mathbb{I}_{\{\eta_h\neq\eta_h^n\}}\;\big|\eta_{h-1},\eta_{h-1}^n,\hat y_{h-1}\big]\big],
\end{aligned}$$

and that, using inequality (87) of Lemma A.2, with $a=\bar\lambda_1$, the above quantity is bounded above by

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}}\big[(\bar\lambda_1)^{\hat y_{h-1}-y}\mathbb{I}_{\{\eta_{h-1}=\eta_{h-1}^n\}}\;\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h-\hat y_{h-1}}\,\big]\max(\|p_0-p_0^n\|_\infty,\|\mu_1-\mu_1^n\|_\infty)/2\big]\\
&\le\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_{h-1}-y}\,\big]\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h-\hat y_{h-1}}\,\big]J^n=\mathbb{E}^{\mathbb{Q}}\big[\,\big(\bar\lambda_1\big)^{\hat y_h-y}\,\big]\;J^n,
\end{aligned}$$

where in the last equality we have used the independence of the increments $\hat y_{h-1}-y$ and $\hat y_h-\hat y_{h-1}$ under $\mathbb{Q}$. □

REFERENCES

[1] A. G. Bhatt, G. Kallianpur, and R. L. Karandikar, *Robustness of the nonlinear filter*, Stochastic Process. Appl., 81 (1999), pp. 247–254.

[2] A. G. Bhatt and R. L. Karandikar, *Robustness of the nonlinear filter: The correlated case*, Stochastic Process. Appl., 97 (2002), pp. 41–58.

[3] A. Budhiraja and H. J. Kushner, *Robustness of nonlinear filters over the infinite time interval*, SIAM J. Control Optim., 36 (1998), pp. 1618–1637.

[4] A. Calzolari, P. Florchinger, and G. Nappo, *Nonlinear Filtering for Markov Systems with Delayed Observations*, preprint, Department of Mathematics, University of Rome "La Sapienza," Rome, Italy, 2004. Available online at http://www.mat.uniroma1.it/people/nappo/papers.pdf/Delay42.pdf.

[5] A. Calzolari, P. Florchinger, and G. Nappo, *Convergence in Nonlinear Filtering for Stochastic Delay Systems*, preprint, Department of Mathematics, University of Rome "La Sapienza," Rome Italy, 2005. Available online at http://www.mat.uniroma1.it/people/nappo/papers.pdf/CFN-conv-delay-rate.pdf.

[6] A. Calzolari and G. Nappo, *A filtering problem with counting observations: Error bounds due to the uncertainty on the infinitesimal parameters*, Stochastics Stochastics Rep., 61 (1997), pp. 1–19.

[7] A. Calzolari and G. Nappo, *Robust approximation in a filtering problem with real state space and counting observations*, Appl. Math. Optim., 42 (2000), pp. 51–71.

[8] A. Calzolari and G. Nappo, *Counting observations: A note on state estimation sensitivity with an L^1-bound*, Appl. Math. Optim., 44 (2001), pp. 177–201.

[9] C. Ceci, A. Gerardi, and P. Tardelli, *An estimate of the approximation error in the filtering of a discrete jump process*, Math. Models Methods Appl. Sci., 11 (2001), pp. 181–198.

[10] M. H. Chang, *Discrete approximation of nonlinear filtering for stochastic delay equations*, Stochastic Anal. Appl., 5 (1987), pp. 267–298.

[11] P. Del Moral, *Measure-valued processes and interacting particle systems. Application to nonlinear filtering problems*, Ann. Appl. Probab., 8 (1998), pp. 438–495.

[12] G. B. Di Masi, M. Pratelli, and W. J. Runggaldier, *An approximation for the nonlinear filtering problem, with error bound*, Stochastics, 14 (1985), pp. 247–271.

[13] R. J. Elliott and R. Glowinski, *Approximations to solutions of the Zakai filtering equation*, Stochastic Anal. Appl., 7 (1989), pp. 145–168.

[14] M. Kirch and W. J. Runggaldier, *Efficient hedging when asset prices follow a geometric Poisson process with unknown intensities*, SIAM J. Control Optim., 43 (2004), pp. 1174–1195.

[15] W. Kliemann, G. Koch, and F. Marchetti, *On the unnormalized solution of the filtering problem with counting observations*, IEEE Trans. Inform. Theory, 316 (1990), pp. 1415–1425.

[16] T. G. Kurtz and D. L. Ocone, *Unique characterization of conditional distributions in nonlinear filtering*, Ann. Probab., 16 (1988), pp. 80–107.

[17] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Math. Sci. Engrg. 129, Academic Press (Harcourt Brace Jovanovich, Publishers), New York, London, 1977.

[18] F. Le Gland, *Time discretization of nonlinear filtering equations*, in Proceedings of the 28th IEEE Conference on Decision and Control (Tampa, FL), Vol. 3, 1989, pp. 2601–2606.

[19] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, *Nonlinear filtering revisited: A spectral approach*, SIAM J. Control Optim., 35 (1997), pp. 435–461.

[20] É. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, in École d'Été de Probabilités de Saint-Flour XIX—1989, Lecture Notes in Math. 1464, Springer, Berlin, 1991, pp. 67–163.

[21] M. Schweizer, *Risk-minimizing hedging strategies under restricted information*, Math. Finance, 4 (1994), pp. 327–342.

[22] K. Twardowska, T. Marnik, and M. Pasławaska-Połuniak, *Approximation of the Zakai equation in a nonlinear filtering problem with delay*, Int. J. Appl. Math. Comput. Sci., 13 (2003), pp. 151–160.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 634–667

THE VERIFICATION AND CONTROL OF INTERACTING SIMILAR DISCRETE-EVENT SYSTEMS*

KURT ROHLOFF† AND STÉPHANE LAFORTUNE‡

Abstract. This paper explore issues related to the control and verification of similar module systems in the discrete-event systems framework. Similar module systems are distributed systems comprised of subsystem modules that exhibit isomorphic local behavior coordinated on global event occurrences. When given a global model of these systems, it is shown how to decompose the global model into the component subsystems in polynomial time. It is also shown how to perform various verification tasks for these interacting systems while mitigating common state explosion difficulties by taking advantage of the special similar module system structure. Control properties of the similar module systems are also discussed. It is assumed that the local modules are supervised by exactly one local controller and the controllers enforce the same local control policy. Necessary and sufficient conditions for achieving local and global control specifications in this setting are identified.

Key words. control, discrete-event systems, isomorphic behavior, distributed systems

AMS subject classifications. 93C65, 93A14

DOI. 10.1137/040610209

1. Introduction.

Many significant, real-world systems are comprised of concurrent processes that are most conveniently modeled in a distributed manner. The current standard methods for solving many important verification and control problems for these distributed systems typically involve converting the distributed system models into equivalent monolithic system models. Unfortunately, these monolithic system conversions are generally computationally expensive. This has prompted researchers to investigate solution methods for distributed system problems that avoid these procedures. However, it has been found that many of these important verification and control problems for distributed systems are still computationally intractable despite the avoidance of the monolithic conversions [8, 9, 14, 26, 28, 30].

This paper expands on prior investigations into the control of distributed systems by exploring the properties of these systems that can be modeled as a set of interacting discrete-event system modules, where the system module behaviors are identical with respect to a relabeling of local behavior. This class of distributed systems is referred to as the class of *similar module systems* and is formally defined below. For these systems, the behaviors of the modules are represented as languages over a set of events partitioned into private and global events generated by finite-state automata. All modules must coordinate their behavior on the occurrence of global events, while private events represent behavior that is relevant to only one module.

The similar module system model used in this paper is designed to be as general as possible in order for it to have wide applicability, yet still have computationally

*Received by the editors June 18, 2004; accepted for publication (in revised form) December 7, 2005; published electronically May 3, 2006. This research was supported in part by NSF grants CCR-0082784 and CCR-0325571. The majority of the work presented in this paper was performed while the first author was at the University of Michigan.

http://www.siam.org/journals/sicon/45-2/61020.html

†BBN Technologies, 10 Moulton St., Cambridge, MA 02138 (krohloff@bbn.com; http://www.dist-systems.bbn.com/people/krohloff).

‡Department of Electrical Engineering and Computer Science, The University of Michigan, 1301 Beal Ave., Ann Arbor, MI 48109-2122 (stephane@eecs.umich.edu; www.eecs.umich.edu/umdes).

reasonable solution methods for otherwise difficult verification and control problems. Consequently, many important real-world processes are included in the class of similar module systems that can be modeled in the framework discussed in this paper. Aspects of the behavior of systems such as communication networks, where the nodes are all running the same communication protocol, distributed sensor networks, and platoons of similar unmanned aerial vehicles, can all be effectively represented by similar module systems for the purpose of verification and control with respect to logical specifications. A number of biological processes can also be effectively represented by similar module systems [1, 21].

The work in this paper is presented using the framework of discrete-event systems and supervisory control theory surveyed in the seminal work [25]. The interested reader may consult the text [2] for a general introduction to this framework. The finite automaton modeling formalism used in this paper is generally considered to be the simplest one for discrete-event systems that is expressive enough to model the behavior of real-world systems in a reasonable manner. Furthermore, as the behavior of the similar module systems is coordinated on the occurrence of global events, an equivalent monolithic global system model of a set of interacting similar module systems is constructed by composing the similar module systems using the parallel composition operation. The parallel composition operation is currently the standard method for modeling the coordination of behavior of distributed discrete-event systems.

Because similar module systems exhibit behavior that is distinctly local and global, specifications on the behavior of similar modular systems can be made at both the global and local levels. Global specifications describe the desired behavior of a system when the modules in the similar module system interact. Local specifications describe the desired locally relevant behavior of a system module when it interacts with other modules. The global and local behavioral specifications used in the paper are specified as the languages marked by finite-state automata.

Properties related to the control of similar modular systems for both local and global specifications given as regular languages are explored when the desired behavior of the controlled similar module system should match the given specification. It is assumed that when the modules are controlled, they are controlled locally with exactly one controller per module such that the controllers make local observations of the behavior of a module and enforce local control actions. As with the system modules, the controllers are similar in that all controllers enforce the same control policy at each of their respective subsystems with respect to a renaming of the relevant local events. It is also assumed that there is no communication between controllers besides the implicit communication due to the occurrence of observable global events.

Beyond the introduction of the similar module systems, the main contributions of this paper include

- a polynomial time decomposition operation for converting a monolithic similar module system model into a more efficiently encoded distributed similar module system model.
- a quotient machine construction for more efficiently testing state reachability and other important properties of composed similar module systems that avoids expensive parallel composition operations.
- a polynomial time method for verifying local behavior specifications and a set of necessary and sufficient conditions for achieving local control objectives in similar module systems.

- a set of necessary and sufficient conditions for achieving global control objectives in similar module systems that implies a polynomial time method for synthesizing similar module systems which satisfy global control specifications.

The work in this paper generalizes and extends that in [27], which contains preliminary versions of some of the results presented herein. The extensions in this paper beyond those shown in [27] include the improved method for performing the verification of important system properties such as state reachability in similar module systems and the polynomial time decomposition operation for composed similar module systems.

The next section gives a survey of the literature relevant to the work presented in this paper. Section 3 introduces the system model along with the notation and basic system properties used in this paper. The quotient automaton is introduced in section 4 for decreasing the computational difficulty of testing state reachability, global nonblockingness, and deadlock-freeness in similar module systems. In section 5 the properties related to the verification of similar module systems for specifications made at local sites are discussed. Section 6 shows the computationally efficient method of decomposing global models of similar module systems into distributed models. The control formalisms used in this paper are presented section 7. Section 8 discusses properties related to the control of similar module systems for specifications made at the local sites, and the control of similar module systems for the satisfaction of global specifications is discussed in section 9. This paper concludes with a discussion of the results and possible avenues for future work in section 10.

2. Related literature survey. The supervision of modular discrete-event systems is currently receiving much attention in the control community; see, e.g., [3, 12, 13, 15, 16, 17, 19, 20, 22, 23, 31, 32, 33]. Some of the earlier results related to modular supervision are shown in [17, 19, 23, 31, 33]. Properties of modular discrete-event systems in which the modules have disjoint event sets are investigated in [22, 23]. Various local specification and concurrent supervision problems, respectively, are investigated in [12, 13]. Properties related to the supervision of modular systems using specific architectures are explored in [15, 16]. A form of modular control in which each controller has a different objective is discussed in [3], and situations when local nonblocking behavior implies global nonblocking behavior are discussed in [20].

Excluding [7, 27], there has been little work in the supervisory control community that exploits system symmetry when controlling discrete-event systems. Group-theoretic methods are used in [7] to define classes of states and transitions in monolithic system that are equivalent under predefined group-theoretic permutation operations. These classes of equivalent states and transitions are used to construct a quotient automaton that has a state space smaller than that of the original system, which is used to perform various centralized controller synthesis operations for the original system. Similar module systems are briefly discussed in [7] and the predefined quotient structure is extended to this setting. However, centralized control is discussed exclusively in [7]. A similar quotient automaton with predefined state equivalence classes is also discussed in [27] for similar module systems and is likewise based on state permutation equivalences.

A type of quotient automaton is presented below to test properties such as state reachability, nonblockingness, and deadlock-freeness that uses a much more efficient predefined state aggregation method for defining state equivalence classes than seen in [7, 27]. This results in a quotient automaton with a much smaller state space than the previously discussed quotient automata with predefined state equivalence classes.

Furthermore, necessary and sufficient conditions are discussed for the existence of decentralized controllers for these systems using both local and global specifications. This control setting is not discussed in [7].

The work in [7, 27] and this paper is inspired by an extensive literature from the formal methods community based on the uses of symmetry for performing model checking over large symmetric systems. Most of the current work in the formal methods community related to symmetry is based on methods presented in [5, 6]. Some of this work is also summarized in the text [4]. Due to the natural similarity of the supervisory control theory models used in this paper and the Kripke structure models used in the formal methods literature, results related to the verification of symmetric systems as presented in this literature are briefly discussed.

Group-theoretic methods and the quotient automaton structures for Kripke structures are presented in [5, 6]. These references present how various temporal logic properties can hold in the original system if and only if those properties hold in the quotient automaton. Defining effective state equivalence classes is computationally at least as difficult as the graph isomorphism problem [10]. This has induced several authors to attempt to design distributed systems with special architectures so that symmetry is guaranteed to occur a priori, as in [6, 7, 27] and this paper. Even when used with distributed systems, the symmetry reductions due to the construction of the quotient automata presented in [5, 6] do not alleviate the state-explosion problem in the worst case. However, several real-world examples are presented in [11], where the quotient models allow for an 80%–90% reduction in computation time and space for several important temporal logic verification problems.

There has also been work in the formal methods community on the use of partial order reductions for the verification of concurrent systems [4]. Although the reduction methods appear similar, partial order reduction methods are fundamentally different from the symmetry reduction methods discussed herein.

3. Similar module systems. As was stated in the introduction, the similar module systems are modeled as a set of finite automata $\{G_1, \ldots, G_n\}$ that interact on the occurrence of common global events. Furthermore, all of the modules in $\{G_1, \ldots, G_n\}$ are exact copies of one another with respect to a renaming of private events.

Let the automaton $G_i = (X, x_0, \Sigma_i, \delta_i, X_m)$ be the ith module in the system, where $i \in \{1, \ldots, n\}$. The module event set Σ_i is partitioned into the distinct subsets Σ_g and Σ_{pi}, the (unique) global event set and the private event set for module i, respectively. The private event sets $\Sigma_{p1}, \ldots, \Sigma_{pn}$ are copies of one another for $G_1, \ldots, G_n$, respectively, such that for all $i, j, i \neq j$, $\Sigma_{pi} \cap \Sigma_{pj} = \emptyset$. It is assumed that $\Sigma_g \cap \Sigma_{pi} = \emptyset$ and Σ denotes $\Sigma_1 \cup \cdots \cup \Sigma_n$.

For any local event $\sigma_i \in \Sigma_{pi}$ there are corresponding local events $\Psi_{i1}(\sigma_i) \in \Sigma_{p1}, \ldots, \Psi_{in}(\sigma_i) \in \Sigma_{pn}$ in the other private event sets. The one-to-one mapping $\Psi_{ij}(\cdot)$ is extended for notational simplicity in this setting to $\Psi_{ij} : \Sigma \to \Sigma$ so that the private event set of module i is mapped to the jth private event set, the private event set of module j is mapped to the ith private event set, and all other events are mapped to themselves. The function $\Psi_{ij}(\cdot)$ is extended in the usual manner to map between strings of events and languages. Note that $\Psi_{ii}(\cdot)$ is the identity function and $\Psi_{ii}(\cdot) = \Psi_{ij}(\Psi_{ij}(\cdot))$. As an example of the operation of the $\Psi_{ij}(\cdot)$ function, suppose there are system event sets $\Sigma_i = \{a_i, b_i, d, g\}$ and $\Sigma_j = \{a_j, b_j, d, g\}$. In this case $\Sigma_g = \{d, g\}$, $\Sigma_{p1} = \{a_i, b_i\}$, and $\Sigma_{p2} = \{a_j, b_j\}$. Then, $\Psi_{ij}(dga_ia_jb_jgb_i) = dga_ja_ib_igb_j$.

With the above framework, similar module systems can now be formally defined.

DEFINITION 1. *The set of modules* $\{G_1, \ldots, G_n\}$ *as described above is a similar module system if for all* $i, j \in \{1, \ldots, n\}$*, for all* $x \in X$ *and* $\gamma \in \Sigma_i$*,* $\delta_i(x, \gamma) = \delta_j(x, \Psi_{ij}(\gamma))$ *if it is defined.*

Example 1 shows a simple example of the systems discussed in this paper.

Example 1. Consider the 2 module similar module system composed of G_1 and G_2 seen in Figure 1. The relevant event sets are $\Sigma_g = \{\gamma, \lambda\}$, $\Sigma_{p1} = \{\alpha_1, \beta_1\}$, and $\Sigma_{p2} = \{\alpha_2, \beta_2\}$. Note that G_1 and G_2 are both locally nonblocking and deadlock free.

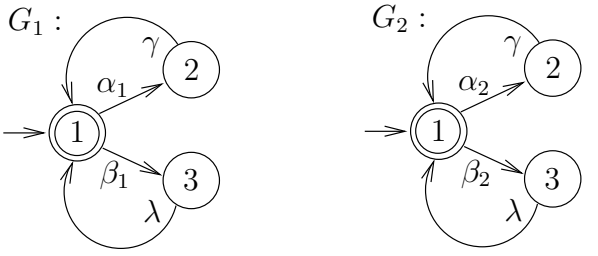


FIG. 1. *The automata* G_1 *and* G_2.

The natural projection $P_i : \Sigma^* \to \Sigma_i^*$ is defined such that $P_i(s)$ is the string s with events in $\Sigma \setminus \Sigma_i$ erased. Similarly, let $P_{pi} : \Sigma^* \to \Sigma_{pi}^*$ be the natural projection that erases events in $\Sigma \setminus \Sigma_{pi}$, and let $P_g : \Sigma^* \to \Sigma_g^*$ be the natural projection that erases events in $\Sigma \setminus \Sigma_g$.

The automata $\{G_1, \ldots, G_n\}$ model the private behaviors of the modules of the similar module system, but the parallel composition $G_1 \| \cdots \| G_n$ is used to model the global behavior of all the interacting subsystems. The behavior of the composed automaton $G_1 \| \cdots \| G_n$ is equivalent to the concurrent behavior of the modules $\{G_1, \ldots, G_n\}$, where the local behaviors of $\{G_1, \ldots, G_n\}$ are coordinated on the occurrence of common global events in Σ_g.

To review the parallel composition operation, suppose the automata G_1 and G_2 are given. The parallel composition of G_1 and G_2, denoted by $G_1 \| G_2$, is defined as follows:

$$G_1 \| G_2 := ((X \times X), (x_o, x_o), \Sigma_1 \cup \Sigma_2, \delta^{G_1 \| G_2}, (X_m \times X_m)),$$

where

(1)

$$\delta^{G_1 \| G_2}((x^{G_1}, x^{G_2}), \sigma) = \begin{cases} (\delta_1(x^{G_1}, \sigma), \delta_2(x^{G_2}, \sigma)) & \text{if } \delta_1(x^{G_1}, \sigma)! \wedge \delta_2(x^{G_2}, \sigma)! \\ (\delta_1(x^{G_1}, \sigma), x^{G_2}) & \text{if } \delta_1(x^{G_1}, \sigma)! \wedge (\sigma \notin \Sigma_2) \\ (x^{G_1}, \delta_2(x^{G_2}, \sigma)) & \text{if } \delta_2(x^{G_2}, \sigma)! \wedge (\sigma \notin \Sigma_1) \\ \text{undefined} & \text{otherwise} \end{cases}.$$

Note that the unary operator ! is used such that $f(\alpha)!$ returns true if $f(\cdot)$ is defined for input α, but false otherwise. The parallel composition $G_1 \| G_2$ is also defined in [2] for the case where G_1 and G_2 are not similar module systems.

Example 2 shows the automaton $G_1 \| G_2$ constructed from the automata in Example 1.

Example 2. Consider the composed system $G_1 \| G_2$ as seen in Figure 2.

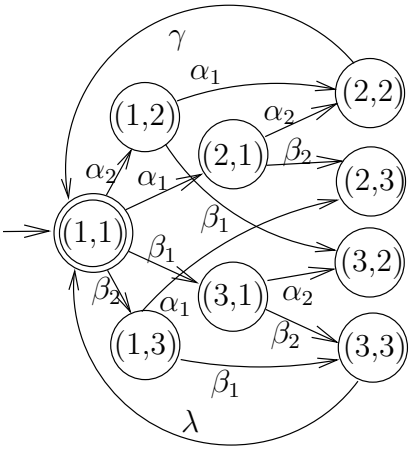


FIG. 2. *The automaton* $G_1 \| G_2$.

The automaton $G_1 \| G_2$ has a large degree of state-based symmetry due to the similarity of the component automata G_1 and G_2. Consider for instance the states $(1,2)$ and $(2,1)$; these states are reached by the occurrence of α_2 and α_1 events, respectively.

Now consider the states $(2,3)$ and $(3,2)$. The state $(2,3)$ can be reached from $(1,1)$ on the occurrence of either $\alpha_1\beta_2$ or $\beta_2\alpha_1$. The state $(3,2)$ can be reached from $(1,1)$ on the occurrence of either $\alpha_2\beta_1$ or $\beta_1\alpha_2$. Notice that if the subscript labels of the events in the strings leading to state $(2,3)$ are swapped, then the resulting string leads to state $(3,2)$.

This is a symptom of properties inherent to similar module systems such that many properties of state $(2,3)$ are also held in a sense by state $(3,2)$. For this similar module system the swapping of event subscript labels is equivalent to reordering the parallel composition of the component automata by swapping state locations. This operation is valid because the parallel composition operation is commutative and therefore the order of parallel composition is arbitrary. The various similar modules are identical with respect to a renaming of private events.

For $G_1 \| G_2$ there are six classes of states that could be considered equivalent with respect to a reordering of the component states. The six sets of equivalent state classes are

$$\{(1,1)\}, \{(1,2),(2,1)\}, \{(1,3),(3,1)\}, \{(2,2)\}, \{(2,3),(3,2)\}, \{(3,3)\}.$$

It is now shown that $P_i\left(\mathcal{L}_m(G_1 \| \cdots \| G_n)\right)$, the local behavior relevant to module i of the interacting modules, is language equivalent to the local behavior of module i when the modules operate in isolation (i.e., $\mathcal{L}_m(G_i)$).

THEOREM 1. *For a similar module system* $\{G_1, \ldots, G_n\}$ *as introduced above with respective local projection operations* $\{P_1, \ldots, P_n\}$ *and for* $i \in \{1, \ldots, n\}$,

$$P_i\left(\mathcal{L}_m(G_1 \| \cdots \| G_n)\right) = \mathcal{L}_m(G_i).$$

Proof. It is known that

$$\mathcal{L}_m(G_1 \| \cdots \| G_n) = \left[P_1^{-1}(\mathcal{L}_m(G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(G_n))\right],$$

and thus it is sufficient to show that $\left[P_i\left[P_1^{-1}(\mathcal{L}_m(G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(G_n))\right] = \mathcal{L}_m(G_i)\right]$.

This proof is composed of two parts.
It is known that $P_1^{-1}(\mathcal{L}_m(G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(G_n)) \subseteq P_i^{-1}(\mathcal{L}_m(G_i))$.
This implies that $P_i\left[P_1^{-1}(\mathcal{L}_m(G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(G_n))\right] \subseteq P_i\left[P_i^{-1}(\mathcal{L}_m(G_i))\right]$.
For any language K it is known that $P_i\left[P_i^{-1}(K)\right] = K$, and thus
$P_i\left[P_1^{-1}(\mathcal{L}_m(G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(G_n))\right] \subseteq \mathcal{L}_m(G_i)$
$\Rightarrow P_i(\mathcal{L}_m(G_1 \| \cdots \| G_n)) \subseteq \mathcal{L}_m(G_i)$.
The other direction is proved by showing
$(t_i \in \mathcal{L}_m(G_i)) \Rightarrow \left(t_i \in P_i\left[\cap_{j=1}^n P_j^{-1}(\mathcal{L}_m(G_j))\right]\right)$.

Let $t_i \in \mathcal{L}_m(G_i)$. This implies that $t_i \in P_i(P_i^{-1}(\mathcal{L}_m(G_i)))$. Let the string $t_{\{1,\ldots,n\}}$ be a copy of t_i except that occurrences of all private events $\sigma_{pi} \in \Sigma_{pi}$ in t_i are replaced with the string of events $\sigma_{p1}\Psi_{12}(\sigma_{p1}), \ldots, \Psi_{1n}(\sigma_{p1})$. Due to the construction of $\{G_1, \ldots, G_n\}$, $t_{\{1,\ldots,n\}} \in P_j^{-1}(\mathcal{L}_m(G_j))$ for all $j \in \{1, \ldots, n\}$. Thus,

$\forall j \in \{1, \ldots, n\}\ t_{\{1,\ldots,n\}} \in P_j^{-1}(\mathcal{L}_m(G_j))$
$\Rightarrow t_{\{1,\ldots,n\}} \in \left(\cap_{j=1}^n P_j^{-1}(\mathcal{L}_m(G_j))\right)$.
$\Rightarrow t_i \in P_i\left[\cap_{j=1}^n P_j^{-1}(\mathcal{L}_m(G_j))\right]$.
Therefore, $\mathcal{L}_m(G_i) \subseteq P_i\left[\cap_{j=1}^n P_j^{-1}(\mathcal{L}_m(G_j))\right]$ which implies

$$\mathcal{L}_m(G_i) \subseteq P_i\left(\mathcal{L}_m(G_1 \| \cdots \| G_n)\right). \qquad \square$$

It is a simple matter to extend the result in Theorem 1 to the case of languages generated instead of languages marked, i.e.,

$$P_i\left[P_1^{-1}(\mathcal{L}(G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}(G_n))\right] = \mathcal{L}(G_i).$$

Some basic properties of distributed discrete-event systems are now presented.

DEFINITION 2 (see [31]). *A language $K \subseteq \Sigma^*$ is* separable *with respect to the projections $\{P_1, \ldots, P_n\}$ if $K = P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))$.*

According to the above definition, a language is separable if, with distributed local knowledge of K at all modules, i.e., $P_1(K), \ldots, P_n(K)$, the language K can be recovered exactly through combining all local knowledge as follows: $P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))$. Note that in [29] a concept similar to separability, called decomposability, is defined where a language $K \subseteq \Sigma^*$ is decomposable with respect to the set of projections $\{P_1, \ldots, P_n\}$ and the automaton G if $K = P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K)) \cap \mathcal{L}(G)$.

In general the current known methods for testing a language for separability take exponential time. These methods compute automata $\{B_1, \ldots, B_n\}$ such that for all $i \in \{1, \ldots, n\}$, $\mathcal{L}(B_i) = P_i^{-1}(P_i(K))$ and then test to see if $K = \mathcal{L}(B_1 \| \cdots \| B_n)$.

DEFINITION 3 (see [33]). *The languages in the set $\{L_1, \ldots, L_n\}$ are called* nonconflicting *if*

$$\overline{L_1 \cap \cdots \cap L_n} = \overline{L_1} \cap \cdots \cap \overline{L_n}.$$

It is known that the parallel composition of a set of nonblocking automata need not be nonblocking unless the respective languages marked by the automata are nonconflicting.

DEFINITION 4. *A language $K \subseteq \Sigma^*$ is* symmetric *with respect to the mappings $\Psi_{11}, \ldots, \Psi_{1n}$ if for all $i \in \{1, \ldots, n\}$, $\Psi_{1i}(K) = K$.*

A language marked by a deterministic automaton B can be generally tested for symmetry in linear time. To do this, for all $i \in \{1, \ldots, n\}$, construct B_{1i} such that $\mathcal{L}(B_{1i}) = \Psi_{1i}(\mathcal{L}(B))$. This can be done by replacing the event labels on the state

transitions in B according to the $\Psi_{1i}(\cdot)$ function and can be performed in linear time. Then, if for all $i \in \{1, \ldots, n\}$, $\mathcal{L}(B_{1i}) = \mathcal{L}(B)$, then by definition $\mathcal{L}(B)$ is symmetric. Note that the language equivalence of deterministic automata can be tested in linear time using standard methods.

The symmetry definition is used to convey the intuition that K is identical with respect to a relabeling of private events for the similar module system $\{G_1, \ldots, G_n\}$. This prompts the following lemmas, whose proofs are straightforward and therefore omitted.

LEMMA 1. *Given a set $\Sigma_i \subseteq \Sigma$, a language $K \subseteq \Sigma^*$ with the natural projection operation $P_i : \Sigma^* \to \Sigma_i^*$ and corresponding inverse projection $P_i^{-1} : \Sigma_i^* \to \Sigma^*$,*

1. $P_i(\overline{K}) = \overline{P_i(K)}$.
2. $P_i^{-1}(\overline{K}) = \overline{P_i^{-1}(K)}$.

LEMMA 2. *Given a language $K \subseteq \Sigma^*$, a set of projection operations $\{P_1, \ldots, P_n\}$, and translation mappings $\{\Psi_{11}, \ldots, \Psi_{1n}\}$,*

$$\forall i \in \{1, \ldots, n\} \, (K = \Psi_{1i}(K)) \Rightarrow (\Psi_{1i}(P_1(K)) = P_i(K)) \, .$$

The next result relates the $\Psi_{ij}(\cdot)$ operator with the prefix-closure of languages.

LEMMA 3. *Given a language K and the $\Psi_{ij}(\cdot)$ operator introduced above,*

$$\Psi_{ij}(\overline{K}) = \overline{\Psi_{ij}(K)}.$$

Proof. Suppose $s \in \overline{\Psi_{ij}(K)}$. Then there exists a string t such that $st \in \Psi_{ij}(K)$.

$$\begin{aligned} &\Rightarrow \Psi_{ij}(st) \in K \\ &\Rightarrow \Psi_{ij}(s)\Psi_{ij}(t) \in K \\ &\Rightarrow \Psi_{ij}(s) \in \overline{K} \\ &\Rightarrow s \in \Psi_{ij}(\overline{K}). \end{aligned}$$

The reverse direction of this proof follows from performing the steps above in the opposite order. □

Lemma 3 can be used to show the following theorem about the symmetry of prefix-closed languages.

THEOREM 2. *Let K be symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$. Then, $\overline{K}$ is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$.*

Proof. Because K is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$, then for all $i, j \in \{1, \ldots, n\}$, $\Psi_{ij}(K) = K$. Hence, by Lemma 3, for all $i, j \in \{1, \ldots, n\}$, $\Psi_{ij}(\overline{K}) = \overline{K}$ and therefore $\overline{K}$ is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$. □

Even though a language K may be symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$ and separable with respect to the sets of projections $\{P_1, \ldots, P_n\}$, it is possible that the sets of languages $\{P_1^{-1}(P_1(K)), \ldots, P_n^{-1}(P_n(K))\}$ are conflicting. Consider the following example.

Example 3. Consider the automaton G that is the trimmed version of $G_1 \| G_2$ from Example 1 with the states renamed for convenience. This automaton can be seen in Figure 3.

Note that $\mathcal{L}_m(G) = ((\alpha_1\alpha_2 + \alpha_2\alpha_1)\,\gamma + (\beta_1\beta_2 + \beta_2\beta_1)\,\lambda)^*$. By inspection, this language is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$ and separable with respect to $\{P_1, \ldots, P_n\}$.

Now consider the G_1 and G_2 automata seen in Figure 1. Note that $P_1(\mathcal{L}(G)) = \mathcal{L}(G_1)$ and $P_2(\mathcal{L}(G)) = \mathcal{L}(G_2)$. As seen in Figure 2, the parallel composition of

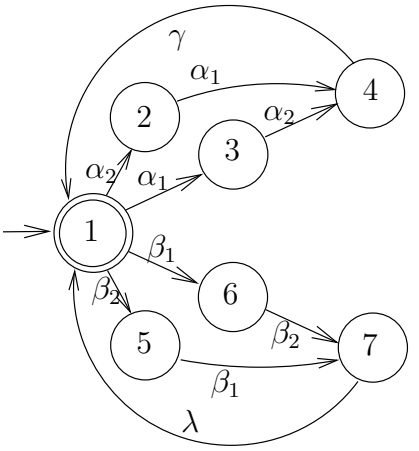


FIG. 3. *The automaton* $G = Trim(G_1 \| G_2)$.

G_1 and G_2 is blocking. Therefore $\{P_1^{-1}(P_1(\mathcal{L}_m(G_1))), \ldots, P_n^{-1}(P_n(\mathcal{L}_m(G_2)))\}$ are conflicting.

Finally, note that even though a language K may be symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$ and $\{P_1^{-1}(P_1(K)), \ldots, P_n^{-1}(P_n(K))\}$ are nonconflicting, it is possible that K is not separable with respect to $\{P_1, \ldots, P_n\}$. Consider the language $K = \{\epsilon, \alpha_1, \alpha_2\}$. K is symmetric with respect to the mapping induced by the event sets $\{\alpha_1\}$ and $\{\alpha_2\}$. Because

$$P_1^{-1}(P_1(K)) = \{\alpha_2^*, \alpha_2^*\alpha_1\alpha_2^*\} = \overline{P_1^{-1}(P_1(K))}$$

and

$$P_2^{-1}(P_2(K)) = \{\alpha_1^*, \alpha_1^*\alpha_2\alpha_1^*\} = \overline{P_2^{-1}(P_2(K))},$$

it holds that

$$P_1^{-1}(P_1(K)) \cap P_2^{-1}(P_2(K)) = \overline{P_1^{-1}(P_1(K))} \cap \overline{P_2^{-1}(P_2(K))} = \{\epsilon, \alpha_1, \alpha_2, \alpha_1\alpha_2, \alpha_2\alpha_1\}.$$

Consequently, $\overline{P_1^{-1}(P_1(K)) \cap P_2^{-1}(P_2(K))} = \overline{P_1^{-1}(P_1(K))} \cap \overline{P_2^{-1}(P_2(K))}$, and thus $P_1^{-1}(P_1(K))$ and $P_2^{-1}(P_2(K))$ are nonconflicting. However, $K = \{\epsilon, \alpha_1, \alpha_2\}$ is not separable with respect to $\{\{\alpha_1\}, \{\alpha_2\}\}$ because $\{\alpha_1\alpha_2, \alpha_2\alpha_1\} \subset P_1^{-1}(P_1(K)) \cap P_2^{-1}(P_2(K))$, but neither $\alpha_1\alpha_2$ nor $\alpha_2\alpha_1$ is in K.

4. Quotient automata for verification. Given a set of automata $\{G_1, \ldots, G_n\}$ such that the size of their respective state spaces is bounded by k, the composed automaton $G_1 \| \cdots \| G_n$ has k^n reachable states in the worst case. Therefore, as n grows, the size of the state space of $G_1 \| \cdots \| G_n$ can become unbearably large. This would consequently cause many verification and control procedures that require an enumeration over all of the reachable states of $G_1 \| \cdots \| G_n$, such as those discussed in [26, 28], to take an unbearable amount of time if n is large.

However, for many types of symmetric systems a quotient automaton construction can be used to greatly decrease computation time when testing if some important properties hold at various system states [4, 27]. The underlying idea behind a

quotient automaton construction is that the system states, X^n for $G_1 \| \cdots \| G_n$, are partitioned into equivalence classes $\{X_1, \ldots, X_z\} \subseteq 2^{(X^n)}$ and each state equivalence class X_j is assigned a unique representative state x'_j from a set of representative states $\{x'_1, \ldots, x'_z\}$. The set of representative states $\{x'_1, \ldots, x'_z\}$ forms the state space of the quotient automaton and the transition structure is defined such that if there is a transition between two states $x_i \in X_i$ and $x_j \in X_j$, then there is a transition between the corresponding representative states x'_i and x'_j in the quotient automaton. Depending on how the states of $G_1 \| \cdots \| G_n$ are partitioned into $\{X_1, \ldots, X_z\}$ and how the labeling of transitions in the quotient automaton are assigned, it could be necessary and sufficient for various system properties to hold in the original system if the properties hold in the quotient automaton.

In general, defining the most efficient state partition for a quotient automaton to verify important system properties such as state reachability is computationally difficult and at least as difficult as the graph isomorphism problem [10]. However, for special systems, such as the similar module systems discussed in this paper, there has been research (such as in [7, 27]) into using predefined efficient state partitionings for the construction of quotient automata constructions that are used to test important system properties in the composed system $G_1 \| \cdots \| G_n$. This predefinition of the quotient structure avoids the difficult computations associated with finding the most efficient state partitioning, but still allows for efficient state partitionings that are generally useful in practice.

This section presents a predefined state partitioning of the system states in the composed similar module system $G_1 \| \cdots \| G_n$ for the construction of a nondeterministic quotient automaton $\tilde{G}$. This partitioning uses sets of states with the same component states as the state equivalence classes used in the quotient automaton construction. The $\tilde{G}$ automaton constructed in this manner can be used to test state reachability properties in the composed similar module system $G_1 \| \cdots \| G_n$ from which it was constructed. A simple tutorial example is shown that demonstrates how the states of the composed similar module system can be partitioned into sets of "equivalent" states.

Example 4. Reconsider the system first introduced in Example 1 with a third component G_3 similar to G_1 and G_2 in Figure 1. The automaton $G_1 \| G_2 \| G_3$ can be seen in Figure 4.

State $(1, 1, 2)$ in $G_1 \| G_2 \| G_3$ is equivalent to $(2, 1, 1)$, $(1, 2, 2)$, $(2, 1, 2)$, etc., with respect to the sets of module states these composed states contain. By inspection, these states are also equivalent with respect to some important system properties such as deadlock, blocking, and reachability. However, not all of these states are reachable from the initial state in the same number of transitions.

The intuition behind Example 4 is that state orderings and duplicated component states do not matter when testing several important global properties. A quotient automaton $\tilde{G}$ for testing these properties in similar module systems is now formally introduced.

4.1. The construction of $\tilde{G}$. Define $G^{\|} = (X^{\|}, x_0^{\|}, \Sigma, \delta^{\|}, X_m^{\|}) = G_1 \| \cdots \| G_n$. A state $x^{\|} \in X^{\|}$ from the composed automaton is an n-tuple of states in X. Let $\tilde{x}$ represent the set of module states that compose the n-tuple $x^{\|}$. For example, the set $\{1, 2, 6\}$ is the set of module states that compose the states $(1, 2, 6, 2)$ and $(1, 6, 6, 2)$ in $X^{\|}$. The function $Comp : X^{\|} \to 2^X$ is defined such that $Comp(x^{\|}) = \tilde{x}$ and the function $Comp^{-1}(\tilde{x})$ returns the set of states that have exactly the states in $\tilde{x}$ in their n-tuple. As an example of the operation of the $Comp(\cdot)$ and $Comp^{-1}(\cdot)$ functions

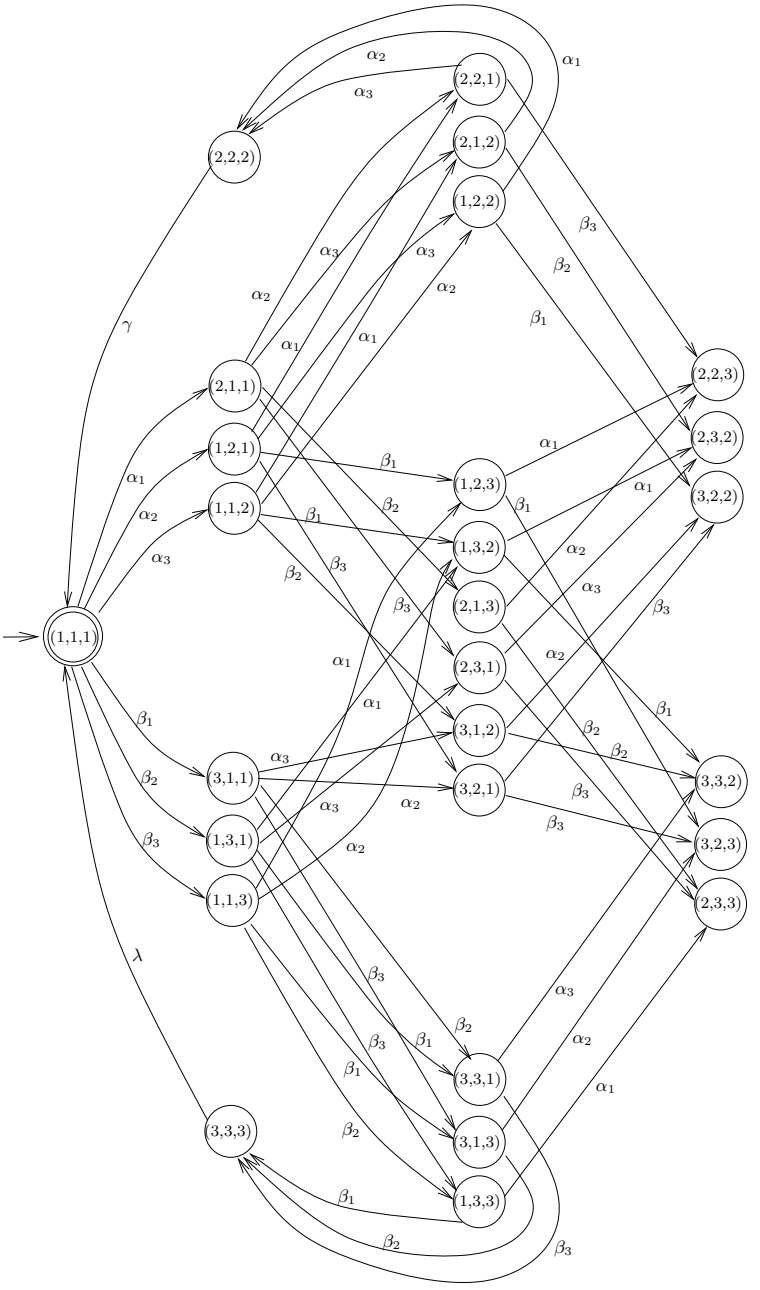


FIG. 4. *The automaton* $G_1 \| G_2 \| G_3$.

for the system in Example 4, $Comp((2,1,2)) = \{1,2\}$, $Comp((2,1,3)) = \{1,2,3\}$, $Comp^{-1}(\{1,2\}) = \{(1,2,2),(2,1,2),(2,2,1),(1,1,2),(1,2,1),(2,1,1)\}$. Note that [6, 7, 27] use state component permutation operators to define the state equivalence classes. It is discussed below that the state equivalence classes defined using the $Comp(\cdot)$ operator are generally much more efficient than those using permutation operators.

The set of state equivalence classes defined by the $Comp(\cdot)$ function is used as the state space of $\tilde{G}$. Let $\tilde{G} = (\tilde{X}, \{x_0\}, \Sigma_1, \tilde{\delta}, \tilde{X}_m)$ where $\tilde{X} = 2^X$, the power set of the component automaton G_i states, and let $\tilde{X}_m$, the set of marked states, be the states of $\tilde{G}$ such that all component states in the set are marked, i.e., $\tilde{X}_m = 2^{X_m}$. The possibly nondeterministic transition operator $\tilde{\delta}(\cdot,\cdot)$ is defined such that if there exists two states $x^{\|} \in Comp^{-1}(\tilde{x})$ and $y^{\|} \in Comp^{-1}(\tilde{y})$ and an event $\sigma_i \in \Sigma_i$ such that $\delta^{\|}(x^{\|}, \sigma_i) = y^{\|}$, then there is a corresponding transition in $\tilde{G}$ from $\tilde{x}$ to $\tilde{y}$ labeled as $\Psi_{i1}(\sigma_i)$. A formal construction algorithm for $\tilde{G}$ is now given.

ALGORITHM 1 (quotient construction algorithm).

Input:
$\{G_1, \ldots, G_n\}$ such that for all $i \in \{1, \ldots, n\}$,
$G_i = \left(X^i, x_0^i, \Sigma_i, \delta_i, X_m^i\right)$.

Output:
$\tilde{G} = (\tilde{X}, \tilde{x}_0, \Sigma, \tilde{\delta}, \tilde{X}_m)$.
(Note that $\tilde{\delta} \subseteq (\tilde{X} \times \Sigma \times \tilde{X})$).

Assumptions:
$\{G_1, \ldots, G_n\}$ is a similar module system with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$.
S is a stack with the normal *push* and *pop* operations.

Initialize:
$\tilde{X} := \{\{x_0\}\}$;
$\tilde{x}_0 := \{x_0\}$;
$\tilde{\delta} := \emptyset$;
$S := [\tilde{x}_0]$;
If $\tilde{x}_0 \subseteq X_m$
Then $\tilde{X}_m := \tilde{x}_0$
Else $\tilde{X}_m := \emptyset$;

Repeat:
{
$\tilde{x}_s := pop(S)$
Do the following for all $\sigma \in \Sigma$:
{
If $(\sigma \in \Sigma_g) \wedge$ (for all $x \in \tilde{x}, \delta_1(x, \sigma)!$)
Then
{
$\tilde{x}_c := \{\delta_1(x, \sigma) | x \in \tilde{x}\}$;
$\tilde{\delta} := \tilde{\delta} \cup (\tilde{x}_s, \sigma, \tilde{x}_c)$;
If $\tilde{x}_c \notin \tilde{X}$
Then
{
$\tilde{X} := \tilde{X} \cup \tilde{x}_c$;
$push(S, \tilde{x}_c)$;
If $\tilde{x}_c \subseteq X_m$
Then $\tilde{X}_m := \tilde{X}_m \cup \{\tilde{x}_c\}$
}
}
}
If $(\sigma \in \Sigma_{p1}) \wedge (|\tilde{x}| < n) \wedge (\delta_1(x, \sigma)!) \wedge (x \in \tilde{x})$
Then
{

```
                    x̃_c := x̃ ∪ {δ_1(x, σ)};
                    δ̃ := δ̃ ∪ (x̃_s, σ, x̃_c);
                    If x̃_c ∉ X̃
                        Then
                            {
                            X̃ := X̃ ∪ x̃_c;
                            push(S, x̃_c);
                           If x̃_c ⊆ X_m
                                Then X̃_m := X̃_m ∪ {x̃_c}
                            }
                }
            }
        }
        If (σ ∈ Σ_p1) ∧ (|x̃| > 1) ∧ (δ_1(x, σ)!) ∧ (x ∈ x̃)
            Then
                {
                x̃_c := [x̃ \ {x}] ∪ {δ_1(x, σ)};
                δ̃ := δ̃ ∪ (x̃_s, σ, x̃_c);
                If x̃_c ∉ X̃
                    Then
                        {
                        X̃ := X̃ ∪ x̃_c;
                        push(S, x̃_c);
                       If x̃_c ⊆ X_m
                            Then X̃_m := X̃_m ∪ {x̃_c}
                        }
                }
            }
        }
    }
Until S is empty;
Return G̃ = (X̃, x̃_0, Σ, δ̃, X̃_m).
```

Note that even if $G^{\|}$ is a deterministic automaton, the corresponding $\tilde{G}$ automaton may be nondeterministic. That is, at a state $\tilde{x} \in \tilde{X}$ there may be several outgoing transitions, each labeled by an event σ. In this case the state transition function $\tilde{\delta}$ is defined such that $\tilde{\delta} : \tilde{X} \times \Sigma_1 \to 2^{\tilde{X}}$, where for a state $\tilde{x} \in \tilde{X}$ and an event $\sigma \in \Sigma_1$, $\tilde{\delta}(\tilde{x}, \sigma) \subseteq \tilde{X}$ represents the set of states in $\tilde{G}$ that lead from $\tilde{x}$ by transitions labeled by σ. Therefore, the state transition function $\tilde{\delta}$ is defined such that

(2)
$$\tilde{\delta}(\tilde{x}, \sigma) = \begin{cases} \{\delta_1(x, \sigma) | x \in \tilde{x}\} & \text{if } (\sigma \in \Sigma_g) \wedge (\forall x \in \tilde{x}, \delta_1(x, \sigma)!) \\ \tilde{x} \cup \{\delta_1(x, \sigma)\} & \text{if } (\sigma \in \Sigma_{p1}) \wedge (|\tilde{x}| < n) \wedge (\delta_1(x, \sigma)!) \wedge (x \in \tilde{x}) \\ [\tilde{x} \setminus \{x\}] \cup \{\delta_1(x, \sigma)\} & \text{if } (\sigma \in \Sigma_{p1}) \wedge (|\tilde{x}| > 1) \wedge (\delta_1(x, \sigma)!) \wedge (x \in \tilde{x}) \end{cases}.$$

The state transition definition is also extended in the usual manner to allow for strings of transitions labeled by strings of events in Σ_1^* instead of single transitions labeled by a single event.

The intuition behind the definition of $\tilde{\delta}(\cdot, \cdot)$ is now shown in three parts for the three cases in which this transition operator is defined.

The first case of the $\tilde{\delta}(\cdot, \cdot)$ definition corresponds to the occurrence of a global

event $\sigma \in \Sigma_g$ such that all elements of $\tilde{x}$ are updated simultaneously on the occurrence of σ. All elements of $\tilde{x}$ need to be updated on the occurrence of σ because all states in $Comp^{-1}(\tilde{x})$ from $G^{\|}$ are updated on the occurrence of σ according to the transition rules of $\delta^{\|}(\cdot,\cdot)$. Therefore, $\delta_1(x,\sigma)$ is defined for all $x \in \tilde{x}$. This implies that $Comp\left(\delta^{\|}(x^{\|},\sigma)\right) \in \tilde{\delta}(\tilde{x},\sigma)$, and in this case $\{\delta_1(x,\sigma)|x \in \tilde{x}\} \in \tilde{\delta}(\tilde{x},\sigma)$ if $\sigma \in \Sigma_g$ and $\delta_1(x,\sigma)$ is defined for all $x \in \tilde{x}$.

The second case of the definition corresponds to the situation in which there exists some composed state $x^{\|} \in Comp^{-1}(\tilde{x})$ such that there exists $i,j \in \{1,\ldots,n\}$ such that $i \neq j$ and $x^{\|^i} = x^{\|^j}$. In other words, $|\tilde{x}| < n$ as required for this case in the definition of $\tilde{\delta}(\cdot,\cdot)$. Therefore, at $x^{\|}$, if an event $\sigma_{pi} \in \Sigma_{pi}$ were to occur such that $\delta_i(x^{\|^i},\sigma_{pi})!$, then the resulting state $\delta^{\|}(x^{\|},\sigma_{pi})$ would exist and

$$Comp(\delta(x^{\|},\sigma_{pi})) = Comp(x^{\|}) \cup \left\{\delta_i(x^{\|^i},\sigma_{pi})\right\}$$

as a replicated state in $x^{\|}$ would update on the occurrence of σ_{pi}. Therefore,

$$Comp(x^{\|}) \cup \left\{\delta_i(x^{\|^i},\sigma_{pi})\right\} \in \tilde{\delta}(x^{\|},\Psi_{i1}(\sigma_{pi})),$$

and for this case there is some $x \in \tilde{x}$, $\sigma \in \Sigma_{p1}$ such that $\tilde{x} \cup \{\delta_1(x,\sigma)\} \in \tilde{\delta}(\tilde{x},\sigma)$ as specified in the definition of $\tilde{\delta(\cdot,\cdot)}$.

The third case of the definition corresponds to the situation in which there exists some state $x^{\|} \in Comp^{-1}(\tilde{x})$ such that there exists $i \in \{1,\ldots,n\}$ and for any $j \in (\{1,\ldots,n\} \setminus \{i\})$, $x^{\|^i} \neq x^{\|^j}$. Thus, $|\tilde{x}| > 1$, as required for this case of the definition of $\tilde{\delta}(\cdot,\cdot)$. Therefore, at $x^{\|}$, if an event $\sigma_{pi} \in \Sigma_{pi}$ were to occur such that $\delta_i(x^{\|^i},\sigma_{pi})!$, then the resulting state $\delta(x^{\|},\sigma_{pi})$ would be in

$$\left(\left[Comp(x^{\|}) \setminus \left\{x^{\|^i}\right\}\right] \cup \left\{\delta_i(x^{\|^i},\sigma_{pi})\right\}\right) \in \tilde{\delta}(Comp(\tilde{x}),\Psi_{i1}(\sigma_{pi})),$$

and for this case there is some $x \in \tilde{x}$ and $\sigma_{pi} \in \Sigma_{pi}$ such that

$$[\tilde{x} \setminus x] \cup \{\delta_1(x,\sigma)\} \in \tilde{\delta}(\tilde{x},\sigma)$$

as specified in the definition of $\tilde{\delta(\cdot,\cdot)}$. For this case a module state that is not replicated in $x^{\|}$ is updated on the occurrence of a private event.

Also, notation is sometimes used in which for two states $\tilde{x},\tilde{y} \in \tilde{X}$ and an event $\sigma \in \Sigma_1$, $\tilde{x}\overset{\sigma}{\mapsto}_{\tilde{G}}\tilde{y}$ denotes that there is a transition in $\tilde{G}$ labeled by σ from $\tilde{x}$ to $\tilde{y}$. Similarly, for two states $x^{\|},y^{\|} \in X^{\|}$ and an event $\sigma \in \Sigma$, $x^{\|}\overset{\sigma}{\mapsto}_{G^{\|}}y^{\|}$ denotes that there is a transition in $G^{\|}$ labeled by σ from $x^{\|}$ to $y^{\|}$.

The similar module system introduced in Example 4 is now used to demonstrate the construction of a simple reduced state space composed automaton $\tilde{G}$.

Example 5. Consider the similar module system comprised of the automata G_1,G_2,G_3 discussed in Example 4 above. The set of state equivalence classes in the composed automaton $G_1\|G_2\|G_3$ with respect to the $Comp(\cdot)$ operation is

$$\{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}.$$

The quotient automaton $\tilde{G}$ constructed from G_1,G_2,G_3 can be seen in Figure 5.

The three different cases for transitions in the definition of $\tilde{\delta}(\cdot,\cdot)$ are all exhibited in this example of $\tilde{G}$.

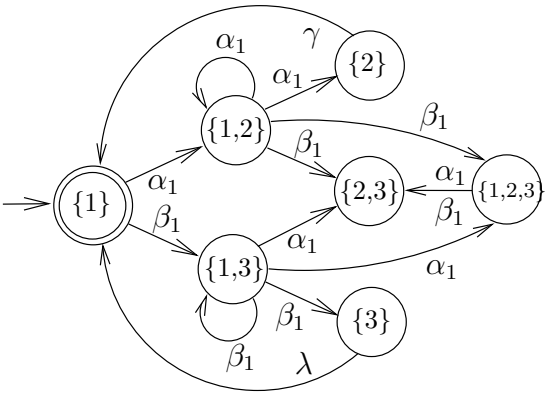


FIG. 5. *The automaton $\tilde{G}$ constructed from G_1, G_2, G_3.*

Consider the $\{2\} \overset{\gamma}{\mapsto}_{\tilde{G}} \{1\}$ transition in $\tilde{G}$. This transition is in the first case of the $\tilde{\delta}(\cdot,\cdot)$ definition and in the original $G_1 \| G_2 \| G_3$ automaton; it corresponds to the $(2,2,2) \overset{\gamma}{\mapsto}_{G_1\|G_2\|G_3} (1,1,1)$ transition. That is, a global event γ occurs and all states of $(2,2,2)$ are updated simultaneously.

Now consider the $\{1,2\} \overset{\alpha_1}{\mapsto}_{\tilde{G}} \{1,2\}$ transition in $\tilde{G}$. This transition corresponds to several transitions in the $G_1 \| G_2 \| G_3$ automaton, including the transition $(1,1,2) \overset{\alpha_2}{\mapsto}_{G_1\|G_2\|G_3} (1,2,2)$. For this transition, module state 1 is replicated and the second module state of $(1,1,2)$ is updated on the occurrence of α_2, but the first module state remains 1. The $\{1,2\} \overset{\alpha_1}{\mapsto}_{\tilde{G}} \{1,2\}$ transition is in the second case for transitions in the definition of $\tilde{\delta}(\cdot,\cdot)$.

The $\{1,2\} \overset{\alpha_1}{\mapsto}_{\tilde{G}} \{2\}$ transition corresponds to the third case in the definition of $\tilde{\delta}(\cdot,\cdot)$. In the original $G_1 \| G_2 \| G_3$ automaton, this transition corresponds to the $(2,2,1) \overset{\alpha_3}{\mapsto}_{G_1\|G_2\|G_3} (2,2,2)$ transition, among others. In this case there is only one module in state 1, and on the occurrence of α_3 this module updates to state 2 and the other modules remain unaltered.

Also notice that $\tilde{G}$ contains a transition $\{1\} \overset{\alpha_1}{\mapsto}_{\tilde{G}} \{1,2\}$. This corresponds to the second case of transitions in the definition of $\tilde{\delta}(\cdot,\cdot)$. It is true that $\{2,2,1\} \in Comp^{-1}(\{1,2\})$, but in $G_1 \| G_2 \| G_3$ there is no event σ such that there is a transition $(1,1,1) \overset{\sigma}{\mapsto}_{G_1\|G_2\|G_3} (2,2,1)$. Therefore, state reachability in $\tilde{G}$ does not exactly imply state reachability in $G_1 \| G_2 \| G_3$ for the same number of transitions. However, there is a string $\alpha_1\alpha_2$ such that $(1,1,1) \overset{\alpha_1\alpha_2}{\mapsto}_{G_1\|G_2\|G_3} (2,2,1)$. It is shown below that state reachability in $\tilde{G}$ is shown to imply a form of reachability in $G_1 \| \cdots \| G_n$ and vice versa.

4.2. Verifying system properties using $\tilde{G}$. Fundamental reachability results related to the $\tilde{G}$ and $G^{\|}$ constructions for similar module systems are now shown.

LEMMA 4. *Suppose a similar module system $\{G_1, \ldots, G_n\}$ is given that is used to construct $G^{\|}$ and $\tilde{G}$. For two states of $G^{\|}$, $x_1^{\|}, x_2^{\|} \in X^{\|}$ such that $Comp(x_1^{\|}) = \tilde{x}_1$ and $Comp(x_2^{\|}) = \tilde{x}_2$ and a string of transitions labeled by $t^{\|} \in \Sigma^*$ such that $x_1^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} = x_2^{\|}$, there exists a string of transitions labeled by $\tilde{t} \in \Sigma_1^*$ such that, according to the transition rules of $\tilde{G}$, $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$.*

Proof. This lemma is shown using a proof by generalized induction on the length of $t^{\|}$. For the base of induction, suppose $|t^{\|}| = 1$. The base of induction is shown in three cases according to the three cases in the definition of $\tilde{\delta}(\cdot,\cdot)$.

For the first case, suppose that $t^{\|} = \sigma_g \in \Sigma_g$. It is assumed that $\delta^{\|}(x_1^{\|}, \sigma_g) = x_2^{\|}$, and thus for all $i \in \{1,\ldots,n\}$, $\delta_i({x_1^{\|}}^i, \sigma_g) = {x_2^{\|}}^i$. Consequently, for all $x \in \tilde{x}_1$,

$$\exists i \in \{1,\ldots,n\} \text{ such that } \delta_1(x, \sigma_g) = {x_2^{\|}}^i \in \tilde{x}_2,$$

and for all $x \in \tilde{x}_2$,

$$\exists i \in \{1,\ldots,n\} \text{ such that } {x_1^{\|}}^i \in \tilde{x}_1 \text{ and } \delta_1({x_1^{\|}}^i, \sigma_g) = x.$$

Therefore, by the first case of the definition of $\tilde{\delta}(\cdot,\cdot)$, $\tilde{x}_1 \overset{\sigma_g}{\mapsto}_{\tilde{G}} \tilde{x}_2$.

For the second case of the base of induction, suppose that $t^{\|} = \sigma_{pi} \in \Sigma_{pi}$ and $\exists j \in \{1,\ldots,n\}$ such that $j \neq i$ and $x_i^{\|} = x_j^{\|}$. Therefore, when module G_i is in state ${x_1^{\|}}^i$, the module can update on the occurrence of σ_{pi} to state ${x_2^{\|}}^i$, and all other modules in $G^{\|}$ do not update; i.e., for all $k \neq i$, ${x_1^{\|}}^k = {x_2^{\|}}^k$. $\delta^{\|}(x_1^{\|}, \sigma_{pi}) = x_2^{\|}$, and thus $\delta_i({x_1^{\|}}^i, \sigma_{pi}) = {x_2^{\|}}^i$ and for $k \in \{1,\ldots,n\}$, $i \neq k$ ${x_1^{\|}}^k = {x_2^{\|}}^k$. Therefore, by the second case of the definition of $\tilde{\delta}(\cdot,\cdot)$, $\tilde{x}_1 \overset{\sigma_{pi}}{\mapsto}_{G^{\|}} \tilde{x}_2$.

For the third case of the base of induction, suppose that $t^{\|} = \sigma_{pi} \in \Sigma_{pi}$ and for all $j \in \{1,\ldots,n\}$ such that $j \neq i$ and $x_i^{\|} \neq x_j^{\|}$. Therefore, when module G_i is in state ${x_1^{\|}}^i$, the module can update on the occurrence of σ_{pi} to state ${x_2^{\|}}^i$, and all other modules in $G^{\|}$ do not update; i.e., for all $k \neq i$, ${x_1^{\|}}^k = {x_2^{\|}}^k$. $\delta^{\|}(x_1^{\|}, \sigma_{pi}) = x_2^{\|}$, and thus $\delta_i({x_1^{\|}}^i, \sigma_{pi}) = {x_2^{\|}}^i$ and for $k \in \{1,\ldots,n\}$, $i \neq k$ ${x_1^{\|}}^k = {x_2^{\|}}^k$. Therefore, by the second case of the definition of $\tilde{\delta}(\cdot,\cdot)$, $\tilde{x}_1 \overset{\sigma_{pi}}{\mapsto}_{G^{\|}} \tilde{x}_2$. This completes the base of induction.

For the induction hypothesis, suppose that there is a string of transitions labeled by $t^{\|} \in \Sigma^n$ such that $x_1^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x_2^{\|}$, $Comp(x_1^{\|}) = \tilde{x}_1$, and $Comp(x_2^{\|}) = \tilde{x}_2$. Then there exists a string of transitions labeled by $\tilde{t} \in \Sigma_1^*$ such that $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$.

For the induction step, suppose there is a string of transitions labeled by $t^{\|} \in \Sigma^n$ and a transition labeled by $\sigma^{\|} \in \Sigma$ such that $x_1^{\|} \overset{t^{\|}\sigma^{\|}}{\mapsto}_{G^{\|}} x_3^{\|}$, $Comp(x_1^{\|}) = \tilde{x}_1$, and $Comp(x_3^{\|}) = \tilde{x}_3$. Because $x_1^{\|} \overset{t^{\|}\sigma^{\|}}{\mapsto}_{G^{\|}} x_3^{\|}$, there must be a state $x_2^{\|}$ such that $x_1^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x_2^{\|}$ and $x_2^{\|} \overset{\sigma^{\|}}{\mapsto}_{G^{\|}} x_3^{\|}$, where $Comp(x_2^{\|}) = \tilde{x_2}$. Because of the induction hypothesis, there is a string of transitions labeled by $\tilde{t} \in \Sigma_1^*$ such that $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$. By the same argument as used in the proof of the base of induction above, there is a transition labeled by $\tilde{\sigma} \in \Sigma_1$ such that $\tilde{x}_2 \overset{\tilde{\sigma}}{\mapsto}_{\tilde{G}} \tilde{x}_3$. Therefore, there is a string of transitions labeled by $\tilde{t}\tilde{\sigma} \in \Sigma_1^*$ such that $\tilde{x}_1 \overset{\tilde{t}\tilde{\sigma}}{\mapsto}_{\tilde{G}} \tilde{x}_3$. ∎

LEMMA 5. *Suppose a similar module system $\{G_1,\ldots,G_n\}$ is given that is used to construct $G^{\|}$ and $\tilde{G}$ with two states of the quotient automaton $\tilde{x}_1, \tilde{x}_2 \in 2^X$, a state $x_1^{\|} \in X^{\|}$ such that $Comp(x_1^{\|}) = \tilde{x}_1$, and a string of transitions labeled by $\tilde{t} \in \Sigma_1^*$ such that $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$. Then there exists a state $x_2^{\|} \in X^{\|}$ and a string of transitions labeled by $t^{\|} \in \Sigma^*$ such that $Comp(x_2^{\|}) = \tilde{x}_2$ and $x_1^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x_2^{\|}$.*

Proof. The $\tilde{G}$ automaton can be thought of as a partial simulation of the behavior of the corresponding $G^{\|}$ automaton. For $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$, suppose that $\tilde{t}' < \tilde{t}$ and $\tilde{x}_1 \overset{\tilde{t}'}{\mapsto}_{\tilde{G}} \tilde{x}$.

If $\sigma_g \in \Sigma_g$ and $\tilde{t}'\sigma_g < \tilde{t}$, then for $x^{\|} \in Comp^{-1}(\tilde{x})$, the transition at $\tilde{x}$ labeled by σ_g corresponds to a transition in $G^{\|}$, where all component states of $x^{\|}$ update simultaneously. If $\sigma_p \in \Sigma_{p1}$ and $\tilde{t}'\sigma_p < \tilde{t}$, then for $x^{\|} \in Comp^{-1}(\tilde{x})$, the transition at $\tilde{x}$ labeled by σ_p corresponds in $G^{\|}$ to exactly one component state ${x^{\|}}^i$ of $x^{\|}$ for some $i \in \{1, \ldots, n\}$ that updates on the occurrence of an event $\sigma_{pi} = \Psi_{1i}(\sigma_p)$.

Therefore, if $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$, then for every component state $x_1 \in \tilde{x}_1$, there must be a string of transitions labeled by events in Σ_1^* that leads to a component state $x_2 \in \tilde{x}_2$ according to the transition rules of G_1, and for every component state $x_2 \in \tilde{x}_2$, there must be a string of transitions labeled by events in Σ_1^* from a component state in $x_1 \in \tilde{x}_1$ to the component state x_2 according to the transition rules of G_1. All of these strings are able to occur synchronously with respect to the occurrence of global events and are copies of the string $\tilde{t}$ with some events from Σ_{p1} removed.

With this reasoning the string of events $\tilde{t} \in \Sigma_1^*$ such that $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$ can be split into a set of strings $T = \{t_{x_1^1 x_2^1}, \ldots, t_{x_1^m x_2^m}\} \subseteq \Sigma_1^*$ such that

- for all $i \in \{1, \ldots, m\}$, $t_{x_1^i x_2^i}$ is a copy of $\tilde{t}$ with some Σ_{p1} events removed.
- $P_g(t_{x_1^1 x_2^1}) = \cdots = P_g(t_{x_1^m x_2^m})$.
- for every component state $x_1^i \in \tilde{x}_1$, there is at least one component state $x_2^i \in \tilde{x}_2$ and a string of events $t_{x_1^i x_2^i} \in T$ such that $\delta^1(x_1^i, t_{x_1^i x_2^i}) = x_2^i$.
- for every component state $x_2^j \in \tilde{x}_2$, there is at least one component state $x_1^j \in \tilde{x}_1$ and a string of events $t_{x_1^j x_2^j} \in T$ such that $\delta^1(x_1^i, t_{x_1^i x_2^i}) = x_2^i$.
- $m = \max\{|\tilde{x}_1|, |\tilde{x}_2|\} \leq n$.

Consider the ordered list of n strings,

$$L_t = [t_{x_1^1 x_2^1}, \ldots, t_{x_1^m x_2^m}, t_{x_1^m x_2^m}, \ldots, t_{x_1^m x_2^m}].$$

Corresponding to L_t, there are two ordered lists of pairs of states in X,

$$L_x^1 = [x_1^1, \ldots, x_1^m, x_1^m, \ldots, x_1^m],$$
$$L_x^2 = [x_2^1, \ldots, x_2^m, x_2^m, \ldots, x_2^m]$$

such that if

$$x_{1\Psi}^{\|} = (x_1^1, \ldots, x_1^m, x_1^m, \ldots, x_1^m),$$
$$x_{2\Psi}^{\|} = (x_2^1, \ldots, x_2^m, x_2^m, \ldots, x_2^m),$$

then

$$Comp(x_{1\Psi}^{\|}) = \tilde{x}_1,$$
$$Comp(x_{2\Psi}^{\|}) = \tilde{x}_2.$$

Moreover, according to the transition rules of G_1, $\delta^1(x_{k,1}, t_k) = x_{k,2}$ where $L_t(k) = t_k$ is the kth string in L_t and $L_x^1(k) = x_{k,1}$ and $L_x^2(k) = x_{k,2}$ are the kth states in L_x^1 and L_x^2, respectively. Suppose that $x_{1\Psi}^{\|} = x_1^{\|}$. It is discussed below how this assumption can be made without loss of generality.

The list of strings L_t are now converted using the $\Psi_{1i}(\cdot)$ operator to the list

$$L_t^{\Psi} = [\Psi_{11}(t_{x_1 x_2}^1), \ldots, \Psi_{1m}(t_{x_1 x_2}^m), \Psi_{1(m+1)}(t_{x_1 x_2}^m), \ldots, \Psi_{1n}(t_{x_1 x_2}^m)].$$

Because all of the strings in this converted list can still be synchronized on the occurrence of events in Σ_g, the intersection

$$\begin{aligned}\big(&P_1^{-1}(\Psi_{11}(t^1_{x_1x_2}))\cap\cdots\cap P_m^{-1}(\Psi_{1m}(t^m_{x_1x_2}))\\&\cap P_{m+1}^{-1}(\Psi_{1(m+1)}(t^m_{x_1x_2}))\cap\cdots\cap P_n^{-1}(\Psi_{1n}(t^m_{x_1x_2}))\big)\end{aligned}$$

is nonempty. Let $t^{\|}_{\Psi}$ be some string in this intersection. From the construction above for L^1_x and L^2_x, if $x^{\|}_{1\Psi}$ is the n-tuple corresponding to the list of states in L^1_x, and $x^{\|}_{2\Psi}$ is the n-tuple corresponding to the list of states in L^2_x, then $\delta^{\|}(x^{\|}_{1\Psi}, t^{\|}_{\Psi}) = x^{\|}_{2\Psi}$. Let $x^{\|}_2$ be a copy of $x^{\|}_{2\Psi}$. Therefore, there exists a state $x^{\|}_2 \in X^{\|}$ and a string of transitions labeled by $t^{\|} \in \Sigma^*$ such that $Comp(x^{\|}_2) = \tilde{x}_2$ and $x^{\|}_1 \overset{t^{\|}}{\mapsto}_{G^{\|}} x^{\|}_2$.

Notice that the lists L_t, L^1_x, and L^2_x could be constructed in an alternative manner by replicating the appropriate elements of these lists in the proper locations to force $x^{\|}_{1\Psi}$ to be a copy of any arbitrary $x^{\|}_1$. □

An example is now given to aid in the visualization of the constructions used in the proof of Lemma 5.

Example 6. It is now shown how to construct lists L_t, L^1_x, L^2_x and the states $x^{\|}_{1\Psi}$, $x^{\|}_{1\Psi}$ in Lemma 5 from given $\tilde{x}_1$, $\tilde{x}_2$, $x^{\|}_1$, and $\tilde{t}$ using the $\tilde{G}$ automaton in Example 4. Suppose $\tilde{x}_1 = \{1,2\}$, $\tilde{x}_2 = \{2,3\}$, $\tilde{t} = \alpha_1\gamma\beta_1\beta_1\alpha_1$, and $x^{\|}_1 = (1,1,2)$.

From $x^{\|}_1 = (1,1,2)$, L^1_x is set to be $[1,1,2]$. Let $L_t = [\alpha_1\gamma\alpha_1, \alpha_1\gamma\beta_1, \gamma\beta_1]$ and let $L^2_x = [2,3,3]$. Therefore, $x^{\|}_{1\Psi} = (1,1,2)$ and $x^{\|}_{1\Psi} = (2,3,3)$. The proper L_t and L^2_x can be found mechanically by a search over all possible candidate strings, which is guaranteed to be finite.

Note that

$$\begin{aligned}\delta_1(1,\alpha_1\gamma\alpha_1) &= 2,\\ \delta_1(1,\alpha_1\gamma\beta_1) &= 3,\\ \delta_1(2,\gamma\beta_1) &= 3.\end{aligned}$$

Also note that $L^{\Psi}_t = [\alpha_1\gamma\alpha_1, \alpha_2\gamma\beta_2, \gamma\beta_3]$ and that it is possible for the string $t^{\|}$ corresponding to $\tilde{t}$ to be $\alpha_1\alpha_2\gamma\alpha_1\beta_2\beta_3$. Therefore, $\delta^{\|}((1,1,2), \alpha_1\alpha_2\gamma\alpha_1\beta_2\beta_3) = (2,3,3)$.

The dual of Lemma 5 is now shown, where $x^{\|}_2$ is specified instead of $x^{\|}_1$.

LEMMA 6. *Suppose a similar module system $\{G_1,\ldots,G_n\}$ is given that is used to construct $G^{\|}$ and $\tilde{G}$ with two states of the quotient automaton $\tilde{x}_1, \tilde{x}_2 \in 2^X$, a state $x^{\|}_2 \in X^{\|}$ such that $Comp(x^{\|}_2) = \tilde{x}_2$, and a string of transitions labeled by $\tilde{t} \in \Sigma^*$ such that $\tilde{x}_1 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_2$. Then there exists a state $x^{\|}_1 \in X^{\|}$ and string of transitions labeled by $t^{\|} \in \Sigma^*_1$ such that $Comp(x^{\|}_1) = \tilde{x}_1$ and $x^{\|}_1 \overset{t^{\|}}{\mapsto}_{G^{\|}} x^{\|}_2$.*

Proof. This proof follows from the same construction as in Lemma 5 except that the lists L_t, L^1_x, and L^2_x could be constructed to correspond to any state $x^{\|}_2$ such that $x^{\|}_2 \in Comp^{-1}(\tilde{x}_2)$ and $x^{\|}_2$ is the n-tuple corresponding to the list of states in L^2_x. □

THEOREM 3. *Suppose a similar module system $\{G_1,\ldots,G_n\}$ is given that is used to construct $G^{\|}$ and $\tilde{G}$. A state $x^{\|} \in X^{\|}$ deadlocks according to the transition rules of $G^{\|}$ if and only if the state $Comp(x^{\|}) = \tilde{x}$ deadlocks according to the transition rules of $\tilde{G}$.*

Proof. The proof of this theorem is demonstrated in two parts. First, it is shown

that if $x^{\|} \in X^{\|}$ deadlocks according to the transition rules of $G^{\|}$, then the state $Comp(x^{\|}) = \tilde{x}$ deadlocks according to the transition rules of $\tilde{G}$.

If $x^{\|} \in X^{\|}$ deadlocks according to the transition rules of $G^{\|}$, then there is no global event $\sigma_g \in \Sigma_g$ that could synchronize a state transition in all component states of $x^{\|}$. Due to the first case of the state transition function $\tilde{\delta}(\cdot,\cdot)$, there is therefore no global event $\sigma_g \in \Sigma_g$ that could synchronize a state transition in all component states of $\tilde{x}$. Also, there are no private events $\sigma_{pi} \in \Sigma_{pi}$ that could occur at any of the component states of $x^{\|}$. There can therefore be no events $\sigma_{p1} \in \Sigma_{p1}$ that could occur at any of the states of $\tilde{x}$ according to the transition rules of G_1. Consequently, there are no transitions driven by an event $\sigma_{p1} \in \Sigma_{p1}$ that could occur at the state $\tilde{x}$ according to the second and third cases of the transition rules of $\tilde{G}$. Therefore, the state $Comp(x^{\|}) = \tilde{x}$ deadlocks according to the transition rules of $\tilde{G}$.

Now it is shown that if $x^{\|} \in X^{\|}$ does not deadlock according to the transition rules of $G^{\|}$, then the state $Comp(x^{\|}) = \tilde{x}$ does not deadlock according to the transition rules of $\tilde{G}$. If $x^{\|} \in X^{\|}$ does not deadlock according to the transition rules of $G^{\|}$, then there is a state $y^{\|} \in X^{\|}$ and an event $\sigma \in \Sigma$ for a transition in $G^{\|}$ such that $x^{\|} \overset{\sigma}{\mapsto}_{G^{\|}} = y^{\|}$. Therefore, by Lemma 4, there exists a string of transitions labeled by $\tilde{t} \in \Sigma_1^*$ such that according to the transition rules of $\tilde{G}$, $Comp(x^{\|}) \overset{\tilde{t}}{\mapsto}_{\tilde{G}} Comp(y^{\|})$. Therefore, the state $Comp(x^{\|}) = \tilde{x}$ does not deadlock according to the transition rules of $\tilde{G}$. □

THEOREM 4. *Suppose a similar module system $\{G_1, \ldots, G_n\}$ is given that is used to construct $G^{\|}$ and $\tilde{G}$. A state $x^{\|} \in X^{\|}$ is reachable according to the transition rules of $G^{\|}$ if and only if the state $Comp(x^{\|}) = \tilde{x}$ is reachable according to the transition rules of $\tilde{G}$.*

Proof. This theorem is demonstrated in two parts. For the first part of this proof, suppose that a state $x^{\|} \in X^{\|}$ is reachable according to the transition rules of $G^{\|}$. Therefore, there is a string of transitions labeled by $t^{\|}$ such that $x_0^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x^{\|}$. Because of Lemma 4, there is a string of transitions labeled by $\tilde{t}$ such that $\tilde{x}_0 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}$, where $Comp(x^{\|}) = \tilde{x}$.

For the second part of this proof, suppose that a state $\tilde{x} \in \tilde{X}$ is reachable according to the transition rules of $\tilde{G}$. Therefore, there is a string of transitions labeled by $\tilde{t}$ such that $\tilde{x}_0 \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}$. Using Lemma 6 there is a string of transitions labeled by $t^{\|}$ such that for some $x_0^{\|'} \in Comp^{-1}(\tilde{x}_0)$, $x_0^{\|'} \overset{t^{\|}}{\mapsto}_{G^{\|}} x^{\|}$. Because $Comp^{-1}(\tilde{x}_0) = \{x_0^{\|}\}$, there is a string of transitions labeled by $t^{\|}$ such that $x_0^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x^{\|}$, where $Comp(x^{\|}) = \tilde{x}$. □

THEOREM 5. *Suppose a similar module system $\{G_1, \ldots, G_n\}$ is given that is used to construct $G^{\|}$ and $\tilde{G}$. A state $x^{\|} \in X^{\|}$ is blocking according to the transition rules of $G^{\|}$ if and only if the state $Comp(x^{\|}) = \tilde{x}$ is blocking according to the transition rules of $\tilde{G}$.*

Proof. Suppose that the state $x^{\|} \in X^{\|}$ is nonblocking according to the transition rules of $G^{\|}$. Therefore, there is a state $x_m^{\|} \in X_m^{\|}$ and a string of transitions labeled by $t^{\|}$ such that $x^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x_m^{\|}$. Because of Lemma 4, there is a string of transitions labeled by $\tilde{t}$ such that $\tilde{x} \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_m$, where $Comp(x_m^{\|}) = \tilde{x}_m$. Because $Comp(x_m^{\|}) = \tilde{x}_m$, $\tilde{x}_m$ must be marked.

Suppose that the state $\tilde{x} \in \tilde{X}$ is nonblocking according to the transition rules of $\tilde{G}$. Therefore, there is a state $\tilde{x}_m \in \tilde{X}_m$ and a string of transitions labeled by $\tilde{t}$ such that $\tilde{x} \overset{\tilde{t}}{\mapsto}_{\tilde{G}} \tilde{x}_m$. By Lemma 5 there is a string of transitions labeled by $t^{\|}$ and a state

$x_2^{\|}$ such that $x^{\|} \overset{t^{\|}}{\mapsto}_{G^{\|}} x_2^{\|}$ and $Comp(x_2^{\|}) = \tilde{x}_m$. Because $Comp(x_2^{\|}) = \tilde{x}_m$, and $\tilde{x}_m$ is marked, then $x_2^{\|}$ must also be marked. Therefore, $x^{\|}$ is nonblocking. □

4.3. The size of the state space of $\tilde{G}$. It should be apparent that the $\tilde{G}$ constructed from the similar module system $\{G_1, \ldots, G_n\}$ has a smaller state space than the $G^{\|}$ automaton constructed from the same set of modules. The worst-case size of the $\tilde{G}$ automaton state space is now quantified. Suppose that $k = |X|$, the size of the state space of the individual modules. Consider the following example that shows how the $\tilde{G}$ construction can have a bounded number of states even if the number of modules is unbounded.

Example 7. Consider the similar module system G_1, G_2 as in Example 1 and the corresponding $\tilde{G}_2$ constructed from these automata, which can be seen in Figure 6. The automata G_1, G_2 use $X = \{1, 2, 3\}$ as their state space and $\tilde{G}_2$ uses $\{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$ as its state space.

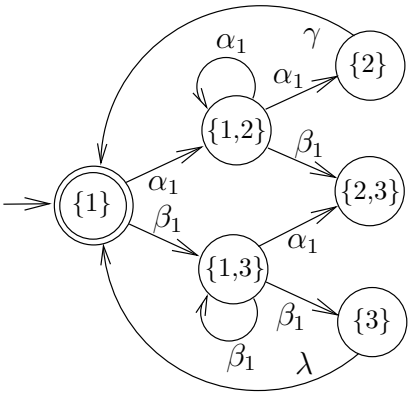


FIG. 6. *The automaton $\tilde{G}_2$ constructed from G_1, G_2.*

Now consider the set of similar modules, except with three components G_1, G_2, G_3 and a corresponding $\tilde{G}$ automaton, as discussed in Example 5. The automaton $\tilde{G}$ can be seen in Figure 5. The automata G_1, G_2, G_3 use $X = \{1, 2, 3\}$ as their state space and $\tilde{G}$ for this set of modules uses $2^X = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$ as its state space. Note that the largest cardinality subset of X is $\{1, 2, 3\}$.

Again consider the same set of similar modules, except with four components G_1, G_2, G_3, G_4 and the corresponding quotient automaton $\tilde{G}_4$, which can be seen in Figure 7. The automata G_1, G_2, G_3, G_4 use $X = \{1, 2, 3\}$ as their state space and $\tilde{G}_4$ uses

$$2^X = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

as its state space. The only difference between the $\tilde{G}_4$ automaton and the $\tilde{G}$ automaton for the three module system seen in Figure 5 is that there are added self-loop transitions at state $\{1, 2, 3\}$ labeled by α_1 and β_1 events.

Now consider the same set of n similar modules $\{G_1, \ldots, G_n\}$ with $n \geq 4$. The automaton $\tilde{G}_n$ constructed from this set of modules still uses

$$2^X = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

as its state space and is structurally equivalent to the $\tilde{G}_4$ automaton. That is, all defined transitions in $\tilde{G}_n$ are also defined in $\tilde{G}_n$. Therefore, the $\tilde{G}_n$ construction is

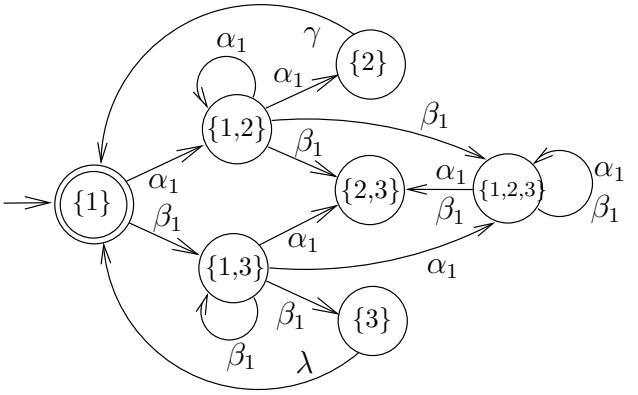


FIG. 7. *The automaton $\tilde{G}_4$ constructed from G_1, G_2, G_3, G_4.*

equivalent to the $\tilde{G}_4$ construction no matter how many modules are used as long as $n \geq 4$. Furthermore, the size of the state space of $\tilde{G}_n$ does not change for $n \geq 3$. This is due to the fact that $|X| = 3$, and for $n \geq 3$, $\tilde{X} = 2^X$.

Given an n-module system, each state of $\tilde{G}$ can have at most n component states, and in each state of $\tilde{G}$ no component states are repeated. Therefore, the maximum number of component states in $\tilde{G}$ is $\max(n, k)$. Consequently, the worst-case size of $\tilde{X}$ needs to be analyzed in two cases, $n \geq k$ and $n < k$.

Case 1: $n < k$. Each of the states of $\tilde{G}$ has at most n component states. There are possibly $\binom{k}{i}$ states with i component states. Therefore, if the sum of the possible number of state sets for each number of components is taken from i to n, then the number of possible state sets for $\tilde{G}$ for $n < k$ is

$$\sum_{i=1}^{n} \binom{k}{i}. \tag{3}$$

Note that for $n < k$, the number of possible state sets for $\tilde{G}$ is bounded by $2^k - 1$.

Case 2: $n \geq k$. For this case, any state set in the power set 2^X is a possible state of $\tilde{G}$, except for the empty set $\emptyset$. Therefore, the size of the state space of $\tilde{G}$ is at most

$$|2^X \setminus \{\emptyset\}| = 2^k - 1. \tag{4}$$

This means that if $n \geq k$, then the size of $\tilde{G}$ is independent of the number of modules, and testing state reachability, blocking, and deadlock-freeness can be done without regard to the number of components.

Therefore, if $\tilde{X}_{\tilde{G}}$ is the set of reachable states in $\tilde{G}$, then

$$|\tilde{X}_{\tilde{G}}| \leq \left\{ \begin{array}{ll} \sum_{i=1}^{n} \binom{k}{i} & \text{if } n < k \\ 2^k - 1 & \text{if } n \geq k \end{array} \right\}. \tag{5}$$

Furthermore, no matter how many modules comprise a similar module system, $|\tilde{X}_{\tilde{G}}| \leq 2^k - 1$. This is much less than the worst-case number of reachable states in $G^{\|}$, k^n. Therefore, the verification of such important system properties as state reachability and blocking can be performed with a large reduction in the number of composed states that need to be searched. Furthermore, using the $\tilde{G}$ construction,

the difficulty of verifying these properties is independent of the number of modules if the number of modules is sufficiently large. This construction shows more of the computational advantages gained when investigating modular systems with symmetry between components. Other quotient automaton constructions shown in [6, 7, 27], which use permutation operators to define the state equivalence classes, have a worst-case size of $\left(\frac{(k+n-1)!}{(k-1)!n!}\right)$ for testing state reachability properties, and the size of this automaton grows as the number of modules grows. Note that with some mathematical manipulation it can be shown that

$$\left\{\begin{array}{ll} \sum_{i=1}^{n}\binom{k}{i} & \text{if } n < k \\ 2^k - 1 & \text{if } n \geq k \end{array}\right\} \leq \left(\frac{(k+n-1)!}{(k-1)!n!}\right). \tag{6}$$

Therefore, $|\tilde{X}_{\tilde{G}}| \leq \left(\frac{(k+n-1)!}{(k-1)!n!}\right)$.

5. Verification for local specifications. Now topics related to the verification of similar module systems with respect to local specifications are discussed. First, the class of behavior language specifications used is defined.

DEFINITION 5. *A set of languages $\{K_1, \ldots, K_n\}$ is called a set of similar languages if for all $i \in \{1, \ldots, n\}$, $K_i \subseteq \Sigma_i^*$ and $\Psi_{1i}(K_1) = K_i$.*

Suppose it needs to be checked if for all i that $P_i\left(\mathcal{L}_m(G_1 \| \cdots \| G_n)\right) = K_i$. This property can be checked solely by verifying that $\mathcal{L}_m(G_i) = K_i$ for similar module systems. The following corollary is a direct extension of Theorem 1.

COROLLARY 1. *Given a local language specification K_i and a similar module system $\{G_1, \ldots, G_n\}$, $P_i\left(\mathcal{L}_m(G_1 \| \cdots \| G_n)\right) = K_i$ if and only if $\mathcal{L}_m(G_i) = K_i$. Likewise, $P_i\left(\mathcal{L}(G_1 \| \cdots \| G_n)\right) = \overline{K_i}$ if and only if $\mathcal{L}(G_i) = \overline{K_i}$.*

The problems of verifying whether $\mathcal{L}_m(G_i) = K_i$ and $\mathcal{L}(G_i) = \overline{K_i}$ are both known to be computationally simple if K_i is specified by a deterministic automaton and G_i is likewise deterministic. This greatly simplifies previously known verification methods based on more general modular systems because local behavior in a composed similar module system can be tested by investigating a single module.

6. Decomposition properties. This section presents one of the main results of this paper: a polynomial time decomposition method for composed similar module systems that runs in polynomial time. Sufficient conditions are first shown such that if a language K is separable, then $\overline{K}$ is also separable.

LEMMA 7. *Given a language K and a set of projections $\{P_1, \ldots, P_n\}$, if K is separable with respect to $\{P_1, \ldots, P_n\}$ and the set of languages*

$$\{P_1^{-1}(P_1(K)), \ldots, P_n^{-1}(P_n(K))\}$$

is nonconflicting, then $\overline{K}$ is separable with respect to $\{P_1, \ldots, P_n\}$.

Proof. Using Lemma 1, it is known that

$$\forall i \in \{1, \ldots, n\}, P_i^{-1}(P_i(\overline{K})) = \overline{P_i^{-1}(P_i(K))}.$$

Consequently,

$$\cap_{i=1}^{n} \overline{P_i^{-1}(P_i(K))} = \cap_{i=1}^{n} P_i^{-1}(P_i(\overline{K})).$$

Due to the assumption that the languages $\{P_1^{-1}(P_1(K)), \ldots, P_n^{-1}(P_n(K))\}$ are nonconflicting,

$$\overline{\cap_{i=1}^{n} P_i^{-1}(P_i(K))} = \cap_{i=1}^{n} P_i^{-1}(P_i(\overline{K})).$$

Therefore, due to the separability of K,

$$K = \cap_{i=1}^{n} P_i^{-1}(P_i(K))$$

and

$$\overline{K} = \cap_{i=1}^{n} P_i^{-1}(P_i(\overline{K})). \qquad \square$$

It is now demonstrated how symmetric and separable languages can be modeled using the similar module system framework.

THEOREM 6. *Suppose that a set of global events Σ_g is given with private events $\{\Sigma_{p1}, \ldots, \Sigma_{pn}\}$ that can be mapped to one another using $\Psi_{ij}(\cdot)$ operations. If an automaton $H = (X, x_0, \Sigma, \delta, X_m)$ is given that is trim and the language marked by the automaton $\mathcal{L}_m(H)$ is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$ and separable with respect to $\{P_1, \ldots, P_n\}$, then there exists a similar module system $\{H_1, \ldots, H_n\}$ such that*

$$\mathcal{L}_m(H_1 \| \cdots \| H_n) = \mathcal{L}_m(H)$$

and

$$\forall i \in \{1, \ldots, n\},\ P_i(\mathcal{L}_m(H)) = \mathcal{L}_m(H_i).$$

Proof. Using standard methods, given Σ_i and H, an automaton H_i can be constructed such that $\mathcal{L}_m(H_i) = P_i(\mathcal{L}_m(H))$ for all i such that $i \in \{1, \ldots, n\}$. Because $\mathcal{L}_m(H)$ is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$,

$\forall i, j \in \{1, \ldots, n\}\ \Psi_{ij}(\mathcal{L}_m(H)) = \mathcal{L}_m(H)$,

$\Rightarrow \Psi_{ij}(P_i(\mathcal{L}_m(H))) = P_j(\mathcal{L}_m(H))$.

$\Rightarrow \Psi_{ij}(\mathcal{L}_m(H_i)) = \mathcal{L}_m(H_j)$.

Therefore, the automata $\{H_1, \ldots, H_n\}$ can be constructed such that they are copies of one another with respect to a renaming of their private events.

Because $\mathcal{L}_m(H)$ is separable with respect to $\{P_1, \ldots, P_n\}$,

$P_1^{-1}(P_1(\mathcal{L}_m(H))) \cap \cdots \cap P_n^{-1}(P_n(\mathcal{L}_m(H))) = \mathcal{L}_m(H)$

$\Rightarrow P_1^{-1}(\mathcal{L}_m(H_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(H_n)) = \mathcal{L}_m(H)$

$\Rightarrow \mathcal{L}_m(H_1 \| \cdots \| H_n) = \mathcal{L}_m(H)$ due to properties of the parallel composition operation.

$\Rightarrow \forall i \in \{1, \ldots, n\} P_i(\mathcal{L}_m(H_1 \| \cdots \| H_n)) = P_i(\mathcal{L}_m(H))$

$\Rightarrow \forall i \in \{1, \ldots, n\} \mathcal{L}_m(H_i) = P_i(\mathcal{L}_m(H))$.

This last step is by the result of Theorem 1 shown above. $\square$

In the proof of Theorem 6 any method can be used to construct the similar module system $\{H_1, \ldots, H_n\}$ from H such that $\mathcal{L}_m(H_1 \| \cdots \| H_n) = \mathcal{L}_m(H)$. Note that there might not even be a unique set of automata $\{H_1, \ldots, H_n\}$ such that $\mathcal{L}_m(H_1 \| \cdots \| H_n) = \mathcal{L}_m(H)$. However, the standard method of computing a set of automata $\{H_1, \ldots, H_n\}$ such that $\mathcal{L}_m(H_1 \| \cdots \| H_n) = \mathcal{L}_m(H)$ consists of converting all transitions labeled by events in $\Sigma \setminus \Sigma_1$ in H to nondeterministic ϵ-transitions. Then, this nondeterministic automaton is converted into a deterministic automaton H_1. Unfortunately, the determinization algorithm takes time and space exponential in the size of H in the worst case. The other automata $\{H_2, \ldots, H_n\}$ are copies of H_1 with respect to $\Psi_{1i}(\cdot)$ event translations.

There has been little discussion in the discrete-event systems literature on ways of more efficiently performing modular decompositions besides ad hoc methods developed on a case-by-case basis. Developing formal methods for performing this operation

is very important for many real-world problems in which a large complicated system model would need to be converted into simpler, modular model blocks. This section presents a computationally efficient algorithm for performing this decomposition on similar module systems. As an example of a situation in which such a decomposition algorithm would be useful, it can be at times difficult to model the behavior of a biological cell without observing its interactions with other biological cells. Notice that at the proper level of abstraction, a biological cellular network can be thought of as a similar module system. Therefore, if a discrete-event system model of a biological cellular network is developed from observations of the behaviors of the cellular network, then the decomposition algorithm would provide an efficient method to develop a model for a cell that comprises the cellular network.

Before the decomposition method is presented, an intuitive introduction to the algorithm's operation is now given. Let $\{H_1, \ldots, H_n\}$ be the desired modules that compose the given symmetric and separable H, i.e., $\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \cdots \| H_n)$. Suppose distributed simulations of the composed behavior of $\{H_1, \ldots, H_n\}$ are run such that the only way for private events to occur would be in strings such as $\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{pn}) = \sigma_{p1}\sigma_{p2} \cdots \sigma_{pn}$ with no interleavings of other events. On the occurrence of these $\sigma_{p1}\sigma_{p2} \cdots \sigma_{pn}$ strings the local states $\{H_1, \ldots, H_n\}$ are forced to update with identical state transitions. Because $\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \ldots \| H_n)$, this lockstep simulation of private behaviors in the global system model can be used to calculate the local system behaviors. Given H, this automaton can be trimmed in a special manner by allowing only global events and strings of private events $\sigma_{p1}\sigma_{p2} \cdots \sigma_{pn}$, as outlined above, to occur. The behavior of this specially trimmed automaton matches the behavior of the lockstep simulation of the $\{H_1, \ldots, H_n\}$ automata. Then, if the event σ_{p1} is substituted for every occurrence of a string $\sigma_{p1}\sigma_{p2} \cdots \sigma_{pn}$, the behavior of the resulting automaton will match the behavior of the H_1 module.

In the following algorithm, a slightly modified set notation for the state transition function is used; i.e., if $\delta(x, s) = y$ for state transition function $\delta(\cdot, \cdot)$, states x, y, and string s, then the notation that $(x, s, y) \in \delta$ is used. For simplicity it is assumed that the inputted automaton H is deterministic and trim.

ALGORITHM 2 (decomposition construction algorithm).

Input:
$H = \left(X^H, x_0^H, \Sigma, \delta^H, X_m^H\right)$
$\Sigma_1, \ldots, \Sigma_n$
(Note that $\delta^H \subseteq (X^H \times \Sigma \times X^H)$).

Output:
$\{H_1, \ldots, H_n\}$ such that for all $i \in \{1, \ldots, n\}$,
$H_i = \left(X^i, x_0^i, \Sigma_i, \delta_i, X_m^i\right)$.
(Note that $\delta_i \subseteq (X^i \times \Sigma_i \times X^i)$).

Assumptions:
H is trim.
S is a stack with the normal *push* and *pop* operations.

Initialize:
$X := \{x_0^H\}$;
$x_0 := x_0^H$;
$\delta^1 := \emptyset$;
$S := [x_0]$;

Repeat:
{
$x_s = pop(S)$

Do the following for all $x_b \in X^H$:
{
Do the following for all $\sigma_g \in \Sigma_g$ with $(x_s, \sigma_g, x_b) \in \delta^H$:
{
$\delta^1 := \delta^1 \cup \{(x_s, \sigma_g, x_b)\}$;
If $x_b \notin X$
Then
{
$X := X \cup \{x_b\}$;
$push(S, x_b)$;
}
}
Do the following for all $\sigma_{p1} \in \Sigma_{p1}, \ldots, \sigma_{pn} \in \Sigma_{pn}$
such that $\sigma_{pi} = \Psi_{1i}(\sigma_{p1})$, $(x_s, \sigma_{p1}\sigma_{p2}\cdots\sigma_{pn}, x_b) \in \delta^H$:
{
$\delta^1 := \delta^1 \cup \{(x_s, \sigma_{p1}, x_b)\}$;
If $x_b \notin X$
Then
{
$X := X \cup \{x_b\}$;
$push(S, x_b)$;
}
}
}
}
Until S is empty;
$X_m = X_m^H \cap X$;
Construct $\{H_2, \ldots, H_n\}$ from H_1 using $\Psi_{1i}(\cdot)$ operations;
Return $\{H_1, \ldots, H_n\}$.

For Algorithm 2 it may be known that there exists a set of automaton $\{H_1', \ldots, H_n'\}$ such that $\mathcal{L}_m(H) = \mathcal{L}_m(H_1' \| \cdots \| H_n')$. However, the automata $\{H_1', \ldots, H_n'\}$ might not be known explicitly, and thus Algorithm 2 gives a polynomial time method to compute a similar module system $\{H_1, \ldots, H_n\}$ such that $\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \cdots \| H_n)$ and for all $i \in \{1, \ldots, n\}$, $P_i(\mathcal{L}_m(H)) = \mathcal{L}_m(H_i)$. The correctness of the decomposition algorithm is demonstrated in the following theorem.

THEOREM 7. *Suppose there is a trim automaton H and there exists a set of similar module automata $\{H_1', \ldots, H_n'\}$ such that $\mathcal{L}_m(H) = \mathcal{L}_m(H_1' \| \cdots \| H_n')$. Then, if Algorithm 2 is run with H as input, a similar module system $\{H_1, \ldots, H_n\}$ is returned such that*

$$\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \cdots \| H_n)$$

and

$$\forall i \in \{1, \ldots, n\},\ P_i(\mathcal{L}_m(H)) = \mathcal{L}_m(H_i).$$

Proof. From H, a specially trimmed version of H called H^{lock} can be constructed by trimming all private event transitions in H except those that correspond to occurrences of chains of private events

$$\sigma_{p1}\Psi_{12}(\sigma_{p1})\cdots\Psi_{1n}(\sigma_{p1})$$

for $\sigma_{p1} \in \Sigma_{p1}$, and no other events are allowed to be interleaved in these strings.

Therefore,

$$\mathcal{L}_m(H^{lock}) = \mathcal{L}_m(H) \cap (\{\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1}) | \sigma_{p1} \in \Sigma_{p1}\} \cup \Sigma_g)^* .$$

Now consider the composition of the $\{H'_1, \ldots, H'_n\}$ automata, where global events are allowed to occur freely and private events are restricted to occur only in strings $\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1})$ for $\sigma_{p1} \in \Sigma_{p1}$ with no other events interleaved, as above. This automaton marks the same language as H^{lock} shown above because $\mathcal{L}_m(H) = \mathcal{L}_m(H'_1 \| \cdots \| H'_n)$. Furthermore, due the construction of H^{lock}, it should be apparent that

$$P_i\left(\mathcal{L}_m(H^{lock})\right) = \mathcal{L}_m(H'_i)$$

because the marking behavior of H_i can be reclaimed from H^{lock} by replacing the $\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1})$ transitions with a single σ_{p1} transition. Therefore,

$$P_i\left[\mathcal{L}_m(H) \cap (\{\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1}) | \sigma_{p1} \in \Sigma_{p1}\} \cup \Sigma_g)^*\right] = \mathcal{L}_m(H'_i).$$

Algorithm 2 constructs H_1 by restricting the behavior of private events in H to strings such as $\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1})$ and then converts these transition strings into σ_{p1} events. Therefore,

$$\mathcal{L}_m(H_1) = P_i\left[\mathcal{L}_m(H) \cap (\{\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1}) | \sigma_{p1} \in \Sigma_{p1}\} \cup \Sigma_g)^*\right].$$

Consequently, $\mathcal{L}_m(H'_1) = \mathcal{L}_m(H_1)$. For all $i \in \{1, \ldots, n\}$, the automata H'_i and H_i are copies of the H'_1 and H_1 automata with respect to a renaming of transition labels by the $\Psi_{1i}(\cdot)$ functions. This implies that

$$\bigcap_{i \in \{1,\ldots,n\}} \mathcal{L}_m(H'_i) = \bigcap_{i \in \{1,\ldots,n\}} \mathcal{L}_m(H_i).$$

Therefore,

$$\mathcal{L}_m(H'_1 \| \cdots \| H'_n) = \mathcal{L}_m(H_1 \| \cdots \| H_n).$$

Hence, $\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \cdots \| H_n)$.

By construction, $\{H_1, \ldots, H_n\}$ is a similar module system, and by Theorem 1, it is known that $P_i\left(\mathcal{L}_m(H_1 \| \cdots \| H_n)\right) = \mathcal{L}_m(H_i)$. Therefore, by substitution,

$$\forall i \in \{1, \ldots, n\},\ P_i(\mathcal{L}_m(H)) = \mathcal{L}_m(H_i). \qquad \square$$

Note that due to Theorems 7 and 1 and the facts that $\mathcal{L}_m(H) = \mathcal{L}_m(H'_1 \| \cdots \| H'_n)$, $\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \cdots \| H_n)$, it is known that $\mathcal{L}_m(H'_1) = \mathcal{L}_m(H_1)$. Hence Algorithm 2 gives a method for computing automata $\{H_1, \ldots, H_n\}$, which mark the same languages as the similar module systems $\{H'_1, \ldots, H'_n\}$, which comprise the composed system H when $\{H'_1, \ldots, H'_n\}$ are known to exist. Corollary 2 also follows immediately from Theorems 7 and 6.

COROLLARY 2. *Suppose there is a trim automaton H such that $\mathcal{L}_m(H)$ is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$ and separable with respect to the set of projections*

$\{P_1, \ldots, P_n\}$. *Then, if Algorithm* 2 *is run with* H *as input, a similar module system* $\{H_1, \ldots, H_n\}$ *is returned such that*

$$\mathcal{L}_m(H) = \mathcal{L}_m(H_1 \| \cdots \| H_n)$$

and

$$\forall i \in \{1, \ldots, n\},\ P_i(\mathcal{L}_m(H)) = \mathcal{L}_m(H_i).$$

The main *Repeat–Until* loop in Algorithm 2 iterates $|X|$ times, which is the size of the state space of H_1. The size of the state space of H_1 is always at least as small as $|X^H|$, the size of the state space of H, and generally much more so. Inside the main iterative loop, there are two types of tests for transition existence: one for individual global events and one for chains of private events. The tests for transition existence are generally efficient and negligible depending on the encoding of H, especially if each state contains a list of its outgoing transitions. In this case, only the existence of $|\Sigma_g|$ global event transitions and $|\Sigma_{p1}|$ private event string transitions at each state need to be tested. If there is a $\sigma_{p1} \in \Sigma_{p1}$ transition at the current state, it needs to be tested if there is a chain of transitions $\sigma_{p1}\sigma_{p2}\cdots\sigma_{pn}$ from the current state. Therefore, for every private event transition detected, at most $n-1$ other transitions need to be tested. Therefore, Algorithm 2 is in $O\left(|X| * (|\Sigma_g| + n * |\Sigma_{p1}|)\right)$.

An example of a run of Algorithm 2 is now given using a trimmed version of the automaton $G_1 \| G_2$ from Example 1.

Example 8. Consider the automaton G that is the trimmed version of $G_1 \| G_2$ from Example 1 with the states renamed for convenience. This automaton can be seen in Figure 3. Recall that $\mathcal{L}_m(G) = \left((\alpha_1\alpha_2 + \alpha_2\alpha_1)\,\gamma + (\beta_1\beta_2 + \beta_2\beta_1)\,\lambda\right)^*$.

Algorithm 2 is now run with G as input for $\Sigma_1 = \{\alpha_1, \beta_1, \gamma, \lambda\}$ and $\Sigma_2 = \{\alpha_2, \beta_2, \gamma, \lambda\}$ as follows:

To initialize: $X := \{1\}$, $x_0 := 1$, $\delta^1 := \emptyset$, $S := [1]$.

The first state is popped off of S. $x_s := 1$.

There are no Σ_g transitions from state 1, but there are two private event chains, $\alpha_1\alpha_2$ and $\beta_1\beta_2$ starting from state 1 and, respectively, leading to states 4 and 7. Therefore, states $4, 7$ are added to X; $(1, \alpha_1, 4)$ and $(1, \beta_1, 7)$ are added to δ^1; and 4, 7 are pushed onto S.

Now, $X = \{1, 4, 7\}$ $\delta^1 = \{(1, \alpha_1, 4), (1, \beta_1, 7)\}$, $S = [4, 7]$.

Next, 4 is popped off of S and $x_s := 4$. There is a Σ_g transition from 4 $(4, \gamma, 1)$, but no private event chains. Therefore, $(4, \gamma, 1)$ is added to δ^1 and no other changes are made.

Next, 7 is popped off of S and $x_s := 7$. There is a Σ_g transition from 7 $(7, \lambda, 1)$, but no private event chains. Therefore, $(7, \lambda, 1)$ is added to δ^1 and no other changes are made. The stack S is empty, so the *Repeat–Until* loop has been completed.

The marked state list is now assigned $X_m = \{1\}$ and this completes the construction of G_1. G_1 is then copied as G_2 by replacing α_1 with α_2 and β_1 with β_2.

This results in the automata seen in Figure 8.

It is now shown that the $\{H_1, \ldots, H_n\}$ automata returned by Algorithm 2 are guaranteed to be trim.

THEOREM 8. *Suppose a trim automaton* H *is given such that* $\mathcal{L}_m(H)$ *is symmetric with respect to* $\{\Psi_{11}, \ldots, \Psi_{1n}\}$ *and separable with respect to the set of projections* $\{P_1, \ldots, P_n\}$. *Let* $\{H_1, \ldots, H_n\}$ *be the automata constructed from* H *using Algorithm* 2. *Then, the automata* $\{H_1, \ldots, H_n\}$ *are all trim.*

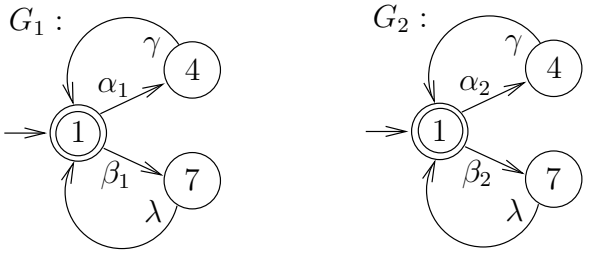


FIG. 8. *The automata G_1 and G_2 decomposed from G in Figure 3.*

Proof. It suffices to show that H_1 is trim.

It is known that all states X in H_1 are reachable from the initial state, and thus to show that H_1 is trim, it must be demonstrated that for every unmarked state in H_1, there must be a string to a marked state.

Suppose $x \in X \setminus X_m$. Because H is trim, there must be two strings $s, t \in \Sigma$ such that s is in

$$\left(\{\sigma_{p1}\Psi_{12}(\sigma_{p1}) \cdots \Psi_{1n}(\sigma_{p1}) | \sigma_{p1} \in \Sigma_{p1}\} \cup \Sigma_g\right)^*,$$

where $\delta^H(x_0^H, s) = x$, and $\delta^H(x, t) \in X_m^H$. Define s_1, t_1 such that $s_1 = P_1(s)$ and $t_1 = P_1(t)$. Because of the construction in Algorithm 2, it must be true that $\delta^1(x_0, s_1) = x$. Furthermore, because $\delta^H(x, t) \in X_m^H$ and t_1 is the string of events that is relevant to H_1 when t occurs, then it must be true that $\delta^1(x, t_1) \in X_m$ because $\mathcal{L}_m(H_1 \| \cdots \| H_n) = \mathcal{L}_m(H)$. □

7. Control operations. Now that verification methods for similar module systems have been discussed, properties related to the control of these systems are explored. The control framework for these systems is first introduced. It is a version of the standard decentralized supervisory control framework specialized for similar module systems.

Given a supervisor S_1 and a system G_1, the composed system of S_1 supervising G_1 is denoted as the supervised system S_1/G_1. Furthermore, because the parallel supervisors are assumed to be realized as finite-state automata, S_1/G_1 is equivalent to $S_1 \| G_1$. Supervisor S_1 is said to be nonblocking for system G_1 if S_1/G_1 is nonblocking, i.e., if $\overline{\mathcal{L}_m(S_1/G_1)} = \mathcal{L}(S_1/G_1)$.

Controller S_i can observe a subset of system events $\Sigma_{oi} \subseteq \Sigma_i$. Furthermore, on the occurrence of observable events, controller S_i may be given sufficient actuation to selectively disable the locally controllable events $\Sigma_{ci} \subseteq \Sigma_i$. Controllers should not be able to disable uncontrollable events, and control actions should not update on the occurrence of unobservable events. Furthermore, let $\Sigma_{uci} = \Sigma_i \setminus \Sigma_{ci}$. Let $P_{oi} : \Sigma^* \to \Sigma_{oi}^*$ be the natural projection that erases events in $\Sigma \setminus \Sigma_{oi}$ and represents the projection operation for the local observations of controller S_i.

For the control systems model, it is assumed that the controller automata $\{S_2, \ldots, S_n\}$ are copies of the generic control module S_1 with the local controller's events replaced according to the $\Psi_{1i}(\cdot)$ mapping. The set of controllers $\{S_1, \ldots, S_n\}$ defined in this way is called a set of *similar controllers*.

DEFINITION 6. *A set of controllers $\{S_1, \ldots, S_n\}$ is called a set of* similar controllers *if all controllers are exact copies of one another with respect to $\Psi_{ij}(\cdot)$ operations. In particular, $\Sigma_{ci} = \Psi_{1i}(\Sigma_{c1})$, $\Sigma_{oi} = \Psi_{1i}(\Sigma_{o1})$, and if controller S_1 disables events $\gamma_1 \subseteq \Sigma_{c1}$ after observing a string s_1, then controller S_i disables $\Psi_{1i}(\gamma_1) \subseteq \Psi_{1i}(\Sigma_{c1})$ after observing string $\Psi_{1i}(s_1)$.*

Note that due to Definition 6, if a global event $\sigma \in \Sigma_g$ is controllable (observable) by S_i, then it is controllable (observable) by all other controllers.

The controllers $\{S_1, \ldots, S_n\}$ are nonblocking for $\{G_1, \ldots, G_n\}$ if

$$\overline{\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n))} = \mathcal{L}((S_1/G_1)\| \cdots \|(S_n/G_n)).$$

8. Control for local specifications. There are potentially great reductions in computational effort for many similar module system control problems with local specifications. For a similar module system $\{G_1, \ldots, G_n\}$ and a set of similar language specifications $\{K_1, \ldots, K_n\}$, suppose it is desirable to know if there exists a set of similar controllers $\{S_1, \ldots, S_n\}$ such that

$$\forall i \in \{1, \ldots, n\}, \quad \begin{pmatrix} P_i\left[\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n))\right] = K_i \\ \wedge\, P_i\left[\mathcal{L}((S_1/G_1)\| \cdots \|(S_n/G_n))\right] = \overline{K_i} \end{pmatrix}.$$

This problem can be solved by looking only at the local behavior of G_1 and the locally observable and controllable event sets, Σ_{o1} and Σ_{c1}, respectively.

THEOREM 9. *For a similar module system $\{G_1, \ldots, G_n\}$ with local event sets $\{\Sigma_1, \ldots, \Sigma_n\}$, observable event sets $\{\Sigma_{o1}, \ldots, \Sigma_{on}\}$, controllable event sets $\{\Sigma_{c1}, \ldots, \Sigma_{cn}\}$, and local behavior specifications $\{K_1, \ldots, K_n\}$ such that for all $i, j \in \{1, \ldots, n\}$, $K_i \neq \emptyset$ and $K_i \subseteq \mathcal{L}_m(G_i)$, $K_j = \Psi_{ij}(K_i)$, there exists a set of similar controllers $\{S_1, \ldots, S_n\}$ such that for all $i \in \{1, \ldots, n\}$,*

$$P_i\left[\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n))\right] = K_i$$

and

$$P_i\left[\mathcal{L}((S_1/G_1)\| \cdots \|(S_n/G_n))\right] = \overline{K_i}$$

if and only if

1. *K_1 is controllable with respect to $\mathcal{L}(G_1)$ and Σ_{uc1}.*
2. *K_1 is observable with respect to $\mathcal{L}(G_1)$, P_{o1}, and Σ_{c1}.*
3. *K_1 is $\mathcal{L}_m(G_1)$-closed.*

Proof. It was shown in Corollary 1 that

$$P_i\left[\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n))\right] = K_i$$

and

$$P_i\left[\mathcal{L}((S_1/G_1)\| \cdots \|(S_n/G_n))\right] = \overline{K_i}$$

if and only if $\mathcal{L}_m(S_i/G_i) = K_i$ and $\mathcal{L}(S_i/G_i) = \overline{K_i}$. Due to the similar behavior of the systems, controllers, and specifications, $\mathcal{L}_m(S_i/G_i) = K_i$ and $\mathcal{L}(S_i/G_i) = \overline{K_i}$ if and only if $\mathcal{L}_m(S_1/G_1) = K_1$ and $\mathcal{L}(S_1/G_1) = \overline{K_1}$. The result then follows using the controllability and observability theorem of supervisory control theory [18]. ∎

Theorem 9 shows that controller existence for local behavior specifications can be decided by testing controller existence locally and apart from the interaction of other modules. The controllability and observability theorem is known to be constructive, and therefore there are known methods for synthesizing the local controllers $\{S_1, \ldots, S_n\}$ such that local nonblocking specifications are satisfied when they interact, and the conditions of Theorem 9 are satisfied. These results also generalize to cases when marking properties are not a concern.

Despite these very positive results, a caveat is in order with respect to the nature of language semantics. All local modules in a similar module system may be nonblocking and deadlock free, but blocking and deadlock may both still occur globally. Consider the following example.

Example 9. Consider the automaton $G_1 \| G_2$ introduced in Example 1. The composed automaton $G_1 \| G_2$ is blocking and contains two deadlock states, $(2,3)$ and $(3,2)$. These states are reached when an α event is followed immediately by a β event or when a β event is followed immediately by an α event. Neither G_1 nor G_2 can observe the interleaving of α and β events due to the restriction that the local systems cannot observe events that are private to other modules. However, when the language generated by the system $G_1 \| G_2$ is projected to either the Σ_1 or Σ_2 event sets, the behavior is equivalent to G_1 or G_2, respectively, and these automata are individually nonblocking and deadlock free.

Example 9 shows one of the major limitations of using local language specifications for similar module systems and motivates why attention should not be restricted solely to local behavior. Suppose a set of trim similar specification automata $\{H_1, \ldots, H_n\}$ is given such that for all $i \in \{1, \ldots, n\}$ $\mathcal{L}_m(H_i) = K_i$. One way to ensure that the global composed system is nonblocking after the local controllers are designed would be to verify that $\{K_1, \ldots, K_n\}$ are nonconflicting by testing if $H^{\|}$ is nonblocking. This can be verified fairly efficiently by testing if the $\tilde{H}$ construction is nonblocking, as seen in Theorem 5. For concurrent systems, blocking and deadlock properties are inherently global in nature and cannot generally be investigated on a local level. Another approach to ensuring that the global controlled system is nonblocking would be to use global control specifications instead of local specifications. This situation is explored in the following section.

9. Control for global specifications. Instead of considering a set of local similar specifications $\{K_1, \ldots, K_n\}$, suppose a global language specification K is given and that it needs to be decided if there exist nonblocking similar controllers $\{S_1, \ldots, S_n\}$ for a similar module system $\{G_1, \ldots, G_n\}$ such that $\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n)) = K$. Due to the similarity of the controllers and system modules, one would think that K would need to exhibit a degree of symmetry with respect to the occurrence of private events. This is exactly the case in Theorem 10 below.

THEOREM 10. *For a similar module system $\{G_1, \ldots, G_n\}$ with respective local projection operations $\{P_1, \ldots, P_n\}$, observation projections $\{P_{o1}, \ldots, P_{on}\}$, controllable event sets $\{\Sigma_{c1}, \ldots, \Sigma_{cn}\}$, and global behavior specification K such that $K \neq \emptyset$ and $K \subseteq \mathcal{L}_m(G_1 \| \cdots \| G_n)$, a set of nonblocking similar controllers $\{S_1, \ldots, S_n\}$ exists such that $\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n)) = K$ if and only if*

1. *$P_1(K)$ is controllable with respect to $\mathcal{L}(G_1)$ and Σ_{uc1}.*
2. *$P_1(K)$ is observable with respect to $\mathcal{L}(G_1)$, P_{o1}, and Σ_{c1}.*
3. *$P_1(K)$ is $\mathcal{L}_m(G_1)$-closed.*
4. *K is symmetric with respect to $\{\Psi_{11}, \ldots, \Psi_{1n}\}$.*
5. *K is separable with respect to $\{P_1, \ldots, P_n\}$.*
6. *$\{P_1^{-1}(P_1(K)), \ldots, P_n^{-1}(P_n(K))\}$ are nonconflicting.*

Proof. This theorem is shown in two parts. Assume there exists a set of controllers $\{S_1, \ldots, S_n\}$ such that

$$K = \mathcal{L}_m((S_1/G_1)\| \cdots \|(S_i/G_i)\| \cdots \|(S_n/G_n)), \tag{7}$$

$$\overline{K} = \mathcal{L}((S_1/G_1)\| \cdots \|(S_i/G_i)\| \cdots \|(S_n/G_n)). \tag{8}$$

By the definition of $\Psi_{1i}(\cdot)$,

$$\begin{aligned}&\Psi_{1i}(\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_i/G_i)\| \cdots \|(S_n/G_n)))\\ &\quad = \mathcal{L}_m((S_i/G_i)\| \cdots \|(S_1/G_1)\| \cdots \|(S_n/G_n)).\end{aligned}$$

The parallel composition operation is commutative, and thus

$$\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_i/G_i)\| \cdots \|(S_n/G_n)) = \mathcal{L}_m((S_i/G_i)\| \cdots \|(S_1/G_1)\| \cdots \|(S_n/G_n)).$$

$\Rightarrow K = \Psi_{1i}(K)$. This proves the fourth part of the implication.

By (7) and (8) there exists a set of controllers $\{S_1, \ldots, S_n\}$ such that
$K = \mathcal{L}_m((S_1/G_1)\| \cdots \|(S_i/G_i)\| \cdots \|(S_n/G_n))$,
$\overline{K} = \mathcal{L}((S_1/G_1)\| \cdots \|(S_i/G_i)\| \cdots \|(S_n/G_n))$
$\Rightarrow$
there exists a set of controllers $\{S_1, \ldots, S_n\}$ such that for all $i \in \{1, \ldots, n\}$,
$P_i(K) = P_i(\mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n)))$,
$P_i(\overline{K}) = P_i(\mathcal{L}((S_1/G_1)\| \cdots \|(S_n/G_n)))$
$\Rightarrow$ using Corollary 1:
there exists a set of controllers $\{S_1, \ldots, S_n\}$ such that for all $i \in \{1, \ldots, n\}$,
$P_i(K) = \mathcal{L}_m(S_i/G_i)$,
$P_i(\overline{K}) = \mathcal{L}(S_i/G_i)$
$\Rightarrow$ from the controllability and observability theorem of supervisory control theory,

1. $P_i(K)$ is controllable with respect to $\mathcal{L}(G_i)$ and Σ_{uci}.
2. $P_i(K)$ is observable with respect to $\mathcal{L}(G_i)$, P_{oi}, and Σ_{ci}.
3. $P_i(K)$ is $\mathcal{L}_m(G_i)$-closed.

This proves the first three parts of the implication.

It is known that
$K = P_1^{-1}(\mathcal{L}_m(S_1/G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(S_n/G_n))$,
$\overline{K} = P_1^{-1}(\mathcal{L}(S_1/G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}(S_n/G_n))$
$\Rightarrow$ using Corollary 1:
$K = P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))$,
$\overline{K} = P_1^{-1}(\overline{P_1(K)}) \cap \cdots \cap P_n^{-1}(\overline{P_n(K)})$
$\Rightarrow$ using Lemma 1:
$K = P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))$,
$\overline{K} = \overline{P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))}$.

This proves the fifth and sixth parts of the implication. This completes the first part of the proof.

It is now shown that the controllers exist if the six conditions hold.

Because of the first three conditions, it is known that there exists a controller S_1 such that $P_1(K) = \mathcal{L}_m(S_1/G_1)$ and $\overline{P_1(K)} = \mathcal{L}(S_1/G_1)$.

Because of the separability condition, $K = P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))$.

Because of the symmetry condition and Lemma 2, it is known that there exists controllers $S_1, \ldots, S_n$ such that for all $i \in \{1, \ldots, n\}$, $(P_i(K) = \mathcal{L}_m(S_i/G_i))$.

Therefore by substitution,
$K = P_1^{-1}(\mathcal{L}_m(S_1/G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}_m(S_n/G_n))$.
This implies that
$K = \mathcal{L}_m((S_1/G_1)\| \cdots \|(S_n/G_n))$.
Because of condition 6 and Lemma 1,
$\overline{P_1^{-1}(P_1(K)) \cap \cdots \cap P_n^{-1}(P_n(K))} = P_1^{-1}(\overline{P_1(K)}) \cap \cdots \cap P_n^{-1}(\overline{P_n(K)})$.
$\Rightarrow$ Because of conditions 5 and 6, $\overline{K}$ is separable, and
$\overline{K} = P_1^{-1}(\overline{P_1(K)}) \cap \cdots \cap P_n^{-1}(\overline{P_n(K)})$.

Because of $\overline{P_1(K)} = \mathcal{L}(S_1/G_1)$, the symmetry condition, and Lemma 2, it is known that there exists controllers $S_1, \ldots, S_n$ such that for all $i \in \{1, \ldots, n\}$,

$(\overline{P_i(K)} = \mathcal{L}(S_i/G_i))$.

Therefore, for the set of similar controllers $\{S_1, \ldots, S_n\}$,

$\overline{K} = P_1^{-1}(\mathcal{L}(S_1/G_1)) \cap \cdots \cap P_n^{-1}(\mathcal{L}(S_n/G_n))$

$\Rightarrow$

$\overline{K} = \mathcal{L}(S_1/G_1 \| \cdots \| S_n/G_n)$.

Overall it has been shown that there exists a set of similar controllers $\{S_1, \ldots, S_n\}$ such that

$K = \mathcal{L}_m((S_1/G_1) \| \cdots \| (S_n/G_n))$,

$\overline{K} = \mathcal{L}((S_1/G_1) \| \cdots \| (S_n/G_n))$. □

The six necessary and sufficient conditions for controller existence in Theorem 10 can be divided into two types. The first three conditions (local controllability, observability, and $\mathcal{L}_m$-closure) are essentially existence conditions for local controllers to achieve local projections of global behavior. These conditions are inherent to many supervisory control problems and have been well known from the early papers in supervisory control theory such as [18, 24]. However, the combination of the last three conditions (symmetry, separability, and nonconflictingness) is unique to this problem setting.

The separability condition ensures that when the specification K is decomposed into sets of desired local behaviors, if automata that mark the desired local behavior are composed, then the behavior of the composed automaton is equal to K. That is, the composition of the automata that mark the $\{P_1(K), \ldots, P_n(K)\}$ (i.e., $\{S_1/G_1, \ldots, S_n/G_n\}$) is equal to K. The nonconflicting condition ensures that if the local automata $(\{S_1/G_1, \ldots, S_n/G_n\})$ are all nonblocking, then the composed automaton $(S_1/G_1 \| \cdots \| S_n/G_n)$ is nonblocking.

The symmetry condition ensures that if K is decomposed with respect to $\{P_1, \ldots, P_n\}$, then the set of modules that mark $\{P_1(K), \ldots, P_n(K)\}$ is a similar module system. This condition in retrospect should be expected when one considers Theorem 6. If the behavior of all controllers operating on the modules is similar with respect to a renaming of local events, then the specification would necessarily be symmetric if it can be achieved.

Taken together, if conditions 1 through 3 are satisfied, then conditions 4 through 6 imply that the global specification K needs to be expressible as a set of similar nonconflicting local specifications $\{P_1(K), \ldots, P_n(K)\}$ if there exists a set of similar controllers $\{S_1, \ldots, S_n\}$ that can be coupled with the similar module system $\{G_1, \ldots, G_n\}$ to achieve the specification K.

Suppose a trim automaton H is given such that $\mathcal{L}_m(H) = K$, where K satisfies the conditions of Theorem 10. Using the decomposition algorithm, Algorithm 2, a set of trim automata $\{H_1, \ldots, H_n\}$ can be constructed from H such that for all $i \in \{1, \ldots, n\}$, $\mathcal{L}_m(H_i) = K_i$ and $P_1^{-1}(K_1) \cap \cdots \cap P_n^{-1}(K_n) = K$ from Theorems 7 and 8. The local similar controllers $\{S_1, \ldots, S_n\}$ can then be synthesized using known centralized control methods.

Note that conditions 1 through 6 together imply that the language K is controllable with respect to $\mathcal{L}(G_1 \| \cdots \| G_n)$, co-observable with respect to $\mathcal{L}(G_1 \| \cdots \| G_n)$, and $\mathcal{L}_m(G_1 \| \cdots \| G_n)$-closed, but the reverse implication does not hold because of the assumptions on the controllers being used. If the controllers were allowed to be asymmetric, a larger class of specifications could be achieved, but this would require an extra amount of coordination in the control synthesis. The considered framework is designed so that once one control module is designed, the implementation of more controllers is merely a matter of copying that first controller.

10. Discussion. A model for a similar module system was given that can be used to model a wide variety of real-world processes. In addition, a method has been shown that can be used to test global blocking and deadlock-freeness without enumerating all possible combinations of system states, and it has been shown that verification of local behavior can be performed in an off-line manner without module interaction. Necessary and sufficient conditions for the existence of local controllers have also been introduced for the similar module system model.

A method has also been shown for decomposing an automaton representing global system behavior into automata representing the local, modular subsystems; this method runs in polynomial time with respect to the size of the local specification automaton. This result is important because the standard methods for performing decompositions currently known in the supervisory control literature take exponential time with respect to the size of the global model in the worst case.

It would be interesting to develop ad hoc methods for controlling similar module systems when the necessary and sufficient conditions for controller existence do not hold. It might be desired to synthesize controllers that achieve behavior that is "maximal" in some sense when the behavioral specifications cannot be matched exactly; that is, to find a set of similar controllers $\{S'_1, \ldots, S'_n\}$ for a similar module system $\{G_1, \ldots, G_n\}$ such that for any other set of similar controllers $\{S_1, \ldots, S_n\}$, $\mathcal{L}(S_1/G_1 \| \cdots \| S_n/G_n) \subseteq \mathcal{L}(S'_1/G_1 \| \cdots \| S'_n/G_n)$.

The case when the control modules are similar has been investigated solely in this paper, but it might be the case in many real-world problems that the control systems may be asymmetric. For instance, one controller may be a "leader" that has more leeway in enforcing global control actions to avoid situations in which blocking and deadlock may occur. It would also be interesting to investigate more general system models. For instance, it is possible that using a method other than parallel composition to model module interaction might lead to other results. Furthermore, it would be interesting to investigate more general system models, where the modules might have some sort of similarity besides being isomorphic with respect to a renaming of private events.

Acknowledgments. The authors would like to thank the associate editor and the reviewers for their useful comments.

REFERENCES

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, Garland Publishing, Inc., New York, 1994.

[2] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*, Kluwer Academic Publishers, Boston, MA, 1999.

[3] Y. L. Chen, S. Lafortune, and F. Lin, *Design of nonblocking modular supervisors using event priority functions*, IEEE Trans. Automat. Control, 45 (2000), pp. 432–452.

[4] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model Checking*, The MIT Press, Cambridge, MA, 2002.

[5] E. M. Clarke, S. Jha, R. Enders, and T. Filkhorn, *Exploiting symmetry in temporal logic model checking*, Formal Methods in System Design, 9 (1996), pp. 77–104.

[6] E. A. Emerson and A. P. Sistla, *Symmetry and model checking*, Formal Methods in System Design, 9 (1996), pp. 105–131.

[7] J. M. Eyzell and J. E. R. Cury, *Exploiting symmetry in the synthesis of supervisors for discrete event systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1500–1505.

[8] P. Gohari and W. M. Wonham, *On the complexity of supervisory control design in the RW framework*, IEEE Trans. Systems Man Cybernet. Part B, 30 (2000), pp. 643–652.

[9] D. Harel, O. Kupferman, and M. Y. Vardi, *On the complexity of verifying concurrent transition systems*, Inform. and Comput., 173 (2002), pp. 143–161.

[10] C. Hoffman, *Graph Isomorphism and Permutation Groups*, Lecture Notes in Comput. Sci. 132, Springer-Verlag, Berlin, 1982.
[11] C. N. Ip and D. L. Dill, *Better verification through symmetry*, Formal Methods in System Design, 9 (1996), pp. 41–75.
[12] S. Jiang, V. Chandra, and R. Kumar, *Decentralized control of discrete event systems with multiple local specifications*, in Proceedings of the 2001 American Control Conference, IEEE, Piscataway, NJ, 2001, pp. 959–966.
[13] S. Jiang and R. Kumar, *Decentralized control of discrete event systems with specializations to local control and concurrent systems*, IEEE Trans. Systems Man Cybernet. Part B, 30 (2000), pp. 653–660.
[14] O. Kupferman, M. Vardi, and P. Wolper, *An automata-theoretic approach to branching-time model checking*, J. ACM, 47 (2000), pp. 312–360.
[15] R. J. Leduc, B. Brandin, M. Lawford, and W. M. Wonham, *Hierarchical interface-based supervisory control: Serial case*, in Proceedings of the 40th IEEE Conference on Decision and Control, 2001, pp. 4116–4121.
[16] R. J. Leduc, M. Lawford, and W. M. Wonham, *Hierarchical interface-based supervisory control: AIP example*, in Proceedings of the 39th Allerton Conference on Communication, Control, and Computing, Coordinated Science Laboratory, University of Illinois, Urbana, IL, 2001, pp. 396–405.
[17] F. Lin and W. M. Wonham, *Decentralized supervisory control of discrete-event systems*, Inform. Sci., 44 (1988), pp. 199–224.
[18] F. Lin and W. M. Wonham, *On observability of discrete-event systems*, Inform. Sci., 44 (1988), pp. 173–198.
[19] F. Lin and W. M. Wonham, *Decentralized control and coordination of discrete-event systems with partial observation*, IEEE Trans. Automat. Control, 35 (1990), pp. 199–224.
[20] F. Lin and W. M. Wonham, *Verification of nonblocking in decentralized supervision*, Control-Theory Adv. Tech., 7 (1991), pp. 223–232.
[21] P. Lincoln and A. Tiwari, *Symbolic systems biology: Hybrid modeling and analysis of biological networks*, in Proceedings of the 7th International Workshop, HSCC, R. Alur and G. Pappas, eds., Lecture Notes in Comput. Sci. 2993, Springer-Verlag, Berlin, 2004, pp. 660–672.
[22] M. H. de Queiroz and J. E. R. Cury, *Modular control of composed systems*, in Proceedings of the American Control Conference, IEEE, Piscataway, NJ, 2000, pp. 4051–4055.
[23] P. J. Ramadge, *Some tractable supervisory control problems for discrete-event systems modeled by Büchi automata*, IEEE Trans. Automat. Control, 34 (1989), pp. 10–19.
[24] P. J. Ramadge and W. M. Wonham, *Supervisory control of a class of discrete-event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
[25] P. J. Ramadge and W. M. Wonham, *The control of discrete-event systems*, Proc. IEEE, 77 (1989), pp. 81–98.
[26] K. Rohloff and S. Lafortune, *On the computational complexity of the verification of modular discrete-event systems*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 16–21.
[27] K. Rohloff and S. Lafortune, *The control and verification of similar agents operating in a broadcast network*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2673–2679.
[28] K. Rohloff and S. Lafortune, *Supervisor existence for modular discrete-event systems*, in Proceedings of the 2nd IFAC Conference on Control Systems Design (Bratislava, Slovakia), IFAC, Laxenburg, Austria, 2003.
[29] K. Rudie and W. M. Wonham, *Think globally, act locally: Decentralized supervisory control*, IEEE Trans. Automat. Control, 37 (1992), pp. 1692–1708.
[30] M. Vardi and P. Wolper, *Reasoning about infinite computations*, Inform. and Comput., 115 (1994), pp. 1–37.
[31] Y. Willner and M. Heyman, *Supervisory control of concurrent discrete event systems*, Internat. J. Control, 54 (1991), pp. 1143–1169.
[32] K. C. Wong and W. M. Wonham, *Modular control and coordination of discrete-event systems*, J. Discrete Event Dyn. Syst. Theory Appl., 8 (1998), pp. 247–297.
[33] W. M. Wonham and P. J. Ramadge, *Modular supervisory control of discrete event systems*, Math. Control Signals Systems, 1 (1988), pp. 13–30.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 668–698

MAXIMUM PRINCIPLE FOR SINGULAR STOCHASTIC CONTROL PROBLEMS*

F. DUFOUR† AND B. MILLER‡

Abstract. In this paper, an optimal singular stochastic control problem is considered. For this model, a general stochastic maximum principle is obtained by using a time transformation. This is the first version of the stochastic maximum principle that covers nonlinear cases.

Key words. nonlinear stochastic systems, optimal control, singular control, maximum principle, time change

AMS subject classifications. 49J30, 49N25, 93E20

DOI. 10.1137/040612403

1. Introduction. In this paper a maximum principle is proved for stochastic singular controls. On a filtered probability space $\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}\big)$ the set of admissible controls is defined as the class of right continuous with left-hand limits (*corlol*) processes $\big\{(u_t, v_t)\big\}$, $\{\mathcal{F}_t\}$-progressively measurable satisfying the following conditions: $\{v_t\}$ is increasing, $v_T \le M$ P-a.s., and $u_t \in K$, $|u_t| \le 1$, where K is a subset of $\mathbb{R}^r$, and $T > 0$ is the finite horizon. Assuming that on the probability space $\big(\Omega, \mathcal{F}, P\big)$ there exists a standard Brownian motion $\{W_t\}$, the state process is defined by the following stochastic differential equation for any admissible control $\big\{(u_t, v_t)\big\}$:

$$x(t) \doteq \zeta + \int_0^t A(s, x(s))ds + \int_{[0,t]} B(s)u(s)dv(s) + \int_0^t D(s, x(s))dW(s), \tag{1.1}$$

and the cost is given by

$$J[C] \doteq E_P\left[G\left(\int_0^T |u(s)|dv(s), x(T)\right)\right], \tag{1.2}$$

where the functions A, B, D, and G are deterministic.

In recent years, singular stochastic control problems have received considerable attention. The connection between singular control problem and optimal stopping problem has been studied by many authors including Alvarez [1, 2], Boetius [5, 6], Boetius and Kohlmann [7], Chow, Menaldi, and Robin [10], Dufour and Miller [14], El Karoui and Karatzas [16, 17], Karatzas [28, 29], and Karatzas and Shreve [31, 32, 33].

*Received by the editors July 28, 2004; accepted for publication (in revised form) November 28, 2005; published electronically May 12, 2006. This research was supported by a CNRS/Russian Academy of Sciences cooperation (PECO/NEI 9570) and in part by the Nonlinear Control Network and by Russian Basic Investigation Foundation grant 02-01-00361.
http://www.siam.org/journals/sicon/45-2/61240.html
†Corresponding author. MAB, Université Bordeaux I, 351 cours de la Liberation, 33405 Talence Cedex, France (dufour@math.u-bordeaux1.fr).
‡Institute for Information Transmission Problems, 19 Bolshoy Karetny per., Moscow 127994, Russia (bmiller@iitp.ru). Current address: Centre of Modelling of Stochastic Systems, Monash University, 3800 Victoria, Australia (boris.miller@sci.monash.edu.au).

Results on the dynamic programming principle can be found in Boetius [6], Haussmann and Suo [25], Fleming and Soner [22], and Zhu [43]. Sufficient conditions for the existence of optimal singular control for general nonlinear models have been obtained in Dufour and Miller [14] and Haussmann and Suo [24]. Explicit problems have been solved, for example, in [11] and [30]. The authors do not pretend to present here an exhaustive panorama of the literature relative to singular control problems. However, the interested reader may consult Boetius [6] for a survey on stochastic singular control problems including theoretical results and applications.

There is a very extensive literature on the maximum principle for stochastic classical control (see, for example, [3], [4], [9], [20], [21], [26], [36], [38], [41], [42], and the references therein; this list of references is not exhaustive).

In this paper we have chosen to define the control strategy as a pair of processes $\{u(t), v(t)\}$, where $|u(t)| \leq 1$, and $v(t)$ is increasing and right continuous. An equivalent description of the singular stochastic control problem would have been to choose a process $\{\zeta(t)\}$ of bounded variation which acts on the dynamic state through the term $\int_{[0,t]} B(s)d\zeta(s)$. The connection between these two definitions of the control process $\{u(t), v(t)\}$ and $\{\zeta(t)\}$ ($\zeta(t) = \int_{[0,t]} u(s)dv(s)$) has been discussed by many authors (see, for example, [11], [22, page 318], [35], and [43]).

To the best knowledge of the authors the stochastic maximum principle for singular controls was only considered by Cadenillas and Haussmann [8]. In their paper Cadenillas and Haussmann used a different approach and different hypotheses that are presented now in order to bring to the fore the main differences between their results and ours. In [8], the control process is described by a process $\{\zeta(t)\}$ (see the above discussion) of bounded variation, and they do not impose any L^p bounds on the control while we assume that the class of admissible controls $\{v(t)\}$ is such that $v(T) \leq M$ for a constant M. However, in [8] the state process must satisfy a linear stochastic differential equation (the functions A and D are assumed to be linear with respect to the state variable $\{x(t)\}$), and the cost function is convex. In our work, we suppose that the state process is defined by a general nonlinear stochastic differential equation, and we do not impose a convexity hypothesis on the cost function (see assumptions A(1)–A(3) in the next section). In many aspects the results obtained in [8] and here are different and complementary.

In general terms, the approach we used to obtain the maximum principle for singular control problems can be divided into three steps. The first step is to convert the original singular control problem into a classical control problem by using a special time transformation. In order to be concise, the description of this method is briefly presented, and only the important properties, which we need here, are given. Although this technique of time change is similar to the one already described in [14], it must be pointed out that the approach used here presents some technical differences that are explained in section 3. The second step is to derive the maximum principle for the auxiliary control problem (see Theorem 4.5). Important properties of the adjoint variables (see Theorem 4.4 and in particular (4.13)) that will be used to obtain the singular maximum principle are derived. The last step consists of recovering from the auxiliary maximum principle the original state and control variables by using a time change, thus giving a maximum principle for the singular control problem (see Theorem 5.9). The form of the maximum principle we obtained turns out to be different from the one derived in [8] since the adjoint variables have a singular part (see Definition 5.1 and in particular the second term of the right-hand side of (5.1)) and since the optimal singular control maximizes an Hamiltonian a.s. with respect to

the Doleans-Dade measure generated by the optimal control $\{v(t)\}$ (see the detailed discussion before Theorem 5.9).

REMARK 1.1. *It must be pointed out that our time change technique cannot be used to solve singular control problems in the context of infinite horizon.*

The finite fuel constraint $(v(T) \leq M)$ *is a crucial assumption in order to derive the maximum principle by using our method of time transformation. Indeed, in* [14]*, it has been shown that our approach of time change enables us to convert a nonlinear singular control problem into an auxiliary control problem where the control variables are of the classical type and where the controller must choose a stopping time. However, since it is very difficult to derive a general maximum principle combining classical control and optimal stopping, we need to impose the finite fuel constraint* $(v(T) \leq M)$ *to convert the singular control problem into the classical control problem where the horizon is fixed. In* [8]*, Cadenillas and Haussmann do not have to impose a finite fuel constraint, but they used different hypotheses (see the discussion above for a comparison between their approach and ours).*

Nevertheless, we have shown in [14] *that our method of time transformation can be applied to study the existence of optimal singular control in a general context where the finite fuel constraint is not a necessary assumption.*

Singular control problems have been studied under various hypotheses (for example, the finite horizon case can be found in [7, 16, 28, 29, 31]*; the infinite horizon case in* [7, 11, 28, 33]*; with a finite fuel constraint in* [16, 28, 33]*).*

The paper is organized as follows. In section 2, we formulate the singular control problem. The time change and the auxiliary control problem is briefly described in section 3. Section 4 deals with the auxiliary maximum principle and its properties. In section 5, the main results are obtained and in particular the stochastic maximum principle for singular controls (see Theorem 5.9). A simple example is presented in section 6 to illustrate the set of necessary conditions that must be satisfied by an optimal solution. In the last section, we make some comments about possible generalizations of our work.

NOTATION.

The Lebesgue measure on $\mathbb{R}$ is denoted by λ.

$\mathbb{N}_N$ is the set of the first N integers, that is, $\mathbb{N}_N = \{1, \ldots, i, \ldots, N\}$.

$\mathbb{N}^* \doteq \{k \in \mathbb{N} : k > 0\}$ and $\mathbb{R}_+ \doteq \{x \in \mathbb{R} : x \geq 0\}$.

If V is a vector, V_i denotes the ith component of V.

If M is a matrix, M_i denotes a vector given by the ith column of the matrix M, and M_{ij} is the element corresponding to the ith row and the jth column.

$(^\top)$ denotes the transpose operation.

For $x \in \mathbb{R}^k$, $|x|$ denotes its Euclidean norm and for a matrix $A \in \mathbb{R}^{k \times d}$ the norm of A is defined by $|A| = \sqrt{\mathrm{tr}\big[AA^\top\big]}$.

For $K \subset \mathbb{R}^k$, $B_1(K) \doteq \{x \in K : |x| \leq 1\}$, and $S_1(K) \doteq \{x \in K : |x| = 1\}$.

$\mathcal{S}_n$ denotes the set of all $(n \times n)$ real symmetric matrices.

$0_n \in \mathbb{R}^n$ is the zero vector.

The indicator function of a set A is defined as $I_A(x)$.

The function δ defined on $\mathbb{N} \times \mathbb{N}$ is such that $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

For $x \in \mathbb{R}$, x^+ is defined by $x^+ = \frac{1}{x}$ if $x \neq 0$ and $x^+ = 0$ if $x = 0$.

If V is a metric space, then $\mathcal{B}(V)$ denotes its associated borel σ-field.

A filtered probability space $\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}\big)$ is said to satisfy the usual hypotheses if the probability space $\big(\Omega, \mathcal{F}, P\big)$ is complete and if the filtration $\{\mathcal{F}_t\}$ is right continuous and if every $\mathcal{F}_t$ contains all P-null sets of $\mathcal{F}$.

Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\})$ be a filtered probability space and $\{\theta_t\}$ be a $[0,1]$-valued, $\{\mathcal{F}_t\}$-progressively measurable process. Suppose that V denotes any of the spaces $\mathbb{R}^k$, $\mathbb{R}^{k\times d}$, or $\mathcal{S}_k$.
Then $L^2\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}; [0,T]; V\big)$ (respectively, $L^2_\theta(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}; [0,T]; V)$) denotes the set of V-valued processes $\{x(t)\big\}$ which are $\{\mathcal{F}_t\}$-progressively measurable and satisfy $E_P[\int_0^T |x(s)|^2 ds] < +\infty$ (respectively, $E_P[\int_0^T |x(s)|^2[1-\theta(s)]ds] < +\infty$).
If $\mathbb{F}$ denotes the filtered probability space $\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}\big)$, then $L^2\big(\mathbb{F}; [0,T]; V\big)$ (respectively, $L^2_\theta(\mathbb{F}; [0,T]; V)$) is used to write in a more compact form $L^2\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}; [0,T]; V\big)$ (respectively, $L^2_\theta(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}; [0,T]; V)$).
Moreover, $L^2\big(\Omega, \mathcal{F}, P; V\big)$ denotes the set of V-valued random variables X defined on the probability space $\big(\Omega, \mathcal{F}, P\big)$ such that $E_P\big[|X|^2\big] < +\infty$.
Let $\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}\big)$ be a filtered probability space; if $\{A(t)\}$ is a *corlol*, adapted process of finite variation on each interval $[0,t]$, then dA denotes the measure associated with the distribution function $\{A(t)\}$. For $\{H(t)\}$, a progressively measurable process, the integral process $\int_0^t H(s)dA(s)$ is denoted by $H \cdot A_t$. If $\{A(t)\}$ is an increasing *corlol*, adapted process, the measure defined on $(\mathbb{R}_+ \times \Omega, \mathcal{B}(\mathbb{R}_+) \otimes \mathcal{F})$ by $E_P[\int_0^{+\infty} I_C(s)dA(s)]$ for $C \in \mathcal{B}(\mathbb{R}_+) \otimes \mathcal{F}$ is denoted by $\mathcal{M}_A$.
Let $\big(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}\big)$ be a filtered probability space satisfying the usual hypotheses and supporting a standard m-dimensional Brownian motion $\{W_t\}$. Then $\{\mathcal{F}_t^W\}$ denotes the augmentation of the natural filtration generated by $\{W_t\}$.

In order to define the state processes, let us introduce the following data:

- T and M are fixed real numbers.
- K is a subset of $\mathbb{R}^r$.
- ζ is a fixed vector in $\mathbb{R}^n$.
- $A : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$.
- $B : \mathbb{R} \to \mathbb{R}^{n\times r}$.
- $D : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$.
- $G : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$.
- $N : \mathbb{R} \to \mathbb{R}^2$ such that $N(t) = \left(\begin{smallmatrix} t-T \\ t^2 - T^2 \end{smallmatrix}\right)$.

Let us introduce the following notation:

$\mathcal{A} : \mathbb{R} \times \mathbb{R}^n \times B_1(K) \times [0,1] \longrightarrow \mathbb{R}^{n+2}$ is defined by

$$\mathcal{A}(t,x,u,z) \doteq \begin{pmatrix} 1-z \\ A(t,x)(1-z) + zB(t)u \\ z|u| \end{pmatrix}$$

and

$\mathcal{D} : \mathbb{R} \times \mathbb{R}^n \times [0,1] \longrightarrow \mathbb{R}^{n+2}$ is defined by

$$\mathcal{D}(t,x,z) \doteq \begin{pmatrix} 0 \\ D(t,x)\sqrt{1-z} \\ 0 \end{pmatrix}$$

$(\forall (t,x,u,z,p,q,P) \in \mathbb{R} \times \mathbb{R}^n \times B_1(K) \times [0,1] \times \mathbb{R}^{2+n} \times \mathbb{R}^{(2+n)\times m} \times \mathbb{R}^{n\times n})$,

$$\mathcal{H}(t,x,u,z,p,q) \doteq \mathcal{A}(t,x,u,z)^\top p + \operatorname{tr}\big[\mathcal{D}(t,x,z)^\top q\big] \tag{1.3}$$

$(\forall (t,x,u,z,p^1,p^2,p^3,q,P,\psi) \in \mathbb{R}\times\mathbb{R}^n\times B_1(K)\times[0,1]\times\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^{n\times m}\times\mathbb{R}^{n\times n}\times\mathbb{R}^3)$,

$$J(t,x,p^2,q) \doteq A(t,x)^\top p^2 + \operatorname{tr}\big[D(t,x)^\top q\big], \tag{1.4}$$

$$\begin{aligned} H(t,x,u,z,p^1,p^2,p^3,q,P,\psi) \doteq{}& (\psi_2 p^1 - \psi_1 - 2T\psi_3)(1-z) \\ &+ \psi_2\big[(1-z)A(t,x) + zB(t)u\big]^\top p^2 \\ &+ \psi_2 z|u|p^3 + \psi_2\sqrt{1-z}\operatorname{tr}\big[D(t,x)^\top q\big] \\ &+ \frac{\psi_2}{2}\operatorname{tr}[D(t,x)^\top P D(t,x)](1-z). \end{aligned} \tag{1.5}$$

The following assumptions will be used in the paper:

(A1) The maps A, B, D, and G are C^2.

(A2) The first and second derivatives of A, B, D and the second derivative of G are bounded. The maps $A(t,x)$, $B(t,x)$, $D(t,x)$ are bounded by $C(1+|t|+|x|)$. The first derivative of $G(w,x)$ is bounded by $C(1+|w|+|x|)$.

(A3) $(\forall x\in\mathbb{R}^n)$, $(\forall (w_1,w_2)\in\mathbb{R}\times\mathbb{R})$ if $w_1\le w_2$, then $G(w_1,x)\le G(w_2,x)$.

In the rest of the paper, the derivative of the function B will be denoted by B_t, and the partial derivatives of the function A (respectively, G, $\mathcal{H}$, and J) with respect to the first variable will be denoted by A_t (respectively, $\mathcal{H}_t$ and J_t) and with respect to the second variable will be denoted by A_x (respectively, $\mathcal{H}_x$ and J_x). For $j\in\mathbb{N}_m$, D_{jt} (respectively, D_{jx}) denotes the partial derivative of the function D_j with respect to the first variable (respectively, the second variable). The partial derivative of G with respect to the first variable will be denoted by G_w and with respect to the second variable will be denoted by G_x. The second partial derivatives of the function G (respectively, $\mathcal{H}$ and J) with respect to the second variable will be denoted by G_{xx} (respectively, $\mathcal{H}_{xx}$ and J_{xx}).

2. Problem statement. In this section, we formulate the original singular stochastic control problem presented in the introduction using the formulation described in [18] and in [23].

DEFINITION 2.1. *A singular control is defined by the following term:*

$$C \doteq (\Omega,\mathcal{F},P,\{\mathcal{F}_t\},\{u(t),v(t)\},\{W(t)\},\{x(t)\}),$$

where

(i) $(\Omega,\mathcal{F},P)$ *is a complete probability space with a right continuous complete filtration* $\{\mathcal{F}_t\}$*;*

(ii) $\{W(t)\}$ *is a standard* m*-dimensional* $\{\mathcal{F}_t\}$*-Brownian motion;*

(iii) $\{u(t),v(t)\}$ *is a* $B_1(K)\times\mathbb{R}_+$*-valued,* corlol, $\{\mathcal{F}_t\}$*-progressively measurable process such that* $\{v(t)\}$ *is increasing and satisfies*

$$v(T)\le M; \tag{2.1}$$

(iv) $\{x(t)\}$ *is an* $\mathbb{R}^n$*-valued,* corlol, $\{\mathcal{F}_t\}$*-progressively measurable process such that* $(\forall t\in[0,T])$

$$\begin{aligned} x(t) \doteq{}& \zeta + \int_0^t A(s,x(s))ds + \int_{[0,t]} B(s)u(s)dv(s) \\ &+ \int_0^t D(s,x(s))dW(s), \end{aligned} \tag{2.2}$$

and $x(0-)=\zeta$.

We write $\mathfrak{C}$ *for the set of controls satisfying the previous conditions.*

The cost is given by

$$J[C] \doteq E_P\left[G\left(\int_0^T |u(s)|dv(s), x(T)\right)\right]. \tag{2.3}$$

The set $\mathfrak{C}^a$ *of admissible controls is defined by*

$$\mathfrak{C}^a \doteq \{C \in \mathfrak{C} : J[C] < \infty\}. \tag{2.4}$$

The singular control problem is defined by the minimization of $J[C]$ on $\mathfrak{C}^a$. Assuming the existence of an optimal singular control $\widetilde{C}$, the aim of the paper is to derive necessary conditions for $\widetilde{C}$ to be optimal in terms of variational inequalities (see the maximum principle presented in Theorem 5.9).

3. The auxiliary control problem. In this section, it is shown that the original singular control problem can be converted into a classical control problem by using a time transformation (see Propositions 3.4, 3.5, and 3.7). We used the technique previously described in [14]. These results are presented here with minimal details in order to be concise. The interested reader may consult [14] to have a complete description of this approach. However, it must be pointed out that the model under consideration in [14] and the one presented here are different. Indeed, in order to derive a stochastic maximum principle one needs to have real-valued state constraints (see (3.26) and the definition of N), whereas the constraints used in [14] to derive sufficient conditions for the existence of an optimal control can take infinite value. Moreover, an important difference is that one needs to have a special measurability property for the auxiliary control problem (see the last part of Proposition 3.7) to ensure the existence of the adjoint variables defined by backward stochastic differential equations (see Corollary 4.2).

Assume the existence of an optimal singular control denoted by

$$\widetilde{C} \doteq \left(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\}, \{\widetilde{u}(t), \widetilde{v}(t)\}, \{\widehat{W}(t)\}, \{\widetilde{x}(t)\}\right)$$

such that $\{\widetilde{u}(t), \widetilde{v}(t)\}$ is $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$-progressively measurable.

REMARK 3.1. *The existence problem for singular stochastic control has already been studied under general hypotheses in many papers (see, for example,* [14] *and* [24] *and the references therein). As just stated, it is assumed that the optimal control* $\widetilde{C}$ *satisfies an assumption related to the measurability of the singular control* $\{\widetilde{u}(t), \widetilde{v}(t)\}$ *with respect to the filtration generated by the noise* $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$. *It must be pointed out that this hypothesis is classical (see, for example, assumption* $(S0)$ *on page* 114 *in* [41] *or the definition of the underlying filtration on page* 967 *in* [38]*).*

With the next proposition, we show how it is possible to construct an optimal singular control $\widehat{C}$ satisfying $\widehat{v}(T) = M$ from the optimal singular control $\widetilde{C}$.

PROPOSITION 3.2. *The control* $\widehat{C}$ *defined by*

$$\widehat{C} \doteq \left(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\}, \{\widehat{u}(t), \widehat{v}(t)\}, \{\widehat{W}(t)\}, \{\widehat{x}(t)\}\right), \tag{3.1}$$

where

$$\widehat{v}(t) = \widetilde{v}(t) I_{[\![0,T[\![} + (M - \widetilde{v}(T) + \widetilde{v}(t)) I_{[\![T,+\infty[\![}, \tag{3.2}$$

$$\widehat{u}(t) = \widetilde{u}(t)\left[\frac{\widetilde{v}(T) - \widetilde{v}(T-)}{M - \widetilde{v}(T-)} I_{[T,+\infty[\times\{\widetilde{v}(T)<M\}} + I_{[T,+\infty[\times\{\widetilde{v}(T)=M\}}\right]$$

$$+ \widetilde{u}(t) I_{[\![0,T[\![}, \tag{3.3}$$

is optimal. Moreover, $\widehat{v}(T) = M$*, and* $\{\widehat{u}(t), \widehat{v}(t)\}$ *is a* $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$*-progressively measurable process.*

Proof. From the definition of $\{\widehat{u}(t)\}$ and $\{\widehat{v}(t)\}$, we have that $\widehat{v}(T) = M$, and

$$(\forall t \in [0, T[), \quad \widehat{u}(t) = \widetilde{u}(t), \quad \widehat{v}(t) = \widetilde{v}(t), \quad \text{and} \quad \widehat{u}(T)\Delta\widehat{v}(T) = \widetilde{u}(T)\Delta\widetilde{v}(T); \tag{3.4}$$

consequently, we obtain that $\widehat{x}(T) = \widetilde{x}(T)$, and $\int_{[0,T]} |\widehat{u}(t)| d\widehat{v}(t) = \int_{[0,T]} |\widetilde{u}(t)| d\widetilde{v}(t)$, showing that $J[\widehat{C}] = J[\widetilde{C}]$. Clearly, $\{\widehat{u}(t), \widehat{v}(t)\}$ is a $B_1(K) \times \mathbb{R}_+$-valued, *corlol*, $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$-progressively measurable process implying that $\widehat{C}$ is optimal. □

REMARK 3.3. *Now we will work with the optimal control* $\widehat{C}$ *for technical reasons. However, by using Propositions* 3.2 *and* 5.8*, a general stochastic maximum principle will be derived in terms of the optimal control* $\widetilde{C}$ *giving the full generality to our result (see Theorem* 5.9*).*

PROPOSITION 3.4. *Denote the process* $\{t + \widehat{v}(t)\}$ *by* $\{\widehat{\Gamma}(t)\}$*. Let* $\{\eta^*(t)\}$ *be the right inverse of* $\{\widehat{\Gamma}(t)\}$*. Then* $\{\eta^*(t)\}$ *is a continuous time change such that the probability space* $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)}\})$ *satisfies the usual hypotheses. Moreover, there exists a* $[0,1]$*-valued,* $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$*-progressively measurable process* $\{\widehat{z}(t)\}$ *such that*

$$\widehat{v}(t) = \int_{[0,t]} \widehat{z}(s) d\widehat{\Gamma}(s). \tag{3.5}$$

Define the $B_1(K) \times [0,1]$*-valued,* $\{\widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)}\}$*-progressively measurable process* $\{(\alpha^*(t), \theta^*(t))\}$ *by*

$$\alpha^*(t) = \widehat{u}(\eta^*(t)) \quad \textit{and} \quad \theta^*(t) = \widehat{z}(\eta^*(t)). \tag{3.6}$$

Then

$$\eta^*(t) = \int_0^t (1 - \theta^*(s)) ds, \tag{3.7}$$

$$J[\widehat{C}] = E_{\widehat{P}}\Big[G\big(\mu^*(T+M), \xi^*(T+M)\big)\Big], \tag{3.8}$$

$$E_{\widehat{P}}\Big[N\big(\eta^*(T+M)\big)\Big] = 0_2, \tag{3.9}$$

where the processes $\{\xi^*(t)\}$ *and* $\{\mu^*(t)\}$ *are solutions of the following equations:*

$$\begin{aligned} \xi^*(t) \doteq \zeta &+ \int_0^t A(\eta^*(s), \xi^*(s))(1 - \theta^*(s)) ds + \int_0^t \theta^*(s) B(\eta^*(s)) \alpha^*(s) ds \\ &+ \int_0^t D(\eta^*(s), \xi^*(s)) \sqrt{(1 - \theta^*(s))(1 - \theta^*(s))^+} d\widehat{W}(\eta^*(s)), \end{aligned} \tag{3.10}$$

$$\mu^*(t) \doteq \int_0^t |\alpha^*(s)| \theta^*(s) ds. \tag{3.11}$$

Moreover,

$$\widehat{x}(t) = \xi^*(\widehat{\Gamma}(t)) \quad \textit{and} \quad \mu^*(T+M) = \int_{[0,T]} |\widehat{u}(s)| d\widehat{v}(s) \tag{3.12}$$

and

$$\widehat{v}(t) = \int_0^{\widehat{\Gamma}_t} \theta^*(s) ds \quad \textit{and} \quad \int_{[0,t]} |\widehat{u}(s)| d\widehat{v}(s) = \int_0^{\widehat{\Gamma}(t)} |\alpha^*(s)| \theta^*(s) ds. \tag{3.13}$$

Proof. Following Proposition 3.1, Proposition 3.2, and Theorem 4.2 in [14], the result can be obtained. □

Let $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{P}, \{\widetilde{\mathcal{F}}_t\})$ be a filtered probability space satisfying the usual hypotheses and supporting a standard m-dimensional Brownian motion $\{\widetilde{V}_t\}$. Define by $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$ the usual augmentation of the filtered probability space $\{\widehat{\Omega} \times \widetilde{\Omega}, \widehat{\mathcal{F}} \otimes \widetilde{\mathcal{F}}, \widehat{P} \otimes \widetilde{P}, \widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)} \otimes \widetilde{\mathcal{F}}_t\}$.

A random variable $\widehat{X}$ defined on $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ may be viewed as defined on $(\Omega, \mathcal{G}, Q)$ by setting $\overline{X}(\widehat{\omega}, \widetilde{\omega}) = \widehat{x}(\widehat{\omega})$ for $(\widehat{\omega}, \widetilde{\omega}) \in \widehat{\Omega} \times \widetilde{\Omega}$ and similarly for a random variable defined on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{P})$. Consequently, let us introduce on $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$ the following processes:

$$\begin{cases} \overline{\alpha}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \alpha^*(t, \widehat{\omega}), \quad \overline{\theta}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \theta^*(t, \widehat{\omega}), \quad \overline{\eta}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \eta^*(t, \widehat{\omega}), \\ \overline{\xi}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \xi^*(t, \widehat{\omega}), \quad \overline{\mu}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \mu^*(t, \widehat{\omega}), \quad \overline{W}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \widehat{W}(t, \widehat{\omega}), \\ \qquad\qquad \widetilde{W}(t, \widehat{\omega}, \widetilde{\omega}) \doteq \widetilde{V}(t, \widetilde{\omega}). \end{cases} \tag{3.14}$$

Proposition 3.5. *On $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$, the process $\{\overline{V}(t)\}$ defined by*

$$\overline{V}(t) \doteq \int_0^t \sqrt{(1 - \overline{\theta}(s))^+}\, d\overline{W}(\overline{\eta}(s)) + \int_0^t \sqrt{1 - (1 - \overline{\theta}(s))(1 - \overline{\theta}(s))^+}\, d\widetilde{W}(s) \tag{3.15}$$

is a standard m-dimensional $\{\mathcal{G}_t\}$-Brownian motion. The process $\{\overline{\xi}(t), \overline{\eta}(t), \overline{\mu}(t)\}$ is the unique solution of the following equations:

$$\begin{aligned} \overline{\xi}(t) &= \zeta + \int_0^t A(\overline{\eta}(s), \overline{\xi}(s))[1 - \overline{\theta}(s)]ds + \int_0^t B(\overline{\eta}(s))\overline{\alpha}(s)\overline{\theta}(s)ds \\ &\quad + \int_0^t D(\overline{\eta}(s), \overline{\xi}(s))\sqrt{1 - \overline{\theta}(s)}\, d\overline{V}(s), \end{aligned} \tag{3.16}$$

$$\overline{\eta}(t) = \int_0^t [1 - \overline{\theta}(s)]ds, \tag{3.17}$$

$$\overline{\mu}(t) = \int_0^t |\overline{\alpha}(s)|\overline{\theta}(s)ds. \tag{3.18}$$

Moreover,

$$J[\overline{C}] = E_Q\Big[G\big(\overline{\mu}(T + M), \overline{\xi}(T + M)\big)\Big], \tag{3.19}$$

$$E_Q\Big[N\big(\overline{\eta}(T + M)\big)\Big] = 0_2. \tag{3.20}$$

Proof. From Theorem 2.75 in [27], it follows that $\{\overline{V}(t)\}$ is a standard m-dimensional $\{\mathcal{G}_t\}$-Brownian motion. Using Theorem 6 in [39, page 194], we obtain that the solution $\{\xi(t), \eta(t), \mu(t))\}$ of (3.16)–(3.18) exists and is unique. From the definition of $\{\overline{V}(t)\}$, it is easy to show that $\{\xi(t)\}$ is the unique solution of the following equation:

$$\begin{aligned} \xi(t) &\doteq \zeta + \int_0^t A(\overline{\eta}(s), \xi(s))(1 - \overline{\theta}(s))ds + \int_0^t B(\overline{\eta}(s))\overline{\alpha}(s)\overline{\theta}(s)ds \\ &\quad + \int_0^t D(\overline{\eta}(s), \xi(s))\sqrt{(1 - \overline{\theta}(s))(1 - \overline{\theta}(s))^+}\, d\overline{W}(\overline{\eta}(s)). \end{aligned} \tag{3.21}$$

However, combining the fact that the process $\{\xi^*(t)\}$ is the solution of (3.10) and Proposition 10.46 in [27], it follows that $\{\overline{\xi}(t)\}$ satisfies (3.16). Therefore, $\{\overline{\xi}(t)\}$ is the unique solution of (3.16). Moreover, it is clear from their definitions that the processes $\{\mu(t)\}$ and $\{\overline{\mu}(t)\}$ (respectively, $\{\eta(t)\}$ and $\{\overline{\eta}(t)\}$) are indistinguishable.

Finally, (3.19) and (3.20) follow easily from (3.8) and (3.9) and the definition of the probability Q. □

On the probability space $(\Omega, \mathcal{G}, Q)$, define the filtration $\mathcal{J}_t \doteq \widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)} \otimes \{\emptyset, \widetilde{\Omega}\}$.

The set of auxiliary control $\mathcal{E}$ is the set of $\{\mathcal{J}_t\}$-progressively measurable processes defined on $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$ and taking their value in $B_1(K) \times [0,1]$. For any $\{(\alpha(t), \theta(t))\}$ in $\mathcal{E}$, the auxiliary state process $(\eta(t), \xi(t), \mu(t))$ is defined on $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$ by

$$\eta(t) \doteq \int_0^t (1-\theta(s))ds, \tag{3.22}$$

$$\begin{aligned}\xi(t) \doteq \zeta &+ \int_0^t A(\eta(s), \xi(s))(1-\theta(s))ds + \int_0^t B(\eta(s))\alpha(s)\theta(s)ds \\ &+ \int_0^t D(\eta(s), \xi(s))\sqrt{(1-\theta(s))}d\overline{V}(s),\end{aligned} \tag{3.23}$$

$$\mu(t) \doteq \int_0^t |\alpha(s)|\theta(s)ds. \tag{3.24}$$

Note that for any $\{(\alpha(t), \theta(t))\}$ in $\mathcal{E}$, the previous system admits a unique solution. Moreover, we have $E_Q[G(\mu(T+M), \xi(T+M))] < \infty$.

The associated cost functional is defined by

$$\mathcal{M}[\alpha, \theta] \doteq E_Q\Big[G\big(\mu(T+M), \xi(T+M)\big)\Big]. \tag{3.25}$$

Definition 3.6. *The set of admissible auxiliary control $\mathcal{E}_{ad}$ is defined by the set of processes $\{(\alpha(t), \theta(t))\} \in \mathcal{E}$ such that the corresponding auxiliary state process $\{(\eta(t), \xi(t), \mu(t))\}$ satisfies the following constraint:*

$$E_Q\Big[N\big(\eta(T+M)\big)\Big] = 0_2. \tag{3.26}$$

The auxiliary control problem is to minimize the cost (3.25) over $\mathcal{E}_{ad}$.

Proposition 3.7. *The auxiliary control process $\{(\overline{\alpha}(t), \overline{\theta}(t))\}$ defined by* (3.14) *is optimal. Moreover, $\{(\overline{\alpha}(t), \overline{\theta}(t))\}$ and the corresponding optimal auxiliary state $\{(\overline{\eta}(t), \overline{\xi}(t), \overline{\mu}(t))\}$ are $\{\mathcal{J}_t\}$-progressively measurable processes.*

Proof. Using Proposition 3.5, and following the same arguments as in the proof of Theorem 4.6 in [14], the result can be obtained. □

Lemma 3.8. *Denote by $\{\tau_n\}_{n\in\mathbb{N}^*}$ the sequence of $\{\widehat{\mathcal{F}}^{\widehat{W}}_t\}$-stopping times which exhausts the jumps of $\{\widehat{v}(t)\}$. Then $\{\widehat{\Gamma}(\tau_n)\}_{n\in\mathbb{N}^*}$ and $\{\widehat{\Gamma}(\tau_n-)\}_{n\in\mathbb{N}^*}$ are sequences of $\{\widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)}\}$-stopping times.*

Proof. Since $\{\widehat{\Gamma}(t)\}$ is a $\{\widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)}\}$-progressively measurable process, then there exists a sequence of $\{\widehat{\mathcal{F}}^{\widehat{W}}_t\}$-stopping times, $\{\tau_n\}_{n\in\mathbb{N}^*}$, exhausting the jumps of $\{\widehat{\Gamma}(t)\}$. Since $\{\eta^*(t)\}$ is a time change on $\big(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}^{\widehat{W}}_t\}\big)$ and using item a) of Lemma 10.5 in [27], it follows that $\{\widehat{\Gamma}(\tau_n)\}_{n\in\mathbb{N}^*}$ is a sequence of $\{\widehat{\mathcal{F}}_{\eta^*(t)}\}$-stopping times. Moreover, using the fact that $\{\widehat{\Gamma}(\tau_n-) > t\} = \{\eta^*(t) < \tau_n\}$, it follows that $\{\widehat{\Gamma}(\tau_n-)\}_{n\in\mathbb{N}^*}$ is a sequence of $\{\widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)}\}$-stopping times. □

For the rest of the paper, we shall use the following notation for the different filtered probability space under consideration:

$$\widehat{\mathbb{F}}_{\eta^*}^{\widehat{W}} \doteq \left(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_{\eta^*(t)}^{\widehat{W}}\}\right), \quad \mathbb{G} \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\}\right), \quad \mathbb{J} \doteq \left(\Omega, \mathcal{G}, Q, \{\mathcal{J}_t\}\right).$$

4. The maximum principle for the auxiliary control problem. In this section, a maximum principle is obtained for the auxiliary singular control problem. Some important properties are derived to render possible the time change in order to obtain the maximum principle for the original singular control problem presented in the next section (see Theorem 5.9). In Proposition 4.1 and Corollary 4.2 the existence of a general backward stochastic differential equations is established. In particular, this result states there exists an $\{\mathcal{J}_t\}$-adapted solution, labeled $(Y(t), X(t))$, to (4.4), although the backward equation is driven by a Wiener process, labeled $\{\overline{V}(t)\}$, such that $\mathcal{G}_t^{\overline{V}} \not\subset \mathcal{J}_t$ and $\mathcal{J}_t \not\subset \mathcal{G}_t^{\overline{V}}$. However, more importantly, a special expression is obtained for $X(t)$ (see (4.5)) that will be crucial in what follows. Using these results, Theorem 4.4 states the existence of the adjoint variables for the auxiliary control variables with an important property shown in (4.13). The stochastic maximum principle for the auxiliary control problem is then obtained in terms of these adjoint variables (see Theorem 4.5).

PROPOSITION 4.1. *On the filtered probability space $\widehat{\mathbb{F}}_{\eta^*}^{\widehat{W}}$, let us assume that the function $f : \widehat{\Omega} \times [0, T+M] \times \mathbb{R}^k \times \mathbb{R}^{k\times m} \to \mathbb{R}^k$ is $\mathcal{M} \otimes \mathcal{B}(\mathbb{R}^k) \otimes \mathcal{B}(\mathbb{R}^{k\times m})$-measurable, where $\mathcal{M}$ denotes the progressive σ-field, and satisfies the following:*

(i) $f(., 0, 0) \in L^2(\widehat{\mathbb{F}}_{\eta^*}^{\widehat{W}}; [0, T+M]; \mathbb{R}^k)$.

(ii) *There exists a constant $L > 0$ such that* $(\forall (y, \overline{y}, z, \overline{z}) \in \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^{k\times m} \times \mathbb{R}^{k\times m})$

$$|f(t, y, z) - f(t, \overline{y}, \overline{z})| \leq L\,[|y - \overline{y}| + |z - \overline{z}|] \qquad \lambda \otimes \widehat{P}\text{-a.s.}$$

Then for any given $Y^* \in L^2(\widehat{\Omega}, \widehat{\mathcal{F}}_{\eta^*(T+M)}^{\widehat{W}}, \widehat{P}; \mathbb{R}^k)$, the backward stochastic differential equation

$$\begin{aligned} Y^*(t) = Y^* &- \int_t^{T+M} f(s, Y^*(s), [1 - \theta^*(s)]Z^*(s))ds \\ &- \int_t^{T+M} Z^*(s)d\widehat{W}(\eta^*(s)) \end{aligned} \tag{4.1}$$

admits a unique solution in the following class of processes:

$$\left(\{Y^*(t)\}, \{Z^*(t)\}\right) \in L^2\left(\widehat{\mathbb{F}}_{\eta^*}^{\widehat{W}}; [0, T+M]; \mathbb{R}^k\right) \times L^2_{\theta^*}\left(\widehat{\mathbb{F}}_{\eta^*}^{\widehat{W}}; [0, T+M]; \mathbb{R}^k\right)$$

with $\{Y^*(t)\}$ continuous.

Proof. Clearly, the function f satisfies hypotheses (5.8) and (5.9) in [15, page 29], where the underlying filtered probability space is given by $\widehat{\mathbb{F}}_{\eta^*}^{\widehat{W}}$. Now, using Theorem 6.1 in [15], it follows that there exists a unique tiple $\left(\{Y^*(t)\}, \{Z^*(t)\}, \{N^*(t)\}\right)$ such that $(\forall t \in [0, T+M])$

$$\begin{aligned} Y^*(t) = Y^* &- \int_t^{T+M} f(s, Y^*(s), [1 - \theta^*(s)]Z^*(s))ds - \int_t^{T+M} Z^*(s)d\widehat{W}(\eta^*(s)) \\ &- \int_t^{T+M} dN^*(s), \end{aligned} \tag{4.2}$$

where

$$\{Y^*(t)\} \in L^2\Big(\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}; [0, T+M]; \mathbb{R}^k\Big),$$
$$\{Z^*(t)\} \in L^2_{\theta^*}\Big(\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}; [0, T+M]; \mathbb{R}^{k\times m}\Big),$$

and $\{N^*(t)\}$ is a *corlol*, $\mathbb{R}^k$-valued, $\{\widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)}\}$-martingale satisfying

$$(\forall (i,j) \in \mathbb{N}_k \times \mathbb{N}_k), \quad (\forall t \in [0, T+M]), \quad [\widehat{W}_i(\eta^*(.)), N^*_j]_t = 0 \tag{4.3}$$

with $E_{\widehat{P}}[\mathrm{tr}[[N^*, N^*]_{T+M} - [N^*, N^*]_0]] < +\infty$.

Combining the optional stopping theorem, the martingale representation theorem (see Theorem 4.15 in [34, p. 182]), and Proposition (1.5) in [40, p. 181], we obtain that there exists a process $\{\Phi(t)\} \in L^2_{\theta^*}(\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}; [0, T+M]; \mathbb{R}^{k\times m})$ such that

$$N^*(t) = E_{\widehat{P}}\left[N^*(T)\right] + \int_0^t \Phi(s) d\widehat{W}(\eta^*(s)).$$

Combining the previous equation, (3.7), and (4.3), we obtain that $\forall (i,j) \in \mathbb{N}_k \times \mathbb{N}_k$, $\forall t \in [0, T+M]$, $\int_0^t \Phi_{ij}(s)(1-\theta^*(s))ds = 0$. Consequently $(\forall t \in [0, T+M])$, $\int_t^{T+M} dN^*(s) = 0$, and the result follows. □

Note that this result is not an immediate consequence of general results for backward stochastic differential equations (see, for example, [15]) since we need in particular to show that the last term in (4.2) cancels.

The next result demonstrates the existence, for a general backward stochastic differential equation driven by the Wiener process $\{\overline{V}(t)\}$ and defined on the probability space $(\Omega, \mathcal{G}, Q, \{\mathcal{G}_t\})$, of an $\{\mathcal{J}_t\}$-adapted solution satisfying (4.5). This result will be used to show the existence of the adjoint variables defined on the same probability space.

COROLLARY 4.2. *On the filtered probability space $\mathbb{G}$, let us assume that the function $g : \Omega \times [0, T+M] \times \mathbb{R}^k \times \mathbb{R}^{k\times m} \to \mathbb{R}^k$ is $\mathcal{N} \otimes \mathcal{B}(\mathbb{R}^k) \otimes \mathcal{B}(\mathbb{R}^{k\times m})$-measurable, where $\mathcal{N}$ denotes the progressive σ-field when the probability space $(\Omega, \mathcal{G}, Q)$ is equipped with the filtration $\{\mathcal{J}_t\}$ and satisfies the following:*

(i) *$g(., 0, 0) \in L^2(\mathbb{J}; [0, T+M]; \mathbb{R}^k)$.*

(ii) *There exists a constant $L > 0$ such that $(\forall (y, \overline{y}, z, \overline{z}) \in \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^{k\times m} \times \mathbb{R}^{k\times m})$*

$$|g(t, y, z) - g(t, \overline{y}, \overline{z})| \le L\left[|y - \overline{y}| + |z - \overline{z}|\right] \qquad \lambda \otimes Q\text{-a.s.}$$

Then for any given $Y \in L^2(\Omega, \mathcal{J}_{T+M}, Q; \mathbb{R}^k)$ the backward stochastic differential equation

$$Y(t) = Y - \int_t^{T+M} g\left(s, Y(s), \sqrt{1-\overline{\theta}(s)}X(s)\right) ds - \int_t^{T+M} X(s) d\overline{V}_s \tag{4.4}$$

admits a unique solution in the following class of processes:

$$\{Y(t), X(t)\} \in L^2\Big(\mathbb{G}; [0, T+M]; \mathbb{R}^k \times \mathbb{R}^{k\times m}\Big)$$

with $\{Y(t)\}$ continuous.

Moreover, $\{Y(t), X(t)\} \in L^2(\mathbb{J}; [0, T+M]; \mathbb{R}^k \times \mathbb{R}^{k\times m})$*, and the process* $\{X(t)\}$ *can be written in the following form:*

$$X(t) = \sqrt{1-\overline{\theta}(t)}Z(t), \tag{4.5}$$

where $\{Z(t)\} \in L^2_{\overline{\theta}}(\mathbb{J}; [0, T+M]; \mathbb{R}^{k\times m})$.

Proof. For (y, z) fixed in $\mathbb{R}^k \times \mathbb{R}^{k\times m}$, the process $\{g(t, x, y)\}$ is $\{\mathcal{J}_t\}$-progressively measurable. Since $\mathcal{J}_t = \widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(t)} \otimes \{\emptyset, \widetilde{\Omega}\}$

$$(\forall \widehat{\omega} \in \widehat{\Omega}), \quad (\forall (\widetilde{\omega}_1, \widetilde{\omega}_2) \in \widetilde{\Omega} \times \widetilde{\Omega}), \quad g(\widehat{\omega}, \widetilde{\omega}_1, t, x, y) = g(\widehat{\omega}, \widetilde{\omega}_2, t, x, y).$$

For $\widetilde{\omega}_1$ fixed in $\widetilde{\Omega}$, let us define the function $f : \widehat{\Omega} \times [0, T+M] \times \mathbb{R}^k \times \mathbb{R}^{k\times m} \to \mathbb{R}^k$ such that $f(\widehat{\omega}, t, x, y) = g(\widehat{\omega}, \widetilde{\omega}_1, t, x, y)$. Then f is $\mathcal{M} \otimes \mathcal{B}(\mathbb{R}^k) \otimes \mathcal{B}(\mathbb{R}^{k\times m})$-measurable since $\mathcal{N}$ is isomorphic to $\mathcal{M} \otimes \{\emptyset, \widetilde{\Omega}\}$, where $\mathcal{M}$ has been defined in Proposition 4.1. Clearly, f so defined satisfies items (i) and (ii) of Proposition 4.1. In the same manner, define Y^* by $Y^*(\widehat{\omega}) \doteq Y(\widehat{\omega}, \widetilde{\omega}_1)$, and so $Y^* \in L^2(\widehat{\Omega}, \widehat{\mathcal{F}}^{\widehat{W}}_{\eta^*(T+M)}, \widehat{P}; \mathbb{R}^k)$.

Therefore, from Proposition 4.1 we can claim that there exists a unique solution to the following equation:

$$\begin{aligned} Y^*(t) = Y^* &- \int_t^{T+M} f(s, Y^*(s), [1-\theta^*(s)]Z^*(s))ds \\ &- \int_t^{T+M} Z^*(s)d\widehat{W}(\eta^*(s)). \end{aligned} \tag{4.6}$$

It is easy to show that

$$\int_t^{T+M} Z^*(s)d\widehat{W}(\eta^*(s)) = \int_t^{T+M} Z^*(s)\sqrt{(1-\theta^*(s))(1-\theta^*(s))^+}d\widehat{W}(\eta^*(s)). \tag{4.7}$$

Now define $Y(\widehat{\omega}, \widetilde{\omega}, t) \doteq Y^*(\widehat{\omega}, t)$, $Z(\widehat{\omega}, \widetilde{\omega}, t) \doteq Z^*(\widehat{\omega}, t)$. Following the same arguments as in the proof of Proposition 3.5 and combining (4.6) and (4.7), we obtain

$$Y(t) = Y - \int_t^{T+M} g(s, Y(s), [1-\overline{\theta}(s)]Z(s))ds - \int_t^{T+M} \sqrt{1-\overline{\theta}(s)}Z(s)d\overline{V}_s. \tag{4.8}$$

Consequently, we have that $(\{Y(t)\}, \{X(t)\})$ is the solution of (4.4), where

$$X(t) \doteq Z(t)\sqrt{1-\overline{\theta}(t)}.$$

Clearly,

$$\{Y(t), X(t)\} \in L^2\Big(\mathbb{J}; [0, T+M]; \mathbb{R}^k \times \mathbb{R}^{k\times m}\Big).$$

Then we have shown the existence of a solution with the desired property (see (4.5)). The uniqueness follows exactly as in [37]. □

We will need the following result in the proof of the next theorem.

LEMMA 4.3. $\forall (t, x, u, z) \in \mathbb{R} \times \mathbb{R}^n \times B_1(K) \times [0, 1]$, $(p^1, p^2, p^3) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$, *and* $(q^1, q^2, q^3) \in \mathbb{R}^{1\times m} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{1\times m}$,

$$\mathcal{H}_t(t, x, u, z, p, q) = \big[A_t(t, x)(1-z) + zB_t(t)u\big]^\top p^2 + \sum_{j=1}^m D_{jt}(t, x)^\top q_j^2\sqrt{1-z},$$

$$\mathcal{H}_x(t, x, u, z, p, q) = \big[A_x(t, x)(1-z)\big]^\top p^2 + \sum_{j=1}^m D_{jx}(t, x)^\top q_j^2\sqrt{1-z},$$

where

$$p = \begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix}$$

and

$$q = \begin{pmatrix} q^1 \\ q^2 \\ q^3 \end{pmatrix}.$$

Proof. One may easily show that these two equalities follow directly from the definition of $\mathcal{H}$ (see (1.3)). □

With the previous results, we may now obtain in the following theorem a crucial property given by (4.13) of the solution of the backward stochastic differential equations (4.9)–(4.12) which renders possible a time change.

THEOREM 4.4. *On the filtered probability space $\mathbb{G}$, the system of backward stochastic differential equations*

$$d\overline{p}(t) = -\begin{pmatrix} \mathcal{H}_t(\overline{\eta}(t), \overline{\xi}(t), \overline{\alpha}(t), \overline{\theta}(t), \overline{p}(t), \overline{q}(t)) \\ \mathcal{H}_x(\overline{\eta}(t), \overline{\xi}(t), \overline{\alpha}(t), \overline{\theta}(t), \overline{p}(t), \overline{q}(t)) \\ 0 \end{pmatrix} dt + \overline{q}(t) d\overline{V}(t), \tag{4.9}$$

with

$$\overline{p}(T+M) = -\begin{pmatrix} 0 \\ G_x\big(\overline{\mu}(T+M), \overline{\xi}(T+M)\big) \\ G_w\big(\overline{\mu}(T+M), \overline{\xi}(T+M)\big) \end{pmatrix} \tag{4.10}$$

and

$$\begin{aligned} d\overline{P}(t) = {} & -A_x(\overline{\eta}(t), \overline{\xi}(t))^\top \overline{P}(t)(1-\overline{\theta}(t))dt - \overline{P}(t)A_x(\overline{\eta}(t), \overline{\xi}(t))(1-\overline{\theta}(t))dt \\ & - \sum_{j=1}^{m} \left[D_{jx}(\overline{\eta}(t), \overline{\xi}(t))\right]^\top \overline{P}(t) D_{jx}(\overline{\eta}(t), \overline{\xi}(t))(1-\overline{\theta}(t))dt \\ & - \sum_{j=1}^{m} \Big(\left[D_{jx}(\overline{\eta}(t), \overline{\xi}(t))\right]^\top \overline{Q}^j(t) + \overline{Q}^j(t) D_{jx}(\overline{\eta}(t), \overline{\xi}(t)) \Big) \sqrt{1-\overline{\theta}(t)}\, dt \\ & - \mathcal{H}_{xx}(\overline{\eta}(t), \overline{\xi}(t), \overline{\alpha}(t), \overline{\theta}(t), \overline{p}(t), \overline{q}(t))dt + \sum_{j=1}^{m} \overline{Q}^j(t) d\overline{V}_j(t), \end{aligned} \tag{4.11}$$

with

$$\overline{P}(T+M) = -G_{xx}\big(\overline{\mu}(T+M), \overline{\xi}(T+M)\big), \tag{4.12}$$

admits a unique solution in the following class of processes:

$$\Big\{ \overline{p}(t), \overline{q}(t), \overline{P}(t), \big(\overline{Q}^j(t)\big)_{j\in\mathbb{N}_m} \Big\} \in L^2\Big(\mathbb{G}; [0, T+M]; \mathbb{R}^{2+n} \times \mathbb{R}^{(2+n)\times m} \times \mathcal{S}^n \times \Big[\mathcal{S}^n\Big]^m \Big)$$

with $\{\overline{p}(t), \overline{P}(t)\}$ continuous.

Moreover,

$$\begin{cases} \overline{p}(t,\widehat{\omega},\widetilde{\omega}) = p^*(t,\widehat{\omega}), & \overline{q}(t,\widehat{\omega},\widetilde{\omega}) = \sqrt{1-\theta^*(t,\widehat{\omega})}q^*(t,\widehat{\omega}), \\ \overline{P}(t,\widehat{\omega},\widetilde{\omega}) = P^*(t,\widehat{\omega}), & \overline{Q}^j(t,\widehat{\omega},\widetilde{\omega}) = \sqrt{1-\theta^*(t,\widehat{\omega})}Q^{*j}(t,\widehat{\omega}) \end{cases} \tag{4.13}$$

$\forall(\widehat{\omega},\widetilde{\omega}) \in \widehat{\Omega}\times\widetilde{\Omega}$ *and* $\forall j \in \mathbb{N}_m$*, where* $\{p^*(t), P^*(t), q^*(t), \big(Q^{*j}(t)\big)_{j\in\mathbb{N}_m}\}$ *is the unique solution in the following class of processes:*

$$\{p^*(t), P^*(t)\} \in L^2\Big(\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}; [0, T+M]; \mathbb{R}^{2+n}\times\mathcal{S}^n\Big),$$

$$\{q^*(t), \big(Q^{*j}(t)\big)_{j\in\mathbb{N}_m}\} \in L^2_{\theta^*}\Big(\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}; [0, T+M]; \mathbb{R}^{(2+n)\times m}\times\Big[\mathcal{S}^n\Big]^m\Big)$$

with $\{p^*(t), P^*(t)\}$ *continuous of the system of backward stochastic differential equations defined on the filtered probability space* $\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}$*:*

$$\begin{aligned} dp^*(t) = &-\begin{pmatrix} \mathcal{H}_t(\eta^*(t),\xi^*(t),\alpha^*(t),\theta^*(t),p^*(t),\sqrt{1-\theta^*(t)}q^*(t)) \\ \mathcal{H}_x(\eta^*(t),\xi^*(t),\alpha^*(t),\theta^*(t),p^*(t),\sqrt{1-\theta^*(t)}q^*(t)) \\ 0 \end{pmatrix} dt \\ &+ q^*(t)d\widehat{W}(\eta^*(t)) \end{aligned} \tag{4.14}$$

with

$$p^*(T+M) = -\begin{pmatrix} 0 \\ G_x\big(\mu^*(T+M),\xi^*(T+M)\big) \\ G_w\big(\mu^*(T+M),\xi^*(T+M)\big) \end{pmatrix} \tag{4.15}$$

and

$$\begin{aligned} dP^*(t) = &-A_x(\eta^*(t),\xi^*(t))^\top P^*(t)(1-\theta^*(t))dt - P^*(t)A_x(\eta^*(t),\xi^*(t))(1-\theta^*(t))dt \\ &- \sum_{j=1}^m \left[D_{jx}(\eta^*(t),\xi^*(t))\right]^\top P^*(t)D_{jx}(\eta^*(t),\xi^*(t))(1-\theta^*(t))dt \\ &- \sum_{j=1}^m \left(\left[D_{jx}(\eta^*(t),\xi^*(t))\right]^\top Q^{*j}(t) + Q^{*j}(t)D_{jx}(\eta^*(t),\xi^*(t))\right)(1-\theta^*(t))dt \\ &- \mathcal{H}_{xx}(\eta^*(t),\xi^*(t),\alpha^*(t),\theta^*(t),p^*(t),\sqrt{1-\theta^*(t)}q^*(t))dt \\ &+ \sum_{j=1}^m Q^{*j}(t)d\widehat{W}(\eta^*(t)) \end{aligned} \tag{4.16}$$

with

$$P^*(T+M) = -G_{xx}\big(\mu^*(T+M),\xi^*(T+M)\big). \tag{4.17}$$

Proof. Using Lemma 4.3, (4.9) can be written in the form

$$\overline{p}(t) = -\overline{p}(T+M) - \int_t^{T+M} g(s,\overline{p}(s),\sqrt{1-\overline{\theta}(s)}\overline{q}(s))ds - \int_t^{T+M}\overline{q}(s)d\overline{V}_s,$$

where the $\mathbb{R}^{n+2}$-valued function g is defined by

$$\begin{aligned} &g(t,p,q) \\ &= -\begin{pmatrix} \left[A_t(\overline{\eta}(t),\overline{\xi}(t))(1-\overline{\theta}(t)) + \overline{\theta}(t)B_t(\overline{\eta}(t))\overline{\alpha}(t)\right]^\top p^2 + \sum_{j=1}^m D_{jt}(\overline{\eta}(t),\overline{\xi}(t))^\top q_j^2 \\ \left[A_x(\overline{\eta}(t),\overline{\xi}(t))(1-\overline{\theta}(t))\right]^\top p^2 + \sum_{j=1}^m D_{jx}(\overline{\eta}(t),\overline{\xi}(t))^\top q_j^2 \\ 0 \end{pmatrix}, \end{aligned}$$

for

$$p = \begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix}$$

with $p^1 \in \mathbb{R}$, $p^2 \in \mathbb{R}^n$, and $p^3 \in \mathbb{R}$ and for

$$q = \begin{pmatrix} q^1 \\ q^2 \\ q^3 \end{pmatrix},$$

where $q^1 \in \mathbb{R}^{1\times n}$, $q^2 \in \mathbb{R}^{n\times n}$, and $q^3 \in \mathbb{R}^{1\times n}$. Moreover, from (4.10), it follows that $\widehat{p}(T+M) \in L^2(\Omega, \mathcal{J}_{T+M}, Q; \mathbb{R}^k)$. Now, using the hypotheses (A1)–(A2) on the data A, B, and D, it is easy to show that (4.9) satisfies the hypotheses of Corollary 4.2. Therefore, the existence and the uniqueness results for (4.9) are a straightforward consequence of Corollary 4.2. Using similar arguments, the existence and the uniqueness results can be obtained for (4.11). By using Proposition 4.1, it can be shown that (4.14)–(4.17) admit a unique solution. The last part of the theorem (see (4.13)) follows easily from Corollary 4.2 and its proof. □

Now we give the maximum principle for the auxiliary control problem.

THEOREM 4.5. *There exist* $\psi \in S_1(\mathbb{R}^3)$ *such that* $(\forall(\alpha,\theta) \in B_1(K) \times [0,1])$

$$\begin{aligned} &H(\eta^*(t), \xi^*(t), \alpha, \theta, p^{*1}(t), p^{*2}(t), p^{*3}(t), r^*(t), P^*(t), \psi) \\ &\qquad \le H(\eta^*(t), \xi^*(t), \alpha^*(t), \theta^*(t), p^{*1}(t), p^{*2}(t), p^{*3}(t), r^*(t), P^*(t), \psi) \end{aligned} \tag{4.18}$$

$\lambda \otimes \widehat{P}$*-a.s. on* $[0, T+M] \times \widehat{\Omega}$*, where* $\{p^{*1}(t)\}$ *(respectively,* $\{p^{*2}(t)\}$*,* $\{p^{*3}(t)\}$*) is an* $\mathbb{R}$ *(respectively,* $\mathbb{R}^n$*,* $\mathbb{R}$*)-valued process, and* $\{q^{*1}(t)\}$ *(respectively,* $\{q^{*2}(t)\}$*,* $\{q^{*3}(t)\}$*) is an* $\mathbb{R}^{1\times m}$ *(respectively,* $\mathbb{R}^{n\times m}$*,* $\mathbb{R}^{1\times m}$*)-valued process with*

$$\left(\left\{ \begin{pmatrix} p^{*1}(t) \\ p^{*2}(t) \\ p^{*3}(t) \end{pmatrix} \right\}, \left\{ \begin{pmatrix} q^{*1}(t) \\ q^{*2}(t) \\ q^{*3}(t) \end{pmatrix} \right\}, \{P^*(t)\}, \left(\{Q^{*j}\}\right)_{j\in\mathbb{N}_m} \right)$$

solutions of (4.14)–(4.17) *and*

$$r^*(t) \doteq \left[q^{*2}(t) - P^*(t) D(\eta^*(t), \xi^*(t)) \right] \sqrt{1 - \theta^*(t)}. \tag{4.19}$$

Proof. From assumptions (A1)–(A2) the hypothesis (3) of Theorem 5 in [38] is satisfied. However, one hypothesis of Theorem 5 in [38] is not satisfied here in the sense that we do not require the control processes $\{\overline{\alpha}(t)\}$, $\{\overline{\theta}(t)\}$ to be $\{\mathcal{F}_t^{\overline{V}}\}$-adapted. As pointed out in [41] (see page 114 and the top of page 116), this hypothesis may appear crucial in order to ensure the existence of the adjoint variables. However, although we required the processes to be $\{\mathcal{J}_t\}$-adapted, the proof of Theorem 5 in [38] remains unchanged since we have shown in Corollary 4.2 the existence and the uniqueness of the adjoint variables with the desired property.

Due to the particular block structure of the matrix $\mathcal{D}$ and from Theorem 5 in [38], it follows that there exist $\psi \in S_1(\mathbb{R}^3)$ such that $(\forall(\alpha,\theta) \in B_1(K) \times [0,1])$ the

variational inequality can be written in the following form:

$$
\begin{aligned}
&(1-\theta)\overline{p}^{1\psi}(t) + (1-\theta)A(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{p}^{2\psi}(t) + \theta\big[B(\overline{\eta}(t))\alpha\big]^{\top}\overline{p}^{2\psi}(t) + \theta|u|\overline{p}^{3\psi}(t)\\
&\qquad + \sqrt{1-\theta}\,\mathrm{tr}\big[D(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{r}^{\psi}(t)\big] + (1-\theta)\,\mathrm{tr}\big[D(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{P}^{\psi}(t)D(\overline{\eta}(t),\overline{\xi}(t))\big]\\
&\leq (1-\overline{\theta}(t))\overline{p}^{1\psi}(t) + (1-\overline{\theta}(t))A(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{p}^{2\psi}(t) + \overline{\theta}(t)\big[B(\overline{\eta}(t))\overline{\alpha}(t)\big]^{\top}\overline{p}^{2\psi}(t)\\
&\qquad + \overline{\theta}(t)|\overline{\alpha}(t)|\overline{p}^{3\psi}(t) + \sqrt{1-\overline{\theta}(t)}\,\mathrm{tr}\big[D(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{r}^{\psi}(t)\big]\\
&\qquad + (1-\overline{\theta}(t))\,\mathrm{tr}\big[D(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{P}^{\psi}(t)D(\overline{\eta}(t),\overline{\xi}(t))\big]
\end{aligned}
\tag{4.20}
$$

$\lambda\otimes P$-a.s. on $[0,T+M]\times\Omega$ with

$$
\overline{r}^{\psi}(t) = \overline{q}^{2\psi}(t) - \overline{P}^{\psi}(t)D(\overline{\eta}(t),\overline{\xi}(t))\sqrt{1-\overline{\theta}(t)},
\tag{4.21}
$$

and where $\{\overline{p}^{1\psi}(t)\}$ (respectively, $\{\overline{p}^{2\psi}(t)\}$, $\{\overline{p}^{3\psi}(t)\}$) is an $\mathbb{R}$ (respectively, $\mathbb{R}^n$, $\mathbb{R}$)-valued process, and $\{\overline{q}^{1\psi}(t)\}$ (respectively, $\{\overline{q}^{2\psi}(t)\}$, $\{\overline{p}^{3\psi}(t)\}$) is an $\mathbb{R}^{1\times m}$ (respectively, $\mathbb{R}^{n\times m}$, $\mathbb{R}^{1\times m}$)-valued process such that

$$
\left(\left\{\begin{pmatrix}\overline{p}^{1\psi}(t)\\ \overline{p}^{2\psi}(t)\\ \overline{p}^{3\psi}(t)\end{pmatrix}\right\},\left\{\begin{pmatrix}\overline{q}^{1\psi}(t)\\ \overline{q}^{2\psi}(t)\\ \overline{q}^{3\psi}(t)\end{pmatrix}\right\}\right)
$$

are the first-order adjoint variables solution of the following backward stochastic differential equation (see equation (21) in [38]):

$$
d\overline{p}^{\psi}(t) = -\begin{pmatrix}\mathcal{H}_t(\overline{\eta}(t),\overline{\xi}(t),\overline{\alpha}(t),\overline{\theta}(t),\overline{p}^{\psi}(t),\overline{q}^{\psi}(t))\\ \mathcal{H}_x(\overline{\eta}(t),\overline{\xi}(t),\overline{\alpha}(t),\overline{\theta}(t),\overline{p}^{\psi}(t),\overline{q}^{\psi}(t))\\ 0\end{pmatrix}dt + \overline{q}^{\psi}(t)d\overline{V}(t)
\tag{4.22}
$$

with

$$
\overline{p}^{\psi}(T+M) = -\begin{pmatrix}\psi_1 N_{1t}\big(\overline{\eta}(T+M)\big) + \psi_3 N_{2t}\big(\overline{\eta}(T+M)\big)\\ \psi_2 G_x\big(\overline{\mu}(T+M),\overline{\xi}(T+M)\big)\\ \psi_2 G_w\big(\overline{\mu}(T+M),\overline{\xi}(T+M)\big)\end{pmatrix}
\tag{4.23}
$$

and where the second-order adjoint variable is the solution of the following equation (see equation (22) in [38]):

$$
\begin{aligned}
d\overline{P}^{\psi}(t) = &-A_x(\overline{\eta}(t),\overline{\xi}(t))^{\top}\overline{P}^{\psi}(t)(1-\overline{\theta}(t))dt - \overline{P}^{\psi}(t)A_x(\overline{\eta}(t),\overline{\xi}(t))(1-\overline{\theta}(t))dt\\
&-\sum_{j=1}^{m}\big[D_{jx}(\overline{\eta}(t),\overline{\xi}(t))\big]^{\top}\overline{P}^{\psi}(t)D_{jx}(\overline{\eta}(t),\overline{\xi}(t))(1-\overline{\theta}(t))dt\\
&-\sum_{j=1}^{m}\Big(\big[D_{jx}(\overline{\eta}(t),\overline{\xi}(t))\big]^{\top}\overline{Q}^{\psi,j}(t) + \overline{Q}^{\psi,j}(t)D_{jx}(\overline{\eta}(t),\overline{\xi}(t))\Big)\sqrt{1-\overline{\theta}(t)}dt\\
&-\mathcal{H}_{xx}(\overline{\eta}(t),\overline{\xi}(t),\overline{\alpha}(t),\overline{\theta}(t),\overline{p}^{\psi}(t),\overline{q}^{\psi}(t))dt + \sum_{j=1}^{m}\overline{Q}^{\psi,j}(t)d\overline{V}_j(t)
\end{aligned}
\tag{4.24}
$$

with

$$
\overline{P}^{\psi}(T+M) = -\psi_2\, G_{xx}\big(\overline{\mu}(T+M),\overline{\xi}(T+M)\big).
\tag{4.25}
$$

From Theorem 3.1 in [37], the solutions of (4.22)–(4.25) exist and are unique. Since $E_P\big[N(\overline{\eta}(T+M))\big] = 0_2$, we have $N_{1t}(\overline{\eta}(T+M)) = 1$, and $N_{2t}(\overline{\eta}(T+M)) = 2T$ (see the definition of N). Therefore, the terminal condition (4.23) is given by

$$\overline{p}^{\psi}(T+M) = - \begin{pmatrix} \psi_1 + 2T\psi_3 \\ \psi_2 G_x\big(\overline{\mu}(T+M), \overline{\xi}(T+M)\big) \\ \psi_2 G_w\big(\overline{\mu}(T+M), \overline{\xi}(T+M)\big) \end{pmatrix}. \tag{4.26}$$

Simple calculations show that

$$\left(\left\{ \psi_2 \overline{p}(t) + \begin{pmatrix} -\psi_1 - 2T\psi_3 \\ 0 \\ 0 \end{pmatrix} \right\}, \{\psi_2 \overline{q}(t)\} \right)$$

is the solution of (4.22) with the terminal condition (4.26). By uniqueness, we obtain that

$$\overline{p}^{\psi}(t) = \psi_2 \overline{p}(t) + \begin{pmatrix} -\psi_1 - 2T\psi_3 \\ 0 \\ 0 \end{pmatrix}, \qquad \overline{q}^{\psi}(t) = \psi_2 \overline{q}(t). \tag{4.27}$$

Moreover, combining (4.27) and the definition of $\mathcal{H}$ (see (1.3)), it is easy to obtain that

$$\mathcal{H}_{xx}(\overline{\eta}(t), \overline{\xi}(t), \overline{\alpha}(t), \overline{\theta}(t), \overline{p}^{\psi}(t), \overline{q}^{\psi}(t)) = \psi_2 \mathcal{H}_{xx}(\overline{\eta}(t), \overline{\xi}(t), \overline{\alpha}(t), \overline{\theta}(t), \overline{p}(t), \overline{q}(t)).$$

Consequently, $(\{\psi_2 \overline{P}(t)\}, (\{\psi_2 \overline{Q}^j(t)\})_{j \in \mathbb{N}_m})$ is the solution of (4.24) with the terminal condition (4.25). By uniqueness, we obtain that

$$\overline{P}^{\psi}(t) = \psi_2 \overline{P}(t), \qquad \overline{Q}^{\psi,j}(t) = \psi_2 \overline{Q}^j(t). \tag{4.28}$$

Now using Theorem 4.4 (see (4.13)) and (4.27), (4.28), it follows that

$$\overline{p}^{\psi}(t, \widehat{\omega}, \widetilde{\omega}) = \psi_2 p^*(t, \widehat{\omega}) + \begin{pmatrix} -\psi_1 - 2T\psi_3 \\ 0 \\ 0 \end{pmatrix}, \qquad \overline{P}^{\psi}(t, \widehat{\omega}, \widetilde{\omega}) = \psi_2 P^*(t, \widehat{\omega}), \tag{4.29}$$

$$\overline{q}^{\psi}(t, \widehat{\omega}, \widetilde{\omega}) = \psi_2 \sqrt{1 - \theta^*(t, \widehat{\omega})} q^*(t, \widehat{\omega}), \qquad \overline{Q}^{\psi,j}(t, \widehat{\omega}, \widetilde{\omega}) = \psi_2 \sqrt{1 - \theta^*(t, \widehat{\omega})} Q^{*j}(t, \widehat{\omega}), \tag{4.30}$$

and from (4.21) it implies that

$$\begin{aligned} \overline{r}^{\psi}(t, \widehat{\omega}, \widetilde{\omega}) &= \psi_2 \sqrt{1 - \theta^*(t, \widehat{\omega})} \Big[q^*(t, \widehat{\omega}) - P^*(t, \widehat{\omega}) D(\eta^*(t, \widehat{\omega}), \xi^*(t, \widehat{\omega})) \Big] \\ &= \psi_2 r^*(t, \widehat{\omega}), \end{aligned} \tag{4.31}$$

where we have used the definitions of $\{\overline{\eta}(t)\}$, $\{\overline{\xi}(t)\}$ (see (3.14)).

Now using the definitions of the processes $\{\overline{\alpha}(t)\}$ and $\{\overline{\theta}(t)\}$, the definition of H (see (1.5)), and (4.30), (4.31), it follows that the variational inequality (4.20) can be written on the filtered probability space $\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}$ as $(\forall (\alpha, \theta) \in B_1(K) \times [0,1])$

$$\begin{aligned} &H(\eta^*(t), \xi^*(t), \alpha, \theta, p^{*1}(t), p^{*2}(t), p^{*3}(t), r^*(t), P^*(t), \psi) \\ &\qquad \le H(\eta^*(t), \xi^*(t), \alpha^*(t), \theta^*(t), p^{*1}(t), p^{*2}(t), p^{*3}(t), r^*(t), P^*(t), \psi) \end{aligned}$$

$\lambda \otimes \widehat{P}$-a.s. on $[0, T+M] \times \widehat{\Omega}$, showing the result. □

5. The singular maximum principle and adjoint variables. In this section we first characterize the adjoint variables for the original control problem. Using the important property given by (4.13), it is possible to obtain the adjoint variables for the original control problem as shown in Theorem 5.2 by using a time change. The interesting feature of these adjoint variables is that the first component is the solution of a singular backward equation (see (5.1)). Then some technical results (Lemma 5.3, Proposition 5.4, Corollary 5.5, Lemma 5.7) are derived. Proposition 5.8 is particularly important to describe the connection between the optimal controls $\widetilde{C}$ and $\widehat{C}$. Finally, the stochastic maximum principle for the original singular control problem in the general case is obtained and presented in Theorem 5.9.

In the following definition, we introduce the backward stochastic differential equations that will be satisfied by the adjoint variables for the original control problem. Note the special form of the second term of the right-hand side of (5.1) that gives a singular part for these adjoint variables.

DEFINITION 5.1. *Let $C \in \mathfrak{C}^a$ be a singular control*

$$C \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}, \{u(t), v(t)\}, \{W(t)\}, \{x(t)\})$$

such that $\{u(t), v(t)\}$ is $\{\mathcal{F}_t^W\}$-progressively measurable. If the system of backward stochastic differential equations

$$\begin{aligned} p^1(t) &= \int_t^T A_t(s, x(s))^\top p^2(s) ds + \int_{]t,T]} p^2(s)^\top B_t(t) u(s) dv(s) \\ &\quad + \int_t^T \sum_{j=1}^m D_{jt}(t, x(t))^\top q_j^2(s) dt - \int_t^T q^1(s) dW(s), \end{aligned} \tag{5.1}$$

$$\begin{aligned} p^2(t) &= -G_x\left(\int_0^T |u(s)| dv(s), x(T)\right) + \int_t^T A_x(s, x(s))^\top p^2(s) ds \\ &\quad + \int_t^T \sum_{j=1}^m D_{jx}(s, x(s))^\top q_j^2(s) dt - \int_t^T q^2(s) dW(s), \end{aligned} \tag{5.2}$$

$$p^3(t) = -G_w\left(\int_0^T |u(s)| dv(s), x(T)\right) - \int_t^T q^3(s) dW(s), \tag{5.3}$$

and

$$\begin{aligned} P(t) &= -G_{xx}\left(\int_0^T |u(s)| dv(s), x(T)\right) + \int_t^T \Big[A_x(s, x(s))^\top P(s) - P(s) A_x(s, x(s))\Big] ds \\ &\quad + \sum_{j=1}^m \int_t^T D_{jx}(s, x(s))^\top P(s) D_{jx}(s, x(s)) ds + \int_t^T J_{xx}(s, x(s), p^2(s), q^2(s)) ds \\ &\quad + \sum_{j=1}^m \int_t^T \big[D_{jx}(s, x(s))^\top Q^j(s) + Q^j(s) D_{jx}(s, x(s))\big] ds - \int_t^T \sum_{j=1}^m Q^j(s) dW(s), \end{aligned}$$

admits a solution in the class of processes

$$\Big\{p^1(t), p^2(t), p^3(t)\Big\} \in L^2\Big(\mathbb{F}^W; [0,T]; \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}\Big),$$

$$\Big\{q^1(t), q^2(t), q^3(t)\Big\} \in L^2\Big(\mathbb{F}^W; [0,T]; \mathbb{R}^{1\times m} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{1\times m}\Big),$$

$$\Big(\{P(t)\}, \big(\{Q^j(t)\}\big)_{j\in\mathbb{N}_m}\Big) \in L^2\Big(\mathbb{F}^W; [0,T]; \mathcal{S}^n\Big) \times \Big[L^2\Big(\mathbb{F}^W; [0,T]; \mathcal{S}^n\Big)\Big]^m,$$

with $\{p^1(t), p^2(t), p^3(t), P(t)\}$ corlol, *and where* $\mathbb{F}^W \doteq (\Omega, \mathcal{F}, P, \{\mathcal{F}_t^W\})$, *then* $((\{p^i(t)\}, \{q^i(t)\})_{i\in\mathbb{N}_3}, \{P(t)\}, (\{Q^j(t)\})_{j\in\mathbb{N}_m})$ *are called the adjoint variables associated with the control* C. *The solution is said to be unique if the solution of the previous system is unique in this class of processes.*

By using a time transformation, we show that we can obtain the adjoint variables for the original optimal control from $(\{p^*(t)\}, \{q^*(t)\}, \{P^*(t)\}, (\{Q^{*j}(t)\})_{j\in\mathbb{N}_m})$.

THEOREM 5.2. *Define the following processes:*

$$\hat{p}(t) \doteq p^*(\widehat{\Gamma}(t)), \quad \hat{q}(t) \doteq q^*(\widehat{\Gamma}(t)), \quad \widehat{P}(t) \doteq P^*(\widehat{\Gamma}(t)), \quad \widehat{Q}^j(t) \doteq Q^{*j}(\widehat{\Gamma}(t)), \tag{5.4}$$

for $j \in \mathbb{N}_m$, *where* $(\{p^*(t)\}, \{q^*(t)\}, \{P^*(t)\}, (\{Q^{*j}\})_{j\in\mathbb{N}_m})$ *are solutions of* (4.14)–(4.17). *Write* $\hat{p}(t)$ *in the form*

$$\hat{p}(t) = \begin{pmatrix} \hat{p}^1(t) \\ \hat{p}^2(t) \\ \hat{p}^3(t) \end{pmatrix},$$

where $\hat{p}^1(t) \in \mathbb{R}$, $\hat{p}^2(t) \in \mathbb{R}^n$, *and* $\hat{p}^3(t) \in \mathbb{R}$ *and similarly*

$$\hat{q}(t) = \begin{pmatrix} \hat{q}^1(t) \\ \hat{q}^2(t) \\ \hat{q}^3(t) \end{pmatrix},$$

where $\hat{q}^1(t) \in \mathbb{R}^{1\times n}$, $\hat{q}^2(t) \in \mathbb{R}^{n\times n}$, *and* $\hat{q}^3(t) \in \mathbb{R}^{1\times n}$.

Then $((\{\hat{p}^i(t)\}, \{\hat{q}^i(t)\})_{i\in\mathbb{N}_3}, \{\widehat{P}(t)\}, (\{\widehat{Q}^j(t)\})_{j\in\mathbb{N}_m})$ *are the unique adjoint variables associated with the control* $\widehat{C}$.

Proof. Combining Lemma 4.3 and (4.14), (3.7), we obtain

$$\begin{aligned} p^{*2}(t) = &- G_x\big(\mu^*(T+M), \xi^*(T+M)\big) + \int_t^{T+M} A_x(\eta^*(s), \xi^*(s))^\top p^{*2}(s) d\eta^*(s) \\ &+ \int_t^{T+M} \sum_{j=1}^m D_{jx}(\eta^*(s), \xi^*(s))^\top q_j^{*2}(s) d\eta^*(s) - \int_t^{T+M} q^{*2}(s) d\widehat{W}(\eta^*(s)). \end{aligned} \tag{5.5}$$

However, from (3.12), it follows that

$$G_x\big(\mu^*(T+M), \xi^*(T+M)\big) = G_x\bigg(\int_{[0,T]} |\widehat{u}(s)| d\widehat{v}(s), \widehat{x}(T)\bigg).$$

Now, using the fact that $\widehat{\Gamma}(T) = T + M$, (5.5) becomes

$$\begin{aligned} p^{*2}(t) = &-G_x\big(\mu^*(T+M), \xi^*(T+M)\big) + \int_t^{\widehat{\Gamma}(T)} A_x(\eta^*(s), \xi^*(s))^\top p^{*2}(s) d\eta^*(s) \\ &+ \int_t^{\widehat{\Gamma}(T)} \sum_{j=1}^m D_{jx}(\eta^*(s), \xi^*(s))^\top q_j^{*2}(s) d\eta^*(s) - \int_t^{\widehat{\Gamma}(T)} q^{*2}(s) d\widehat{W}(\eta^*(s)). \end{aligned}$$

Consequently, the value of the process $\{p^{*2}(t)\}$ at time $\widehat{\Gamma}(t)$ is given by $\{p^{*2}(\widehat{\Gamma}(t))\}$ and satisfies

$$\begin{aligned} p^{*2}(\widehat{\Gamma}(t)) = &- G_x\bigg(\int_{[0,T]} |\widehat{u}(s)| d\widehat{v}(s), \widehat{x}(T)\bigg) + \int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} A_x(\eta^*(s), \xi^*(s))^\top p^{*2}(s) d\eta^*(s) \\ &+ \int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} \sum_{j=1}^m D_{jx}(\eta^*(s), \xi^*(s))^\top q_j^{*2}(s) d\eta^*(s) - \int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} q^{*2}(s) d\widehat{W}(\eta^*(s)). \end{aligned} \tag{5.6}$$

Clearly, $\{\widehat{\Gamma}(t)\}$ is a time change on the filtered probability space $\widehat{\mathbb{F}}^{\widehat{W}}_{\eta^*}$. According to Definition (1.3) in [40, p. 180], $\{\eta^*(t)\}$ is clearly a $\{\widehat{\Gamma}(t)\}$-continuous process since $\{\eta^*(t)\}$ is the right inverse of $\{\widehat{\Gamma}(t)\}$. Recalling $\eta^*(\widehat{\Gamma}(s)) = s$, $\xi^*(\widehat{\Gamma}(s)) = \widehat{x}(s)$ and applying Proposition (1.4) in [40, p. 180] to the second and the third terms of the right-hand side of (5.6), it follows that

$$\begin{aligned} p^{*2}(\widehat{\Gamma}(t)) = &-G_x\left(\int_{[0,T]} |\widehat{u}(s)|d\widehat{v}(s), \widehat{x}(T)\right) + \int_t^T A_x(s,\widehat{x}(s))^\top p^{*2}(\widehat{\Gamma}(s))ds \\ &+ \int_t^T \sum_{j=1}^m D_{jx}(s,\widehat{x}(s))^\top q_j^{*2}(\widehat{\Gamma}(s))ds - \int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} q^{*2}(s)d\widehat{W}(\eta^*(s)). \end{aligned}$$

Finally, applying Proposition (1.5) in [40, p. 181] to the last term of the right-hand side of the previous equation gives

$$\begin{aligned} p^{*2}(\widehat{\Gamma}(t)) = &-G_x\left(\int_{[0,T]} |\widehat{u}(s)|d\widehat{v}(s), \widehat{x}(T)\right) + \int_t^T A_x(s,\widehat{x}(s))^\top p^{*2}(\widehat{\Gamma}(s))ds \\ &+ \int_t^T \sum_{j=1}^m D_{jx}(s,\widehat{x}(s))^\top q_j^{*2}(\widehat{\Gamma}(s))ds - \int_t^T q^{*2}(\widehat{\Gamma}(s))d\widehat{W}(s). \end{aligned} \tag{5.7}$$

Using similar arguments, it is easy to obtain that

(5.8)

$$p^{*3}(\widehat{\Gamma}(t)) = -G_w\left(\int_{[0,T]} |\widehat{u}(s)|d\widehat{v}(s), \widehat{x}(T)\right) - \int_t^T q^3(\widehat{\Gamma}(s))d\widehat{W}(s),$$

$$\begin{aligned} P^*(\widehat{\Gamma}(t)) = &-G_{xx}\left(\int_{[0,T]} |\widehat{u}(s)|d\widehat{v}(s), \widehat{x}(T)\right) + \int_t^T J_{xx}(s,\widehat{x}(s), p^{*2}(\widehat{\Gamma}(s)), q^{*2}(\widehat{\Gamma}(s)))ds \\ &+ \int_t^T \Big[A_x(s,\widehat{x}(s))^\top P^*(\widehat{\Gamma}(s)) - P^*(\widehat{\Gamma}(s))A_x(s,\widehat{x}(s))\Big]ds \\ &+ \sum_{j=1}^m \int_t^T D_{jx}(s,\widehat{x}(s))^\top P^*(\widehat{\Gamma}(s))D_{jx}(s,\widehat{x}(s))ds \\ &+ \sum_{j=1}^m \int_t^T \big[D_{jx}(s,\widehat{x}(s))^\top Q^{*j}(\widehat{\Gamma}(s)) + Q^{*j}(\widehat{\Gamma}(s))D_{jx}(s,\widehat{x}(s))\big]ds \\ &- \int_t^T \sum_{j=1}^m Q^{*j}(\widehat{\Gamma}(s))d\widehat{W}(s). \end{aligned} \tag{5.9}$$

Now for $\{p^{*1}(\widehat{\Gamma}(t))\}$, we have from Lemma 4.3 and (4.14)

$$\begin{aligned} p^{*1}(\widehat{\Gamma}(t)) = &\int_{\widehat{\Gamma}(t)}^{T+M} A_t(\eta^*(s),\xi^*(s))^\top p^{*2}(s)d\eta^*(s) - \int_{\widehat{\Gamma}(t)}^{T+M} q^{*1}(s)d\widehat{W}(\eta^*(s)) \\ &+ \int_{\widehat{\Gamma}(t)}^{T+M} \sum_{j=1}^m D_{jt}(\eta^*(s),\xi^*(s))^\top q_j^{*2}(s)d\eta^*(s) \\ &+ \int_{\widehat{\Gamma}(t)}^{T+M} p^{*2}(s)^\top B_t(\eta^*(s))\alpha^*(s)\theta^*(s)ds. \end{aligned} \tag{5.10}$$

Except for the last term, the same reasoning establishes

$$p^{*1}(\widehat{\Gamma}(t)) = \int_t^T A_t(s,\widehat{x}(s))^\top p^{*2}(\widehat{\Gamma}(s))ds + \int_t^T \sum_{j=1}^m D_{jt}(s,\widehat{x}(s))^\top q_j^{*2}(\widehat{\Gamma}(s))ds$$
$$- \int_t^T q^{*1}(\widehat{\Gamma}(s))d\widehat{W}(s) + \int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} p^{*2}(s)^\top B_t(\eta^*(s))\alpha^*(s)\theta^*(s)ds. \tag{5.11}$$

Note from (5.5) that $(\forall t \in [\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)])\ p^{*2}(t)$ is constant $\widehat{P}$-a.s. on the set $\{\tau_i \le T\}$, and so $(\forall t \in [0,T+M])\ p^{*2}(t) = p^{*2}(\widehat{\Gamma}(\eta^*(t)))$. Consequently, combining the previous equality and (3.6), the last term of (5.11) becomes

$$\int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} p^{*2}(s)^\top B_t(\eta^*(s))\alpha^*(s)\theta^*(s)ds = \int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} p^{*2}(\widehat{\Gamma}(\eta^*(s)))^\top B_t(\eta^*(s))\widehat{u}(\eta^*(s))d\widetilde{v}(s),$$

where $\widetilde{v}(t) = \int_0^t \theta^*(s)ds$.

Now using item (55.2) in Theorem 55 in [13, p. 132], it follows that

$$\int_{\widehat{\Gamma}(t)}^{\widehat{\Gamma}(T)} p^{*2}(s)^\top B_t(\eta^*(s))\alpha^*(s)\theta^*(s)ds = \int_{]t,T]} p^{*2}(\widehat{\Gamma}(s))^\top B_t(s)\widehat{u}(s)d\widetilde{v}(\widehat{\Gamma}(s)). \tag{5.12}$$

However, $\widehat{v}(t) = \int_0^{\widehat{\Gamma}_t} \theta^*(s)ds = \widetilde{v}(\widehat{\Gamma}(t))$ (see (3.13)). Consequently, combining (5.12) and (5.11) gives

$$p^{*1}(\widehat{\Gamma}(t)) = \int_t^T A_t(s,\widehat{x}(s))^\top p^{*2}(\widehat{\Gamma}(s))ds + \int_t^T \sum_{j=1}^m D_{jt}(s,\widehat{x}(s))^\top q_j^{*2}(\widehat{\Gamma}(s))ds$$
$$- \int_t^T q^{*1}(\widehat{\Gamma}(s))d\widehat{W}(s) + \int_{]t,T]} p^{*2}(\widehat{\Gamma}(s))^\top B_t(s)\widehat{u}(s)d\widehat{v}(s). \tag{5.13}$$

In conclusion, from (5.7), (5.8), (5.9), and (5.13), we obtain that

$$\Big(\big(\{p^{*i}(\widehat{\Gamma}(t))\},\{q^{*i}(\widehat{\Gamma}(t))\}\big)_{i\in\mathbb{N}_3},\{P(\widehat{\Gamma}(t))\},\big(\{Q^{*j}(\widehat{\Gamma}(t))\}\big)_{j\in\mathbb{N}_m}\Big),$$

the adjoint variables associated with the control $\widehat{C}$. By linearity, the uniqueness is easily obtained for $(\{\widehat{P}(t)\},(\{\widehat{Q}^j(t)\})_{j\in\mathbb{N}_m})$, and $(\{\widehat{p}^2(t)\},\{\widehat{q}^2(t)\})$ implying the uniqueness for $(\{\widehat{p}^1(t)\},\{\widehat{q}^1(t)\})$, and $(\{\widehat{p}^3(t)\},\{\widehat{q}^3(t)\})$, which shows the result. □

We present now some technical results that we will need to obtain the maximum principle.

LEMMA 5.3. *For $\psi \in \mathbb{R}^3$, and $(\alpha,\theta) \in B_1(K)\times[0,1]$, write*

$$L(t,\alpha,\theta) = H(\eta^*(t),\xi^*(t),\alpha,\theta,p^{*1}(t),p^{*2}(t),p^{*3}(t),r^*(t),P^*(t),\psi)$$
$$- H(\eta^*(t),\xi^*(t),\alpha^*(t),\theta^*(t),p^{*1}(t),p^{*2}(t),p^{*3}(t),r^*(t),P^*(t),\psi). \tag{5.14}$$

Then, $\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]$,

$$
\begin{aligned}
L(t,\alpha,\theta) &= \psi_2\big[B(\tau_i)(\theta\alpha - \widehat{u}(\tau_i))\big]^\top \hat{p}^2(\tau_i) + \psi_2(\theta|\alpha| - |\widehat{u}(\tau_i)|)\hat{p}^3(\tau_i) \\
&\quad + (1-\theta)\Bigg\{-\psi_1 - 2T\psi_3 + \psi_2\Big[\hat{p}^1(\tau_i-) - [B(\tau_i)\widehat{u}(\tau_i)]^\top (t - \widehat{\Gamma}(\tau_i-))\hat{p}^2(\tau_i) \\
&\quad + A\Big(\tau_i, \widehat{x}(\tau_i-) + B(\tau_i)\widehat{u}(\tau_i)(t-\widehat{\Gamma}(\tau_i-))\Big)^\top \hat{p}^2(\tau_i) \\
&\quad + \frac{1}{2}\operatorname{tr}\Big[D\Big(\tau_i, \widehat{x}(\tau_i-) + B(\tau_i)\widehat{u}(\tau_i)(t-\widehat{\Gamma}(\tau_i-))\Big)^\top \widehat{P}(\tau_i) \\
&\quad \times D\Big(\tau_i, \widehat{x}(\tau_i-) + B(\tau_i)\widehat{u}(\tau_i)(t-\widehat{\Gamma}(\tau_i-))\Big)\Big]\Big]\Bigg\}
\end{aligned}
$$

$\widehat{P}$*-a.s. on* $\{\tau_i \le T\}$.

Proof. Clearly, we have $\widehat{\Gamma}(\tau_i-) < \widehat{\Gamma}(\tau_i)$ $\widehat{P}$-a.s. on $\{\tau_i \le T\}$. Since $\{\eta^*(t)\}$ is the right inverse of $\{\widehat{\Gamma}(t)\}$, it gives that

$$
(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]), \quad \eta^*(t) = \tau_i \quad \widehat{P}\text{-a.s. on } \{\tau_i \le T\}, \tag{5.15}
$$

and so with (3.6) $(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)])$,

$$
\theta^*(t) = \widehat{z}(\tau_i) \quad \text{and} \quad \alpha^*(t) = \widehat{u}(\tau_i) \quad \widehat{P}\text{-a.s. on } \{\tau_i \le T\}. \tag{5.16}
$$

Therefore, using (3.7)

$$
(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]), \quad \theta^*(t) = 1 \quad \widehat{P}\text{-a.s. on } \{\tau_i \le T\}, \tag{5.17}
$$

and by (4.19)

$$
(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]), \quad r^*(t) = 0 \quad \widehat{P}\text{-a.s. on } \{\tau_i \le T\}. \tag{5.18}
$$

Combining Lemma 3.8 and (5.15) we have $(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)])$

$$
\int_{\widehat{\Gamma}(\tau_i-)}^{t} q(s)dW(\eta^*(s)) = 0 \quad \widehat{P}\text{-a.s. on } \{\tau_i \le T\}. \tag{5.19}
$$

Write $p^*(t)$ in the form

$$
p^*(t) = \begin{pmatrix} p^{*1}(t) \\ p^{*2}(t) \\ p^{*3}(t) \end{pmatrix},
$$

where $p^{*1}(t) \in \mathbb{R}$, $p^{*2}(t) \in \mathbb{R}^n$, and $p^{*3}(t) \in \mathbb{R}$.

For $\theta = 1$, simple calculations give that $\forall(\eta,\xi,\alpha,p,q) \in \mathbb{R}\times\mathbb{R}^n \times B_1(K) \times \mathbb{R}^{2+n} \times \mathbb{R}^{(2+n)\times m}$

$$
\mathcal{H}_x(\eta,\xi,\alpha,\theta,p,q) = 0 \quad \text{and} \quad \mathcal{H}_t(\eta,\xi,\alpha,\theta,p,q) = \Big(0|\big[B_t(\eta)\alpha\big]^\top|0\Big)p.
$$

It follows from (4.14), (5.15)–(5.17), and (5.19) that $\widehat{P}$-a.s. surely on $\{\tau_i \le T\}$, $\forall t$ in $[\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]$,

$$
p^*(t) = \begin{pmatrix} p^{*1}(\widehat{\Gamma}(\tau_i-)) - \displaystyle\int_{\widehat{\Gamma}(\tau_i-)}^{t} \big[B_t(\tau_i)\widehat{u}(\tau_i)\big]^\top p^{*2}(s)ds \\ p^{*2}(\widehat{\Gamma}(\tau_i)) \\ p^{*3}(\widehat{\Gamma}(\tau_i)) \end{pmatrix}
$$

and so

$$p^*(t) = \begin{pmatrix} \hat{p}^1(\tau_i-) - \left[B_t(\tau_i)\widehat{u}(\tau_i)\right]^\top \hat{p}^2(\tau_i)(t-\widehat{\Gamma}(\tau_i-)) \\ \hat{p}^2(\tau_i) \\ \hat{p}^3(\tau_i) \end{pmatrix}, \tag{5.20}$$

by using the definition of $\hat{p}(t)$.

Using similar arguments, we obtain that $\widehat{P}$-a.s. on $\{\tau_i \leq T\}$

$$(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]), \quad P^*(t) = P^*(\widehat{\Gamma}(\tau_i)) = \widehat{P}(\tau_i) \tag{5.21}$$

and from (3.10) and (3.12)

$$\begin{aligned} (\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]), \quad \xi^*(t) &= \xi^*(\widehat{\Gamma}(\tau_i-)) + \int_{\widehat{\Gamma}(\tau_i-)}^t \theta^*(s)B(\eta^*(s))\alpha^*(s)ds \\ &= \widehat{x}(\tau_i-) + B(\tau_i)\widehat{u}(\tau_i)\big(t-\widehat{\Gamma}(\tau_i-)\big). \end{aligned} \tag{5.22}$$

Therefore, combining (1.3), (1.5), (5.15), and (5.18), we obtain that $\forall t$ in $[\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)])$,

$$\begin{aligned} H(\eta^*(t), \xi^*(t), \alpha, \theta, p^{*1}(t), p^{*2}(t), p^{*3}(t), r^*(t), P^*(t), \psi) &= (1-\theta)\Bigg\{-\psi_1 - 2T\psi_3 \\ &+ \psi_2\Big[p^{*1}(t) + A\big(\tau_i, \xi^*(t)\big)^\top p^{*2}(t) + \frac{1}{2}\operatorname{tr}\Big[D\big(\tau_i, \xi^*(t)\big)^\top P^*(t) D\big(\tau_i, \xi^*(t)\big)\Big]\Big]\Bigg\} \\ &+ \psi_2\theta\big[B(\tau_i)\alpha\big]^\top p^{*2}(t) + \psi_2\theta|\alpha|p^{*3}(t) \end{aligned}$$

$\widehat{P}$-a.s. on $\{\tau_i \leq T\}$. Finally, the result follows using the previous equation and (5.14), (5.20), (5.21), and (5.22). □

PROPOSITION 5.4. *Let $\psi \in S_1(\mathbb{R}^3)$ such that the variational inequality* (4.18) *is satisfied. Then, for $(\alpha, \theta) \in B_1(K) \times [0,1]$*

$$(\forall t \in [\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]), \quad L(t, \alpha, \theta) \leq 0 \quad \widehat{P}\text{-a.s. on } \{\tau_i \leq T\}. \tag{5.23}$$

Proof. From Lemma 5.3, it follows that $L(t, \alpha, \theta)$ is continuous on $[\widehat{\Gamma}(\tau_i-), \widehat{\Gamma}(\tau_i)]$ $\widehat{P}$-a.s. on $\{\tau_i \leq T\}$. However, using Fubini's theorem and Theorem 4.5, we obtain that

$$\int_{\widehat{\Gamma}(\tau_i-)}^{\widehat{\Gamma}(\tau_i)} I_{\{(t,\widehat{\omega}) \in [0,T+M] \times \widehat{\Omega} : L(t,\alpha,\theta) > 0\}}(s, \widehat{\omega})ds = 0$$

$\widehat{P}$-a.s. on $\{\tau_i \leq T\}$, giving the result. □

COROLLARY 5.5. *Let $\psi \in S_1(\mathbb{R}^3)$ such that the variational inequality* (4.18) *is satisfied. Then, for $(\alpha, \theta) \in B_1(K) \times [0,1]$,*

$$L(\widehat{\Gamma}(t), \alpha, \theta) \leq 0 \quad \mathcal{M}_{\widehat{\Gamma}}\text{-a.s. on } [0,T] \times \widehat{\Omega}. \tag{5.24}$$

Proof. Write

$$\begin{aligned} N(\alpha, \theta) &\doteq \{(t, \widehat{\omega}) \in [0, T+M] \times \widehat{\Omega} : L(t, \alpha, \theta) > 0\}, \\ \widetilde{N}(\alpha, \theta) &\doteq \{(t, \widehat{\omega}) \in [0,T] \times \widehat{\Omega} : (\widehat{\Gamma}(t), \omega) \in N(\alpha, \theta)\}. \end{aligned}$$

Then

$$\begin{aligned}\mathcal{M}_{\widehat{\Gamma}}(\widetilde{N}(\alpha,\theta)) &= E_{\widehat{P}}\left[\int_{[0,T]} I_{\widetilde{N}(\alpha,\theta)}(s,\widehat{\omega})d\widehat{\Gamma}(s)\right]\\ &= E_{\widehat{P}}\left[\int_0^{\widehat{\Gamma}(T)} I_{\widetilde{N}(\alpha,\theta)}(\eta^*(s),\widehat{\omega})ds\right]\\ &= \lambda\otimes\widehat{P}(N^*(\alpha,\theta)),\end{aligned} \tag{5.25}$$

where $N^*(\alpha,\theta)=\{(t,\widehat{\omega})\in[0,T+M]\times\widehat{\Omega}:(\widehat{\Gamma}(\eta^*(t)),\omega)\in N(\alpha,\theta)\}$.

Now, if $(t,\widehat{\omega})\in[\overset{\infty}{\underset{i=1}{\cup}}[\![\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)]\!]]^c$, then $\widehat{\Gamma}(\eta^*(t))=t$. Consequently,

$$N^*(\alpha,\theta)\cap\left[\overset{\infty}{\underset{i=1}{\cup}}[\![\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)]\!]\right]^c\subset N(\alpha,\theta)\cap\left[\overset{\infty}{\underset{i=1}{\cup}}[\![\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)]\!]\right]^c.$$

However, from Theorem 4.5, $\lambda\otimes\widehat{P}(N(\alpha,\theta))=0$. Therefore,

$$\lambda\otimes\widehat{P}(N^*(\alpha,\theta))=\lambda\otimes\widehat{P}\Big(N^*(\alpha,\theta)\cap\overset{\infty}{\underset{i=1}{\cup}}[\![\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)]\!]\Big). \tag{5.26}$$

Furthermore, writing π for the projection of $[0,T+M]\times\widehat{\Omega}$ to $\widehat{\Omega}$, we have

$$\pi\Big(N^*(\alpha,\theta)\cap[\![\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)]\!]\Big)\subset\{\tau_i\le T\}\cap\{\widehat{\omega}\in\widehat{\Omega}:L(\widehat{\Gamma}(\tau_i),\alpha,\theta)>0\}.$$

From Proposition 5.4, it follows that $\widehat{P}(\{\tau_i\le T\}\cap\{\widehat{\omega}\in\widehat{\Omega}:L(\widehat{\Gamma}(\tau_i),\alpha,\theta)>0\})=0$, and so $\lambda\otimes\widehat{P}(N^*(\alpha,\theta)\cap\cup_{i=1}^{\infty}[\![\widehat{\Gamma}(\tau_i-),\widehat{\Gamma}(\tau_i)]\!])=0$. Using (5.26), we have $\lambda\otimes\widehat{P}(N^*(\alpha,\theta))=0$, and from (5.25), it follows that $\mathcal{M}_{\widehat{\Gamma}}(\widetilde{N}(\alpha,\theta))=0$, giving the result. □

LEMMA 5.6. *Let $(\Omega,\mathcal{F},P)$ be a complete probability space with a right continuous complete filtration $\{\mathcal{F}_t\}$ and let $\{v(t)\}$ be an $\mathbb{R}_+$-valued,* corlol, *$\{\mathcal{F}_t\}$-progressively measurable increasing process. Define*

$$v'(t)=\begin{cases}\overline{\lim}_{\substack{\epsilon\to0\\ \epsilon>0}}\dfrac{v(t+\epsilon)-v(t-\epsilon)}{2\epsilon} & \text{if the limit exists in } \mathbb{R},\\ 0 & \text{otherwise.}\end{cases}$$

Then $\{v'(t)\}$ is an $\mathbb{R}_+$-valued, $\{\mathcal{F}_t\}$-progressively measurable process.

Proof. Clearly, the $\mathbb{R}_+$-valued process $\{v'(t)\}$ exists. By using the result 14 in [12, page 141], $\{v'(t)\}$ is an $\{\mathcal{F}_t\}$-progressively measurable process. □

LEMMA 5.7. *Let $(\Omega,\mathcal{F},P)$ be a complete probability space with a right continuous complete filtration $\{\mathcal{F}_t\}$ and let $\{v(t)\}$ be an $\mathbb{R}_+$-valued,* corlol, *$\{\mathcal{F}_t\}$-progressively measurable increasing process such that $v(T)\le M$. Write $\Gamma(t)=t+v(t)$ and let $\{z(t)\}$ be an $\{\mathcal{F}_t\}$-progressively measurable process such that $v(t)=\int_{[0,t]}z(s)d\Gamma(s)$. Then $z(t)=1$ $\mathcal{M}_{v^s}$-a.s. on $[0,T]\times\Omega$, and $z(t)=\frac{v'(t)}{1+v'(t)}$ $\mathcal{M}_{v^{ac}}$-a.s. on $[0,T]\times\Omega$.*

Proof. Clearly, we have $(\forall A\in\mathcal{B}([0,T+M]))$ $dv^{ac}(A)=\int_A z(s)d\Gamma^{ac}(s)$, and so $\int_A v'(s)ds=\int_A z(s)\Gamma'(s)ds$. Now using the definition of $\{\Gamma(t)\}$ the last part of the result follows. Denote by F a set such that $\lambda\otimes P(F)=0$ and $(\forall A\in\mathcal{B}([0,T])\otimes\mathcal{F}_T)$ $\mathcal{M}_{\Gamma^s}(A)=\mathcal{M}_{\Gamma}(A\cap F)$. Note that $\mathcal{M}_{v^s}=\mathcal{M}_{\Gamma^s}$. Consequently, $\forall A\in F$,

$$E_P\left[\int_{[0,T]}I_A z(s)d\Gamma(s)\right]=E_P\left[\int_{[0,T]}I_A d\Gamma(s)\right],$$

and so $z(t)=1$ $\mathcal{M}_{v^s}$-a.s. on $[0,T]\times\Omega$, showing the last result. □

The next proposition presents some relations between the optimal singular control $\widetilde{C}$ and $\widehat{C}$. This result is important, as it gives the full generality to our maximum principle.

PROPOSITION 5.8. *Assume the existence of an optimal singular control denoted by*

$$\widetilde{C} \doteq \left(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\}, \{\widetilde{u}(t), \widetilde{v}(t)\}, \{\widehat{W}(t)\}, \{\widetilde{x}(t)\}\right)$$

such that $\{\widetilde{u}(t), \widetilde{v}(t)\}$ *is* $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$*-progressively measurable.*

Then the optimal control

$$\widehat{C} \doteq \left(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\}, \{\widehat{u}(t), \widehat{v}(t)\}, \{\widehat{W}(t)\}, \{\widetilde{x}(t)\}\right)$$

defined by (3.2) *and* (3.3) *is such that on the filtered probability space* $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\})$ *the processes* $\{\widetilde{x}(t)\}$ *and* $\{\widehat{x}(t)\}$ *are indistinguishable, and*

$$\mathcal{M}_{\widetilde{v}^s} \ll \mathcal{M}_{\widehat{v}^s}, \quad \mathcal{M}_{\widetilde{v}^{ac}} = \mathcal{M}_{\widehat{v}^{ac}}, \tag{5.27}$$

$$\widetilde{u}(t) = \widehat{u}(t), \quad \textit{and} \quad \widetilde{v}(t) = \widehat{v}(t) \quad \mathcal{M}_{\widetilde{v}^{ac}}\textit{-a.s. on } [0,T] \times \widehat{\Omega}. \tag{5.28}$$

The adjoint variables $((\{\widetilde{p}^i(t)\}, \{\widetilde{q}^i(t)\})_{i\in\mathbb{N}_3}, \{\widetilde{P}(t)\}, (\{\widetilde{Q}^j(t)\})_{j\in\mathbb{N}_m})$ *associated with* $\widetilde{C}$ *exist and are unique. Moreover, for* $i \in \mathbb{N}_3$, $j \in \mathbb{N}_m$,

$$\widetilde{p}^i(t) = \hat{p}^i(t), \quad \widetilde{q}^i(t) = \hat{q}^i(t), \quad \widetilde{P}(t) = \widehat{P}(t), \quad \textit{and} \quad \widetilde{Q}^j(t) = \widehat{Q}^j(t), \tag{5.29}$$

where $((\{\hat{p}^i(t)\}, \{\hat{q}^i(t)\})_{i\in\mathbb{N}_3}, \{\widehat{P}(t)\}, (\{\widehat{Q}^j(t)\})_{j\in\mathbb{N}_m})$ *are the adjoint variables associated with* $\widehat{C}$.

Proof. From the definition of $\{\widehat{u}(t)\}$ and $\{\widehat{v}(t)\}$ (see (3.2) and (3.3)), we obtain (5.28) and the fact that $\{\widehat{x}(t)\}$ and $\{\widetilde{x}(t)\}$ are indistinguishable. Moreover ($\forall t \in [0,T]$), $\widehat{v}(t) \geq \widetilde{v}(t)$ and $\widehat{v}^{ac}(t) = \widetilde{v}^{ac}(t)$, giving (5.27). Theorem 5.2 shows the existence of the adjoint variables $((\hat{p}^i(t), \hat{q}^i(t))_{i\in\mathbb{N}_3}, \widehat{P}(t), (\widehat{Q}^j(t))_{j\in\mathbb{N}_m})$ associated with $\widehat{C}$. Now (3.4) gives the existence of the adjoint variables $((\widetilde{p}^i(t), \widetilde{q}^i(t))_{i\in\mathbb{N}_3}, \widetilde{P}(t), (\widetilde{Q}^j(t))_{j\in\mathbb{N}_m})$ associated with $\widetilde{C}$ satisfying (5.29). As in the proof of Theorem 5.2, the uniqueness follows from Theorem 3.1 in [37], showing the result. □

We can now establish a necessary condition for any control $\widetilde{C} \in \mathfrak{C}^a$ to be optimal. In fact similarly to the maximum principle for nonsingular control problem (see, for example, [38, 41] and the references therein), we show that if a control $\widetilde{C} = (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\}, \{\widetilde{u}(t), \widetilde{v}(t)\}, \{\widehat{W}(t)\}, \{\widetilde{x}(t)\})$ is optimal, it maximizes a certain Hamiltonian a.s. with respect to the measure of Doleans-Dade generated by the optimal control $\{\widetilde{v}(t)\}$. This last property, which represents an important difference from the classical (nonsingular) control problem where the optimal control maximizes the Hamiltonian a.s. surely with respect to the product measure of the Lebesgue measure and the underlying probability, is already present in the expression of the maximum principle derived by Cadenillas and Haussmann [8] (see the last term of the inequality (66) on page 227 in [8]). Here we present a maximum principle given in terms of three variational inequalities but not in the integral from as in the work by Cadenillas and Haussmann [8]. The first two variational inequalities result from a time change of the variational inequality of the auxiliary control problem. The first one is given with respect to the measure of Doleans-Dade generated by the absolutely continuous

part of $\{\widetilde{v}(t)\}$ (see (5.30)). Note that it does not depend directly on $\{\widetilde{v}(t)\}$ but on the derivative $\frac{d\widetilde{v}}{dt}(t)$. The second one (see (5.32)) is given with respect to the measure of Doleans-Dade generated by the singular part of $\{\widetilde{v}_t\}$. The last inequality can be interpreted as a necessary condition for the size of the jumps of $\{\widetilde{v}_t\}$ (see (5.33)). It is different from the first two ones because it is not obtained from a time change of the variational inequality of the auxiliary control problem (see the proof of the next theorem).

THEOREM 5.9. *Assume the existence of an optimal singular control denoted by*

$$\widetilde{C} \doteq \Big(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}, \{\widehat{\mathcal{F}}_t\}, \{\widetilde{u}(t), \widetilde{v}(t)\}, \{\widehat{W}(t)\}, \{\widetilde{x}(t)\}\Big)$$

such that $\{\widetilde{u}(t), \widetilde{v}(t)\}$ *is* $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$*-progressively measurable. Denote by* $\{\widetilde{\tau}_i\}_{i\in\mathbb{N}^*}$ *the sequence of* $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$*-stopping times which exhausts the jumps of* $\{\widetilde{v}(t)\}$*. Then there exist* $\psi \in S_1(\mathbb{R}^3)$ *such that* $\forall (u,z) \in B_1(K)\times[0,1]$

$$\begin{aligned}
&[\widetilde{z}(t)-z]\Big\{-\psi_1 - 2T\psi_3 + \psi_2\Big[\widetilde{p}^1(t) + \frac{1}{2}\operatorname{tr}\big[D(t,\widetilde{x}(t))^\top \widetilde{P}(t) D(t,\widetilde{x}(t))\big] \\
&\qquad + A(t,\widetilde{x}(t))^\top \widetilde{p}^2(t)\Big]\Big\} \\
&\qquad + \psi_2\Big[\big[B(t)[zu-\widetilde{z}(t)\widetilde{u}(t)]\big]^\top \widetilde{p}^2(t) + [z|u| - \widetilde{z}(t)|\widetilde{u}(t)|]\widetilde{p}^3(t) \\
&\qquad + \operatorname{tr}\big[D(t,\widetilde{x}(t))^\top \widetilde{r}(t)\big]\Big(\sqrt{1-z} - \sqrt{1-\widetilde{z}(t)}\Big)\Big] \le 0
\end{aligned} \tag{5.30}$$

$\mathcal{M}_{\widetilde{v}^{ac}}$*-a.s. on* $[0,T]\times\widehat{\Omega}$*, where*

$$\widetilde{r}(t) \doteq \Big[\widetilde{q}^2(t) - \widetilde{P}(t)D(t,\widetilde{x}(t))\Big]\sqrt{1-\widetilde{z}(t)}, \qquad \widetilde{z}(t) \doteq \frac{\widetilde{v}'(t)}{1+\widetilde{v}'(t)}, \tag{5.31}$$

and

$$\begin{aligned}
&\psi_2\Big[[zu-\widetilde{u}(t)]^\top B(t)^\top \widetilde{p}^2(t) + [z|u| - |\widetilde{u}(t)|]\widetilde{p}^3(t)\Big] + (1-z)\Big\{-\psi_1 - 2T\psi_3 \\
&\quad + \psi_2\Big[\widetilde{p}^1(t) + \frac{1}{2}\operatorname{tr}\Big[D(t,\widetilde{x}(t))^\top \widetilde{P}(t) D(t,\widetilde{x}(t))\Big] + A(t,\widetilde{x}(t))^\top \widetilde{p}^2(t)\Big]\Big\} \le 0
\end{aligned} \tag{5.32}$$

$\mathcal{M}_{\widetilde{v}^s}$*-a.s. on* $[0,T]\times\widehat{\Omega}$*, and* $\forall i\in\mathbb{N}$ *and* $\gamma\in[0,1]$

$$\begin{aligned}
&\psi_2\Big[\big[B(\widetilde{\tau}_i)(zu-\widetilde{u}(\widetilde{\tau}_i))\big]^\top \widetilde{p}^2(\widetilde{\tau}_i) + (z|u| - |\widetilde{u}(\widetilde{\tau}_i)|)\widetilde{p}^3(\widetilde{\tau}_i)\Big] + (1-z)\Big\{-\psi_1 - 2T\psi_3 \\
&\qquad + \psi_2\Big[\widetilde{p}^1(\widetilde{\tau}_i-) - \gamma\Big[B_t(\widetilde{\tau}_i)\widetilde{u}(\widetilde{\tau}_i)\Delta\widetilde{v}(\widetilde{\tau}_i)\Big]^\top \widetilde{p}^2(\widetilde{\tau}_i) \\
&\qquad + A\Big(\widetilde{\tau}_i, \widetilde{x}(\widetilde{\tau}_i-) + \gamma B(\widetilde{\tau}_i)\widetilde{u}(\widetilde{\tau}_i)\Delta\widetilde{v}(\widetilde{\tau}_i)\Big)^\top \widetilde{p}^2(\widetilde{\tau}_i) \\
&\qquad + \frac{1}{2}\operatorname{tr}\Big[D\Big(\widetilde{\tau}_i, \widetilde{x}(\widetilde{\tau}_i-) + \gamma B(\widetilde{\tau}_i)\widetilde{u}(\widetilde{\tau}_i)\Delta\widetilde{v}(\widetilde{\tau}_i)\Big)^\top \widetilde{P}(\widetilde{\tau}_i) \\
&\qquad \times D\Big(\widetilde{\tau}_i, \widetilde{x}(\widetilde{\tau}_i-) + \gamma B(\widetilde{\tau}_i)\widetilde{u}(\widetilde{\tau}_i)\Delta\widetilde{v}(\widetilde{\tau}_i)\Big)\Big]\Big]\Big\} \le 0
\end{aligned} \tag{5.33}$$

$\widehat{P}$-a.s. on $\{\widetilde{\tau}_i \leq T\}$, where $((\{\widetilde{p}^i(t)\}, \{\widetilde{q}^i(t)\})_{i\in\mathbb{N}_3}, \{\widetilde{P}(t)\}, (\{\widetilde{Q}^j(t)\})_{j\in\mathbb{N}_m})$ are the adjoint variables associated with $\widetilde{C}$.

Proof. From Proposition 3.2, it follows that the control $\widehat{C}$ defined by (3.1), (3.2), and (3.3) is such that $\widehat{v}(T) = M$. Therefore, we can apply the results of this section to $\widehat{C}$ and in particular Corollary 5.5. The stochastic maximum principle can now be expressed in the following form: there exist $\psi \in S_1(\mathbb{R}^3)$ such that $\forall (u,z) \in B_1(K) \times [0,1]$

$$
\begin{aligned}
&H(\eta^*(\widehat{\Gamma}(t)), \xi^*(\widehat{\Gamma}(t)), \alpha, \theta, p^{*1}(\widehat{\Gamma}(t)), p^{*2}(\widehat{\Gamma}(t)), p^{*3}(\widehat{\Gamma}(t)), r^*(\widehat{\Gamma}(t)), P^*(\widehat{\Gamma}(t)), \psi)\\
&-H(\eta^*(\widehat{\Gamma}(t)), \xi^*(\widehat{\Gamma}(t)), \alpha^*(\widehat{\Gamma}(t)), \theta^*(\widehat{\Gamma}(t)), p^{*1}(\widehat{\Gamma}(t)), p^{*2}(\widehat{\Gamma}(t)), p^{*3}(\widehat{\Gamma}(t)), r^*(\widehat{\Gamma}(t)),\\
&P^*(\widehat{\Gamma}(t)), \psi) \leq 0 \qquad (5.34)
\end{aligned}
$$

$\mathcal{M}_{\widehat{\Gamma}}$-a.s. on $[0,T] \times \widehat{\Omega}$. Since $\{\eta^*(t)\}$ is the right inverse of $\{\widehat{\Gamma}(t)\}$ which is strictly increasing, we have that $\eta^*(\widehat{\Gamma}(t)) = t$. Moreover, combining (3.6), (3.12), (4.19), and the definition of $\{\hat{p}(t)\}$, $\{\hat{q}(t)\}$, $\{\widehat{P}(t)\}$ (see (5.4)), we obtain that $r^*(\widehat{\Gamma}(t)) = \hat{q}^2(t) - \widehat{P}(t)D(t,\widehat{x}(t))\sqrt{1-\widehat{z}(t)}$. Defining $\hat{r}(t) = r^*(\widehat{\Gamma}(t))$ it follows that $(\forall (u,z) \in B_1(K) \times [0,1])$

$$
\begin{aligned}
&H(t, \widehat{x}(t), u, z, \hat{p}^1(t), \hat{p}^2(t), \hat{p}^3(t), \hat{r}(t), \widehat{P}(t), \psi)\\
&\qquad - H(t, \widehat{x}(t), \widehat{u}(t), \widehat{z}(t), \hat{p}^1(t), \hat{p}^2(t), \hat{p}^3(t), \hat{r}(t), \widehat{P}(t), \psi) \leq 0
\end{aligned}
$$

$\mathcal{M}_{\widehat{\Gamma}}$-a.s. on $[0,T] \times \widehat{\Omega}$.

From the definition of $\{\widehat{\Gamma}(t)\}$, we obtain that $\mathcal{M}_{\widehat{v}} \ll \mathcal{M}_{\widehat{\Gamma}}$, and so by Proposition 5.8 it follows that $(\forall (u,z) \in B_1(K) \times [0,1])$

$$
\begin{aligned}
&H(t, \widetilde{x}(t), u, z, \widetilde{p}^1(t), \widetilde{p}^2(t), \widetilde{p}^3(t), \hat{r}(t), \widetilde{P}(t), \psi)\\
&\qquad - H(t, \widetilde{x}(t), \widetilde{u}(t), \widetilde{z}(t), \widetilde{p}^1(t), \widetilde{p}^2(t), \widetilde{p}^3(t), \hat{r}(t), \widetilde{P}(t), \psi) \leq 0 \qquad (5.35)
\end{aligned}
$$

$\mathcal{M}_{\widehat{v}}$-a.s. on $[0,T] \times \widehat{\Omega}$.

From (3.4), it follows that $\widetilde{v}'(t) = \widetilde{v}'(t)$ $\mathcal{M}_{\widetilde{v}^{ac}}$-a.s. on $[0,T] \times \widehat{\Omega}$. Combining Lemma 5.7 and (5.27) and (5.28), we have that $\widehat{z}(t) = \frac{\widetilde{v}'(t)}{1+\widetilde{v}'(t)}$ $\mathcal{M}_{\widetilde{v}^{ac}}$-a.s. on $[0,T] \times \widehat{\Omega}$, and according to the definition of $\widetilde{r}(t)$ (see (5.31)), it implies that $\hat{r}(t) = \widetilde{r}(t)$ $\mathcal{M}_{\widetilde{v}^{ac}}$-a.s. on $[0,T] \times \widehat{\Omega}$. Therefore, we obtain

$$
\begin{aligned}
&H(t, \widetilde{x}(t), u, z, \widetilde{p}^1(t), \widetilde{p}^2(t), \widetilde{p}^3(t), \widetilde{r}(t), \widetilde{P}(t), \psi)\\
&\qquad - H(t, \widetilde{x}(t), \widetilde{u}(t), \widetilde{z}(t), \widetilde{p}^1(t), \widetilde{p}^2(t), \widetilde{p}^3(t), \widetilde{r}(t), \widetilde{P}(t), \psi) \leq 0
\end{aligned}
$$

$\mathcal{M}_{\widetilde{v}^{ac}}$-a.s. on $[0,T] \times \widehat{\Omega}$. With the definition of H (see (1.5)), simple calculations give the first inequality (5.30).

From Lemma 5.7, we have that $\widehat{z}(t) = 1$ and $\hat{r}(t) = 0$ $\mathcal{M}_{\widehat{v}^s}$-a.s. on $[0,T] \times \widehat{\Omega}$. Consequently, using (5.27), (5.35) reduces to $(\forall (u,z) \in B_1(K) \times [0,1])$

$$
\begin{aligned}
&\psi_2\Big[[zu - \widehat{u}(t)]^\top B(t)^\top \widetilde{p}^2(t) + [z|u| - |\widehat{u}(t)|]\widetilde{p}^3(t)\Big] + (1-z)\Big\{-\psi_1 - 2T\psi_3\\
&+ \psi_2\left[\widetilde{p}^1(t) + \frac{1}{2}\operatorname{tr}\big[D(t,\widetilde{x}(t))^\top \widetilde{P}(t) D(t,\widetilde{x}(t))\big] + A(t,\widetilde{x}(t))^\top \widetilde{p}^2(t)\right]\Big\} \leq 0 \qquad (5.36)
\end{aligned}
$$

$\mathcal{M}_{\widetilde{v}^s}$-a.s. on $[0,T]\times\widehat{\Omega}$. Since the left-hand side of the previous inequality is continuous in (z,u), there exists a set $\mathcal{V}\in\mathcal{B}([0,T])\otimes\widehat{\mathcal{F}}_T^{\widehat{W}}$ such that $\mathcal{M}_{\widetilde{v}^s}(\mathcal{V})=0$, and $\forall(t,\widehat{\omega})\in[0,T]\times\widehat{\Omega}-\mathcal{V}$,

$$\psi_2\Big[[zu-\widehat{u}(t)]^\top B(t)^\top\widetilde{p}^2(t)+[z|u|-|\widehat{u}(t)|]\widetilde{p}^3(t)\Big]+(1-z)\Big\{-\psi_1-2T\psi_3$$

$$+\psi_2\left[\widetilde{p}^1(t)+\frac{1}{2}\operatorname{tr}\Big[D(t,\widetilde{x}(t))^\top\widetilde{P}(t)D(t,\widetilde{x}(t))\Big]+A(t,\widetilde{x}(t))^\top\widetilde{p}^2(t)\right]\Big\}\le 0 \tag{5.37}$$

$\forall(u,z)\in B_1(K)\times[0,1]$.

From the definition of $\widehat{u}(t)$ (see (3.3)), we can write $\widehat{u}(t)=\widetilde{u}(t)\widetilde{c}(t)$, where $\widetilde{c}(t)=I_{[\![0,T[\![}+[\frac{\widetilde{v}(T)-\widetilde{v}(T-)}{M-\widetilde{v}(T-)}I_{[T,+\infty[\times\{\widetilde{v}(T)<M\}}+I_{[T,+\infty[\times\{\widetilde{v}(T)=M\}}]$.

Note that $0\le\widetilde{c}(t)\le 1$.

Consequently, from (5.37) we obtain

$$\psi_2\Big[[zu-\widetilde{c}(t)\widetilde{u}(t)]^\top B(t)^\top\widetilde{p}^2(t)+[z|u|-\widetilde{c}(t)|\widetilde{u}(t)|]\widetilde{p}^3(t)\Big]+(1-z)\Big\{-\psi_1-2T\psi_3$$

$$+\psi_2\left[\widetilde{p}^1(t)+\frac{1}{2}\operatorname{tr}\big[D(t,\widetilde{x}(t))^\top\widetilde{P}(t)D(t,\widetilde{x}(t))\big]+A(t,\widetilde{x}(t))^\top\widetilde{p}^2(t)\right]\Big\}\le 0 \tag{5.38}$$

$\forall(u,z)\in B_1(K)\times[0,1]$. Taking $z=1$ and $u=0$ in the previous inequality, it follows that

$$-\psi_2\widetilde{c}(t)\Big[\widetilde{u}(t)^\top B(t,\widetilde{x}(t))^\top\widetilde{p}^2(t)+|\widetilde{u}(t)|\widetilde{p}^3(t)\Big]\le 0.$$

However, for $(t,\widehat{\omega})\in[0,T]\times\widehat{\Omega}-[T,T]\times\{\Delta\widetilde{v}(T)=0\}$, $\widetilde{c}(t)>0$, and so $\widetilde{c}(t)>0$ $\mathcal{M}_{\widetilde{v}^s}$-a.s. on $[0,T]\times\widehat{\Omega}$ since $\mathcal{M}_{\widetilde{v}^s}([T,T]\times\{\Delta\widetilde{v}(T)=0\})=0$.

Consequently,

$$-\psi_2\Big[\widetilde{u}(t)^\top B(t,\widetilde{x}(t))^\top\widetilde{p}^2(t)+|\widetilde{u}(t)|\widetilde{p}^3(t)\Big]$$

$$\le-\psi_2\widetilde{c}(t)\Big[\widetilde{u}(t)^\top B(t,\widetilde{x}(t))^\top\widetilde{p}^2(t)+|\widetilde{u}(t)|\widetilde{p}^3(t)\Big]. \tag{5.39}$$

Combining (5.38) and (5.39), we obtain the inequality (5.32).

From the definition of $\{\widehat{v}(t)\}$ (see (3.2)), we have that $\{\Delta\widehat{v}(T)\neq 0\}=\{\Delta\widetilde{v}(T)\neq 0\}\cup[\{\Delta\widetilde{v}(T)=0\}\cap\{\widetilde{v}(T)<M\}]$. Clearly, $\{\Delta\widetilde{v}(T)=0\}\cap\{\widetilde{v}(T)<M\}\in\widehat{\mathcal{F}}_T^{\widehat{W}}$, and so from Lemma 5.11 in [19], $T_{\{\Delta\widetilde{v}(T)=0\}\cap\{\widetilde{v}(T)<M\}}$ is an $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$-stopping time. Moreover, $[\![\widetilde{\tau}_i]\!]\cap[\![T_{\{\Delta\widetilde{v}(T)=0\}\cap\{\widetilde{v}(T)<M\}}]\!]=\emptyset$.

The sequence $\{\{\widetilde{\tau}_i\}_{i\in\mathbb{N}^*},T_{\{\Delta\widetilde{v}(T)=0\}\cap\{\widetilde{v}(T)<M\}}\}$ of $\{\widehat{\mathcal{F}}_t^{\widehat{W}}\}$-stopping times exhausts the jumps of $\{\widehat{v}(t)\}$. Denote by $\{\tau_i\}_{i\in\mathbb{N}^*}$ the sequence $\{\{\widetilde{\tau}_i\}_{i\in\mathbb{N}^*},T_{\{\Delta\widetilde{v}(T)=0\}\cap\{\widetilde{v}(T)<M\}}\}$. We can apply Proposition 5.4 to the optimal control $\widehat{C}$. Taking $t=\widehat{\Gamma}(\tau_i-)+\gamma\Delta\widehat{\Gamma}(\tau_i)$

for $\gamma \in [0,1]$, we have for $(u,z) \in B_1(K) \times [0,1]$

$$\psi_2\Big[\big[B(\tau_i)(zu-\widehat{u}(\tau_i))\big]^\top \hat{p}^2(\tau_i) + (z|u| - |\widehat{u}(\tau_i)|)\hat{p}^3(\tau_i)\Big] + (1-z)\Big\{-\psi_1 - 2T\psi_3$$
$$+ \psi_2\Big[\hat{p}^1(\tau_i-) - \gamma\,[B_t(\tau_i)\widehat{u}(\tau_i)]^\top \Delta\widehat{\Gamma}(\tau_i)\hat{p}^2(\tau_i)$$
$$+ A\Big(\tau_i, \widehat{x}(\tau_i-) + \gamma B(\tau_i)\widehat{u}(\tau_i)\Delta\widehat{\Gamma}(\tau_i)\Big)^\top \hat{p}^2(\tau_i)$$
$$+ \frac{1}{2}\operatorname{tr}\Big[D\Big(\tau_i, \widehat{x}(\tau_i-) + \gamma B(\tau_i)\widehat{u}(\tau_i)\Delta\widehat{\Gamma}(\tau_i)\Big)^\top \widehat{P}(\tau_i)$$
$$\times D\Big(\tau_i, \widehat{x}(\tau_i-) + \gamma B(\tau_i)\widehat{u}(\tau_i)\Delta\widehat{\Gamma}(\tau_i)\Big)\Big]\Big\} \le 0$$

$\widehat{P}$-a.s. on $\{\tau_i \le T\}$. However, from the definition of $\widehat{\Gamma}(t)$ and using (3.2) and (3.3) we have $\widehat{u}(\tau_i)\Delta\widehat{\Gamma}(\tau_i) = \widehat{u}(\tau_i)\Delta\widehat{v}(\tau_i) = \widetilde{u}(\tau_i)\Delta\widetilde{v}(\tau_i)$. Finally, the definition of τ_i and Proposition 5.8 give the last inequality (5.33). □

6. Example. In this section, it is shown how the set of necessary conditions given by the inequalities (5.30), (5.32), and (5.33) can be derived and used for a simple worked-out example. Let us consider the following scalar system:

$$dx(t) = Ax(t) + u(t)dv(t) + dW(t), \tag{6.1}$$

with $A > 0$ and $x(0-) = 1$. The aim is to minimize the cost given by $J[C] = E[x(T)]$, subject to $v(T) \le M$ with $M > 0$. It is easy to show that the control defined by

$$\widetilde{v}(t) = \begin{cases} 0 \text{ for } t < 0, \\ M \text{ for } t \ge 0 \end{cases} \quad \text{and} \quad \widetilde{u}(t) = \begin{cases} -1 \text{ for } t = 0, \\ 0 \text{ for } t > 0 \end{cases} \tag{6.2}$$

is an optimal control.

Consequently, we have

$$\widetilde{p}^1(t) = \widetilde{q}^1(t) = 0, \quad \widetilde{p}^3(t) = \widetilde{q}^3(t) = 0, \quad \widetilde{P}(t) = \widetilde{Q}(t) = 0, \tag{6.3}$$

and

$$d\widetilde{p}^2(t) = -A\widetilde{p}^2(t)dt, \text{ with } \widetilde{p}^2(T) = -1, \text{ and } \widetilde{q}^2(t) = 0. \tag{6.4}$$

Therefore, it follows that $\widetilde{p}^2(t) = -e^{A(T-t)}$. Clearly, we have $\mathcal{M}_{\widetilde{v}^{ac}} = 0$. Since $\mathcal{M}_{\widetilde{v}^s}(.) = I_0(.)$, it is easy to show that (5.32) and (5.33) reduce to

$$\psi_2\Big(zu - \widetilde{u}(0)\Big)\widetilde{p}^2(0) + \Big(1-z\Big)\Big[-\psi_1 - 2T\psi_3$$
$$+ \psi_2 A\Big\{\widetilde{x}(0-) + \gamma\widetilde{u}(0)\Delta\widetilde{v}(0)\Big\}\widetilde{p}^2(0)\Big] \le 0 \tag{6.5}$$

for $u \in [-1,1]$, $z \in [0,1]$, and $\gamma \in [0,1]$.

Consequently, applying Theorem 5.9 to this example, it follows that there exist $\psi \in S_1(\mathbb{R}^3)$ such that $\forall u \in [-1,1]$, $z \in [0,1]$, and $\gamma \in [0,1]$, the previous inequality is satisfied. The variational inequality (6.5) becomes

$$-\psi_2\Big(zu+1\Big)e^{AT} + \Big(1-z\Big)\Big[-\psi_1 - 2T\psi_3 - \psi_2 A\Big\{1-\gamma M\Big\}e^{AT}\Big] \le 0.$$

When $M \leq 1$, a straightforward calculation shows that for any $\psi_2 \in [0,1]$, there exist $\psi_1 \in [-1,1]$ and $\psi_3 \in [-1,1]$ satisfying $\psi_1 + 2T\psi_3 = \sqrt{1-\psi_2^2}$ such that the variational inequality is satisfied $\forall u \in [-1,1]$, $z \in [0,1]$, and $\gamma \in [0,1]$.

When $M > 1$, by choosing ψ_2 such that $0 < \psi_2 \leq 1$, and $A\{M-1\}e^{AT} \leq \frac{\sqrt{1-\psi_2^2}}{\psi_2}$, and again $\psi_1 + 2T\psi_3 = \sqrt{1-\psi_2^2}$, it follows that the variational inequality is satisfied $\forall u \in [-1,1]$, $z \in [0,1]$, and $\gamma \in [0,1]$.

7. Conclusion. Our work can be generalized in several directions: a running cost can be added to the definition of $J[C]$ (with no convexity hypothesis) and a classical control process can be added in the dynamic of the state (for example in A, D and into the running cost if it exists) as in [8]. Soft constraints with the same form of the cost G may also be added to the model (see [23, page 855] for constraints of these types in classical control problems).

REFERENCES

[1] L.H.R. Alvarez, *A class of solvable singular stochastic control problems*, Stochastics Stochastics Rep., 67 (1999), pp. 83–122.

[2] L.H.R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[3] A. Bensoussan, *Lectures on stochastic control*, in Nonlinear Filtering and Stochastic Control, Lecture Notes in Math. 972, Springer-Verlag, Berlin, 1981, pp. 1–62.

[4] J.-M. Bismut, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62–78.

[5] F. Boetius, *Bounded variation singular stochastic control and associated Dynkin game*, in Mathematical Finance, Trends in Mathematics, M. Kohlmann and S. Tang, eds., Birkhäuser, Basel, 2001, pp. 111–120.

[6] F. Boetius, *Singular Stochastic Control and Its Relations to Dynkin Game and Entry-Exit Problems*, Ph.D. thesis, University of Konstanz, Konstanz, Germany, 2001.

[7] F. Boetius and M. Kohlmann, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.

[8] A. Cadenillas and U. Haussmann, *The stochastic maximum principle for a singular control problem*, Stochastics Stochastics Rep., 49 (1994), pp. 211–237.

[9] A. Cadenillas and I. Karatzas, *The stochastic maximum principle for linear, convex optimal control with random coefficients*, SIAM J. Control Optim., 33 (1995), pp. 590–624.

[10] P.-L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1995), pp. 858–899.

[11] M.H.A. Davis and M. Zervos, *A pair of explicitly solvable singular stochastic control problems*, Appl. Math. Optim., 38 (1998), pp. 327–352.

[12] C. Dellacherie and P.A. Meyer, *Probabilités et potentiel. Chapitre* I *à* V, Hermann, Paris, 1975.

[13] C. Dellacherie and P.A. Meyer, *Probabilités et potentiel. Théorie des Martingales. Chapitre* V *à* VIII, Hermann, Paris, 1980.

[14] F. Dufour and B. Miller, *Singular stochastic control problems*, SIAM J. Control Optim., 43 (2004), pp. 708–730.

[15] N. El Karoui and S.-J. Huang, *A general result of existence and uniqueness of backward stochastic differential equations*, in Backward Stochastic Differential Equations, Pitman Res. Notes Math. Ser. 364, N. El Karoui and L. Mazliak, eds., Longman, Harlow, UK, 1997, pp. 27–36.

[16] N. El Karoui and I. Karatzas, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.

[17] N. El Karoui and I. Karatzas, *A new approach to the Skorohod problem, and its applications*, Stochastics Stochastics Rep., 34 (1991), pp. 57–82.

[18] N. El Karoui, H. Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

[19] R.J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, New York, 1982.

[20] R.J. Elliott, *The optimal control of diffusions*, Appl. Math. Optim., 22 (1990), pp. 229–240.

[21] R.J. Elliott and M. Kohlmann, *The second order minimum principle and adjoint process*, Stochastics Stochastics Rep., 46 (1994), pp. 25–39.
[22] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[23] U.G. Haussmann and J.P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.
[24] U.G. Haussmann and W. Suo, *Singular optimal stochastic controls* I. *Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.
[25] U.G. Haussmann and W. Suo, *Singular optimal stochastic controls* II. *Dynamic programming*, SIAM J. Control Optim., 33 (1995), pp. 937–959.
[26] U.G. Haussmann, *A Stochastic Maximum Principle for Optimal Control of Diffusions*, Pitman Res. Notes Math. Ser. 151, Longman Scientific and Technical, Harlow, UK, John Wiley Sons, New York, 1986.
[27] J. Jacod, *Calcul stochastique et problèmes de martingales*, Lecture Notes in Math. 714, Springer-Verlag, Berlin, 1979.
[28] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
[29] I. Karatzas, *Probabilistic aspects of finite-fuel stochastic control*, Proc. Nat. Acad. Sci. U.S.A., 82 (1985), pp. 5579–5581.
[30] I. Karatzas, D. Ocone, H. Wang, and M. Zervos, *Finite-fuel singular control with discretionary stopping*, Stochastics Stochastic Rep., 71 (2000), pp. 1–50.
[31] I. Karatzas and S.E. Shreve, *Connections between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[32] I. Karatzas and S.E. Shreve, *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
[33] I. Karatzas and S. Shreve, *Equivalent models for finite-fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.
[34] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.
[35] L. Kruk, *Optimal policies for n-dimensional singular stochastic control problems. Part* II: *The radially symmetric case. Ergodic control*, SIAM J. Control Optim., 39 (2000), pp. 635–659.
[36] H.J. Kushner, *On the stochastic maximum principle: Fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.
[37] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.
[38] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966–979.
[39] P. Protter, *Stochastic Integration and Differential Equations*, Springer-Verlag, New York, 1990.
[40] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Springer-Verlag, New York, 1999.
[41] J. Yong and X. Zhou, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[42] X.Y. Zhou, *A unified treatement of maximum principle and dynaamic programming in stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.
[43] H. Zhu, *Generalized solution in singular control: The nondegenerate problem*, Appl. Math. Optim., 25 (1992), pp. 225–245.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 699–721

OPTIMAL CONTROL OF A SEMILINEAR PDE WITH NONLOCAL RADIATION INTERFACE CONDITIONS*

C. MEYER†, P. PHILIP‡, AND F. TRÖLTZSCH†

Abstract. We consider a control constrained optimal control problem governed by a semilinear elliptic equation with nonlocal interface conditions. These conditions occur during the modeling of diffuse-gray conductive-radiative heat transfer. The problem arises when attempting to optimize the temperature gradient within crystal growth by the physical vapor transport (PVT) method. Based on a minimum principle for the semilinear equation as well as L^∞-estimates for the weak solution, we establish the existence of an optimal solution as well as necessary optimality conditions. The theoretical results are illustrated by numerical computations.

Key words. optimal control, semilinear elliptic equations, nonlocal interface conditions, boundedness of solutions

AMS subject classifications. 49K20, 35J65, 49J20, 80M50

DOI. 10.1137/040617753

1. Introduction.

In this paper, we investigate an optimal control problem related to the sublimation growth of silicon carbide (SiC) single crystals by the physical vapor transport (PVT) method. The semiconductor material SiC is used in numerous industrial applications, e.g., the production of optoelectronic devices such as blue and green light emitting diodes (LEDs) and lasers. For the PVT method, polycrystalline SiC powder is placed in a low-pressure inert gas atmosphere at the bottom of a cavity inside a graphite crucible. The crucible is heated to a temperature between 2000 and 3000 K by induction. Due to the high temperature and the low pressure, the SiC powder sublimates and crystallizes at a single-crystalline SiC seed located at the cooled top of the cavity, such that the single crystal grows into the reaction chamber. See [9, 10] for more details.

Here, we neglect the electromagnetic induction problem and focus on the conductive-radiative heat transfer in the growth apparatus. Therefore, we consider a simplified setup of the growth apparatus, shown in Figure 1.1, where Ω_s denotes the domain of the solid graphite crucible, whereas Ω_g is the domain of gas phase inside. A very important determining factor for the crystal's quality and growth rate is the temperature distribution in the gas phase and, especially, the temperature gradient close to the surface of the growing crystal [16]. Since we do not consider the electromagnetic induction, we will optimize the temperature gradient in the gas phase Ω_g by directly controlling the heat source u in Ω_s.

The temperature y inside the growth apparatus is the solution of the conductive-radiative heat transfer problem in the growth apparatus. Accounting for radiative contributions is essential, owing to the high temperatures. Thus, the problem

*Received by the editors October 27, 2004; accepted for publication (in revised form) December 8, 2005; published electronically May 12, 2006. This research was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.
http://www.siam.org/journals/sicon/45-2/61775.html
†Institut für Mathematik, Technische Universität Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany (cmeyer@math.tu-berlin.de, troeltz@math.tu-berlin.de).
‡Institute for Mathematics and Its Applications (IMA), University of Minnesota, 400 Lind Hall, 207 Church Street S.E., Minneapolis, MN 55455-0436 (philip@ima.umn.edu). This author was supported by the Weierstrass Institute for Applied Analysis and Stochastics (WIAS) in Berlin.

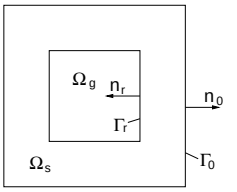


FIG. 1.1. *Two-dimensional section through an exemplary domain for nonlocal radiative heat transfer.*

is described by the stationary heat equation with radiation interface and boundary conditions on Γ_{r} and Γ_0, respectively. We take Ω_{s} to be entirely opaque, whereas Ω_{g} represents a transparent medium which does not interact with radiation. Furthermore, the radiative surfaces $\Gamma_0 := \partial\Omega$ and $\Gamma_{\mathrm{r}} := \overline{\Omega}_{\mathrm{s}} \cap \overline{\Omega}_{\mathrm{g}}$ are presumed to be diffuse-gray; i.e., the emissivity ε is independent of both the direction and the wavelength of the radiation. In particular, the local radiative heat exchange on Γ_0 can be modeled by the Boltzmann radiation condition with an external temperature y_0. Due to the heat exchange between points on Γ_{r}, we obtain an additional radiative heat flux on Γ_{r}, denoted by q_{r}.

In addition to the stationary semilinear heat equation with radiation interface and boundary conditions, we consider box constraints for the control function u. Thus, the optimal control problem, considered here, reads as follows:

$$
\text{(P)} \quad \begin{cases} \text{Minimize} \quad J(y,u) := \dfrac{1}{2}\displaystyle\int_{\Omega_{\mathrm{g}}} |\nabla y - z|^2\,dx + \dfrac{\nu}{2}\int_{\Omega_{\mathrm{s}}} u^2\,dx \\ \text{subject to} \qquad -\mathrm{div}(\kappa_{\mathrm{s}}\nabla y) = u \quad \text{in } \Omega_{\mathrm{s}}, \\ \qquad\qquad -\mathrm{div}(\kappa_{\mathrm{g}}\nabla y) = 0 \quad \text{in } \Omega_{\mathrm{g}}, \\ \qquad\qquad \kappa_{\mathrm{g}}\left(\dfrac{\partial y}{\partial n_r}\right)_{\mathrm{g}} - \kappa_{\mathrm{s}}\left(\dfrac{\partial y}{\partial n_r}\right)_{\mathrm{s}} = q_{\mathrm{r}} \quad \text{on } \Gamma_{\mathrm{r}}, \\ \qquad\qquad \kappa_{\mathrm{s}}\dfrac{\partial y}{\partial n_0} + \varepsilon\sigma\,|y|^3 y = \varepsilon\sigma\, y_0^4 \text{on } \Gamma_0, \\ \qquad \text{and} \quad u_a \le u(x) \le u_b \quad \text{a.e. in } \Omega_{\mathrm{s}}, \end{cases}
$$

where n_0 is the outward unit normal on Γ_0, and n_{r} is the unit normal on Γ_{r} facing outward with respect to Ω_{s} (cf. Figure 1.1). Furthermore, z denotes the desired temperature gradient and $\nu > 0$ is a Tikhonov regularization parameter. In the state equation, σ represents the Boltzmann radiation constant, and κ_{s}, κ_{g} denote the thermal conductivities in Ω_{s}, Ω_{g}, respectively.

In contrast to the boundary condition on Γ_0, the radiative heat transfer on Γ_{r} is nonlocal. The corresponding mathematical model used here is described in detail, e.g., in [15, 18]. It provides the additional radiative heat flux q_{r} on Γ_{r} given by

$$
q_{\mathrm{r}} = (I-K)(I-(1-\varepsilon)K)^{-1}\varepsilon\,\sigma|y|^3 y := G\,\sigma|y|^3 y, \tag{1.1}
$$

where K is an integral operator representing the irradiation on Γ_{r}. The nonlocal operators K and G will be specified in section 3.1. The nonlocal radiation on Γ_{r} represents the main characteristic of our problem, since the nonlinearity in the state equation in (P) is in general not monotone due to nonpositivity of G (see section 3 and [18]).

The state equation in (P) is of semilinear elliptic type. Therefore, the optimal control problem can be viewed as a semilinear elliptic optimal control problem. The list of publications in this field is already quite extensive. We mention only Casas [6], Alibert and Raymond [2], Casas, Tröltzsch, and Unger [7], and Bonnans [5]. This list might be extended considerably by including associated papers on different aspects of numerical analysis.

Our paper differs from all these contributions by the interface conditions containing the nonmonotone and nonlocal operator G. Therefore, our boundary value problem is of nonmonotone type so that special techniques must be applied.

The paper is organized as follows. After stating the mathematical setting in section 2, we provide auxiliary results on the nonlocal operator G as well as a general boundedness result for a class of nonlinear equations; see section 3. In section 4, we prove a weak maximum principle for the semilinear state equation and an L^∞-estimate for its solutions, followed by the existence of an optimal solution in section 5. Section 6 is devoted to the existence and boundedness of a solution to the linearized equation. In section 7, we establish first order necessary optimality conditions based on the differentiability of the solution operator associated with the semilinear equation. The corresponding adjoint state is introduced at the end of section 7. Finally, section 8 presents some numerical results.

2. The mathematical setting. Throughout this paper, we assume the following conditions (A1)–(A3) on the domain Ω and on the quantities and functions occurring in (P):

(A1) We assume that $\Omega \subset \mathbb{R}^N$, $N = 2, 3$, is a bounded simply connected domain with Lipschitz boundary Γ_0. The boundary of the simply connected subdomain $\overline{\Omega}_{\rm g} \subset \Omega$, denoted by $\Gamma_{\rm r}$, is assumed to be a closed Lipschitz surface that is piecewise $C^{1,\delta}$. Notice that the distance from $\Gamma_{\rm r}$ to Γ_0 is positive. Then, $\Omega_{\rm s}$ is defined by $\Omega_{\rm s} = \Omega \backslash \overline{\Omega}_{\rm g}$ (cf. Figure 1.1).

(A2) The Boltzmann radiation constant is assumed to be positive, i.e., $\sigma \in \mathbb{R}^+$. For the thermal conductivity, we assume $\kappa \in L^\infty(\Omega)$ with

$$\kappa(x) = \begin{cases} \kappa_{\rm s}(x) \text{ in } \Omega_{\rm s}, \\ \kappa_{\rm g}(x) \text{ in } \Omega_{\rm g}, \end{cases}$$

and $\kappa(x) \geq \kappa_{\min} > 0$ a.e. on Ω. Furthermore, the emissivity $\varepsilon \in L^\infty(\Gamma_0 \cup \Gamma_{\rm r})$ is bounded by $1 \geq \varepsilon \geq \varepsilon_{\min} > 0$ a.e. on $\Gamma_0 \cup \Gamma_{\rm r}$.

(A3) The desired temperature gradient z is given in $L^2(\Omega_{\rm g})^N$ and ν is a positive constant. For the box constraints, we assume $u_a, u_b \in L^\infty(\Omega_{\rm s})$ and $0 \leq u_a(x) \leq u_b(x)$ a.e. in $\Omega_{\rm s}$.

Notation. For a given p with $1 \leq p \leq \infty$, an operator $B : L^p(\Gamma_{\rm r}) \to L^p(\Gamma_{\rm r})$ is said to be positive if $v \in L^p(\Gamma_{\rm r})$ and $v \geq 0$ a.e. on $\Gamma_{\rm r}$ imply $B\,v \geq 0$ a.e. on $\Gamma_{\rm r}$. Furthermore, $\mathbf{1}$ denotes the function $e(x) \equiv 1$ a.e., and I is the identity operator in the respective function spaces. We introduce the set of admissible controls by $U_{\rm ad} := \{u \in L^\infty(\Omega_{\rm s}) \,|\, u_a(x) \leq u(x) \leq u_b(x) \text{ a.e. in } \Omega_{\rm s}\}$. The measure of a certain set $A \subset \mathbb{R}^n$, $n \geq 1$, is denoted by $|A|$, and $\tau_{\rm r}$ denotes the trace operator on $\Gamma_{\rm r}$, whereas τ_0 is the trace on Γ_0. Throughout this paper, c is a generic constant.

Let W be a Banach space with its dual space W^*. Then, for $f \in W$ and $g \in W^*$, $\langle f\,,\, g\rangle$ denotes the associated pairing.

3. Auxiliary results. In section 3.1, we present some properties of the nonlocal radiation operator G. Since G is in general not positive, the nonlinearity in the

semilinear state equation is nonmonotone. However, G still has sufficiently benign properties such that the nonlinear parts in the state equation are pseudomonotone (see [12]). These properties of G are also used to prove the boundedness of the solution in section 4 below.

In Lemma 3.8 of section 3.2, we will prove the boundedness of the solution y for a general class of nonlinear equations that applies to both the semilinear case and its linearized version.

3.1. The nonlocal radiation operator. The operators K and G arising from the nonlocal radiation on Γ_r were investigated in detail by Laitinen and Tiihonen [12] and Tiihonen [18, 19]. We recall some of their results for convenience.

DEFINITION 3.1. *The integral operator K, representing the irradiation on Γ_r, is given by*

$$(K\,y)(x) = \int_{\Gamma_r} \omega(x,z)\,y(z)\,ds_z \tag{3.1}$$

with a symmetric kernel ω which is, in the two-dimensional case, defined by

$$\omega(x,z) = \Xi(x,z)\,\frac{[n_r(z)\cdot(x-z)][n_r(x)\cdot(z-x)]}{2|z-x|^3}, \tag{3.2a}$$

and, in the three-dimensional case, defined by

$$\omega(x,z) = \Xi(x,z)\,\frac{[n_r(z)\cdot(x-z)][n_r(x)\cdot(z-x)]}{\pi|z-x|^4}, \tag{3.2b}$$

where x, z denote two points on Γ_r, and $n_r(x)$ is the unit normal at x facing outward with respect to Ω_s (see Figure 1.1). Here, Ξ represents the visibility factor, which is given by

$$\Xi(x,z) = \begin{cases} 0 & \text{if } \overline{xz}\cap\Omega_s \neq \emptyset, \\ 1 & \text{if } \overline{xz}\cap\Omega_s = \emptyset. \end{cases}$$

Remark 3.2. Notice that, in case of convex domains, such as in Figure 1.1, we have $\Xi(x,z)\equiv 1$ for all $x,z\in\Gamma_r$.

In [19], it is proven that $\omega(x,z)$ can have a singularity at x of type $|x-z|^{-(1-\delta)}$ in the two-dimensional case and $|x-z|^{-2(1-\delta)}$ in the three-dimensional case, which is, in both cases, integrable. This is the key point to the following lemma also derived in [19].

LEMMA 3.3.

(i) *K maps $L^p(\Gamma_r)$ to $L^p(\Gamma_r)$ for all $1\le p\le\infty$.*
(ii) *If Γ_r is a closed Lipschitz surface that is piecewise $C^{1,\delta}$, then $K\,\mathbf{1}=\mathbf{1}$ holds a.e. on Γ_r.*
(iii) *The operator $I-(1-\varepsilon)K : L^p(\Gamma_r)\to L^p(\Gamma_r)$ is continuously invertible.*

With the help of Lemma 3.3, Laitinen and Tiihonen proved the following properties of $G=(I-K)(I-(1-\varepsilon)K)^{-1}\varepsilon$ (cf. [18, Lemma 6] and [12, Lemma 8]).

LEMMA 3.4.

(i) *G is a bounded linear operator from $L^p(\Gamma_r)$ to itself for all $1\le p\le\infty$.*
(ii) *G can be written as $G=I-H$ with a positive operator H.*
(iii) *For all $y\in L^5(\Gamma_r)$, $\int_{\Gamma_r} G(\sigma|y|^3y)y\,ds\ge 0$ holds true.*

Here, we show another property of G that we will use subsequently.

LEMMA 3.5. *For every function $v(x) \equiv k = const.$ a.e. on Γ_r, we have $G\,v = 0$ and $G^*\,v = 0$.*

Proof. Since K has a symmetric kernel, K is formally self-adjoint, i.e., $\langle v\,,\,K\,w\rangle = \langle K\,v\,,\,w\rangle$ for all $v \in L^p(\Gamma_r)$ and $w \in L^q(\Gamma_r)$, $1/p + 1/q = 1$. Thus, together with Lemma 3.3(ii), we obtain $K^*\,\mathbf{1} = \mathbf{1}$ a.e. on Γ_r. Therefore, with the definition of G, we find

$$G^*v = \varepsilon(I - K^*(1-\varepsilon))^{-1}(I - K^*)v = \varepsilon(I - K^*(1-\varepsilon))^{-1}(k-k) = 0.$$

Furthermore, for an arbitrary constant k, Lemma 3.3(ii) implies

$$(I - (1-\varepsilon)K)k = k - (1-\varepsilon)k = \varepsilon\,k.$$

Thus, by Lemma 3.3(iii), $k = (I - (1-\varepsilon)K)^{-1}\varepsilon\,v$ holds for every function v with $v(x) \equiv k$ a.e. on Γ_r. Hence, we obtain

$$G\,v = (I-K)(I-(1-\varepsilon)K)^{-1}\varepsilon\,v = (I-K)k = 0. \qquad \square$$

3.2. Boundedness for a nonlinear equation. In sections 4 and 6, we show the existence of solutions of the semilinear state equation and its linearization in suitable functions spaces. Moreover, we have to show the boundedness of the solutions of both equations. To unify the proofs, we first prove an auxiliary result for a nonlinear equation of the form

$$\langle F(y)\,,\,v\rangle = \langle f\,,\,v\rangle \qquad \forall\, v \in W, \tag{3.3}$$

where $F : W \to W^*$ is a certain mapping, and y is a given solution in a space $W \subseteq H^1(\Omega)$. It will be shown that, under suitable assumptions, a solution y belongs to the function space

$$V^\infty := H^1(\Omega) \cap L^\infty(\Omega), \tag{3.4}$$

equipped with the norm

$$\|v\|_{V^\infty} = \|v\|_{H^1(\Omega)} + \|v\|_{L^\infty(\Omega)}.$$

Remark 3.6. If $y \in V^\infty$, then $\tau_r\,y \in L^\infty(\Gamma_r)$, and $\tau_0\,y \in L^\infty(\Gamma_0)$.

This is an immediate consequence of Lemma 3.7 below, which follows from Proposition 5.2, part (ii) in [8].

LEMMA 3.7. *For a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$, $n \in \mathbb{N}$, the trace operator $\tau : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ is positive, i.e., $\tau v \geq 0$ a.e. on $\partial\Omega$ for $v \geq 0$ a.e. on Ω.*

To show the boundedness of a solution to the nonlinear equation (3.3), we use the following hypotheses.

(H1) (*regularity of the inhomogeneity*). The right-hand side of (3.3) can be expressed as

$$\langle f\,,\,v\rangle = \int_\Omega f_\Omega\,v\,dx + \int_{\Gamma_r} f_r\,v\,ds + \int_{\Gamma_0} f_0\,v\,ds,$$

where f_Ω, f_r, and f_0 satisfy $f_\Omega \in L^{p_1}(\Omega)$, $f_r \in L^{p_2}(\Gamma_r)$, and $f_0 \in L^{p_2}(\Gamma_0)$, with $p_1 = 6/(5-s)$, $p_2 = 4/(3-s)$, and $1 < s < 3$.

(H2) (*coercivity*). There is a constant $k_0 \geq 0$ such that, for a given solution $y \in W$ of (3.3) and for each $k > k_0$, the functions

$$\varphi_k(x) := \begin{cases} y(x) - k, & y(x) \geq k, \\ 0, & |y(x)| < k, \\ y(x) + k, & y(x) \leq -k \end{cases} \tag{3.5}$$

are elements of W, and there is a constant $c > 0$ such that the nonlinearity in (3.3) satisfies, for each $k > k_0$,

$$\langle F(y)\,,\, \varphi_k \rangle \geq c\,\|\varphi_k\|^2_{H^1(\Omega)}.$$

LEMMA 3.8. *Suppose that $\Omega \subset \mathbb{R}^3$ satisfies* (A1)*, that $y \in W$ is a solution to* (3.3)*, and that* (H1) *and* (H2) *are fulfilled. Then $y \in V^\infty$, and there exists a constant c_0 depending only on Ω, and not on y, f_Ω, f_r, f_0, such that*

$$\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_\mathrm{r}\cup\Gamma_0)} \leq c_0(k_0 + \|f_\Omega\|_{L^{p_1}(\Omega)} + \|f_\mathrm{r}\|_{L^{p_2}(\Gamma_\mathrm{r})} + \|f_0\|_{L^{p_2}(\Gamma_0)}) \tag{3.6}$$

holds true.

Proof. The proof is based on a technique introduced by Kinderlehrer and Stampacchia [8] and Stampacchia [17] in the linear case. For given $k > k_0$, we define $A_\Omega(k) := \{x \in \Omega \mid |y(x)| \geq k\}$ and introduce $A_\mathrm{r}(k)$ and $A_0(k)$ analogously. Our aim is to show that there is a $k > 0$ with $|A_\Omega(k)| = |A_\mathrm{r}(k)| = |A_0(k)| = 0$. We start with (H2), which yields

$$\begin{aligned}
&\langle F(y)\,,\, \varphi_k \rangle \geq c\,\|\varphi_k\|^2_{H^1(\Omega)} \geq c\,(\|\varphi_k\|^2_{L^6(\Omega)} + \|\varphi_k\|^2_{L^4(\Gamma_\mathrm{r})} + \|\varphi_k\|^2_{L^4(\Gamma_0)}) \\
&\geq c\biggl(\biggl(\int_{A_\Omega(k)} (|y| - k)^6 dx\biggr)^{1/3} + \biggl(\int_{A_\mathrm{r}(k)} (|y| - k)^4 ds\biggr)^{1/2} + \biggl(\int_{A_0(k)} (|y| - k)^4 ds\biggr)^{1/2}\biggr),
\end{aligned}$$

since φ_k equals 0 on $\Omega \setminus A_\Omega(k)$, on $\Gamma_\mathrm{r} \setminus A_\mathrm{r}(k)$, and on $\Gamma_0 \setminus A_0(k)$. If $h > k > k_0 \geq 0$, then $A(h) \subseteq A(k)$, where A stands generally for A_Ω, A_r, and A_0, and we have for an arbitrary $m \in \mathbb{N}$,

$$\biggl(\int_{A(k)} (|y| - k)^m dx\biggr)^{2/m} \geq \biggl(\int_{A(h)} (h - k)^m dx\biggr)^{2/m} = (h - k)^2 |A(h)|^{2/m}.$$

Defining

$$\psi(h) := |A_\Omega(h)|^{1/3} + |A_0(h)|^{1/2} + |A_\mathrm{r}(h)|^{1/2}, \tag{3.7}$$

one obtains

$$\|\varphi_k\|^2_{H^1(\Omega)} \geq c\,(h - k)^2 \psi(h). \tag{3.8}$$

Now, we investigate the right-hand side in (3.3). Hypothesis (H1), Hölder's inequality, and embedding theorems imply

$$\begin{aligned}
\langle f\,,\, \varphi_k \rangle &\leq \|f_\Omega\|_{L^{6/5}(A_\Omega(k))} \|\varphi_k\|_{L^6(\Omega)} + \|f_\mathrm{r}\|_{L^{4/3}(A_\mathrm{r}(k))} \|\varphi_k\|_{L^4(\Gamma_\mathrm{r})} \\
&\quad + \|f_0\|_{L^{4/3}(A_0(k))} \|\varphi_k\|_{L^4(\Gamma_0)} \\
&\leq c\,\bigl(\|f_\Omega\|_{L^{6/5}(A_\Omega(k))} + \|f_\mathrm{r}\|_{L^{4/3}(A_\mathrm{r}(k))} + \|f_0\|_{L^{4/3}(A_0(k))}\bigr) \|\varphi_k\|_{H^1(\Omega)}.
\end{aligned}$$

Applying (H2), (3.3), and Young's inequality, we obtain

$$\|\varphi_k\|^2_{H^1(\Omega)} \le c\left(\|f_\Omega\|^2_{L^{6/5}(A_\Omega(k))} + \|f_{\mathrm{r}}\|^2_{L^{4/3}(A_{\mathrm{r}}(k))} + \|f_0\|^2_{L^{4/3}(A_0(k))}\right). \tag{3.9}$$

Using again Hölder's inequality, the first norm on the right-hand side of (3.9) can be estimated by

$$\begin{aligned}\left(\int_{A_\Omega(k)} f_\Omega^{6/5} dx\right)^{5/3} &\le \left(\|f_\Omega^{6/5}\|_{L^{5/(5-s)}(A_\Omega(k))}|A_\Omega(k)|^{s/5}\right)^{5/3}\\ &\le \|f_\Omega\|^2_{L^{p_1}(\Omega)}|A_\Omega(k)|^{s/3}\end{aligned} \tag{3.10}$$

with $p_1 = 6/(5-s)$ and $1 < s < 3$ as defined above. Similarly, we derive

$$\left(\int_{A_{\mathrm{r}}(k)} f_{\mathrm{r}}^{4/3} dx\right)^{3/2} \le \|f_{\mathrm{r}}\|^2_{L^{p_2}(\Gamma_{\mathrm{r}})}|A_{\mathrm{r}}(k)|^{s/2} \tag{3.11}$$

with $p_2 = 4/(3-s)$ (and analogously on A_0).

Due to the Taylor expansion, we have for three arbitrary real nonnegative numbers a, b, and c that $(a^s + b^s + c^s) \le (a+b+c)^s$ for all $s > 1$. Choosing $a = |A_\Omega(k)|^{1/3}$, $b = |A_{\mathrm{r}}(k)|^{1/2}$, and $c = |A_0(k)|^{1/2}$, combining (3.10), (3.11), and (3.9) yields

$$\|\varphi_k\|^2_{H^1(\Omega)} \le c\,\varrho_f\,\psi(k)^s$$

with $\varrho_f = \|f_\Omega\|^2_{L^{p_1}(\Omega)} + \|f_{\mathrm{r}}\|^2_{L^{p_2}(\Gamma_{\mathrm{r}})} + \|f_0\|^2_{L^{p_2}(\Gamma_0)}$, and ψ as defined in (3.7). Together with (3.8), it follows that

$$\psi(h) \le \frac{c\,\varrho_f}{(h-k)^2}\psi(k)^s \quad \text{for } h > k > k_0 \ge 0. \tag{3.12}$$

Stampacchia proved in [17] that each nonnegative and nonincreasing function $\psi = \psi(t)$ satisfying (3.12) with some $s > 1$ has a zero at some $t = d$, $d > 0$, where

$$d = k_0 + 2^{s/(s-1)}(c|\psi(k_0)|^{s-1})^{1/2}\sqrt{\varrho_f},$$

and, in our case, $|\psi(k_0)| = \psi(k_0)$ can be estimated by

$$\psi(k_0) \le \psi(0) = |\Omega|^{1/3} + |\Gamma_0|^{1/2} + |\Gamma_{\mathrm{r}}|^{1/2}$$

(see also Kinderlehrer and Stampacchia [8, Lemma B.1]). Due to the definition of ψ in (3.7), this implies $|y(x)| \le d$ a.e. on Ω and on $\Gamma_{\mathrm{r}} \cup \Gamma_0$. Thus, with the definition of ϱ_f, we obtain

$$\begin{aligned}\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_{\mathrm{r}}\cup\Gamma_0)} &\le 2\,d\\ &\le 2\left(k_0 + 2^{s/(s-1)}(c|\psi(0)|^{s-1})^{1/2}\sqrt{\varrho_f}\right)\\ &\le c\left(k_0 + \sqrt{\|f_\Omega\|^2_{L^{p_1}(\Omega)} + \|f_{\mathrm{r}}\|^2_{L^{p_2}(\Gamma_{\mathrm{r}})} + \|f_0\|^2_{L^{p_2}(\Gamma_0)}}\right)\\ &\le c_0\left(k_0 + \|f_\Omega\|_{L^{p_1}(\Omega)} + \|f_{\mathrm{r}}\|_{L^{p_2}(\Gamma_{\mathrm{r}})} + \|f_0\|_{L^{p_2}(\Gamma_0)}\right)\end{aligned}$$

by the equivalence of the Euclidian norm and the L^1-norm in $\mathbb{R}^3$. ∎

Remark 3.9. Due to the positivity of the trace operator (cf. Lemma 3.7) in Lemma 3.8, it would have sufficed to prove the boundedness of y on Ω. However, this would not have simplified the proof, as we still needed the direct estimates of the inhomogeneities on Γ_0 and Γ_{r} as carried out above.

4. The semilinear equation. Before we apply Lemma 3.8 to the semilinear equation in (P), we will introduce the variational form of the state equation and recall a theorem of Laitinen and Tiihonen [12] that covers its solvability.

Based on Lemma 3.4, we are now able to derive the weak formulation of the state equation in (P). To fix G, we specify p in Lemma 3.4(i) by $p = 5/4$ and obtain $G : L^{5/4}(\Gamma_r) \to L^{5/4}(\Gamma_r)$. Then, formal integration by parts over Γ_0 and Γ_r yields

$$\begin{aligned} &\int_\Omega \kappa \nabla y \cdot \nabla v \, dx + \int_{\Gamma_r} G(\sigma|y|^3 y) v \, ds + \int_{\Gamma_0} \varepsilon\sigma \, |y|^3 y \, v \, ds \\ &\quad = \int_{\Omega_s} u \, v \, dx + \int_{\Gamma_0} \varepsilon\sigma \, y_0^4 v \, ds \qquad \forall\, v \in V \end{aligned} \tag{4.1}$$

with $V = \{v \in H^1(\Omega) \mid \tau_r v \in L^5(\Gamma_r)\,,\ \tau_0 v \in L^5(\Gamma_0)\}$. Clearly, due to $G : L^{5/4}(\Gamma_r) \to L^{5/4}(\Gamma_r)$, we have $G(\sigma|y|^3 y)v \in L^1(\Gamma_r)$ for all $y, v \in V$. The state space V is equipped with the norm

$$\|v\|_V = \|v\|_{H^1(\Omega)} + \|v\|_{L^5(\Gamma_r)} + \|v\|_{L^5(\Gamma_0)}.$$

A function $y \in V$ is said to be a weak solution of the state equation in (P) if (4.1) is fulfilled for every $v \in V$.

THEOREM 4.1 (see [12, Theorem 2]). *For every $u \in H^1(\Omega_s)^*$ and $y_0 \in L^5(\Gamma_0)$, the semilinear equation* (4.1) *admits a unique solution in V.*

The proof is mainly based on Brezis' theorem [20, Theorem 27.A] for pseudomonotone operators. Laitinen and Tiihonen showed in [12] that the semilinear differential operator defined by the left-hand side in (4.1) fulfills all assumptions of Brezis' theorem; i.e., it is pseudomonotone, bounded, and coercive on V. Thus, (4.1) admits at least one solution in V. The uniqueness then follows from a comparison principle [12, Theorem 4].

Next, we show the boundedness of solutions to (4.1). This result has been obtained by Laitinen and Tiihonen (see [12, Theorem 5]) in a slightly different setting by another method. Here, we need it for the nonlinear equation (4.1) as well as for its linearized version (see section 6). In both cases, Lemma 3.8 can be applied.

THEOREM 4.2. *Assume that $u \in L^2(\Omega_s)$ and $y_0 \in L^{16}(\Gamma_0)$. Then, there exists a constant c_0 depending only on Ω such that*

$$\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_r \cup \Gamma_0)} \le c_0 (1 + \|u\|_{L^2(\Omega_s)} + \|y_0\|^4_{L^{16}(\Gamma_0)}) \tag{4.2}$$

is valid.

Proof. To apply Lemma 3.8, we have to verify (H1) and (H2). Here, we choose the state space $V \subset H^1(\Omega)$ for the space W. Since $y_0 \in L^{16}(\Gamma_0)$ by assumption, we have $y_0^4 \in L^4(\Gamma_0)$, and thus together with the assumed regularity of u, (H1) is satisfied with $f_0 = y_0^4$ and $s = 2$. To verify (H2), we show

$$\int_\Omega \kappa \nabla y \cdot \nabla \varphi_k \, dx + \int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k \, ds + \int_{\Gamma_0} \varepsilon\sigma \, |y|^3 y \, \varphi_k \, ds \ge c \, \|\varphi_k\|^2_{H^1(\Omega)} \tag{4.3}$$

for each $k \ge 1$, which we will use in the estimate of the Γ_0-term. At first, we show the positivity of the radiation term on Γ_r. To this aim, we split $\varphi_k = \varphi_k^+ + \varphi_k^-$ with

$$\varphi_k^+(x) := \begin{cases} y(x) - k, & y(x) \ge k, \\ 0, & y(x) < k, \end{cases} \qquad \varphi_k^-(x) := \begin{cases} y(x) + k, & y(x) \le -k, \\ 0, & y(x) > -k. \end{cases}$$

It follows that

$$\int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k \, ds = \int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k^+ \, ds + \int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k^- \, ds. \tag{4.4}$$

Due to the linearity of G, the first integral on the right-hand side can be expressed by

$$\int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k^+ \, ds = \int_{\Gamma_r} G(\sigma\Phi)\varphi_k^+ \, ds + \int_{\Gamma_r} G(\sigma\Psi)\varphi_k^+ \, ds, \tag{4.5}$$

where

$$\Phi(x) := \begin{cases} y(x)^4, & y(x) \geq k, \\ k^4, & y(x) < k, \end{cases} \qquad \Psi(x) := \begin{cases} 0, & y(x) \geq k, \\ |y(x)|^3 y(x) - k^4, & y(x) < k. \end{cases}$$

This partition is necessary since G is nonlocal. Notice that $\Phi + \Psi = |y|^3 y$ and $\Psi(x) \leq 0$ on Γ_r. For the first addend in (4.5), we have, with $y_k(x) := \max\{y(x), k\}$,

$$\int_{\Gamma_r} G(\sigma\Phi)\varphi_k^+ \, ds = \int_{\Gamma_r} G(\sigma\, y_k^4)(y_k - k) \, ds = \int_{\Gamma_r} G(\sigma\, y_k^4) y_k \, ds - \int_{\Gamma_r} \sigma\, y_k^4 G^* k \, ds \geq 0$$

because of Lemma 3.4(iii) and Lemma 3.5. Due to $G = I - H$ and the positivity of H (see Lemma 3.4(ii)), the second integral in (4.5) results in

$$\int_{\Gamma_r} G(\sigma\Psi)\varphi_k^+ \, ds = \sigma \int_{\Gamma_r} \Psi\varphi_k^+ \, ds + \sigma \int_{\Gamma_r} H(-\Psi)\varphi_k^+ \, ds \geq 0,$$

since $\Psi(x) \neq 0$ implies $\varphi_k^+(x) = 0$ and $-\Psi, \varphi_k^+ \geq 0$ holds on Γ_r.

Similarly, we prove the positivity of the second integral on the right-hand side of (4.4) with $y_k(x) := \min\{y(x), -k\}$, defining

$$\Phi(x) := \begin{cases} |y(x)|^3 y(x), & y(x) \leq -k, \\ -k^4, & y(x) > -k, \end{cases} \qquad \Psi(x) := \begin{cases} 0, & y(x) \leq -k, \\ |y(x)|^3 y(x) + k^4, & y(x) > -k. \end{cases}$$

In this way, we have shown that $\int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k \, ds \geq 0$. It remains to analyze the other integrals in (4.3). The first integral is estimated by

$$\int_\Omega \kappa \nabla y \cdot \nabla \varphi_k \, dx \geq \kappa_{\min} \|\nabla \varphi_k\|^2_{L^2(\Omega)},$$

since $\nabla\varphi_k(x) = 0$ if $\nabla y(x) \neq \nabla\varphi_k(x)$. On Γ_0, using $k \geq k_0 := 1$, we obtain

$$\int_{\Gamma_0} \varepsilon\sigma\, |y|^3 y\, \varphi_k \, ds \geq \varepsilon_{\min}\sigma\, k^3 \|\varphi_k\|^2_{L^2(A_0(k))} \geq \varepsilon_{\min}\sigma\, \|\varphi_k\|^2_{L^2(\Gamma_0)},$$

as $y\,\varphi_k = (\operatorname{sign}\varphi_k)(|\varphi_k| + k)\varphi_k \geq \varphi_k^2$ is valid on $A_0(k) = \{x \in \Omega \mid \varphi_k(x) \neq 0\}$.

Due to the positivity of $\int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k \, ds$, we finally have

$$\begin{aligned}\int_\Omega \kappa \nabla y \cdot \nabla \varphi_k \, dx + \int_{\Gamma_r} G(\sigma|y|^3 y)\varphi_k \, ds + \int_{\Gamma_0} \varepsilon\sigma \, |y|^3 y \, \varphi_k \, ds \\ \geq c \, (\|\nabla \varphi_k\|^2_{L^2(\Omega)} + \|\varphi_k\|^2_{L^2(\Gamma_0)}) \geq c \, \|\varphi_k\|^2_{H^1(\Omega)}.\end{aligned}$$

Therefore, we can apply Lemma 3.8, and estimate (3.6) gives, with $s = 2$, $k_0 = 1$, $f_\Omega = u$, and $f_0 = y_0^4$,

$$\begin{aligned}\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_r \cup \Gamma_0)} &\leq c_0 \left(1 + \|u\|_{L^2(\Omega_s)} + \|y_0^4\|_{L^4(\Gamma_0)}\right) \\ &\leq c_0 \big(1 + \|u\|_{L^2(\Omega_s)} + \|y_0\|^4_{L^{16}(\Gamma_0)}\big). \qquad \square\end{aligned}$$

For the discussion of an associated linearized equation (see section 6), we need another property of the semilinear solution, namely, the following maximum principle.

THEOREM 4.3. *Suppose that $u(x) \geq 0$ a.e. in Ω_s and $y_0(x) \geq \vartheta > 0$ a.e. on Γ_0. If y is the solution of (4.1), then $y(x) \geq \vartheta$ holds a.e. on Ω and a.e. on $\Gamma_r \cup \Gamma_0$.*

Proof. This time, we use the following test function in (4.1):

$$\rho_\vartheta(x) := \begin{cases} y(x) - \vartheta, & y(x) \leq \vartheta, \\ 0, & y(x) > \vartheta. \end{cases}$$

Furthermore, we define $\Omega^- = \{x \in \Omega \mid y(x) \leq \vartheta\}$ and introduce analogous definitions for Ω_s^-, Γ_0^-, and Γ_r^-. Thus, we have $\rho_\vartheta|_{\Omega \setminus \Omega^-} = 0$, and (4.1) reads

$$\begin{aligned}\int_{\Omega^-} \kappa \nabla y \cdot \nabla \rho_\vartheta \, dx + \int_{\Gamma_r} G(\sigma|y|^3 y)\rho_\vartheta \, ds \\ + \int_{\Gamma_0^-} \varepsilon\sigma(|y|^3 y - y_0^4) \, \rho_\vartheta \, ds = \int_{\Omega_s^-} u \, \rho_\vartheta \, dx.\end{aligned} \tag{4.6}$$

Next, as in the proof of Theorem 4.2, the positivity of the integral over Γ_r can be shown by decomposition:

$$\int_{\Gamma_r} G(\sigma|y|^3 y)\rho_\vartheta \, ds = \int_{\Gamma_r} G(\sigma\Phi)\rho_\vartheta \, ds + \int_{\Gamma_r} G(\sigma\Psi)\rho_\vartheta \, ds$$

with

$$\Phi(x) = \begin{cases} |y(x)|^3 y(x), & y(x) \leq \vartheta, \\ \vartheta^4, & y(x) > \vartheta, \end{cases} \qquad \Psi(x) = \begin{cases} 0, & y(x) \leq \vartheta, \\ |y(x)|^3 y(x) - \vartheta^4, & y(x) > \vartheta. \end{cases}$$

Therefore, (4.6) results in

$$\int_{\Gamma_0^-} \varepsilon\sigma \, (y_0^4 - |y|^3 y) \, \rho_\vartheta \, ds \geq \int_{\Omega^-} \kappa \nabla y \cdot \nabla \rho_\vartheta \, dx - \int_{\Omega_s^-} u \, \rho_\vartheta \, dx \geq 0, \tag{4.7}$$

since $\nabla\rho_\vartheta = \nabla y$ on Ω^-, $\rho_\vartheta \le 0$ by definition, and $u \ge 0$ by assumption. On Γ_0^-, we also have $y - \vartheta \le 0$, implying $y_0^4 - |y|^3 y \ge 0$. Thus, we obtain

$$0 \le \int_{\Gamma_0^-} \varepsilon\sigma\,(y_0^4 - |y|^3 y)\,\rho_\vartheta\,ds = \int_{\Gamma_0^-} \varepsilon\sigma\,(y_0^4 - |y|^3 y)(y-\vartheta)\,ds \le 0,$$

and, consequently, $|\Gamma_0^-| = 0$ or $y = y_0$ or $y = \vartheta$ a.e. on Γ_0, which, in each case, yields the assertion on Γ_0. Therefore, we have $\rho_\vartheta \in H_0^1(\Omega)$.

Now, from (4.7), it follows that

$$\kappa_{\min}\|\nabla\rho_\vartheta\|_{L^2(\Omega)} \le \int_{\Gamma_0^-} \varepsilon\sigma\,(y_0^4 - |y|^3 y)\,\rho_\vartheta\,ds + \int_{\Omega_s^-} u\,\rho_\vartheta\,dx \le 0,$$

since $\nabla\rho_\vartheta = 0$ in $\Omega\backslash\Omega^-$. Hence $\|\nabla\rho_\vartheta\|_{L^2(\Omega)} = \|\rho_\vartheta\|_{H_0^1(\Omega)} = 0$ (notice that we have already shown $\rho_\vartheta = 0$ a.e. on Γ_0). Thus, $\rho_\vartheta = 0$ a.e. in Ω. Hence, we have shown $y \ge \vartheta$ a.e. on both Ω and Γ_0, and Lemma 3.7 then ensures $y \ge \vartheta$ a.e. on Γ_r. ∎

Notice that, within our scope of heat transport modeling, the assumption $y_0(x) \ge \vartheta > 0$ a.e. on Γ_0 constitutes no restriction: Since y_0 represents an external temperature, it is naturally positive.

5. Existence of an optimal solution. With the results of section 4 at hand, the proof of existence of an optimal solution for the optimal control problem (P) is rather standard. We start with the following lemma.

LEMMA 5.1. *Let $u \in L^{6/5}(\Omega_s)$, $y_0 \in L^{16/3}(\Gamma_0)$ be given and assume that the hypotheses of Theorem* 4.3 *are fulfilled. Then a constant c_1 exists depending only on Ω with*

$$\|y\|_{H^1(\Omega)} \le c_1\left(\|u\|_{L^{6/5}(\Omega_s)} + \|y_0\|^4_{L^{16/3}(\Gamma_0)}\right).$$

Proof. Thanks to Theorem 4.1, a unique solution y of (4.1) exists. Using it as a test function in (4.1) yields

$$\int_\Omega \kappa|\nabla y|^2\,dx + \int_{\Gamma_r} G(\sigma|y|^3 y)y\,ds + \int_{\Gamma_0} \varepsilon\sigma\,|y|^3 y^2\,ds = \int_{\Omega_s} u\,y\,dx + \int_{\Gamma_0} \varepsilon\sigma\,y_0^4 y\,ds.$$

Due to Lemma 3.4(iii) and the maximum principle in Theorem 4.3, we have

$$\begin{aligned}&\int_\Omega \kappa|\nabla y|^2\,dx + \int_{\Gamma_r} G(\sigma|y|^3 y)y\,ds + \int_{\Gamma_0} \varepsilon\sigma\,|y|^3 y^2\,ds\\ &\ge \kappa_{\min}\,\|\nabla y\|^2_{L^2(\Omega)} + \varepsilon_{\min}\,\sigma\,\vartheta^3\|y\|^2_{L^2(\Gamma_0)} \ge c\,\|y\|^2_{H^1(\Omega)}.\end{aligned}$$

Therefore, by trace and embedding theorems,

$$\begin{aligned}\|y\|^2_{H^1(\Omega)} &\le c\left(\int_{\Omega_s} u\,y\,dx + \int_{\Gamma_0} \varepsilon\sigma\,y_0^4 y\,ds\right)\\ &\le c\,\left(\|u\|_{L^{6/5}(\Omega_s)}\|y\|_{L^6(\Omega_s)} + \|y_0^4\|_{L^{4/3}(\Gamma_0)}\|y\|_{L^4(\Gamma_0)}\right)\\ &\le c_1\left(\|u\|_{L^{6/5}(\Omega_s)} + \|y_0\|^4_{L^{16/3}(\Gamma_0)}\right)\|y\|_{H^1(\Omega)},\end{aligned}$$

which establishes the case. ∎

THEOREM 5.2. *Assume that* $u_a \geq 0$, $y_0 \in L^{16}(\Gamma_0)$, *and* $y_0 \geq \vartheta > 0$. *Then there exists a solution* $(\bar{u}, \bar{y}) \in L^\infty(\Omega_s) \times V^\infty$ *to* (P).

Proof. As mentioned above, the proof follows standard arguments. We start with a sequence $\{(u_n, y_n)\}_{n=1}^\infty$ converging to the infimum $\bar{J} \geq 0$ of the objective functional in (P), i.e., $J(u_n, y_n) \to \bar{J}$, $n \to \infty$. We will now show the convergence of a subsequence of $\{(u_n, y_n)\}_{n=1}^\infty$ to an optimal solution $(\bar{u}, \bar{y})$.

The box constraints ensure that u is bounded in $L^2(\Omega_s)$. Thus, we can select a weakly converging subsequence w.l.o.g. and again denoted by u_n, $u_n \rightharpoonup \bar{u}$, $n \to \infty$. Since U_{ad} is a closed and convex subset of $L^2(\Omega_s)$, we have $\bar{u} \in U_{ad}$.

In addition to the boundedness, the u_n are also nonnegative because of $u_a \geq 0$. Together with the assumptions on y_0, this yields the boundedness of $\|y_n\|_{H^1(\Omega)}$ thanks to Lemma 5.1. Hence, we can select a weakly converging subsequence and w.l.o.g. y_n itself, i.e., $y_n \rightharpoonup \bar{y}$, $n \to \infty$, in $H^1(\Omega)$. The trace theorem and the compact embedding of $H^{1/2}(\Gamma_r \cup \Gamma_0)$ in $L^2(\Gamma_r \cup \Gamma_0)$ then give

$$\tau y_n \to \tau \bar{y}, \ n \to \infty \quad \text{in } L^2(\Gamma_r \cup \Gamma_0).$$

Now, the convergence of the nonlinearities can be derived on Γ_r and Γ_0. Since u_n, $n \in \mathbb{N}$, is uniformly bounded in $L^2(\Omega_s)$, and $y_0 \in L^{16}(\Gamma_0)$ by assumption, Theorem 4.2 ensures that a constant d exists with $|y(x)| \leq d$ a.e. in Ω and a.e. in $\Gamma_r \cup \Gamma_0$. One can easily verify that the Nemytskii operator $\Phi(y) := |y|^3 y$ satisfies

$$\|\Phi(y_n) - \Phi(y_m)\|_{L^2(\Gamma_r \cup \Gamma_0)} \leq L(d) \|y_n - y_m\|_{L^2(\Gamma_r \cup \Gamma_0)} \tag{5.1}$$

for all $y_n, y_m \in \{y \in L^\infty(\Gamma_0 \cup \Gamma_r) \,|\, |y(x)| \leq d \text{ a.e. on } \Gamma_0 \cup \Gamma_r\}$. Since this set is closed in $L^2(\Gamma_r \cup \Gamma_0)$, $|\bar{y}(x)| \leq d$ also holds true, and the convergence in $L^2(\Gamma_r \cup \Gamma_0)$ together with (5.1) yields

$$|\tau y_n|^3 \tau y_n \to |\tau \bar{y}|^3 \tau \bar{y} \quad \text{in } L^2(\Gamma_r \cup \Gamma_0). \tag{5.2}$$

Consider now the variational equation (4.1) when passing to the limit. For (u_n, y_n), it reads

$$\begin{aligned} &\int_\Omega \kappa \nabla y_n \cdot \nabla v \, dx + \int_{\Gamma_r} G(\sigma |y_n|^3 y_n) v \, ds + \int_{\Gamma_0} \varepsilon\sigma \, |y_n|^3 y_n \, v \, ds \\ &\quad = \int_{\Omega_s} u_n \, v \, dx + \int_{\Gamma_0} \varepsilon\sigma \, y_0^4 v \, ds. \end{aligned}$$

Due to $y_n \rightharpoonup \bar{y}$ in $H^1(\Omega)$, the first integral on the left-hand side converges to $\int_\Omega \kappa \nabla \bar{y} \cdot \nabla v \, dx$. The boundary integrals on the left-hand side converge because of (5.2) and the continuity of G. Finally, the inhomogeneity converges, owing to $u_n \rightharpoonup \bar{u}$ in $L^2(\Omega_s)$. Therefore, the limit $(\bar{u}, \bar{y})$ satisfies the weak formulation (4.1). The optimality of $(\bar{y}, \bar{u})$ follows in a standard way by the lower semicontinuity of J. ∎

6. The linearized equation. In this section, we investigate the linearization of the state equation (4.1) at a fixed reference pair $(\bar{y}, \bar{u}) \in V^\infty \times L^2(\Omega_s)$.[1] The linearized equation arises from the differentiation of the solution operator associated with the semilinear state equation $S : L^2(\Omega_s) \to V^\infty$ mapping u to y. This derivative

[1]Here and in the following, we will denote a fixed solution of the semilinear equation as well as an optimal solution by $(\bar{u}, \bar{y})$.

of S at $(\bar y, \bar u)$ appears in the variational inequality (see section 7). More precisely, we consider the linearized equation with a more general inhomogeneity in the form

$$
\begin{aligned}
-\mathrm{div}(\kappa\,\nabla y) &= f_\Omega \quad \text{in } \Omega,\\
\kappa_{\mathrm{s}}\left(\frac{\partial y}{\partial n_{\mathrm{r}}}\right)_{\mathrm{s}} - \kappa_{\mathrm{g}}\left(\frac{\partial y}{\partial n_{\mathrm{r}}}\right)_{\mathrm{g}} + 4\,G(\sigma|\bar y|^3 y) &= f_{\mathrm{r}} \quad \text{on } \Gamma_{\mathrm{r}},\\
\kappa_{\mathrm{s}}\frac{\partial y}{\partial n_0} + 4\,\varepsilon\sigma|\bar y|^3 y &= f_0 \quad \text{on } \Gamma_0
\end{aligned}
\tag{6.1}
$$

with $\bar y = S\,\bar u \in V^\infty$ and arbitrary functions $(f_\Omega, f_{\mathrm{r}}, f_0)$ in $L^2(\Omega)\times L^2(\Gamma_{\mathrm{r}})\times L^2(\Gamma_0)$. In the next section, we will show that, under a natural assumption, S is indeed Fréchet differentiable and its derivative $y = S'(\bar u)u$ corresponds to the solution of (6.1) with $f_\Omega = u$, $f_{\mathrm{r}} = 0$, and $f_0 = 0$. However, we first focus on (6.1), and, in the present section, prove the existence of a unique solution in V^∞. The existence theory is based on the theory of Fredholm operators and has to account for eigenvalues. This is due to the fact that the Lax–Milgram lemma cannot be directly applied because of the lack of coercivity caused by the radiation operator G in the linearized case. The boundedness of the solution is again shown by Lemma 3.8.

The variational equation of (6.1) reads

$$
\begin{aligned}
&\int_\Omega \kappa\nabla y\cdot\nabla v\,dx + \int_{\Gamma_0} 4\,\varepsilon\sigma\,|\bar y|^3 y\,v\,ds\\
&\qquad = \int_\Omega f_\Omega\,v\,dx + \int_{\Gamma_{\mathrm{r}}} (f_{\mathrm{r}} - 4\,G(\sigma|\bar y|^3 y))\,v\,ds + \int_{\Gamma_0} f_0\,v\;ds \quad \forall\, v\in H^1(\Omega).
\end{aligned}
\tag{6.2}
$$

Assume for awhile that y is given on the right-hand side. In view of $\bar y\in V^\infty$ and $\bar y\ge \vartheta > 0$ (Theorem 4.3), the bilinear form defined by the left-hand side of (6.2) is bounded and coercive in $H^1(\Omega)$. Therefore, the Lax–Milgram lemma yields continuous linear operators $B_\Omega : L^2(\Omega)\to H^1(\Omega)$, $B_{\mathrm{r}} : L^2(\Gamma_{\mathrm{r}})\to H^1(\Omega)$, and $B_0 : L^2(\Gamma_0)\to H^1(\Omega)$ such that, with $\tilde y := \tau_{\mathrm{r}}\,y$,

$$
y = B_\Omega\,f_\Omega + B_{\mathrm{r}}\,(f_{\mathrm{r}} - 4\,G(\sigma|\tau_{\mathrm{r}}\bar y|^3\tilde y)) + B_0\,f_0.
\tag{6.3}
$$

We are now in a position to formulate the existence theorem for (6.1).

THEOREM 6.1. *Assume $\bar y\in V^\infty$, $\bar y\ge\vartheta>0$, and that $\lambda = 1$ is not an eigenvalue of $B(\bar y)(\,\cdot\,) := -\tau_{\mathrm{r}}\,B_{\mathrm{r}}(4\,G(\sigma|\tau_{\mathrm{r}}\bar y|^3\,\cdot\,))$ with $B(\bar y) : L^2(\Gamma_{\mathrm{r}})\to L^2(\Gamma_{\mathrm{r}})$. Then, to every $(f_\Omega, f_{\mathrm{r}}, f_0)\in L^2(\Omega)\times L^2(\Gamma_{\mathrm{r}})\times L^2(\Gamma_0)$, there exists a unique solution y of (6.1) in $H^1(\Omega)$.*

Proof. In the following, we will suppress the dependency of B on $\bar y$, since $\bar y$ is fixed. Thus, we simply write B instead of $B(\bar y)$. Applying the trace operator τ_{r} to both sides, (6.3) results in

$$
(I-B)\tilde y = \tilde y + \tau_{\mathrm{r}}B_{\mathrm{r}}(4\,G(\sigma|\tau_{\mathrm{r}}\bar y|^3\tilde y)) = \tau_{\mathrm{r}}\,(B_\Omega\,f_\Omega + B_{\mathrm{r}}\,f_{\mathrm{r}} + B_0\,f_0).
\tag{6.4}
$$

Since $B_{\mathrm{r}} : L^2(\Gamma_{\mathrm{r}})\to H^1(\Omega)$, and $\bar y\in V^\infty$ by assumption, and, thus, $\tau_{\mathrm{r}}\bar y\in L^\infty(\Gamma_{\mathrm{r}})$, it follows that B maps $L^2(\Gamma_{\mathrm{r}})$ in $H^{1/2}(\Gamma_{\mathrm{r}})$. Due to compact embedding, B is compact from $L^2(\Gamma_{\mathrm{r}})$ to $L^2(\Gamma_{\mathrm{r}})$. Therefore, $\lambda = 1$ is either one of the countably many eigenvalues of B, or $(I-B)$ is continuously invertible. Thus, we obtain a unique solution $\tilde y\in L^2(\Gamma_{\mathrm{r}})$ of (6.4), provided that $\lambda = 1$ is not an eigenvalue of B.

Furthermore, for every $\tilde{y} \in L^2(\Gamma_\mathrm{r})$, we have $G(\sigma|\tau_\mathrm{r}\bar{y}|^3\tilde{y}) \in L^2(\Gamma_\mathrm{r})$ by means of $\bar{y} \in V^\infty$, and, hence, (6.3) admits a unique solution in $H^1(\Omega)$. ∎

As indicated above, we will employ Theorem 6.1 to prove the Fréchet differentiability of S at the reference solution $(\bar{u}, \bar{y})$, which, in turn, is used to establish first order necessary optimality conditions (see section 7 below). It is well known that regularity assumptions on the solution $\bar{y}$, like those present in Theorem 6.1, are typical for the Fréchet differentiability of solution operators associated with semilinear PDEs. In particular, the assumption that $\lambda = 1$ is not an eigenvalue of $B(\bar{y})$ can be interpreted as such a regularity assumption on the solution $\bar{y}$.

THEOREM 6.2. *Suppose that the assumptions of Theorem* 6.1 *hold true. Then, for all $f_\Omega \in L^2(\Omega)$, $f_\mathrm{r} \in L^4(\Gamma_\mathrm{r})$, and $f_0 \in L^4(\Gamma_\mathrm{r})$, there exists a constant c_2 depending only on Ω such that*

$$\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_\mathrm{r}\cup\Gamma_0)} \le c_2 \left(\|f_\Omega\|_{L^2(\Omega)} + \|f_\mathrm{r}\|_{L^4(\Gamma_\mathrm{r})} + \|f_0\|_{L^4(\Gamma_0)}\right) \tag{6.5}$$

holds true.

Proof. We again use Lemma 3.8 to prove the boundedness, this time with $W = H^1(\Omega)$. We apply it to (6.2) for F defined by

$$\langle F(y)\,,\, v\rangle = \int_\Omega \kappa \nabla y \cdot \nabla v \, dx + \int_{\Gamma_0} 4\,\varepsilon\sigma\,|\bar{y}|^3 y\, v \, ds.$$

Hypothesis (H2) clearly holds with $k_0 = 0$, owing to the coercivity of the bilinear form.

Now, (H1), i.e., the regularity of the right-hand side, is the critical point, since the inhomogeneity on Γ_r in (6.2) depends on the solution y. As before, we choose $s = 2$ in (H1), and thus, the required regularities coincide with the assumptions on f_Ω, f_r, and f_0.

The part of the inhomogeneity depending on y and $\bar{y}$ is given by $-4\,G(\sigma|\bar{y}|^3 y)$. It belongs to $L^4(\Gamma_\mathrm{r})$, since $y \in H^1(\Omega)$, and therefore $\tau_\mathrm{r} y \in L^4(\Gamma_\mathrm{r})$ in addition to $\bar{y} \in V^\infty$. Consequently, (H1) also is satisfied, and Lemma 3.8 can be applied.

It remains to verify a bound that is independent of y. With $k_0 = 0$, estimate (3.6) gives

$$\begin{aligned} &\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_\mathrm{r}\cup\Gamma_0)} \\ &\qquad \le c_0 \left(\|f_\Omega\|_{L^2(\Omega)} + \|f_\mathrm{r}\|_{L^4(\Gamma_\mathrm{r})} + \|4\,G(\sigma|\bar{y}|^3 y)\|_{L^4(\Gamma_\mathrm{r})} + \|f_0\|_{L^4(\Gamma_0)}\right). \end{aligned} \tag{6.6}$$

We estimate the term with y by

$$\|G(\sigma|\bar{y}|^3 y)\|_{L^4(\Gamma_\mathrm{r})} \le c\|G\|_{\mathcal{L}(L^4(\Gamma_\mathrm{r}))}\, \|\bar{y}\|^3_{L^\infty(\Gamma_\mathrm{r})}\, \|y\|_{H^1(\Omega)}. \tag{6.7}$$

With (6.3) and the continuity of B_Ω, B_r, and B_0, we obtain by the boundedness of $\bar{y}$,

$$\begin{aligned} \|y\|_{H^1(\Omega)} \le\ & \|B_\Omega\, f_\Omega + B_\mathrm{r}\, f_\mathrm{r} + B_0\, f_0\|_{H^1(\Omega)} \\ &+ c\,\|B_\mathrm{r}\|_{\mathcal{L}(L^2(\Gamma_\mathrm{r}),H^1(\Omega))}\|G\|_{\mathcal{L}(L^2(\Gamma_\mathrm{r}))}\|\bar{y}\|^3_{L^\infty(\Gamma_\mathrm{r})}\|y\|_{L^2(\Gamma_\mathrm{r})} \\ &\le c\,(\|f_\Omega\|_{L^2(\Omega)} + \|f_\mathrm{r}\|_{L^4(\Gamma_\mathrm{r})} + \|f_0\|_{L^4(\Gamma_0)} + \|y\|_{L^2(\Gamma_\mathrm{r})}). \end{aligned} \tag{6.8}$$

Due to (6.4), $\|\tilde{y}\|_{L^2(\Gamma_\mathrm{r})}$ is bounded by

$$\begin{aligned} \|\tilde{y}\|_{L^2(\Gamma_\mathrm{r})} &\le \|(I-B)^{-1}\|_{\mathcal{L}(L^2(\Gamma_\mathrm{r}))} \cdot \|B_\Omega\, f_\Omega + B_\mathrm{r}\, f_\mathrm{r} + B_0\, f_0\|_{L^2(\Gamma_\mathrm{r})} \\ &\le c(\|f_\Omega\|_{L^2(\Omega)} + \|f_\mathrm{r}\|_{L^4(\Gamma_\mathrm{r})} + \|f_0\|_{L^4(\Gamma_0)}), \end{aligned} \tag{6.9}$$

where Fredholm theory grants $\|(I-B)^{-1}\|_{\mathcal{L}(L^2(\Gamma_r))} < \infty$. Combining (6.7)–(6.9) yields

$$\|G(\sigma|\bar{y}|^3 y)\|_{L^4(\Gamma_r)} \le \tilde{c}\,(\|f_\Omega\|_{L^2(\Omega)} + \|f_r\|_{L^4(\Gamma_r)} + \|f_0\|_{L^4(\Gamma_0)}) \tag{6.10}$$

with a constant $\tilde{c}$ depending only on Ω. Inserting (6.10) in (6.6) finally gives

$$\begin{aligned}\|y\|_{L^\infty(\Omega)} + \|y\|_{L^\infty(\Gamma_r\cup\Gamma_0)} &\le c_0(1+\tilde{c})\left(\|f_\Omega\|_{L^2(\Omega)} + \|f_r\|_{L^4(\Gamma_r)} + \|f_0\|_{L^4(\Gamma_0)}\right)\\ &\le c_2\left(\|f_\Omega\|_{L^2(\Omega)} + \|f_r\|_{L^4(\Gamma_r)} + \|f_0\|_{L^4(\Gamma_0)}\right). \qquad \square\end{aligned}$$

7. First order necessary optimality conditions. With the results of sections 4 and 6, we are now able to show the Fréchet differentiability of the semilinear PDE solution operator $S : u \mapsto y$ from $L^2(\Omega_s)$ to V^∞; see section 7.1. Defining $J(y,u) = J(S\,u,u) =: j(u)$, a standard argument then yields the variational inequality for the optimal pair $(\bar{u},\bar{y})$:

$$j'(\bar{u})(u-\bar{u}) = (\nabla\bar{y} - z, \nabla y)_{L^2(\Omega_g)} + \nu(\bar{u},(u-\bar{u}))_{L^2(\Omega_s)} \ge 0 \quad \forall\, u \in U_{\rm ad}, \tag{7.1}$$

with $\bar{y} = S\,\bar{u}$ and $y = S'(\bar{u})(u-\bar{u})$. The latter relation means that y satisfies the PDE

$$\begin{aligned} -\mathrm{div}(\kappa_s\,\nabla y) &= u-\bar{u} && \text{in } \Omega_s,\\ -\mathrm{div}(\kappa_g\,\nabla y) &= 0 && \text{in } \Omega_g,\\ \kappa_s\left(\frac{\partial y}{\partial n_r}\right)_s - \kappa_g\left(\frac{\partial y}{\partial n_r}\right)_g + 4\,G(\sigma|\bar{y}|^3 y) &= 0 && \text{on } \Gamma_r,\\ \kappa_s\frac{\partial y}{\partial n_0} + 4\,\varepsilon\sigma|\bar{y}|^3 y &= 0 && \text{on } \Gamma_0. \end{aligned} \tag{7.2}$$

Note that (7.2) constitutes a special case of (6.1).

In section 7.2, we transform the variational inequality (7.1) into the standard projection formula depending on the adjoint state p.

7.1. Differentiability of the control-to-state operator. We show the differentiability of S by the implicit function theorem. To that end, let us introduce an auxiliary operator T such that S is implicitly defined by $T(u,S(u)) = 0$. Preparing the definition of T, we consider the following equation that is equivalent to the weak formulation of the semilinear PDE (4.1):

$$\begin{aligned}&\int_\Omega \kappa\nabla y\cdot\nabla v\,dx + \int_{\Gamma_0}\lambda\,y\,v\,ds\\ &\qquad = -\int_{\Gamma_r} G(\sigma|y|^3 y)v\,ds + \int_{\Gamma_0}(\lambda\,y + \varepsilon\sigma\,y_0^4 - \varepsilon\sigma\,|y|^3 y)\,v\,ds + \int_{\Omega_s} u\,v\,dx\end{aligned} \tag{7.3}$$

with a fixed $\lambda > 0$. Due to the positivity of λ, the left-hand side in (7.3) represents a bounded, coercive bilinear form in $H^1(\Omega)$. Thus, for every right-hand side in $H^1(\Omega)^*$, we have a unique solution in $H^1(\Omega)$. Furthermore, similar to the proof of Theorem 6.2, it follows that this solution is bounded if the right-hand side is sufficiently regular, i.e., in $L^2(\Omega)\times L^4(\Gamma_r)\times L^4(\Gamma_0)$. Thus, continuous operators $\tilde{B}_{\Omega_s} : L^2(\Omega_s) \to V^\infty$, $\tilde{B}_0 : L^4(\Gamma_0) \to V^\infty$, and $\tilde{B}_r : L^4(\Gamma_r) \to V^\infty$ exist with

$$y = \tilde{B}_{\Omega_s}\,u - \tilde{B}_r\,(G(\sigma|y|^3 y)) + \tilde{B}_0\,(\lambda\,y + \varepsilon\sigma\,y_0^4 - \varepsilon\sigma\,|y|^3 y),$$

and our auxiliary operator is given by

$$T(u,y) := y - \tilde{B}_{\Omega_s}\, u + \tilde{B}_r\,(G(\sigma|y|^3 y)) - \tilde{B}_0\,(\lambda\, y + \varepsilon\sigma\, y_0^4 - \varepsilon\sigma\,|y|^3 y) \tag{7.4}$$

with $T : L^2(\Omega_s) \times V^\infty \to V^\infty$.

THEOREM 7.1. *Assume that $u_a \geq 0$, $y_0 \in L^{16}(\Gamma_0)$, $y_0 \geq \vartheta > 0$, and $\lambda = 1$ is not an eigenvalue of B, where $B = B(\bar{y})$ is as defined in Theorem* 6.1. *Then, $S : L^2(\Omega_s) \to V^\infty$ is Fréchet differentiable at $(\bar{u}, \bar{y})$, and $y := S'(\bar{u})(u - \bar{u})$ is given by the solution of the linearized equation* (7.2).

Proof. According to the definition of T in (7.4) and because of Theorems 4.1 and 4.2, $y = S(u)$ if and only if $T(u,y) = T(u, S\,u) = 0$, since this equation corresponds to the semilinear equation (4.1). To prove the differentiability of S, it thus suffices to verify the hypotheses of the implicit function theorem for T. This is continuous Fréchet differentiability of T and continuous invertibility of $\frac{\partial T}{\partial y}(\bar{u}, \bar{y})$.

The Nemytskii operator $\Phi(y) = |y|^3 y$ is continuously Fréchet differentiable from $L^\infty(\Gamma_r \cup \Gamma_0)$ to $L^\infty(\Gamma_r \cup \Gamma_0)$; see [11]. The other operators in the definition of T are all continuous and linear operators, and, thus, trivially continuously Fréchet differentiable on their particular spaces, i.e., G from $L^\infty(\Gamma_r)$ to $L^\infty(\Gamma_r)$, $\tilde{B}_{\Omega_s}$ from $L^2(\Omega_s)$ to V^∞, $\tilde{B}_r$ from $L^\infty(\Gamma_r)$ to V^∞, and $\tilde{B}_0$ from $L^\infty(\Gamma_r)$ to V^∞. By the chain rule, T is continuously Fréchet differentiable from $L^2(\Omega_s) \times V^\infty$ to V^∞, and, in particular, continuous.

It remains to show the invertibility of $\frac{\partial T}{\partial y}(\bar{u}, \bar{y})$. For a given $f \in V^\infty$, we have to prove the existence of a unique $y \in V^\infty$ satisfying $\frac{\partial T}{\partial y}(\bar{u}, \bar{y})y = f$, i.e., in view of (7.4),

$$y + 4\tilde{B}_r\,(G(\sigma|\bar{y}|^3 y)) = f + \tilde{B}_0\,(\lambda\, y - 4\varepsilon\sigma\,|\bar{y}|^3 y).$$

With the substitution $y = f - w \in V^\infty$, this is equivalent to the weak formulation

$$\begin{aligned}\int_\Omega \kappa \nabla w \cdot \nabla v\, dx + \int_{\Gamma_r} 4\, G(\sigma|\bar{y}|^3 w) v\, ds + \int_{\Gamma_0} 4\,\varepsilon\sigma\,|\bar{y}|^3 w\, v\, ds \\ = \int_{\Gamma_r} 4\, G(\sigma|\bar{y}|^3 f) v\, ds - \int_{\Gamma_0} (\lambda - 4\,\varepsilon\sigma\,|\bar{y}|^3) f\, v\, ds \quad \forall\, v \in H^1(\Omega),\end{aligned}$$

whose bilinear form coincides with the form of (6.2). Here, the inhomogeneity is given by $f_\Omega = 0$, $f_0 := (\varepsilon\sigma\,|\bar{y}|^3 - \lambda) f \in L^\infty(\Gamma_0)$, and $f_r := 4\, G(\sigma|\bar{y}|^3 f) \in L^\infty(\Gamma_r)$. Thus, it fulfills the hypotheses of Theorems 6.1 and 6.2. We therefore have $w \in V^\infty$, and, hence, to every $f \in V^\infty$, there exists a unique solution $y \in V^\infty$ that yields the desired bijectivity of $\frac{\partial T}{\partial y}(\bar{u}, \bar{y})$.

Now, since all hypotheses are satisfied, we can apply the implicit function theorem to (7.4) and obtain, as a derivative of S,

$$y := S'(\bar{u})(u - \bar{u}) = -\left(\frac{\partial T}{\partial y}(\bar{u}, \bar{y})\right)^{-1} \frac{\partial T}{\partial u}(\bar{u}, \bar{y})(u - \bar{u}). \tag{7.5}$$

Finally, a straightforward computation shows that y satisfies (7.5) if and only if y is the solution of the linearized equation (7.2). ∎

7.2. The adjoint equation. A standard technique formally gives the following adjoint equation associated with (P):

$$
\begin{aligned}
\operatorname{div}(\kappa_{\mathrm{g}}\,\nabla p) &= \Delta\bar{y} - \operatorname{div} z \quad \text{in } \Omega_{\mathrm{g}},\\
\operatorname{div}(\kappa_{\mathrm{s}}\,\nabla p) &= 0 \qquad\qquad\quad \text{in } \Omega_{\mathrm{s}},\\
\kappa_{\mathrm{s}}\left(\frac{\partial p}{\partial n_{\mathrm{r}}}\right)_{\mathrm{s}} - \kappa_{\mathrm{g}}\left(\frac{\partial p}{\partial n_{\mathrm{r}}}\right)_{\mathrm{g}} + 4\sigma\,|\bar{y}|^3 G^* p &= \frac{\partial y}{\partial n_{\mathrm{r}}} - z\cdot n_{\mathrm{r}} \quad \text{on } \Gamma_{\mathrm{r}},\\
\kappa_s\,\frac{\partial p}{\partial n_0} + 4\varepsilon\sigma\,|\bar{y}|^3 p &= 0 \qquad\qquad\quad \text{on } \Gamma_0.
\end{aligned}
\tag{7.6}
$$

Formal integration by parts, also on the right-hand side, yields the corresponding weak formulation of (7.6):

$$
\begin{aligned}
&\int_\Omega \kappa\nabla p\cdot\nabla v\,dx \;+\; 4\int_{\Gamma_{\mathrm{r}}} \sigma\,|\bar{y}|^3\,G^*(p)\,v\,ds + 4\int_{\Gamma_0}\varepsilon\sigma\,|\bar{y}|^3\,p\,v\,ds\\
&\quad = \int_{\Omega_{\mathrm{g}}} (\nabla\bar{y} - z)\cdot\nabla v\,dx =: \langle w, v\rangle \quad \forall\, v\in H^1(\Omega)
\end{aligned}
\tag{7.7}
$$

with $w\in H^1(\Omega)^*$, since $\bar{y}\in V^\infty$ and $z\in L^2(\Omega_{\mathrm{g}})$ by (A3).

To show the existence of a unique solution to (7.7), we use a technique similar to that for the linearized equation (6.1) based on the Fredholm alternative (see section 6). To that end, we transform (7.7) into

$$
\int_\Omega \kappa\nabla p\cdot\nabla v\,dx + 4\int_{\Gamma_0}\varepsilon\sigma\,|\bar{y}|^3\,p\,v\,ds = \langle w, v\rangle - 4\int_{\Gamma_{\mathrm{r}}}\sigma\,|\bar{y}|^3\,G^*(p)\,v\,ds.
$$

Due to the positivity of $\bar{y}$, the bilinear form defined by the left-hand side is bounded and coercive in $H^1(\Omega)$. Thus, we again obtain linear continuous operators $B_\Omega : H^1(\Omega)^*\to H^1(\Omega)$ and $B_{\mathrm{r}} : L^2(\Gamma_r)\to H^1(\Omega)$ such that

$$
p = B_\Omega\, w + B_{\mathrm{r}}\,(-4\,\sigma|\bar{y}|^3\,G^*(\tau_{\mathrm{r}}p))
$$

(similar to (6.3)). Now we can argue as in the proof of Theorem 6.1 to obtain the following result.

THEOREM 7.2. *Assume $\bar{y}\in V^\infty$, $\bar{y}\ge\vartheta>0$, and that $\lambda=1$ is not an eigenvalue of $B(\bar{y})(\,\cdot\,) := -\tau_{\mathrm{r}}\,B_{\mathrm{r}}(4\,\sigma|\bar{y}|^3\,G^*(\,\cdot\,))$ with $B(\bar{y}) : L^2(\Gamma_{\mathrm{r}})\to L^2(\Gamma_{\mathrm{r}})$. Then, to every $w\in H^1(\Omega)^*$, there exists a unique solution p of (7.7) in $H^1(\Omega)$.*

Now, if we choose $v=p$ as a test function in the weak formulation of the linearized equation (7.2), we obtain

$$
\int_\Omega \kappa\nabla y\cdot\nabla p\,dx \;+\; 4\int_{\Gamma_{\mathrm{r}}} G(\sigma\,|\bar{y}|^3 y)p\,ds + 4\int_{\Gamma_0}\varepsilon\sigma\,|\bar{y}|^3\,y\,p\,ds = \int_{\Omega_{\mathrm{s}}}(u-\bar{u})p\,dx.
$$

On the other hand, we insert $v=y$ into the weak formulation of the adjoint equation:

$$
\int_\Omega \kappa\nabla p\cdot\nabla y\,dx \;+\; 4\int_{\Gamma_{\mathrm{r}}} \sigma\,|\bar{y}|^3\,y\,G^*(p)\,ds + 4\int_{\Gamma_0}\varepsilon\sigma\,|\bar{y}|^3\,p\,y\,ds = \int_{\Omega_{\mathrm{g}}}(\nabla\bar{y}-z)\cdot\nabla y\,dx.
$$

Substracting one equation from the other yields $(\nabla\bar{y} - z, \nabla y)_{L^2(\Omega_g)} = (u - \bar{u}, p)_{L^2(\Omega_s)}$ for the first expression in the variational inequality (7.1) . Thus, (7.1) can be transformed into

$$j'(\bar{u})(u - \bar{u}) = \int_{\Omega_s} (u - \bar{u})(p + \nu\bar{u})\, dx \geq 0 \quad \forall\, u \in U_{\mathrm{ad}},$$

and, finally, a standard pointwise discussion leads to the projection formula

$$\bar{u}(x) = \mathcal{P}_{[u_a, u_b]} \left\{ -\frac{1}{\nu}\, p(x) \right\} \quad \text{a.e. on } \Omega_s, \tag{7.8}$$

where $\mathcal{P}_{[u_a,u_b]}$ denotes the pointwise projection operator on $[u_a, u_b]$ (cf. Lions [13]).

8. Numerical tests. In this section, we report on some two-dimensional numerical tests. For the computational domain, we chose the square presented in Figure 1.1, which is naturally academic. In contrast to this, the material parameters were close to approximating the realistic distributions given in [15]. Two different temperature levels were investigated in our calculations. At the lower level, the temperature in the gas phase is in the range of the external temperature $y_0 \equiv 293.0$ K, whereas, at the higher level, it constitutes approximately 2000 K and, thus, is sufficiently high for the PVT method mentioned in section 1. In the low-temperature example, the thermal conductivity[2] in the gas phase is fixed at $\kappa_g \equiv 0.0175$ and, having in mind a graphite crucible, at $\kappa_s \equiv 37.0$ in the solid. At the higher temperatures, we obtain $\kappa_g \equiv 0.08$ and $\kappa_s \equiv 24.0$. The emissivity is set to $\varepsilon \equiv 0.65$ at the lower temperature level and $\varepsilon \equiv 0.8$ in the high-temperature case. In all cases, the Boltzmann radiation constant[3] is given by $\sigma = 5.6696 \cdot 10^{-8}$.

To solve the optimal control problems numerically, we used a sequential quadratic programming (SQP) method (see, for instance, [1]). The linearized subproblems were solved with a primal-dual active set algorithm (cf. [3] or [4]). The overall method is described in detail in [14]. Using a projected gradient method with the Armijo step size rule, suitable initial values were generated. The linearized PDEs were approximated by linear finite elements. Since the kernel ω exhibits a singularity as mentioned in section 4, the numerical treatment of K and G, respectively, needs some care, especially in the corners of Ω_g. However, thanks to the simple shape of Γ_r in our case, it is possible to integrate ω exactly, thereby circumventing these problems (see [14] for details). For our numerical investigations, we used a triangle mesh consisting of 98340 points and 196358 elements. Due to the nonlocal radiation boundary condition, it is refined four times on the inner boundary Γ_r. Furthermore, since the corners of Γ_r are nonconvex with respect to the outer domain Ω_s (see Figure 1.1), the mesh is additionally refined eight times in each of these corners up to a local mesh size of approximately $1 \cdot 10^{-5}$.

We present three numerical examples. The first example, at the lower temperature level, represents an exact solution of (P) that is defined in the following. For the Tikhonov parameter, we choose $\nu = 0.01$. The desired temperature gradient[4] is given by $z \equiv (0, 0)^T$, and we take $u_a \equiv 0.0$ and $u_b \equiv 1000$ for the control constraints.[5] Then, it is clear that the global optimal solution to (P) is given by

$$\bar{u} \equiv 0, \qquad p \equiv 0, \qquad \bar{y} = y_0 \equiv 293.0\,\mathrm{K}.$$

[2] in W/(m K).
[3] in W/(m^{2}K^4).
[4] in K/m.
[5] in W/m^3.

TABLE 8.1
Convergence history for the first example.

it	$J(y,u)$	$1/2\|\nabla y - z\|^2_{L^2(\Omega_g)}$	$\nu/2\|u\|^2_{L^2(\Omega_s)}$
1	2.343477e+02	5.696541e-03	2.343420e+02
2	1.464461e+01	3.567923e-04	1.464425e+01
3	9.151546e-01	2.223112e-05	9.151324e-01
4	5.718883e-02	1.381736e-06	5.718745e-02
1	7.772308e-09	7.769828e-09	2.480914e-12
2	7.945074e-09	7.942384e-09	2.689580e-12
3	8.067200e-09	8.064853e-09	2.346440e-12

TABLE 8.2
Comparison with the exact solution.

$\|u_h - \bar{u}\|_{L^2(\Omega_s)}$	$\|y_h - \bar{y}\|_{L^2(\Omega)}$	$\|p_h - \bar{p}\|_{L^2(\Omega)}$
2.166306e-05	2.582871e-06	9.120640e-06

In view of Lemma 3.5, it is easy to verify that, with this optimal solution, the state equation (4.1), the adjoint equation (7.7), and the projection formula (7.8) are satisfied. Moreover, we have $J(\bar{y},\bar{u}) = 0$, and thus the solution represents a global optimum. To demonstrate the convergence behavior of the SQP method, the values of the numerically evaluated objective functional are shown in Table 8.1. Here, the rows above the double line refer to the projected gradient method for the initial value search (rows 1–4). The values that correspond to the SQP method are presented in rows 5–7. Note that the initial value was given by $u_h \equiv 500.0$ with an objective functional of 3.750087e+03. As one can see, the SQP method produces a speed-up of the convergence towards the global optimum $J(\bar{y},\bar{u}) = 0$. In what follows, we denote by the subscript h the approximated solution computed by the SQP method. Table 8.2 shows the difference between the discrete solution and the given exact solution. Due to $\bar{u} = \bar{p} = 0$, absolute errors were used. The relative error of the state is $\|y_h - \bar{y}\|_{L^2(\Omega)}/\|\bar{y}\|_{L^2(\Omega)} = 4.407631\text{e}{-09}$.

In contrast to the first example, no exact solution is known in the two following examples. To deal with higher temperatures in the gas phase, we now set $u_a = 125000$ and $u_b = 725000$. To compensate for the comparatively large values of the control, we choose a small Tikhonov regularization parameter $\nu = 5 \cdot 10^{-10}$. The desired temperature gradient is given by $z = (0,-20)^T$. Figures 8.1–8.4 show the optimal solution for this setting. The optimal control shown in Figure 8.1 is defined only in the solid parts Ω_s, and, therefore, its graph shows a hole in the inner square, where the gas phase is located.

As Figure 8.4 illustrates, there are significant differences between the optimal temperature distribution in the gas phase and the desired temperature gradient. With a value of approximately 17 K, the temperature difference between the lower and upper edges of Γ_r is smaller than the desired 20 K. Furthermore, at the upper edge, the temperature is not constant in the x_1-direction. A possible explanation for this result is the strong cooling effect due to the relatively low external temperature. Because of the comparatively large difference of approximately 1700 K between the temperature in the gas phase and the external temperature outside the crucible, one obtains quite steep gradients in the solid part Ω_s, as one can see in Figure 8.3. Therefore, it is not possible to generate the desired flat temperature distribution. In Figure 8.1, we

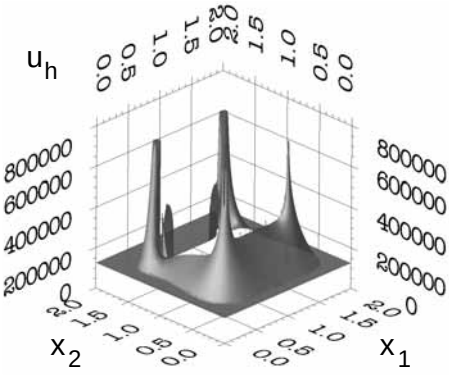


FIG. 8.1. *Control u_h in the second example.*

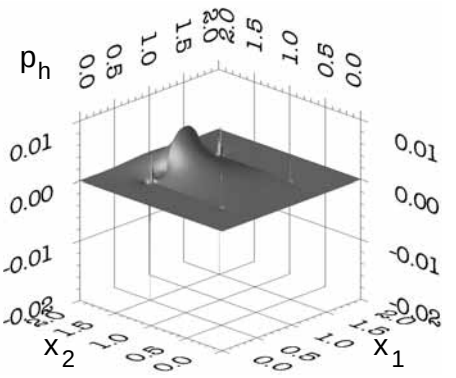


FIG. 8.2. *Adjoint state p_h in the second example.*

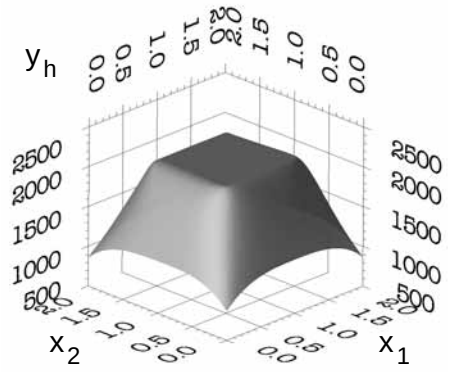


FIG. 8.3. *State y_h in the second example.*

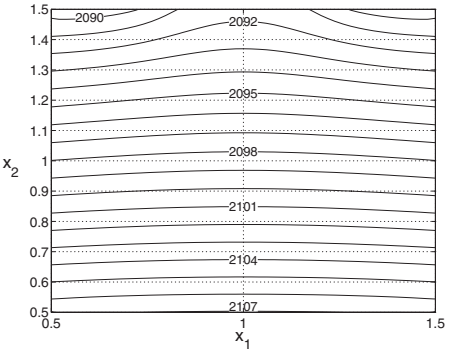


FIG. 8.4. *Isotherms in the gas phase Ω_g in the second example.*

observe noticeable peaks of u_h in the corners of Ω_g. Since the mesh is refined in these corners several times, as mentioned above, this does not seem to show a numerical effect.

Table 8.3 illustrates the development of the objective functional during the iteration. The initial value was given by $u_h \equiv 312500$. The corresponding initial value of the objective functional is 4.921569e+02. As one can see, the algorithm needed more iterations of the projected gradient method than in the first example to find a suitable initial value, i.e., to reach the area of convergence of the SQP method. Again, a strong decrease of the objective functional is observed after the first SQP iteration. To validate the numerical solution, we introduce the following indicators for the relative errors in the optimality system of (P) consisting of the state equation (4.1), the adjoint equation (7.7), and the projection formula (7.8). The error in the state equation is measured by

$$e_y = \frac{\|\hat{y} - y_h\|_{L^2(\Omega)}}{\|y_h\|_{L^2(\Omega)}} \quad \text{with } \hat{y} = S_h(u_h),$$

where S_h is the numerical solution operator of the semilinear equation (4.1). To be more precise, (4.1) is solved numerically by Newton's method on the continuous level such that only the linear PDEs in each Newton step are discretized using finite

TABLE 8.3
Convergence history for the second example.

it	$J(y,u)$	$1/2\|\nabla y - z\|^2_{L^2(\Omega_g)}$	$\nu/2\|u\|^2_{L^2(\Omega_s)}$
1	2.261592e+02	1.551524e+02	7.100687e+01
2	1.754913e+02	1.052975e+02	7.019378e+01
3	1.461291e+02	7.696015e+01	6.916894e+01
4	1.268671e+02	5.873924e+01	6.812786e+01
5	1.136532e+02	4.654510e+01	6.710807e+01
6	1.043283e+02	3.820065e+01	6.612768e+01
7	9.754344e+01	3.235032e+01	6.519312e+01
8	9.241677e+01	2.811423e+01	6.430253e+01
9	8.840054e+01	2.494841e+01	6.345213e+01
1	3.576389e+01	1.106707e+01	2.469682e+01
2	3.735623e+01	9.940953e+00	2.741527e+01
3	3.743465e+01	9.845256e+00	2.758939e+01
4	3.743563e+01	9.844839e+00	2.759079e+01

TABLE 8.4
Errors for the second example.

e_u	e_y	e_p
1.871817e-05	3.475214e-11	8.845293e-05

elements (see [14] for details). Next, we introduce the solution operator associated with the discrete version of the adjoint equation (7.7) denoted by $S_h'(u_h)^*\,\hat{w}$ with $\langle \hat{w}\,,\, v\rangle = \int_{\Omega_g}(\nabla\hat{y} - z)\cdot\nabla v\,dx$. Then, to verify the discrete adjoint equation, we compute

$$e_p = \frac{\|\hat{p} - p_h\|_{L^2(\Omega)}}{\|p_h\|_{L^2(\Omega)}} \quad \text{with } \hat{p} = S_h'(u_h)^*\hat{w}.$$

Finally, as an indicator for the error in the projection formula (7.8), the SQP solution u_h is compared with the projection of $\hat{p}$, i.e.,

$$e_u = \frac{\|\hat{u} - u_h\|_{L^2(\Omega_s)}}{\|u_h\|_{L^2(\Omega_s)}} \quad \text{with } \hat{u} = \mathcal{P}_{ad}\left\{-\frac{1}{\nu}\hat{p}\right\}.$$

The results are shown in Table 8.4.

In the third example, we set $u_a = 200000$, $u_b = 300000$, and $z = (0, -100)^T$. As in the second example, ν is fixed at $5\cdot 10^{-10}$.

Since the x_2-direction of z is comparatively steep, one cannot expect the desired temperature gradient to be reached with these bounds on the control. Thus, although the optimal control is almost bang-bang (cf. Figure 8.5), the optimal temperature distribution differs considerably from the desired one, as Figure 8.7 and, in particular, Figure 8.8 demonstrate. Figure 8.6 shows the adjoint state in this example. The convergence behavior in this case is illustrated by Table 8.5. We choose $u_h = 230000$ as the initial value. The corresponding value of the objective functional amounts to 5.314663e+03. In this case, there is no noteworthy decrease of the objective functional after starting the SQP method. An explanation for this observation is that the bang-bang structure of the control is already reached after the first projected gradient step. Table 8.6 shows the errors in the optimality system, introduced above, for this example.

TABLE 8.5
Convergence history for the third example.

it	$J(y,u)$	$1/2\|\nabla y - z\|^2_{L^2(\Omega_g)}$	$\nu/2\|u\|^2_{L^2(\Omega_s)}$
1	4.214458e+03	4.169869e+03	4.458834e+01
2	4.194483e+03	4.149988e+03	4.449523e+01
1	4.183678e+03	4.138581e+03	4.509680e+01
2	4.183709e+03	4.138611e+03	4.509842e+01
3	4.183710e+03	4.138612e+03	4.509841e+01
4	4.183710e+03	4.138612e+03	4.509843e+01

TABLE 8.6
Errors for the third example.

e_u	e_y	e_p
3.376173e-07	2.028424e-10	1.485374e-04

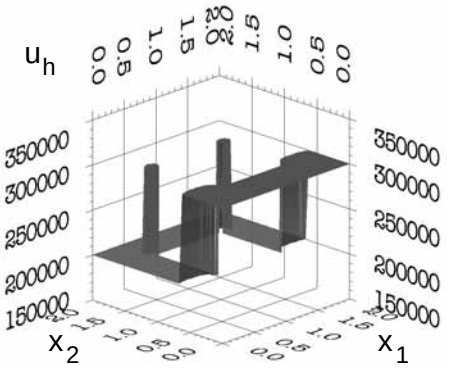


FIG. 8.5. *Control u_h in the third example.*

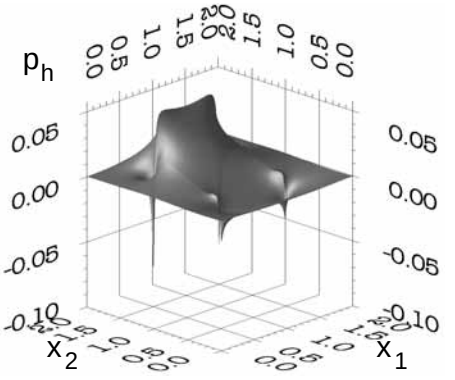


FIG. 8.6. *Adjoint state p_h in the third example.*

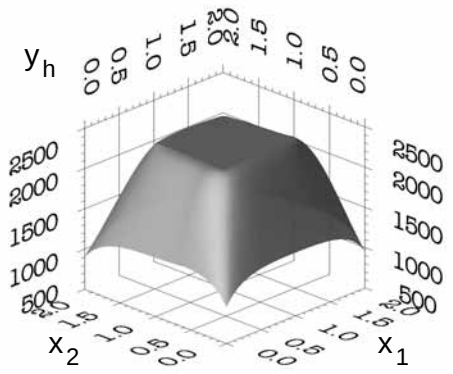


FIG. 8.7. *State y_h in the third example.*

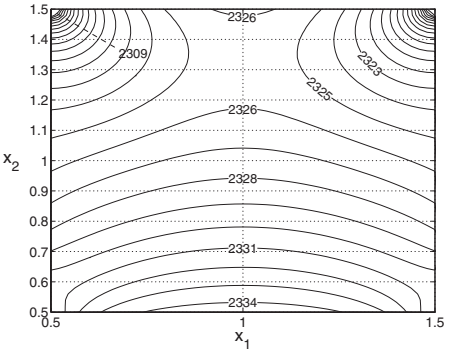


FIG. 8.8. *Isotherms in the gas phase Ω_g in the third example.*

REFERENCES

[1] W. Alt and K. Malanowski, *The Lagrange-Newton method for nonlinear optimal control problems*, Comput. Optim. Appl., 2 (1993), pp. 77–100.

[2] J.-J. Alibert and J.-P. Raymond, *Boundary control of semilinear elliptic equations with discontinuous leading coefficients and unbounded controls*, Numer. Funct. Anal. Optim., 18 (1997), pp. 235–250.

[3] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[4] M. Bergounioux and K. Kunisch, *Primal-dual strategy for state-constrained optimal control problems*, Comput. Optim. Appl., 22 (2002), pp. 193–224.

[5] J. Bonnans, *Second-order analysis for control constrained optimal control problems of semilinear elliptic systems*, Appl. Math. Optim., 38 (1998), pp. 303–325.

[6] E. Casas, *Pontryagin's principle for optimal control problems governed by semilinear elliptic equations*, in Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, Internat. Ser. Numer. Math. 118, Birkhäuser, Basel, 1994, pp. 97–114.

[7] E. Casas, F. Tröltzsch, and A. Unger, *Second order sufficient optimality conditions for a nonlinear elliptic control problem*, J. Anal. Appl., 15 (1996), pp. 687–707.

[8] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Classics Appl. Math. 31, SIAM, Philadelphia, 2000.

[9] O. Klein, P. Philip, and J. Sprekels, *Modeling and simulation of sublimation growth of SiC bulk single crystals*, Interfaces Free Bound., 6 (2004), pp. 295–314.

[10] A. Konstantinov, *Sublimation growth of SiC*, in Properties of Silicon Carbide, EMIS Datareview Series 13, G. Harris, ed., Institution of Electrical Engineers (INSPEC), London, UK, 1995, pp. 170–203.

[11] M. A. Krasnoselskii, P. P. Zabreiko, E. N. Pustylnik, and P. E. Sobolevskii, *Integral Operators in Spaces of Summable Functions*, Leyden, Noordhoff, 1976.

[12] M. Laitinen and T. Tiihonen, *Conductive-radiative heat transfer in grey materials*, Quart. Appl. Math., 59 (2001), pp. 737–768.

[13] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[14] C. Meyer, *An SQP active set method for a semilinear optimal control problem with nonlocal radiation interface conditions*, in Proceedings of the Oberwolfach Conference on Optimal Control of Coupled Systems of PDEs, accepted. Available online at http://www.math.tu-berlin.de/~cmeyer//Publicat/paper/ppr2005_33.pdf

[15] P. Philip, *Transient Numerical Simulation of Sublimation Growth of SiC Bulk Single Crystals. Modeling, Finite Volume Method, Results*, Ph.D. thesis, Department of Mathematics, Humboldt University of Berlin, Germany; Report no. 22, Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, 2003. Available online at http://www.wias-berlin.de/publications/reports/22/wias_reports_22.pdf

[16] H.-J. Rost, D. Siche, J. Dolle, W. Eiserbeck, T. Müller, D. Schulz, G. Wagner, and J. Wollweber, *Influence of different growth parameters and related conditions on 6H-SiC crystals grown by the modified Lely method*, Mater. Sci. Eng. B, 61/62 (1999), pp. 68–72.

[17] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[18] T. Tiihonen, *A nonlocal problem arising from heat radiation on non-convex surfaces*, European J. Appl. Math., 8 (1997), pp. 403–416.

[19] T. Tiihonen, *Stefan-Boltzmann radiation on non-convex surfaces*, Math. Methods Appl. Sci., 20 (1997), pp. 47–57.

[20] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, *Vol.* II/B. *Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 722–761

OPTIMAL CONTROL OF HYBRID SYSTEMS AND A SYSTEM OF QUASI-VARIATIONAL INEQUALITIES*

HUAN ZHANG† AND MATTHEW R. JAMES‡

Abstract. Optimal control of a class of hybrid dynamical systems is studied using the method of dynamic programming. It is proved that the value function is a discontinuous viscosity solution of the corresponding Hamilton–Jacobi–Bellman (HJB) equation which appears as a system of quasi-variational inequalities (SQVI) coupled by a nonlocal operator with a variable boundary condition. The comparison theorems are established for the sub- and super-solutions of the SQVI.

Key words. hybrid dynamical systems, optimal control, dynamic programming, system of quasi-variational inequalities, viscosity solutions, comparison theorem

AMS subject classifications. 34H05, 34K35, 49L20, 49L25

DOI. 10.1137/050630131

1. Introduction.

In recent years, there has been increasing attention given to the study of hybrid dynamical systems. This has been stimulated by increasingly complex applications that involve a combination of discrete and continuous dynamics. The name "hybrid" reflects the mixed nature of the dynamics. The continuous dynamics generally represents the underlying time-driven, continuous-variable physical processes which are usually modeled by a family of differential (or difference) equations, while the high level event-driven discrete dynamics represents some kind of decision making logic regarding the continuous dynamics. The discrete dynamics is usually modeled as timed automata or petri nets. The continuous dynamics and the discrete dynamics interact at some "event" times when the continuous state hits some prescribed sets in the continuous state space. Examples of these systems are found in many areas including production systems (see Cassandras, Pepyne, and Wardi [13]), traffic systems (see Varaiya [25] and Lygeros, Godbole, and Sastry [20]), and air traffic management systems (see Tomlin, Pappas, and Sastry [24]).

Different models have been proposed to model hybrid systems at different levels of generality; among them we mention Branicky, Borkar, and Mitter [11]; Heemels, Schutter, and Bemporad [18]; and Lygeros, Tomlin, and Sastry [21]. In this paper we follow the model of Bensoussan and Menaldi [9], which is actually a simplified version—but with no loss of generality—of the unified model developed in [11]. In both [9] and [11], it is shown that these models are quite general models and include most of the hybrid models in the literature as their special cases.

Optimal control problems of general hybrid systems have been studied by using the maximum principle (see Shaikh and Caines [22] and Sussmann [23]) and dynamic programming (see Branicky, Borkar, and Mitter [11], Bensoussan and Menaldi [9],

*Received by the editors April 28, 2005; accepted for publication (in revised form) November 23, 2005; published electronically June 2, 2006. Support from NICTA is gratefully acknowledged. National ICT Australia is funded by the Australian government's Backing Australia's Ability initiative, in part through the Australian Research Council.

http://www.siam.org/journals/sicon/45-2/63013.html

†Department of Engineering, The Australian National University, Canberra, ACT 0200, Australia (Huan.Zhang@anu.edu.au).

‡Corresponding author. Department of Engineering, The Australian National University, Canberra, ACT 0200, Australia (Matthew.James@anu.edu.au).

and, more recently, Dharmatti and Ramaswamy [16]). In [11], the Hamilton–Jacobi–Bellman (HJB) equations corresponding to the optimal control problem were derived, but no further analysis on the equations has been done. In a more recent work by Ball, Chudoung, and Day [1], the continuous value function corresponding to robust ($\mathcal{H}^\infty$-type) control is proved to satisfy the dynamic programming equations in the continuous viscosity sense; however, their system dynamics model is more special, and they did not provide a uniqueness proof on the solutions of these equations, which, in fact, is crucial to the synthesis of optimal controllers in the approach of dynamic programming. There is a rather complete treatment of the optimal control problem using the dynamic programming approach in [9]. The authors proved that the continuous value function is the unique viscosity solution of the corresponding system of quasi-variational inequalities (SQVI) in some special cases. In [16], the authors proved the uniqueness problem under more general conditions using a different approach. But as our example shows, discontinuous value functions arise frequently for general hybrid optimal control problems. Thus it is desirable to have a theory of optimal hybrid control in the framework of discontinuous value functions. In this paper, we study the optimal control problem for a general class of hybrid systems in the context of discontinuous value functions. In the case of a continuous value function, our results reduce to those of [9] and [16].

The main problem in applying the dynamic programming approach to solve the optimal control problem is how to interpret and, more importantly, characterize the value function as the unique solution of the corresponding HJB equation. In our problem, the HJB equation is

$$\begin{cases} \max\{\mathbf{v}(x,q) - (\mathcal{M}\mathbf{v})(x,q), \lambda\mathbf{v}(x,q) + \mathcal{H}(x,q,\nabla_x\mathbf{v}(x,q))\} = 0 & \text{in } F^c, \\ \mathbf{v}(x,q) = (\mathcal{M}\mathbf{v})(x,q) & \text{on } \partial F^c, \end{cases} \tag{1}$$

where $q \in Q = \{q_1, q_2, \dots, q_p\}$, F^c is the complement of a closed set F, and $\mathcal{M}$ is a nonlocal operator defined by

$$(\mathcal{M}v)(x,q) = \min_{w\in W}\{\kappa(x,q,w) + v(g(x,q,w))\}. \tag{2}$$

Thus it is an SQVI coupled by a nonlocal operator $\mathcal{M}$ with a variable boundary condition. Several issues regarding the SQVI make the characterization process difficult. The first difficulty is the fact that the value function is in general discontinuous due to the structure of the hybrid dynamics and the possible discontinuities of the jumping costs; hence we can interpret it as a solution of the SQVI only in some weak sense (here we employ the notion of discontinuous viscosity solutions introduced by Ishii [19]). It is known that under this notion of weak solution, uniqueness cannot be guaranteed in general (see Barles and Perthame [4], [5] and Blanc [10]); rather, only the maximum sub-solution and minimum super-solution can be characterized. The second difficulty is that the SQVI is coupled by a nonlocal operator which makes it hard to apply the approach of Capuzzo Dolcetta and Evans in [12] when we come to characterize the solution even in the special case of a continuous value function, since the traditional techniques in viscosity solution framework are local in nature. Extra difficulties arise in the boundary behavior of this SQVI, which we see is a variable boundary condition rather than a prescribed function.

In this paper, we generalize the approach of Barles [3] and Bensoussan and Menaldi [9] to the discontinuous case to resolve the difficulties. The essential point of the approach is to introduce two concave operators, whose unique fixed points turn

out to be the maximum sub-solution and minimum super-solution, respectively, and it actually can be shown that they are solutions. Using the two operators, we further show that the two extreme solutions admit the same low semi-continuous envelope, which is the minimum solution. By this we establish the so-called uniqueness in a weak sense; i.e., all solutions have the same lower semi-continuous envelope or the equation has a unique lower semi-continuous solution. Note that our approach in establishing the comparison theorems is essentially different from those of Barles and Perthame [4], [5] and Blanc [10].

There are some similarities between our results and those of the optimal impulse control problem (see Barles [3]), the optimal switching problem (see Capuzzo Dolcetta and Evans [12]), the optimal stopping problem (see Barles and Perthame [4]), and optimal exit problems (see Barles and Perthame [5] and Blanc [10]). Actually, since the structure of the hybrid dynamics involves multiple exits and multiple impulses, the optimal control problem of hybrid dynamical systems can be regarded as the generalization and integration of the above control problems, and under appropriate conditions, the results and approaches fit these special cases.

We begin section 2 by specifying the hybrid model we use and the optimal control problem to be solved. In section 3 we present the dynamic programming principle for the problem. Then in section 4 we prove that the value function is a discontinuous viscosity solution of the associated SQVI. The comparison theorems are given in section 5. Some special cases, including the continuous case, are discussed in section 6. As another application of the approach, we briefly treat the hybrid optimal stopping problem in section 7.

Notation. Given $X \subset \mathbf{R}^n$, a finite set Q, and a set $S \subset (X \times Q)$, denote $S^c = \{(x,q)|(x,q) \in ((X \times Q)\backslash S)\}$ and $S_q = \{x|(x,q) \in S\}$. $\overline{S_q}$ is the closure of S_q in X, ∂S_q is the boundary of S_q and $\partial S = \{(x,q)|x \in \partial S_q\}$, and $\widetilde{S}_q$ is the interior of S_q and $\widetilde{S} = \{(x,q)|x \in \widetilde{S}_q\}$. Given a bounded function $\mathbf{v} : S \to R$, denote by $\mathbf{v}^*$ ($\mathbf{v}_*$) its upper (lower) envelope defined as $\mathbf{v}^*(x,q) = \limsup_{y \in S_q, y \to x} \mathbf{v}(y,q)$,($\mathbf{v}_*(x,q) = \liminf_{y \in S_q, y \to x} \mathbf{v}(y,q)$). $\mathbf{v}$ is called upper semi-continuous (lower semi-continuous) in S if $\mathbf{v}^* = \mathbf{v}$ ($\mathbf{v}_* = \mathbf{v}$) in S. Denote by $USC(S)$ ($LSC(S)$) the set of upper semi-continuous (lower semi-continuous) functions in S, and for a function $\mathbf{v} : X \times Q \to R$, denote by $\mathbf{v}|_S$ the restriction of $\mathbf{v}$ on S.

2. Problem formulation. In this section, we first propose a model of general hybrid dynamical systems and then describe in some detail its trajectory properties and conditions to guarantee that it is well behaved. After this we formulate the infinite horizon discounted optimal control problem.

2.1. Controlled hybrid dynamical systems. Hybrid systems are loosely defined as dynamical systems involving the interaction of continuous state and discrete state dynamics. More precisely, we have the following definition.

DEFINITION 2.1. *A controlled hybrid dynamical system* $\mathbf{H}$ *is a collection*

$$\mathbf{H} = \{X, Q, U, W, f, g, F, O\}, \tag{3}$$

where the items are defined as follows:

- *Hybrid state space $X \times Q$: $X \subset \mathbf{R}^n$ is the continuous state space, and Q is the discrete state space which is a finite set with discrete topology.*
- *Hybrid input space $U \times W$: $U \subset \mathbf{R}^m$ is the continuous input space which is a compact subset of $\mathbf{R}^m$, and W is the discrete input space which is a finite set with discrete topology.*

- *Continuous dynamics f:* $f : X \times Q \times U \to R^n$ *is a family of vector fields describing a family of continuous dynamics.*
- *Discrete dynamics g:* $g : X \times Q \times W \to X \times Q$ *is the discrete transition map.*
- *Forced and optional discrete transition sets* F, O*:* $F, O \subset (X \times Q)$ *are closed subsets of the state space which designate the domain of the discrete dynamics. Let* $D = F \cup O$ *be the discrete transition set.*

To ensure that the dynamics is well behaved, we impose the following conditions on the hybrid systems.

Assumption 2.2. The continuous dynamics f satisfies the following:

1. There exists $C_f > 0$ such that

$$\begin{cases} |f(x,q,u)| \le C_f \quad \forall (x,q,u) \in (X \times Q \times U), \\ |f(x,q,u) - f(y,q,u)| \le C_f |x-y| \quad \forall x, y \in X, \quad \forall q \in Q, \quad \forall u \in U. \end{cases} \tag{4}$$

2. ∂X is smooth and

$$f(x,q,u) \cdot n(x) \ge 0 \quad \forall x \in (\partial X \backslash \partial F_q), \quad \forall q \in Q, \quad \forall u \in U,$$

where $n(x)$ is the unit inner normal to X.

3. ∂F_q is smooth for all $q \in Q$ and

$$f(x,q,u) \cdot n(x) > \alpha > 0 \quad \forall x \in \partial F_q, \quad \forall q \in Q, \quad \forall u \in U,$$

where $n(x)$ is the unit inner normal to F_q.

Assumption 2.3. The discrete dynamics g satisfies

$$\begin{cases} |g(x,q,w)| \le C_g \quad \forall (x,q,w) \in (X \times Q \times W), \\ g(\cdot,q,w) \text{ is continuous in } X \quad \forall q \in Q, \quad \forall w \in W. \end{cases} \tag{5}$$

Remark 2.4. Item 1 in Assumption 2.2 guarantees the existence of the solutions of the continuous dynamics for any initial condition. Item 2 simplifies the behavior of the continuous dynamics on the boundaries of the continuous state space, implying that the state will never exit the state space X by continuous evolution. When X is closed and convex, we can achieve this by projecting outward pointing vector fields to the boundary as suggested in [15].

Remark 2.5. The transversality condition 3 in Assumption 2.2 eliminates the pathological behavior of the continuous dynamics on the boundary of forced jump set F; see [11] for discussion and relaxation. Since the forced discrete transition set F_q acts as an exit set in the hybrid dynamics, this assumption simplifies the exit behavior through the boundary of F_q. Some of our subsequent results and proofs would be more complicated without this assumption. We don't assume a transversality condition for the optional discrete transition set O which is made in [16]. A transversality condition on O implies that when the state enters O, it can escape O only by doing a discrete transition. This is a condition inappropriate for O in a sense that discrete transitions in O are "optional." However, as our example shows, discontinuous value functions can arise without this assumption.

Now, let us see how a controlled hybrid dynamical system works. The execution of the controlled hybrid dynamical system operates as follows:

1. At any time, one of the continuous dynamics $f(\cdot, q, \cdot)$ must be active, while the discrete dynamics can be activated only in some isolated times. When the discrete dynamics is active, it has higher priority than the continuous dynamics.

2. When the state is inside F, the discrete dynamics must be activated, and it must not be activated when the state is outside D.
3. When the state is inside O, the discrete dynamics may be but does not have to be activated.

Under the rules, in order to distinguish the two different types (forced and optional) of discrete transitions, we impose the following assumption.

Assumption 2.6. F and O are strictly separated; i.e.,

$$d(F_q, O_q) > 0 \quad \forall q \in Q.$$

The continuous dynamics changes the continuous state continuously under a continuous input signal and does not change the discrete state. The activation of the discrete dynamics causes a discrete transition which changes the hybrid state to a new state instantaneously. A discrete transition can be a "JUMP" which means only a change in the continuous state variable or a "SWITCH" which means only discrete state changes, or a combination of "JUMP" and "SWITCH." Since we assume here that it takes no time to have a discrete transition, we need to eliminate the pathological executions of multiple discrete transitions at one single time and an infinite number of discrete transitions in any finite period of time that are called "Zeno executions." According to Bensoussan and Menaldi [9], the following assumption is sufficient to eliminate both due to the forced jumps.

Assumption 2.7. g satisfies

$$\inf_{w\in W,(x,q),(x'q')\in F}\{|x' - g|_X(x,q,w)| + |q' - g|_Q(x,q,w)|\} \geq c > 0,$$

where $g|_X : X \times Q \times W \to X$, $g|_Q : X \times Q \times W \to Q$ are the projections of g to X and Q, respectively; i.e., $(g|_X(x,q,w), g|_Q(x,q,w)) = g(x,q,w)$ for all $x \in X$, for all $q \in Q$, for all $w \in W$.

Following the rules, the trajectories of the continuous state are piecewise continuous where discontinuities appear (if the discrete transition is not a SWITCH) at the times of discrete transitions. At every discrete transition time τ_i, let $x(\tau_i^-)$ be the prejump state and let $x(\tau_i^+)$ be the postjump state. The discrete state trajectories are piecewise constant and have discontinuities in the discrete transition time (if it is not a JUMP).

We can see that a control signal of a hybrid system consists of two parts: continuous input and discrete input. A continuous input is a measurable function $u : \mathbf{R}^+ \to U$, which takes effect through the continuous dynamics. Let $\mathcal{U}$ denote the set of all continuous control input. The discrete control takes the form $\{(\tau_i, w_i)\}_1^N$, where

$$0 = \tau_0 \leq \tau_1 \leq \cdots \tau_i \leq \tau_{i+1} \cdots$$

is the sequence of times selected to activate the discrete dynamics to cause discrete transitions. At time τ_i the destination of the discrete transition is parameterized by the discrete input value $w_i \in W$. The total number of activations N can be finite or infinite.

Let us call $\beta = (u(\cdot), \{(\tau_i, w_i)\}_1^N)$ a hybrid control. Obviously, in order to follow the rules of the hybrid dynamical system, not all hybrid controls are acceptable. The two kinds of controls (continuous and discrete) are not independent of each other. Given initial states $(x,q) \in (X \times Q)$ and a continuous control $u(\cdot) \in \mathcal{U}$, we have to activate the discrete dynamics if the state enters F, and we must not activate the

discrete dynamics when the state is outside D. So we have to select those acceptable hybrid controls which we call admissible.

Given any $(x,q) \in (X \times Q)$ and $u(\cdot) \in \mathcal{U}$, define hitting times of F and D to be

$$T^F(x,q,u(\cdot)) = \inf\{t \geq 0 | (x(t),q) \in F\}, \quad T^D(x,q,u(\cdot)) = \inf\{t \geq 0 | (x(t),q) \in D\}, \tag{6}$$

where $x(\cdot)$ is the solution of

$$\dot{x}(t) = f(x(t),q,u(t))$$

with initial state x.

DEFINITION 2.8. *Given initial state $(x,q) \in (X \times Q)$, a hybrid control $\beta = (u(\cdot), \{(\tau_i, w_i)\}_1^N)$ is called an admissible hybrid control with respect to (x,q) if*

1.

$$0 = \tau_0 \leq \tau_1 \leq \cdots \tau_i \leq \tau_{i+1} \cdots, \quad \tau_i \to \infty \text{ if } i \to \infty;$$

2.

$$T^D(x(\tau_i^+), q_i, u(\cdot)) \leq \tau_{i+1} - \tau_i \leq T^F(x(\tau_i^+), q_i, u(\cdot)) \quad \forall i \geq 0.$$

Let $\beta(x,q)$ denote an admissible hybrid control with respect to (x,q) and let $\mathcal{B}(x,q)$ denote the set of all admissible hybrid controls with respect to (x,q).

Any trajectory that starts from an initial hybrid state driven by an admissible hybrid control with respect to the initial state is called an admissible hybrid trajectory. Based on the assumptions on the dynamics and the rules, we know that there exists one and only one admissible hybrid trajectory for any initial hybrid state under any admissible hybrid control input. We now construct a typical hybrid trajectory inductively. Given any initial state $(x,q) \in (X \times Q)$ and $\beta(x,q) \in \mathcal{B}(x,q)$, a hybrid state trajectory is produced as follows:

1. $(x(\tau_0^+), q(\tau_0^+)) = (x,q)$;
2. for $i = 1,2,\ldots,$
   if $\tau_i > \tau_{i-1}$, then for $t \in [\tau_{i-1}, \tau_i)$,
   (a) $\begin{cases} q(t) = q(\tau_{i-1}^+) = q_{i-1} \\ \dot{x}(t) = f(x(t), q(t), u(t)) \end{cases}$ with initial condition $(x(\tau_{i-1}^+), q(\tau_{i-1}^+))$;
   (b) $\begin{cases} q(\tau_i^-) = q(\tau_{i-1}) = q_{i-1} \\ x(\tau_i^-) = \lim_{t \to \tau_i^-} x(t) \end{cases}$;
   if $\tau_i = \tau_{i-1}$, then $(x(\tau_i^-), q(\tau_i^-)) = (x(\tau_{i-1}^+), q(\tau_{i-1}^+))$;
3. $(x(\tau_i^+), q(\tau_i^+)) = g(x(\tau_i^-), q(\tau_i^-), w_i)$.

Notice that if we use $Q = \{1,2,\ldots,p\}$ to represent discrete state space Q, the hybrid dynamics can be written in compact form as

$$\begin{cases} \dot{x}(t) = f(x(t), n(t), u(t)) + \sum_i^N (g|_X(x(\tau_i^-), q(\tau_i^-), w_i) - x(\tau_i^-))\delta(t - \tau_i), \\ \dot{n}(t) = \sum_i^N (g|_Q(x(\tau_i^-), q(\tau_i^-), w_i) - n(\tau_i^-))\delta(t - \tau_i), \end{cases} \tag{7}$$

where δ is the Dirac measure.

2.2. Optimal control problem. Now we associate with every admissible hybrid trajectory a cost which, not surprisingly, is also hybrid, i.e., consists of continuous costs and discrete costs. The continuous costs penalize the continuous running of the hybrid dynamical system and take the form $\ell : X \times Q \times U \to \mathbf{R}^+$. The discrete part is with the form of $\rho : X \times Q \times W \to \mathbf{R}^+$, which represents how much you need to pay for a discrete transition.

We assume the costs satisfy the following assumptions.

Assumption 2.9. The running cost ℓ satisfies

$$\begin{cases} \ell(x,q,u) \le C_\ell \quad \forall (x,q,u) \in (X \times Q \times W), \\ |\ell(x,q,u) - \ell(y,q,u)| \le C_\ell |x-y| \quad \forall x,y \in X, \quad \forall q \in Q, \quad \forall u \in U. \end{cases} \tag{8}$$

Assumption 2.10.

$$\begin{cases} 0 < C_\rho < \rho(x,q,w) < C'_\rho \quad \forall (x,q,w) \in (X \times Q \times W), \\ (\rho^*)_* = \rho_*. \end{cases} \tag{9}$$

Given any initial hybrid state $(x,q) \in (X \times Q)$ and any admissible hybrid control input $\beta(x,q) \in \mathcal{B}(x,q)$, an admissible hybrid trajectory is produced. Then define the (discounted) total cost for the trajectory as

$$\mathbf{J}_{(x,q)}(\beta) = \sum_{i=1}^{N} \left(\int_{\tau_{i-1}}^{\tau_i} \ell(x(s), q_{i-1}, u(s)) e^{-\lambda s} ds + \rho(x(\tau_i^-), q_{i-1}, w_i) e^{-\lambda \tau_i} \right), \tag{10}$$

where $\lambda > 0$ is the discount factor.

The *optimal control of hybrid dynamical systems* is to find an admissible hybrid control input $(\beta^*(x,q))$ for every hybrid state $(x,q) \in (X \times Q)$, such that

$$\mathbf{J}_{(x,q)}(\beta^*) \le \mathbf{J}_{(x,q)}(\beta)$$

for any $\beta(x,q) \in \mathcal{B}(x,q)$.

We will use the approach of dynamic programming to study this problem. In dynamic programming, the key concept is the so-called value function defined for this problem as

$$\mathbf{V}(x,q) = \inf_{\beta(x,q) \in \mathcal{B}(x,q)} \{\mathbf{J}_{(x,q)}(\beta)\} \tag{11}$$

for every initial hybrid state $(x,q) \in (X \times Q)$. In the remainder of the paper we characterize it (if possible) as the unique solution (in some weak sense) of a partial differential equation called an HJB equation.

The following result shows that the value function is nonnegative and bounded.

PROPOSITION 2.11. *The value function is nonnegative and bounded; i.e., there exists $C_v > 0$ such that*

$$0 \le \mathbf{V}(x,q) \le C_v \quad \forall (x,q) \in (X \times Q).$$

Proof. It is obvious that the value function should be nonnegative since all the costs are nonnegative. To prove the boundedness, fix any initial state $(x,q) \in (\mathbf{R}^n \times Q)$. Choose a special admissible hybrid control $\hat{\beta}(x,q) \in \mathcal{B}(x,q)$, such that every discrete transition time $\tau_i, i \ge 1$ is a forced discrete transition time, i.e., $(x(\tau_i^-), q(\tau_i^-)) \in F$ for all $i \ge 1$. Then based on Assumption 2.7, and Theorem

2.1 of Bensoussan and Menaldi [9], we know that there exists an $h > 0$ independent of β such that

$$\tau_{i+1} - \tau_i \geq h$$

for all $i \geq 1$. Hence

$$\begin{aligned}\mathbf{V}(x,q) &\leq \mathbf{J}_{(x,q)}(\hat{\beta}) \\ &\leq \int_0^\infty C_\ell e^{-\lambda s} ds + \sum_{i=1}^{N} C'_\rho e^{-\lambda i h} \\ &= C_\ell \frac{1}{\lambda} + C'_\rho \frac{e^{-\lambda h}(1 - e^{-\lambda N h})}{1 - e^{-\lambda h}} \\ &\leq C_\ell \frac{1}{\lambda} + C'_\rho \frac{e^{-\lambda h}}{1 - e^{-\lambda h}}. \qquad \square\end{aligned}$$

Due to the complex structure of general hybrid dynamical systems, in the general case, the value function could be discontinuous. For general hybrid dynamical systems without the transversality condition in item 3 of Assumption 2.2, three factors contribute to the discontinuities of the value functions. The first is the existence of the forced jump set F_q. F_q acts as the role of exit set for the continuous dynamics. From Barles and Perthame [5] and Blanc [10], we know that the value function is discontinuous for the optimal exit problem unless some compatibility like that of the transversality condition is imposed. The second factor is the restriction of the optional discrete transitions on the domain O, which actually can be turned into the equivalent case of a discontinuous discrete transitional cost without the constraint of O. The last factor is that we allow a discontinuous discrete transition cost ρ. The first factor was eliminated due to the tranversality condition; hence the discontinuity of the value functions are the results of O and ρ. Next we give a simple example to make it clear that the existence of O does lead to a discontinuous value function.

Example 2.12. Consider a hybrid system with $X = \mathbf{R}$, $Q = \{0, 1\}$, $U = [1, 2]$, $W = \{w\}$, $F = \varnothing$, $O_0 = O_1 = [1, 2]$, and continuous dynamics

$$f(x,q,u) = \begin{cases} u & \text{when } q = 0, \\ -u & \text{when } q = 1 \end{cases}$$

and discrete dynamics

$$g(x,q,w) = \begin{cases} (x,1) & \text{when } q = 0, \\ (x,0) & \text{when } q = 1. \end{cases}$$

The costs are as follows:

$$\ell(x,q,u) = \begin{cases} \frac{1}{u} & \text{when } q = 0, \\ \frac{1}{6} & \text{when } q = 1 \end{cases}$$

and discrete transition cost

$$\rho(x,q,w) = \frac{1}{6}.$$

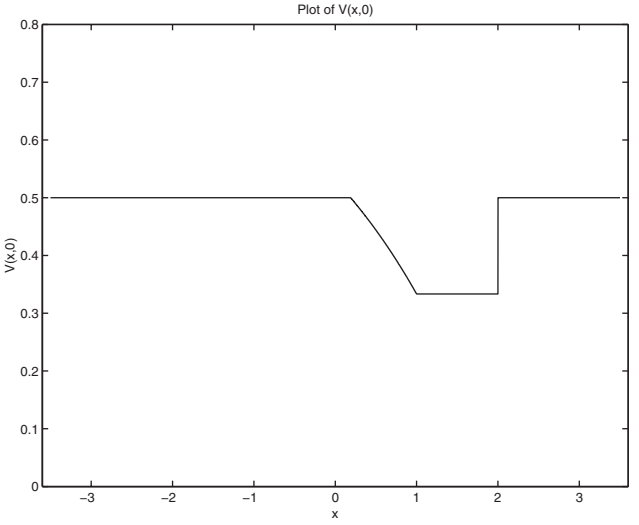


FIG. 1. *Value function* $\mathbf{V}(x,0)$.

Let the discount factor $\lambda = 1$. Then all the data satisfies the assumptions, the value functions can be computed by hand as

$$\mathbf{V}(x,0) = \begin{cases} \frac{1}{2} & \text{when} \quad x > 2, \\ \frac{1}{3} & \text{when} \quad 1 \le x \le 2, \\ \frac{5}{6} - \frac{1}{2}e^{\frac{x-1}{2}} & \text{when} \quad 1 + 2\ln\frac{2}{3} \le x < 1, \\ \frac{1}{2} & \text{when} \quad -1 < x < 1 + 2\ln\frac{2}{3}, \end{cases}$$

and

$$\mathbf{V}(x,1) = \frac{1}{6}, \quad x \in \mathbf{R}.$$

From Figure 1, we can easily see that $\mathbf{V}(x,0)$ is discontinuous at $x = 2$. In this simple example, since $F = \varnothing$ and the discrete transition cost ρ is continuous, the discontinuity of the value function is the result of the existence of O without a transversality condition.

In this paper, we study the optimal control of a hybrid dynamical system without assuming a continuous value function. Hence the analysis of this paper can be regarded as the discontinuous generalization of [9]. In the case of a continuous value function, all of our arguments reduce to the case of paper [9].

3. Dynamic programming principle. Now we turn to study the value function by the approach of dynamic programming. In this section, we develop the so-called dynamic programming principle (DPP) satisfied by the value function, which leads to a system of first order partial differential equations in section 4.

First, let us introduce a function

$$\kappa(x,q,w) = \begin{cases} \rho(x,q,w), & (x,q) \in D, \\ C_\kappa & \text{otherwise}, \end{cases} \tag{12}$$

where $C_\kappa > C_\ell \frac{1}{\lambda} + C'_\rho \frac{e^{-\lambda h}}{1-e^{-\lambda h}}$ is a constant, and a set

$$\mathcal{V} = \{\mathbf{v} : X \times Q \to \mathbf{R} | \mathbf{v} \text{ bounded}\}.$$

Define nonlocal operators $\mathcal{M} : \mathcal{V} \to \mathcal{V}$:

$$(\mathcal{M}\mathbf{v})(x,q) = \min_{w\in W}\{\kappa(x,q,w) + \mathbf{v}(g(x,q,w))\}; \tag{13}$$

$\mathcal{M}^+ : \mathcal{V} \to \mathcal{V}$:

$$(\mathcal{M}^+\mathbf{v})(x,q) = \min_{w\in W}\{\kappa^*(x,q,w) + \mathbf{v}^*(g(x,q,w))\}; \tag{14}$$

$\mathcal{M}^- : \mathcal{V} \to \mathcal{V}$:

$$(\mathcal{M}^-\mathbf{v})(x,q) = \min_{w\in W}\{\kappa_*(x,q,w) + \mathbf{v}_*(g(x,q,w))\}. \tag{15}$$

We give properties of the operators $\mathcal{M}$, $\mathcal{M}^+$, and $\mathcal{M}^-$ in the following lemma.

LEMMA 3.1. *For any function $\mathbf{v} \in \mathcal{V}$, we have*

1. *$\mathcal{M}^+\mathbf{v}^* \in USC(X \times Q)$ and $\mathcal{M}^+\mathbf{v}^* \geq (\mathcal{M}\mathbf{v})^*$;*
2. *$\mathcal{M}^-\mathbf{v}_* \in LSC(X \times Q)$ and $\mathcal{M}^-\mathbf{v}_* \leq (\mathcal{M}\mathbf{v})_*$.*

Proof. Item 1. Fix any $q \in Q$. For any $x \in X$, let w^* be $(\mathcal{M}^+\mathbf{v}^*)(x,q) = \kappa^*(x,q,w^*) + \mathbf{v}^*(g(x,q,w^*))$. For any sequence $x_n \to x$, since $g(\cdot,q,w^*)$ is continuous, we know that $g(x_n,q,w^*) \to g(x,q,w^*)$. Then

$$\begin{aligned}(\mathcal{M}^+\mathbf{v}^*)(x,q) &= \kappa^*(x,q,w^*) + \mathbf{v}^*(g(x,q,w^*))\\ &\geq \limsup_{n\to\infty}\kappa^*(x_n,q,w^*) + \limsup_{n\to\infty}\mathbf{v}^*(g(x_n,q,w^*))\\ &\geq \limsup_{n\to\infty}(\kappa^*(x_n,q,w^*) + \mathbf{v}^*(g(x_n,q,w^*)))\\ &\geq \limsup_{n\to\infty}(\mathcal{M}^+\mathbf{v}^*)(x_n,q),\end{aligned}$$

which proves the upper semicontinuity of $\mathcal{M}^+\mathbf{v}^*$. Since $\mathcal{M}\mathbf{v} \leq \mathcal{M}^+\mathbf{v}^*$,

$$(\mathcal{M}\mathbf{v})^* \leq (\mathcal{M}^+\mathbf{v}^*)^* = \mathcal{M}^+\mathbf{v}^*.$$

Item 2. Fix any $q \in Q$. For any $x \in X$ and any sequence $x_n \to x$, let $w_n^* \in W$ be a sequence such that $(\mathcal{M}^-\mathbf{v}_*)(x_n,q) = \kappa_*(x_n,q,w_n^*) + \mathbf{v}_*(g(x_n,q,w_n^*))$. Since W is a finite set, we can select a constant subsequence, i.e., $w_n^* = w_*$. Similarly, we have $g(x_n,q,w_*) \to g(x,q,w_*)$. Then

$$\begin{aligned}(\mathcal{M}^-\mathbf{v}_*)(x,q) &\leq \kappa_*(x,q,w_*) + \mathbf{v}_*(g(x,q,w_*))\\ &\leq \liminf_{n\to\infty}\kappa_*(x_n,q,w_*) + \liminf_{n\to\infty}\mathbf{v}_*(g(x_n,q,w_*))\\ &\leq \liminf_{n\to\infty}(\kappa_*(x_n,q,w_*) + \mathbf{v}_*(g(x_n,q,w_*)))\\ &= \liminf_{n\to\infty}(\mathcal{M}^-\mathbf{v}_*)(x_n,q),\end{aligned}$$

which proves the lower semicontinuity of $\mathcal{M}^-\mathbf{v}_*$. Since $\mathcal{M}v \geq \mathcal{M}^-\mathbf{v}_*$, then

$$(\mathcal{M}\mathbf{v})_* \geq (\mathcal{M}^-\mathbf{v}_*)_* = \mathcal{M}^-\mathbf{v}_*. \qquad \square$$

The next three theorems consist of the DPP for the value function.

THEOREM 3.2. *The value function* $\mathbf{V}$ *satisfies*

$$\mathbf{V}(x,q) = (\mathcal{M}\mathbf{V})(x,q) \quad \forall (x,q) \in F. \tag{16}$$

Proof. "$\leq$": Fix any $(x,q) \in F$ and $w \in W$, and denote $(x',q') = g(x,q,w)$. For any admissible hybrid control with respect to (x',q'), $\bar{\beta}(g(x,q,w)) = (\bar{u}(\cdot), \{\bar{\tau}_i, \bar{w}_i\}_1^N)$, construct a new hybrid control $\beta = (u(\cdot), \{\tau_i, w_i\}_1^{N+1})$ by

$$\begin{cases} u(t) = \bar{u}(t), \\ \tau_1 = 0, w_1 = w, \tau_i = \bar{\tau}_{i-1}, w_i = \bar{w}_{i-1}, \quad i \geq 2. \end{cases}$$

It is not difficult to see that β is admissible with respect to (x,q); i.e., $\beta(x,q) \in \mathcal{B}(x,q)$. Then

$$\begin{aligned} \mathbf{V}(x,q) &\leq \mathbf{J}_{(x,q)}(\beta) \\ &= \kappa(x,q,w) + \mathbf{J}_{(g(x,q,w))}(\bar{\beta}), \end{aligned}$$

and since $\bar{\beta}(g(x,q,w)) \in \mathcal{B}(g(x,q,w))$ is arbitrary,

$$\begin{aligned} \mathbf{V}(x,q) &\leq \kappa(x,q,w) + \inf_{\bar{\beta}(g(x,q,w)) \in \mathcal{B}(g(x,q,w))} \{\mathbf{J}_{(g(x,q,w))}(\bar{\beta})\} \\ &= \kappa(x,q,w) + \mathbf{V}(g(x,q,w)). \end{aligned}$$

Since we are choosing w arbitrarily, we have

$$\begin{aligned} \mathbf{V}(x,q) &\leq \min_{w \in W} \{\kappa(x,q,w) + \mathbf{V}(g(x,q,w))\} \\ &= (\mathcal{M}\mathbf{V})(x,q). \end{aligned}$$

"$\geq$": For any $(x,q) \in F$ and $\epsilon > 0$, there exists an admissible hybrid control to (x,q), say $\beta(x,q) = (u(\cdot), \{\tau_i, w_i\}_1^N)$, such that

$$\mathbf{V}(x,q) + \epsilon > \mathbf{J}_{(x,q)}(\beta).$$

According to the definition of the admissibility of the hybrid control, $\tau_1 = 0$, and thus

$$\begin{aligned} \mathbf{J}_{(x,q)}(\beta) &= \kappa(x,q,w) + \mathbf{J}_{(g(x,q,w))}(\bar{\beta}) \\ &\geq \kappa(x,q,w) + \mathbf{V}(g(x,q,w)) \\ &\geq (\mathcal{M}\mathbf{V})(x,q), \end{aligned}$$

where $\bar{\beta}(g(x,q,w)) = (\bar{u}(\cdot), \{\bar{\tau}_i, \bar{w}_i\}_1^{N-1})$ is defined as

$$\begin{cases} \bar{u}(t) = u(t), \\ \bar{\tau}_i = \tau_{i+1}, \ \bar{w}_i = w_{i+1}, \quad i \geq 1. \end{cases}$$

Hence

$$\mathbf{V}(x,q) \geq (\mathcal{M}\mathbf{V})(x,q). \qquad \square$$

THEOREM 3.3. *The value function* $\mathbf{V}$ *satisfies the following:*

1. *For any* $(x,q) \in F^c$,

$$\mathbf{V}(x,q) \leq (\mathcal{M}\mathbf{V})(x,q). \tag{17}$$

2. *For any* $u(\cdot) \in \mathcal{U}$ *and any* $0 \le T < T^F(x,q,u(\cdot))$,

$$\mathbf{V}(x,q) \le \int_0^T \ell(x(s),q,u(s))e^{-\lambda s}ds + \mathbf{V}(x(T),q)e^{-\lambda T}, \tag{18}$$

where $x(t) = \int_0^t f(x(t),q,u(t))dt - x$.

Proof. Proof of (17): If $(x,q) \in D^c$, then (17) is trivially true since

$$\mathbf{V}(x,q) \le C_\ell \frac{1}{\lambda} + C'_\rho \frac{e^{-\lambda h}}{1-e^{-\lambda h}} \le C_\kappa = \kappa(x,q,w) \le (\mathcal{M}\mathbf{V})(x,q).$$

If $(x,q) \in O$, then (17) can be proved in exactly the same way as the "$\le$" part of Theorem 3.2.

Proof of (18): Fix any $u(\cdot) \in \mathcal{U}$; then we can get $(x(T),q)$ by $x(T) = \int_0^T f(x(t), q, u(t))dt - x$. Since F is a closed subset, we know that $T^F(x,q,u(\cdot)) > 0$ for $(x,q) \in F^c$ and for any $0 \le T < T^F(x,q,u(\cdot))$, $(x(T),q) \in F^c$. Given any admissible control $\bar\beta(x(T),q) = (\bar u(\cdot), \{\bar\tau_i, \bar w_i\}_1^N)$, construct a new control $\tilde\beta = (\tilde u(\cdot), \{\tilde\tau_i, \tilde w_i\}_1^N)$ by

$$\begin{cases} \tilde u(t) = \begin{cases} u(t), & 0 \le t \le T, \\ \bar u(t-T), & t > T, \end{cases} \\ \tilde\tau_i = \bar\tau_i + T, \ \tilde w_i = \bar w_i, \quad i \ge 1. \end{cases}$$

Then we know that $\tilde\beta \in \mathcal{B}(x,q)$, i.e., the new control is admissible to (x,q), and hence

$$\begin{aligned} \mathbf{V}(x,q) &\le \mathbf{J}_{(x,q)}(\tilde\beta) \\ &= \int_0^T \ell(x(s),q,u(s))e^{-\lambda s}ds + \mathbf{J}_{(x(T),q)}(\bar\beta)e^{-\lambda T}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{V}(x,q) &\le \int_0^T \ell(x(s),q,u(s))e^{-\lambda s}ds + \inf_{\bar\beta(x(T),q)\in\mathcal{B}(x(T),q)} \{\mathbf{J}_{(x(T),q)}(\bar\beta)\}e^{-\lambda T} \\ &= \int_0^T \ell(x(s),q,u(s))e^{-\lambda s}ds + \mathbf{V}(x(T),q)e^{-\lambda T}. \qquad \square \end{aligned}$$

Next, we establish the super-optimality principle for the lower envelope of the value function in the framework of relaxed control theory. Let $\mathbf{P}(U)$ be the space of probability measures on the continuous input space U. A relaxed continuous control is a map $\mu : [0,\infty) \to \mathbf{P}(U)$. Denote $\widehat{\mathcal{U}} = \{\mu_{(\cdot)}\}$. If we change the continuous dynamics to

$$\hat f(x,q,\mu) = \int_U f(x,q,u)d\mu \tag{19}$$

and the continuous running cost to

$$\hat\ell(x,q,\mu) = \int_U \ell(x,q,u)d\mu \tag{20}$$

and keep other data unchanged, then we get the optimal control problem in the context of a relaxed control framework. Let us denote by $\hat\beta(x,q) = (\mu_{(\cdot)}, \{\tau_i, w_i\}_1^N)$ a relaxed admissible hybrid control with respect to (x,q). The admissibility of a relaxed hybrid

control is defined in the same way as that of admissible hybrid controls except that the trajectory is replaced by the corresponding relaxed one. Let $\widehat{\mathcal{B}}(x,q) = \{\hat{\beta}(x,q)\}$.

Now we give the super-optimality principle for the value function.

THEOREM 3.4. *If for some $(x,q) \in F^c$, $\mathbf{V}_*(x,q) < (\mathcal{M}^-\mathbf{V}_*)(x,q)$, then there exists $t(x,q) > 0$ such that for all $0 \le T < t(x,q)$, we have*

$$\mathbf{V}_*(x,q) \ge \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}} \left\{ \int_0^T \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds + \mathbf{V}_*(\hat{x}(T),q) e^{-\lambda T} \right\}, \tag{21}$$

where $\mu_{(\cdot)}$ is the relaxed control and $\hat{x}(\cdot)$ is the solution of $\dot{\hat{x}}(t) = \int_U f(\hat{x}(t),q,u) d\mu_t$ with initial state x.

Proof. For any $(x,q) \in F^c$, choose $x_n \to x$ such that $\mathbf{V}(x_n,q) \to \mathbf{V}_*(x,q)$. Since F is closed, $x_n \in F_q^c$ for n large enough; thus we can assume that $x_n \in F_q^c$ is true for all n. For every n, let $\beta^n(x_n,q) \in \mathcal{B}(x_n,q)$ such that

$$\mathbf{V}(x_n,q) + 1/n \ge \mathbf{J}_{(x_n,q)}(\beta^n).$$

We now show that there must exist $t(x,q) > 0$ such that for n large enough, $(\tau^n)_1 > t(x,q)$. If $(x,q) \in D^c$, it is trivially true since the first discrete transition can happen only after the state enters D, and we notice that D is closed. So we need only show it is true for the case of $(x,q) \in O$. Assume for the sake of contradiction that there exists a subsequence (still denoted as $(\tau^n)_1$) such that $(\tau^n)_1 \to 0$. Since W is a finite set, we can assume there exists $w_1 \in W$ such that $(w^n)_1 = w_1$ for $n \ge 1$. Denote by $x^n(\cdot)$ the continuous state trajectory with initial condition (x_n,q) and control β^n. We know that

$$x^n((\tau^n)_1^-) \to x \quad \text{and} \quad g(x^n((\tau^n)_1^-),q,w_1) \to g(x,q,w_1)$$

by the continuity of the discrete transition dynamics. Hence

$$\begin{aligned}
\mathbf{V}_*(x,q) &= \lim_{n\to\infty} (\mathbf{V}(x_n,q) + 1/n) \\
&\ge \liminf_{n\to\infty} \Bigg(\int_0^{(\tau^n)_1} \ell(x^n(s),q,u^n(s)) e^{-\lambda s} ds + \kappa(x^n((\tau^n)_1^-),q,w_1) e^{-\lambda(\tau^n)_1} \\
&\qquad + \mathbf{V}(g(x^n((\tau^n)_1^-),q,w_1) e^{-\lambda(\tau^n)_1}) \Bigg) \\
&\ge \kappa_*(x,q,w_1) + \mathbf{V}_*(g(x,q,w_1)) \\
&\ge (\mathcal{M}^-\mathbf{V}_*)(x,q),
\end{aligned}$$

which contradicts the assumption. We have shown that for $0 \le T < t(x,q)$ and n large enough

$$\begin{aligned}
\mathbf{V}(x_n,q) + 1/n &\ge \mathbf{J}(x_n,q,\beta_n) \\
&\ge \int_0^T \ell(x^n(s),q,u^n(s)) e^{-\lambda s} ds + \mathbf{V}(x^n(T),q) e^{-\lambda T} \\
&\ge \int_0^T \ell(x^n(s),q,u^n(s)) e^{-\lambda s} ds + \mathbf{V}_*(x^n(T),q) e^{-\lambda T}.
\end{aligned}$$

If we regard the control sequence $u^n(\cdot)$ on $[0,T]$ as the relaxed controls, then there is a relaxed control $\mu_{(\cdot)}$ such that $u^n(\cdot) \to \mu_{(\cdot)}$ weakly. Noticing the fact that $x_n \to x$,

we know $x^n(s) \to \hat{x}(s)$ uniformly on $[0,T]$ and

$$\int_0^T \ell(x^n(s),q,u^n(s))e^{-\lambda s}ds \to \int_0^T \int_U \ell(\hat{x}(s),q,u)d\mu_s e^{-\lambda s}ds.$$

Hence it holds that

$$\begin{aligned}\mathbf{V}_*(x,q) &\geq \int_0^T \int_U \ell(\hat{x}(s),q,u)e^{-\lambda s}d\mu_s ds + \mathbf{V}_*(\hat{x}(T),q)e^{-\lambda T} \\ &\geq \inf_{\mu_{(\cdot)}\in\widehat{\mathcal{U}}}\left(\int_0^T \int_U \ell(\hat{x}(s),q,u)e^{-\lambda s}d\mu_s ds + \mathbf{V}_*(\hat{x}(T),q)e^{-\lambda T}\right). \qquad \square\end{aligned}$$

4. SQVI and discontinuous viscosity solution. In this section, we prove the value function to be a discontinuous viscosity solution of the HJB equations based on the DPP proved in the previous section. In our problem these equations appear as the following SQVI due to the existence of the discrete phenomena; furthermore, the existence of forced discrete transition set F gives the SQVI a variable boundary condition which is different from the usual boundary condition.

Consider a system of first order partial differential equations

$$\max\{\mathbf{v}(x,q) - (\mathcal{M}\mathbf{v})(x,q), \lambda\mathbf{v}(x,q) + \mathcal{H}(x,q,\nabla_x\mathbf{v}(x,q))\} = 0 \quad \text{in } F^c \tag{22}$$

with boundary condition

$$\mathbf{v}(x,q) = (\mathcal{M}\mathbf{v})(x,q) \quad \text{on } \partial F^c, \tag{23}$$

where the Hamiltonian $\mathcal{H} : X \times Q \times \mathbf{R}^n \to \mathbf{R}$ is defined as

$$\mathcal{H}(x,q,p) = \sup_{u\in U}\{-\langle f(x,q,u),p\rangle - \ell(x,q,u)\}. \tag{24}$$

The idea is to extract the information about the value function from the SQVI (22) and boundary condition (23). In the best possible case, (22) and (23) determine the value function completely in the region $(X \times Q)\backslash F$. Then from (16), we can compute the value function in area F, thus determining the whole value function.

Notice that the above SQVI constitutes a collection of $|Q| = p$ partial differential equations coupled together by the nonlocal operator $\mathcal{M}$. So the solutions are in the form of a vector

$$\mathbf{v} = (\mathbf{v}((\cdot),q_1),\mathbf{v}((\cdot),q_2),\dots,\mathbf{v}((\cdot),q_p))$$

or a function $\mathbf{v} : (X \times Q) \to \mathbf{R}$. A similar form of SQVI also appears in the optimal switching problem (see Capuzzo Dolcetta and Evans [12]), but the difference is that in their case, the SQVI is coupled by a local operator. It is also interesting to see that our boundary condition is not a prescribed function but is determined by the values of the solution inside F^c.

It is known that in general the value function can satisfy the HJB equations only in a weak sense such as in the sense of viscosity solutions rather than in the classic sense. In our problem, since the value function is not necessarily even continuous, we need to use the following notion of a discontinuous viscosity solution originally introduced by Ishii [19].

DEFINITION 4.1. *A bounded function* $\mathbf{v} : (\overline{F^c}) \to \mathbf{R}$ *such that* $\mathbf{v} \in USC(\overline{F^c})$ $(\mathbf{v} \in LSC(\overline{F^c}))$ *is a viscosity sub-solution (super-solution) of SQVI* (22) *with boundary*

condition (23) *in the viscosity sense if, for any* $\phi : \overline{F^c} \to \mathbf{R}$ *such that* $\phi \in C^1(\overline{F^c})$ *and for any* x_0 (x_1)*, the local maximum (local minimum) of* $\mathbf{v}(\cdot, q) - \phi(\cdot, q)$ *satisfies*

(25)
$$\begin{cases} \max\{\mathbf{v}(x_0,q) - (\mathcal{M}^+\mathbf{v})(x_0,q), \lambda\mathbf{v}(x_0,q) + \mathcal{H}(x_0,q,\nabla_x\phi(x_0,q))\} \le 0 & \text{if } x_0 \in F_q^c, \\ \mathbf{v}(x_0,q) \le (\mathcal{M}^+\mathbf{v})(x_0,q) & \text{if } x_0 \in \partial F_q^c, \end{cases}$$

respectively,

$$\max\{\mathbf{v}(x_1,q) - (\mathcal{M}^-\mathbf{v})(x_1,q), \lambda\mathbf{v}(x_1,q) + \mathcal{H}(x_1,q,\nabla_x\phi(x_1,q))\} \ge 0 \quad \forall\ x_1. \tag{26}$$

A bounded function $\mathbf{v} : \overline{F^c} \to \mathbf{R}$ *is a (discontinuous) viscosity solution of* (22) *with boundary condition* (23) *if* $\mathbf{v}^*$ ($\mathbf{v}_*$) *is a sub-solution (super-solution) of* (22) *with boundary condition* (23) *in the viscosity sense.*

Remark 4.2. Following the usual definition of discontinuous viscosity solution with boundary condition satisfied in the viscosity sense (see, for example, [5]), the inequality of our sub-solution property should be

$$\begin{cases} \max\{\mathbf{v}(x_0,q) - (\mathcal{M}^+\mathbf{v})(x_0,q), \lambda\mathbf{v}(x_0,q) + \mathcal{H}(x_0,q,\nabla_x\phi(x_0,q))\} \le 0 & \text{if } x_0 \in F_q^c, \\ \min\left\{ \begin{matrix} \mathbf{v}(x_0,q) - (\mathcal{M}^+\mathbf{v})(x_0,q), \\ \max\{\mathbf{v}(x_0,q) - (\mathcal{M}^+\mathbf{v})(x_0,q), \lambda\mathbf{v}(x_0,q) + \mathcal{H}(x_0,q,\nabla_x\phi(x_0,q))\}\} \le 0 \end{matrix} \right\} & \text{if } x_0 \in \partial F_q^c, \end{cases}$$

which is equivalent to (25). Similarly, for our super-solution property inequality, it should be

$$\begin{cases} \max\{\mathbf{v}(x_1,q) - (\mathcal{M}^-\mathbf{v})(x_1,q), \lambda\mathbf{v}(x_1,q) + \mathcal{H}(x_1,q,\nabla_x\phi(x_1,q))\} \ge 0 & \text{if } x_1 \in F_q^c, \\ \max\left\{ \begin{matrix} \mathbf{v}(x_1,q) - (\mathcal{M}^-\mathbf{v})(x_1,q), \\ \max\{\mathbf{v}(x_1,q) - (\mathcal{M}^-\mathbf{v})(x_1,q), \lambda\mathbf{v}(x_1,q) + \mathcal{H}(x_1,q,\nabla_x\phi(x_1,q))\}\} \ge 0 \end{matrix} \right\} & \text{if } x_1 \in \partial F_q^c, \end{cases}$$

which is the same as (26).

Now we come to the main result of this section.

THEOREM 4.3. *The value function* $\mathbf{V}|_{\overline{F^c}}$ *is a discontinuous viscosity solution of the SQVI* (22) *with boundary condition* (23) *in the sense of Definition* 4.1.

Proof of the sub-solution property. We need to show that for any ϕ and x_0 in Definition 4.1 such that $\mathbf{V}^*(\cdot, q) - \phi(\cdot, q)$ has a local maximum at $x_0 \in \overline{F^c}$, (25) holds for $\mathbf{V}^*(x_0, q)$. From (16), (17), and Lemma 3.1 we know that

$$\mathbf{V}^* \le (\mathcal{M}\mathbf{V})^* \le \mathcal{M}^+\mathbf{V}^* \quad \text{in } \overline{F^c};$$

then we need only show that

$$\lambda\mathbf{V}^*(x_0,q) + \mathcal{H}(x_0,q,\nabla_x\phi(x_0,q)) \le 0 \quad \text{in } F^c.$$

Fix any $u \in U$, and let $u(t) = u,\ t \ge 0$. Since F_q^c is open, there is a $\gamma > 0$ such that

$$r(x_0,q,u) = \inf_{x \in B(x_0,\gamma)} T^F(x,q,u(\cdot)) > 0.$$

Then, thanks to (18), for any $0 \le T < r(x_0, q, u)$, we have

$$\mathbf{V}(x,q) \le \int_0^T \ell(x(s),q,u)e^{-\lambda s}ds + \mathbf{V}(x(T),q)e^{-\lambda T}$$

for any $x \in B(x_0, \gamma)$, which implies

$$\mathbf{V}^*(x_0, q) \le \int_0^T \ell(x_0(s), q, u) e^{-\lambda s} ds + \mathbf{V}^*(x_0(T), q) e^{-\lambda T},$$

where $x_0(\cdot)$ is the continuous state trajectory with initial condition (x_0, q). On the other hand, for T small enough, we have

$$\mathbf{V}^*(x_0(T), q) \le \phi(x_0(T), q) + (\mathbf{V}^*(x_0, q) - \phi(x_0, q)),$$

and thus we obtain

$$\mathbf{V}^*(x_0, q)(1 - e^{-\lambda T})/T \le 1/T \Big(\int_0^T \ell(x_0(s), q, u) e^{-\lambda s} ds + (\phi(x_0(T), q) - \phi(x_0, q)) e^{-\lambda T} \Big).$$

Let $T \to 0$. We get

$$\lambda \mathbf{V}^*(x_0, q) - \nabla_x \phi(x_0, q) f(x_0, q, u) - \ell(x_0, q, u) \le 0.$$

Now we maximize over u and obtain

$$\lambda \mathbf{V}^*(x_0, q) + \sup_{u \in U} \{ -\nabla_x \phi(x_0, q) f(x_0, q, u) - \ell(x_0, q, u) \} \le 0.$$

This proves that $\mathbf{V}^*$ is a sub-solution.

To prove that $\mathbf{V}_*$ is also a super-solution, we need a known result in relaxed control theory; see, for example, section 2.2 of Chapter III in [2]. In our problem, let

$$\widehat{\mathcal{H}}(x, q, p) = \sup_{\mu \in \mathbf{P}(U)} \{ -\langle \hat{f}(x, q, \mu), p \rangle - \hat{\ell}(x, q, \mu) \}. \tag{27}$$

Then we have the following lemma.

LEMMA 4.4. *For any* $(x, q) \in (X \times Q)$ *and* $p \in \mathbf{R}^n$,

$$\mathcal{H}(x, q, p) = \widehat{\mathcal{H}}(x, q, p). \tag{28}$$

Proof. Choose relaxed controls $\mu = \delta_u$; then

$$\widehat{\mathcal{H}}(x, q, p) \ge \sup_{u \in U} \{ -\langle \hat{f}(x, q, \delta_u), p \rangle - \hat{\ell}(x, q, \delta_u) \} = \mathcal{H}(x, q, p).$$

On the other hand,

$$\begin{aligned} -\langle \hat{f}(x, q, \mu), p \rangle - \hat{\ell}(x, q, \mu) &= \int_U (-\langle f(x, q, u), p \rangle - \ell(x, q, u)) d\mu \\ &\le \int_U \sup_{u \in U} \{ -\langle f(x, q, u), p \rangle - \ell(x, q, u) \} d\mu \\ &= \mathcal{H}(x, q, p). \end{aligned}$$

Then $\widehat{\mathcal{H}}(x, q, p) \le \mathcal{H}(x, q, p)$. ∎

Now we can prove the super-solution property for $\mathbf{V}_*$.

Proof of the super-solution property. What we need to show is that for any ϕ in Definition 4.1 and any local minimum of $\mathbf{V}_*(\cdot,q)-\phi(\cdot,q)$ at $x_1\in\overline{F_q^c}$, (26) holds for $\mathbf{V}_*$. It is known in viscosity solution theory (see, for example, Remark 1.2 on p. 26 of [2]) that we can assume without loss of generality that $\mathbf{V}_*(x_1,q)=\phi(x_1,q)$; hence there exists $\gamma_1>0$ such that

$$\mathbf{V}_*(x,q)-\phi(x,q)\geq 0 \quad \forall x\in B(x_1,\gamma_1). \tag{29}$$

If

$$\mathbf{V}_*(x_1,q)=(\mathcal{M}^-\mathbf{V}_*)(x_1,q),$$

there is nothing to prove; then we assume that

$$\mathbf{V}_*(x_1,q)<(\mathcal{M}^-\mathbf{V}_*)(x_1,q)$$

and show that

$$\lambda\mathbf{V}_*(x_1,q)+\mathcal{H}(x_1,q,\nabla_x\phi(x_1,q))\geq 0.$$

Case 1. $(x_1,q)\in F^c$. We assume for the sake of contradiction that there exists $\delta>0$ such that

$$\lambda\mathbf{V}_*(x_1,q)+\mathcal{H}(x_1,q,\nabla_x\phi(x_1,q))<-2\delta.$$

From the assumption $\mathbf{V}_*(x_1,q)=\phi(x_1,q)$ and Lemma 4.4, we get

$$\lambda\phi(x_1,q)+\widehat{\mathcal{H}}(x_1,q,\nabla_x\phi(x_1,q))<-2\delta,$$

which means

$$\lambda\phi(x_1,q)-\langle\hat{f}(x_1,q,\mu),\nabla_x\phi(x_1,q)\rangle-\hat{\ell}(x_1,q,\mu)<-2\delta \quad \forall\mu\in\mathbf{P}(U).$$

Hence we know there exists $\gamma_2>0$ such that

$$\lambda\phi(x,q)-\langle\hat{f}(x,q,\mu),\nabla_x\phi(x,q)\rangle-\hat{\ell}(x,q,\mu)<-\delta,\ \ \forall x\in B(x_1,\gamma_2),\ \ \mu\in\mathbf{P}(U). \tag{30}$$

Based on the standard estimation of ordinary differential equations, we know there exists $t_0>0$ such that

$$\hat{x}(t)\in B(x_1,\gamma_1\wedge\gamma_2) \quad \forall 0\leq t<t_0,$$

where $\hat{x}(\cdot)$ is the solution of $\hat{x}(t)=\hat{f}((\hat{x})(t),q,\mu_t),\hat{x}(0)=x_1$ for any relaxed control $\mu\in\widehat{\mathcal{U}}$.

On the other hand, from Theorem 3.4, let $\hat{\mu}\in\widehat{\mathcal{U}}$ be such that for sufficiently small $T>0$

$$\mathbf{V}_*(x_1,q)\geq\int_0^T\hat{\ell}(\hat{x}(t),q,\hat{\mu}_t)e^{-\lambda t}dt+\mathbf{V}_*(\hat{x}(T),q)e^{-\lambda T}-\frac{\delta}{2\lambda}(1-e^{-\lambda T}). \tag{31}$$

For this relaxed control $\hat{\mu}$ and its relaxed state trajectory $\hat{x}(\cdot)$, from (30)

$$\lambda\phi(\hat{x}(t),q)-\langle\hat{f}(\hat{x}(t),q,\hat{\mu}_t),\nabla_x\phi(\hat{x}(t),q)\rangle-\hat{\ell}(\hat{x}(t),q,\hat{\mu}_t)<-\delta \quad \text{a.e. } 0\leq t<T.$$

Then we have

$$\lambda e^{-\lambda t}\phi(\hat{x}(t),q) - e^{-\lambda t}\langle \hat{f}(\hat{x}(t),q,\hat{\mu}_t), \nabla_x \phi(\hat{x}(t),q)\rangle \\ - e^{-\lambda t}\hat{\ell}(\hat{x}(t),q,\hat{\mu}_t) < -e^{-\lambda t}\delta \quad \text{a.e. } 0 \le t < T.$$

Integrating from 0 to T, we get

$$\phi(x_1,q) < \int_0^T \hat{\ell}(\hat{x}(t),q,\hat{\mu}_t)e^{-\lambda t}dt + \phi(\hat{x}(T),q)e^{-\lambda T} - \frac{\delta}{\lambda}(1-e^{-\lambda T}).$$

Subtracting (31) and noting the fact $\mathbf{V}_*(x_1,q) = \phi(x_1,q)$ and (29), we get a contradiction:

$$0 < (\phi(\hat{x}(T),q) - \mathbf{V}_*(\hat{x}(T),q))e^{-\lambda T} - \frac{\delta}{2\lambda}(1-e^{-\lambda T}) \\ \le -\frac{\delta}{2\lambda}(1-e^{-\lambda T}).$$

Case 2. $x_1 \in \partial F_q^c$. Letting $x^n \to x_1$, $x^n \in \overline{F_q^c}$, and $\mathbf{V}(x^n,q) \to \mathbf{V}_*(x_1,q)$, we show that there exists a subsequence (still denoted $\{x^n\}$) such that $x^n \in F_q^c$. Otherwise, there exists $\hat{N}$ such that for $n > \hat{N}$, $x^n \in \partial F_q^c$. Since F_q is closed, we know that $x^n \in F_q$. Then from Theorem 3.2, we know that $\mathbf{V}(x^n,q) = (\mathcal{M}\mathbf{V})(x^n,q)$; this implies

$$\begin{aligned}\mathbf{V}_*(x_1,q) &= \lim_{n\to\infty} \mathbf{V}(x^n,q) \\ &\ge \liminf_{n\to\infty} (\mathcal{M}\mathbf{V})(x^n,q) \\ &\ge (\mathcal{M}\mathbf{V})_*(x_1,q) \\ &\ge (\mathcal{M}^-\mathbf{V}_*)(x_1,q),\end{aligned}$$

which contradicts our assumption. So we assume $x^n \in F_q^c$ for all n. Let $\beta^n(x^n,q) \in \mathcal{B}(x^n,q)$ such that

$$\mathbf{V}(x^n,q) + 1/n \ge \mathbf{J}_{(x^n,q)}(\beta^n).$$

Then we can proceed as in the proof of Theorem 3.4 to show that there exists a $t(x,q) > 0$ such that $t(x_1,q) < (\tau^n)_1$ for large n, and by using the weak convergence argument, we can show for any $0 \le t < t(x_1,q)$ that

$$\mathbf{V}_*(x_1,q) \ge \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}} \left\{ \int_0^t \int_U \ell(\hat{x}(s),q,u)e^{-\lambda s} d\mu_s ds + \mathbf{V}_*(\hat{x}(t),q)e^{-\lambda t} \right\}.$$

This is the super-optimality principle for the point (x_1,q). Then proceeding as in the proof of Case 1, we conclude that

$$\lambda \mathbf{V}_*(x_1,q) + \mathcal{H}_q(x_1, \nabla_x \phi(x_1,q)) \ge 0,$$

which finally shows that $\mathbf{V}_*$ is a super-solution. □

5. Comparison theorems. In general, we cannot guarantee that $\mathbf{V}$ is the unique viscosity solution of SQVI (22) with boundary condition (23) in the sense of Definition 4.1. As in the case of an optimal stopping problem with discontinuous stopping cost [4] and an optimal exit control problem [5], [10], we can define several slightly different optimal control problems which have different values but all satisfy SQVI (22) with boundary condition (23) in the sense of Definition 4.1. The reason is that since only the upper and lower envelopes are involved in the definition of discontinuous viscosity solutions, there is no way to completely recover a discontinuous function from the information provided by its envelopes. However, among all of these solutions we can identify the maximum and minimum solutions.

The main task of this section is to construct and prove the existence of these two extreme solutions in Theorems 5.3 and 5.6, and to show that the minimum solution actually is the unique lower semi-continuous solution by showing that all solutions have the same lower semi-continuous envelope. We accomplish all of this by using a different approach from that of Barles and Perthame [5] and Blanc [10]. The approach used here is a generalization of the idea developed by Bensoussan and Lions [8] and Barles [3] that they use to prove the uniqueness results for continuous solutions.

5.1. Characterization of the maximum solution. Let $\mathcal{V}^+ = \{\mathbf{v} : X \times Q \to \mathbf{R} | \mathbf{v} \geq 0 \text{ and } \mathbf{v} \in USC(X \times Q)\}$. Define a nonlinear operator $\mathcal{T}^+ : \mathcal{V}^+ \to \mathcal{V}^+$:

$$(\mathcal{T}^+\mathbf{v})(x,q) = \inf_{u(\cdot)\in\mathcal{U}, 0\leq\theta\leq T^F(x,q,u(\cdot))} \left\{ \int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds + (\mathcal{M}^+\mathbf{v})(x(\theta),q)e^{-\lambda\theta} \right\}, \tag{32}$$

where $x(t)$ is the solution of

$$\dot{x}(t) = f(x(t),q,u(t))$$

with initial state $x(0) = x$.

The next result summarizes the properties of the operator $\mathcal{T}^+$, which is similar to the operator $\mathcal{T}$ in [3], and the proofs are deferred to Appendix A.2.

LEMMA 5.1. *The operator $\mathcal{T}^+$ satisfies the following:*

1. $\mathcal{T}^+\mathbf{v}_1 \geq \mathcal{T}^+\mathbf{v}_2$ *for* $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^+$ *such that* $\mathbf{v}_1 \geq \mathbf{v}_2$.
2. $\mathcal{T}^+(\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2) \geq \alpha\mathcal{T}^+\mathbf{v}_1 + (1-\alpha)\mathcal{T}^+\mathbf{v}_2$ *for all* $\alpha \in [0,1]$ *and* $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^+$.
3. $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^+$ *and* $\mathbf{v}_1 - \mathbf{v}_2 \leq \gamma\mathbf{v}_1$ *for some* $\gamma \in [0,1]$. *Then there exits a constant* $0 < C_T < 1$ *independent of* γ *such that*

$$\mathcal{T}^+\mathbf{v}_1 - \mathcal{T}^+\mathbf{v}_2 \leq \gamma(1-\alpha)\mathcal{T}^+\mathbf{v}_1$$

   for all $\alpha \in [0, C_T]$.
4. *If* $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^+$ *such that* $\mathbf{v}_1 = \mathcal{T}^+\mathbf{v}_1$ *and* $\mathbf{v}_2 = \mathcal{T}^+\mathbf{v}_2$, *then* $\mathbf{v}_1 = \mathbf{v}_2$.

Property 4 in Lemma 5.1 says that the operator $\mathcal{T}^+$ has only one fixed point in the set $\mathcal{V}^+$. Next we show that this fixed point is actually the value function of an optimal control problem with modified costs for the same hybrid system dynamics. In Theorem 5.3, we prove that this fixed point is the maximum solution of SQVI (22) with boundary condition (23).

For the same hybrid system (3), given any hybrid initial state $(x,q) \in (X \times Q)$ and any admissible hybrid control $\beta(x,q) \in \mathcal{B}(x,q)$, define a new cost to be

$$\mathbf{J}^+_{(x,q)}(\beta) = \sum_{i=1}^{N} \left(\int_{\tau_{i-1}}^{\tau_i} \ell(x(s), q_{i-1}, u(s)) e^{-\lambda s} ds + \kappa^*(x(\tau_i^-), q_{i-1}, w_i) e^{-\lambda \tau_i} \right) \tag{33}$$

and the new value function to be

$$\mathbf{V}^+(x,q) = \inf_{\beta(x,q) \in \mathcal{B}(x,q)} \{\mathbf{J}^+_{(x,q)}(\beta)\} \tag{34}$$

for all (x,q).

PROPOSITION 5.2. *$\mathbf{V}^+$ satisfies $\mathbf{V}^+ = \mathcal{T}^+\mathbf{V}^+$.*

Proof. By an argument similar to that in the proof of (16) in Theorem 3.2, we can show for any hybrid state $(x,q) \in (X \times Q)$ that

$$\mathbf{V}^+(x,q) \le (\mathcal{M}^+\mathbf{V}^+)(x,q).$$

Following the proof of (18) in Theorem 3.3, for any $u(\cdot) \in \mathcal{U}$ and $0 \le \theta \le T^F(x,q,u(\cdot))$, we can obtain

$$\begin{aligned} \mathbf{V}^+(x,q) &\le \int_0^\theta \ell(x(s), q, u(s)) e^{-\lambda s} ds + \mathbf{V}^+(x(\theta), q) e^{-\lambda \theta} \\ &\le \int_0^\theta \ell(x(s), q, u(s)) e^{-\lambda s} ds + (\mathcal{M}^+\mathbf{V}^+)(x(\theta), q) e^{-\lambda \theta}, \end{aligned}$$

where $x(t)$ is the solution of

$$\dot{x}(t) = f(x(t), q, u(t))$$

with initial state $x(0) = x$. Then we have

$$\begin{aligned} \mathbf{V}^+(x,q) &\le \inf_{u(\cdot) \in \mathcal{U}, 0 \le \theta \le T^F(x,q,u(\cdot))} \left\{ \int_0^\theta \ell(x(s), q, u(s)) e^{-\lambda s} ds + (\mathcal{M}^+\mathbf{V}^+)(x(\theta), q) e^{-\lambda \theta} \right\} \\ &= (\mathcal{T}^+\mathbf{V}^+)(x,q). \end{aligned}$$

To prove the reverse inequality, for any $\epsilon > 0$, there exists control $\beta^\epsilon(x,q) \in \mathcal{B}(x,q)$ such that

$$\begin{aligned} \mathbf{V}^+(x,q) + \epsilon &\ge \mathbf{J}^+_{(x,q)}(\beta^\epsilon) \\ &\ge \int_0^{\tau_1^\epsilon} \ell(x^\epsilon(s), q, u^\epsilon(s)) e^{-\lambda s} ds + (\kappa^*(x^\epsilon((\tau_1^\epsilon)^-), q, w_1^\epsilon) \\ &\qquad + \mathbf{V}^+(g(x^\epsilon((\tau_1^\epsilon)^-), q, w_1^\epsilon))) e^{-\lambda \tau_1^\epsilon} \\ &\ge \int_0^{\tau_1^\epsilon} \ell(x^\epsilon(s), q, u^\epsilon(s)) e^{-\lambda s} ds + (\mathcal{M}^+\mathbf{V}^+)(x^\epsilon(x^\epsilon((\tau_1^\epsilon)^-), q)) e^{-\lambda(\tau_1^\epsilon)}. \end{aligned}$$

Since for any $(x,q) \in (X \times Q)$ and any $\beta(x,q) = (u(\cdot), \{(\tau_i, w_i)\}_1^N)$

$$0 \le \tau_1 \le T^F(x,q,u(\cdot)),$$

it holds that

$$\begin{aligned}\mathbf{V}^+(x,q)+\epsilon &\geq \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds \\ &\qquad + (\mathcal{M}^+\mathbf{V}^+)(x(\theta),q)e^{-\lambda\theta}\Big\} \\ &= (\mathcal{T}^+\mathbf{V}^+)(x,q),\end{aligned}$$

where $x(T^F(x,q,u(\cdot))) = \int_0^{T^F(x,q,u(\cdot))} f(x(t,q,u(t)))dt - x$. □

Now we come to the main result of the section, showing that the restriction of $\mathbf{V}^+$ on set $\overline{F^c}$ is the maximum solution.

THEOREM 5.3. *$\mathbf{V}^+|_{\overline{F^c}}$ is the maximum sub-solution and solution of SQVI* (22) *with boundary condition* (23) *in the sense of Definition* 4.1.

Proof. Take any bounded nonnegative sub-solution $\mathbf{v} : \overline{F^c} \to \mathbf{R}$; this determines $(\mathcal{M}^+\mathbf{v})(x,q)$ for all $(x,q) \in \widetilde{F}$ according to the definition of $\mathcal{M}^+$. Define function $V : X \times Q \to \mathbf{R}$ by

$$V(x,q) = \begin{cases}\mathbf{v}(x,q) & \text{if } (x,q)\in\overline{F^c},\\ 0 & \text{if } (x,q)\in\widetilde{F}.\end{cases}$$

It is easy to see that $V \in \mathcal{V}^+$, that V is bounded, i.e., there exists $\widehat{C}$ such that $0 \leq V \leq \widehat{C}$, and that V satisfies the following:

1.

(35) $V|_{\overline{F^c}}$ is a bounded sub-solution of SQVI (22) with boundary condition (23);

2.

$$V(x,q) \leq (\mathcal{M}^+V)(x,q) \quad \forall\ (x,q)\in\widetilde{F}. \tag{36}$$

For any fixed $q \in Q$, according to the definition of operator $\mathcal{T}^+$ (32)

$$\begin{aligned}(\mathcal{T}^+V)(x,q) = \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{&\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds \\ &+ (\mathcal{M}^+V)(x(\theta),q)e^{-\lambda\theta}\Big\};\end{aligned}$$

thus $(\mathcal{T}^+V)|_{\overline{F^c}}(\cdot,q)$ is the value function of the optimal stopping control in restricted area F_q^c of the continuous system

$$\dot{x} = f(x,q,u)$$

with discontinuous stopping cost $(\mathcal{M}^+V)(\cdot,q)$, and since $(\mathcal{M}^+V)(\cdot,q)$ is upper semi-continuous, we know (see Appendix A.1 for a discussion of the problem of an optimal stopping control problem in a restricted area) that $(\mathcal{T}^+V)|_{\overline{F^c}}(\cdot,q)$ is the maximum sub-solution of the QVI

$$\max\{\mathbf{u}(x,q) - (\mathcal{M}^+V)(x,q), \lambda\mathbf{u}(x,q) - \mathcal{H}(x,q,\nabla_x\mathbf{u}(x,q))\} = 0 \quad \text{in } F_q^c \tag{37}$$

with boundary condition

$$\mathbf{u}(x,q) = (\mathcal{M}^+V)(x,q) \quad \text{in } \partial F_q^c. \tag{38}$$

While by assumption $V|_{\overline{F^c}}(\cdot, q)$ is a sub-solution of (37) and boundary condition (38),

$$V(x,q) \le (\mathcal{T}^+ V)(x,q) \quad \forall x \in \overline{F_q^c}.$$

Since it is true for any $q \in Q$, then

$$V \le (\mathcal{T}^+ V) \quad \text{in } \overline{F^c}. \tag{39}$$

When $(x,q) \in \widetilde{F}$, according to the definition of $\mathcal{T}^+$ (32),

$$(\mathcal{T}^+ V)(x,q) = (\mathcal{M}^+ V)(x,q);$$

hence combining this with (36) and (39) we get

$$V \le (\mathcal{T}^+ V) \text{ in } X \times Q.$$

On the other hand, from the definition of operator $\mathcal{T}^+$ (32), it can be shown in a manner similar to that of the proof of Theorem 3.3 that

$$(\mathcal{T}^+ V)(x,q) \le (\mathcal{M}^+ V)(x,q) \le (\mathcal{M}^+(\mathcal{T}^+ V))(x,q) \quad \forall (x,q) \in (X \times Q) \tag{40}$$

and

$$(\mathcal{T}^+ V)(x,q) \le \int_0^T \ell(x(s), q, u(s)) e^{-\lambda s} ds + (\mathcal{T}^+ V)(x(T), q) e^{-\lambda T} \tag{41}$$

for any $(x,q) \in F^c$, $u(\cdot) \in \mathcal{U}$, and $0 \le T < T^F(x,q,u(\cdot))$. Inequalities (40) and (41) are the sub-optimality principles for $\mathcal{T}^+ V$. Then by the same procedure as in the proof of Theorem 4.3, we can show that $(\mathcal{T}^+ V)_{\overline{F^c}}$ is also a sub-solution of SQVI (22) with boundary condition (23). Hence we know that $\mathcal{T}^+ V$ also satisfies conditions (35) and (36); replacing V with $\mathcal{T}^+ V$ in the above analysis, we have

$$\mathcal{T}^+ V \le \mathcal{T}^+(\mathcal{T}^+ V).$$

The above analysis suggests that if we start from any function $V \in \mathcal{V}^+$ satisfying conditions (35) and (36), we get a sequence of nondecreasing functions

$$\{V, \mathcal{T}^+ V, (\mathcal{T}^+)^2 V, \dots, (\mathcal{T}^+)^n V, \dots\}, \tag{42}$$

and every function in this sequence satisfies conditions (35) and (36).

Next we show that this sequence converges uniformly to a function in $\mathcal{V}^+$.

Since for any $(x,q) \in (X \times Q)$

$$\begin{aligned}(\mathcal{T}^+ V)(x,q) &\le (\mathcal{M}^+ V)(x,q) \\ &= \min_{w \in W} \{\kappa^*(x,q,w) + V(g(x,q,w))\} \\ &\le \widehat{C}' + \widehat{C},\end{aligned}$$

where $\widehat{C}'$ is the upper bound for κ^*, from the relation

$$(\mathcal{T}^+)^n V = \mathcal{T}^+((\mathcal{T}^+)^{n-1} V) \le \mathcal{M}((\mathcal{T}^+)^{n-1} V),$$

we can obtain the following bound for $(\mathcal{T}^+)^n V$ inductively:

$$(\mathcal{T}^+)^n V \le n\widehat{C}' + \widehat{C}$$

for any $n \geq 1$. Since $V \geq 0$, we have

$$\mathcal{T}^+ V - V \leq \mathcal{T}^+ V.$$

Using item 3 in Lemma 5.1, we know there exists $0 < \alpha < 1$ such that

$$(\mathcal{T}^+)^2 V - \mathcal{T}^+ V \leq (1-\alpha)((\mathcal{T}^+)^2 V).$$

Then it is easy to show inductively by using item 3 in Lemma 5.1 that for any $n \geq 2$

$$\begin{aligned}(\mathcal{T}^+)^n V - (\mathcal{T}^+)^{n-1} V &\leq (1-\alpha)^{n-1}((\mathcal{T}^+)^n V) \\ &\leq (1-\alpha)^{n-1}(n\widehat{C}' + \widehat{C}).\end{aligned}$$

Since $\sum_1^\infty (1-\alpha)^{n-1}(n\widehat{C}' + \widehat{C}) < \infty$, we know that for any $q \in Q$, the sequence

$$\{V(\cdot,q), (\mathcal{T}^+ V)(\cdot,q), ((\mathcal{T}^+)^2 V)(\cdot,q), \dots, ((\mathcal{T}^+)^n V)(\cdot,q), \dots\}$$

is uniformly convergent. Let $\widehat{\mathbf{V}}(\cdot,q) = \lim_{n\to\infty}((\mathcal{T}^+)^n V)(\cdot,q)$.

Finally, we show that $\widehat{\mathbf{V}} = \mathbf{V}^+$.

First for any n,

$$(\mathcal{T}^+)^n V \leq \mathcal{M}^+((\mathcal{T}^+)^{n-1} V) \leq \mathcal{M}^+ \widehat{\mathbf{V}};$$

then

$$\widehat{\mathbf{V}} \leq \mathcal{M}^+ \widehat{\mathbf{V}} \text{ in } X \times Q. \tag{43}$$

Furthermore, we can follow the same argument as in the proof of (18) to prove that for any $(x,q) \in F^c$, $u(\cdot) \in \mathcal{U}$, and $0 \leq T < T^F(x,q,u(\cdot))$,

$$((\mathcal{T}^+)^n V)(x,q) \leq \int_0^T \ell(x(s),q,u(s)) e^{-\lambda s} ds + ((\mathcal{T}^+)^n V)(x(T),q) e^{-\lambda T},$$

which implies

$$\widehat{\mathbf{V}}(x,q) \leq \int_0^T \ell(x(s),q,u(s)) e^{-\lambda s} ds + \widehat{\mathbf{V}}(x(T),q) e^{-\lambda T}. \tag{44}$$

Equations (43) and (44) are the sub-optimality principles for $\widehat{\mathbf{V}}$, which means that $\widehat{\mathbf{V}}|_{\overline{F^c}}$ is also a sub-solution of SQVI (22) with boundary condition (23). Hence $\widehat{\mathbf{V}}$ also satisfies conditions (35) and (36); thus

$$\widehat{\mathbf{V}} \leq \mathcal{T}^+ \widehat{\mathbf{V}}.$$

On the other hand, let

$$\sigma_n = \sup_{x,q}(\widehat{\mathbf{V}} - (\mathcal{T}^+)^n V)(x,q);$$

then $\sigma_n \geq 0$ and $\lim_n \sigma_n = 0$ by the uniform convergence of sequence (42). For any n,

$$\begin{aligned}\widehat{\mathbf{V}} &\geq (\mathcal{T}^+)^{n+1} V \\ &= \mathcal{T}^+((\mathcal{T}^+)^n V) \\ &\geq \mathcal{T}^+(\widehat{\mathbf{V}} - \sigma_n) \\ &\geq \mathcal{T}^+ \widehat{\mathbf{V}} - \sigma_n.\end{aligned}$$

Hence

$$\widehat{\mathbf{V}} \geq \mathcal{T}^+ \widehat{\mathbf{V}}.$$

Then we conclude that

$$\widehat{\mathbf{V}} = \mathcal{T}^+ \widehat{\mathbf{V}}.$$

We have already shown that $\mathbf{V}^+ = \mathcal{T}^+ \mathbf{V}^+$. Thanks to item 4 in Lemma 5.1, we know that

$$\mathbf{V}^+ = \widehat{\mathbf{V}} \geq V.$$

By restricting $\mathbf{V}^+$ and V to $\overline{F^c}$, we get

$$\mathbf{V}^+|_{\overline{F^c}} \geq V|_{\overline{F^c}} = \mathbf{v}.$$

What is left to be shown is that $(\mathbf{V}^+)_*|_{\overline{F^c}}$ is also a super-solution of SQVI (22) with boundary condition (23). By the same argument as in Theorem 3.4, we can prove the super-optimality principle for $(V^+)_*$ by using the fact that

$$(\kappa^*)_* = \kappa_*.$$

Thus following the second part of the proof of Theorem 4.3, we can show that it is a super-solution of SQVI (22) with boundary condition (23).

The proof is complete. □

5.2. Characterization of the minimum solution. In this section, we use the same approach to identify the minimum super-solution for the SQVI (22) with boundary condition (23). It is no surprise that the minimum super-solution is also a solution (hence, minimum solution). By the same argument, the minimum super-solution is the value function of the optimal control problem in a relaxed control framework for the same hybrid systems but with the costs replaced by their lower envelopes. Also, it is the unique fixed point of the nonlinear operator $\mathcal{T}^- : \mathcal{V}^- \to \mathcal{V}^-$ defined as

$$(\mathcal{T}^- v)(x,q) = \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}, 0 \leq \theta \leq \widehat{T}^F(x,q,\mu_{(\cdot)})} \left\{ \int_0^\theta \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds + (\mathcal{M}^- v)(\hat{x}(\theta),q) e^{-\lambda\theta} \right\}, \tag{45}$$

where $\hat{x}(t)$ is the solution of

$$\dot{\hat{x}}(t) = \int_U f(\hat{x}(t),q,u) d\mu_t$$

with initial state $\hat{x}(0) = x$ and $\mathcal{V}^-$ the set of all bounded and nonnegative lower semi-continuous functions; i.e., $\mathcal{V}^- = \{\mathbf{v} : X \times Q \to \mathbf{R} | \mathbf{v} \geq 0 \text{ and } \mathbf{v} \in LSC(X \times Q)\}$.

As in Lemma 5.1, we can prove properties of $\mathcal{T}^-$ similar to those of $\mathcal{T}^+$.

LEMMA 5.4. *The operator $\mathcal{T}^-$ satisfies the following:*

1. $\mathcal{T}^- \mathbf{v}_1 \geq \mathcal{T}^- \mathbf{v}_2$ *for* $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^-$ *such that* $\mathbf{v}_1 \geq \mathbf{v}_2$.
2. $\mathcal{T}^-(\alpha \mathbf{v}_1 + (1-\alpha)\mathbf{v}_2) \geq \alpha \mathcal{T}^- \mathbf{v}_1 + (1-\alpha)\mathcal{T}^- \mathbf{v}_2$ *for all* $\alpha \in [0,1]$ *and* $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^-$.

3. $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^-$ *and* $\mathbf{v}_1 - \mathbf{v}_2 \leq \gamma \mathbf{v}_1$ *for some* $\gamma \in [0,1]$. *Then there exists constant* $0 < C_T < 1$ *independent of* γ *such that*

$$\mathcal{T}^- \mathbf{v}_1 - \mathcal{T}^- \mathbf{v}_2 \leq \gamma(1-\alpha)\mathcal{T}^- \mathbf{v}_1$$

for all $\alpha \in [0, C_T]$.

4. *If* $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}^-$ *such that* $\mathbf{v}_1 = \mathcal{T}^- \mathbf{v}_1$ *and* $\mathbf{v}_2 = \mathcal{T}^- \mathbf{v}_2$, *then* $\mathbf{v}_1 = \mathbf{v}_2$.

Now, let us define the optimal control problem for the same hybrid systems whose value function is the minimum super-solution we want. For the same hybrid system (3), given any hybrid initial state $(x,q) \in (X \times Q)$ and any relaxed admissible hybrid control $\hat{\beta}(x,q) \in \widehat{\mathcal{B}}(x,q)$, define a new cost to be

$$\mathbf{J}^-_{(x,q)}(\hat{\beta}) = \sum_{i=1}^{N} \Big(\int_{\tau_{i-1}}^{\tau_i} \int_U \ell(\hat{x}(s), q_{i-1}, u) e^{-\lambda s} d\mu_s ds + \kappa_*(\hat{x}(\tau_i^-), q_{i-1}, w_i) e^{-\lambda \tau_i} \Big) \tag{46}$$

and the value function to be

$$\mathbf{V}^-(x,q) = \inf_{\hat{\beta}(x,q) \in \widehat{\mathcal{B}}(x,q)} \{ \mathbf{J}^-_{(x,q)}(\hat{\beta}) \}. \tag{47}$$

By an argument similar to that used in the proof of Proposition 5.2, we have the following proposition.

PROPOSITION 5.5. $\mathbf{V}^-$ *satisfies* $\mathbf{V}^- = \mathcal{T}^- \mathbf{V}^-$.

Now we show that $\mathbf{V}^-|_{\overline{F^c}}$ is the minimum solution of SQVI (22) with boundary condition (23).

THEOREM 5.6. $\mathbf{V}^-|_{\overline{F^c}}$ *is the minimum super-solution and solution of SQVI* (22) *with boundary condition* (23) *in the sense of definition* 4.1.

Proof. Take any bounded and nonnegative super-solution $\mathbf{v} : \overline{F^c} \to \mathbf{R}$. According to the definition of $\mathcal{M}^-$, we can compute $(\mathcal{M}^- \mathbf{v})(x,q)$ for all $(x,q) \in \widetilde{F}$. Define function $V : X \times Q \to \mathbf{R}$ by

$$V(x,q) = \begin{cases} \mathbf{v}(x,q) & \text{if } (x,q) \in \overline{F^c}, \\ \widetilde{C} & \text{if } (x,q) \in \widetilde{F}, \end{cases}$$

where $\widetilde{C}$ is large enough. It is easy to see that $V \in \mathcal{V}^-$ and that V satisfies

1.

$$V|_{\overline{F^c}} \text{ is a super-solution of (22) with boundary condition (23);} \tag{48}$$

2.

$$V(x,q) \geq (\mathcal{M}^- V)(x,q) \quad \forall (x,q) \in \widetilde{F}. \tag{49}$$

Then for any $q \in Q$, from the definition of $\mathcal{T}^-$,

$$(\mathcal{T}^- V)(x,q) = \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}, 0 \leq \theta \leq \widehat{T}^F(x,q,\mu_{(\cdot)})} \Big\{ \int_0^\theta \int_U \ell(\hat{x}(s), q, u) e^{-\lambda s} d\mu_s ds + (\mathcal{M}^- V)(\hat{x}(\theta), q) e^{-\lambda \theta} \Big\}.$$

We know that $\mathcal{T}^- V$ can be regarded as the value function of the optimal stopping control problem in restricted area F_q^c of the nonhybrid system

$$\dot{x} = f(x,q,u)$$

with discontinuous stopping costs $(\mathcal{M}^- V)(\cdot, q)$. Since $(\mathcal{M}^- V)(\cdot, q)$ is lower semi-continuous (see appendix), $(\mathcal{T}^- V)(\cdot, q)|_{\overline{F^c}}$ is the minimum super-solution of QVI (37) with boundary condition (38). By assumption, $V(\cdot, q)|_{\overline{F^c}}$ is a super-solution of it, and thus

$$V(x, q) \geq (\mathcal{T}^- V)(x, q) \quad \forall x \in \overline{F_q^c}.$$

When $(x, q) \in \widetilde{F}$, by (49) and the definition of $\mathcal{T}^-$, we know that

$$V(x, q) \geq (\mathcal{M}^- V)(x, q) = \mathcal{T}^- V(x, q);$$

hence

$$V \geq (\mathcal{T}^- V) \text{ in } X \times Q.$$

On the other hand, by an argument similar to that of Theorem 3.4, we can prove the following super-optimality principle for $\mathcal{T}^- V$ in F^c: for any $(x, q) \in F^c$ where

$$(\mathcal{T}^- V)(x, q) < (\mathcal{M}^-(\mathcal{T}^- V))(x, q),$$

there exists $\tilde{t}(x, q) > 0$ such that

$$(\mathcal{T}^- V)(x, q) \geq \inf_{\mu(\cdot) \in \widehat{\mathcal{U}}} \left\{ \int_0^t \int_U \ell(\hat{x}(s), q, u) e^{-\lambda s} ds + (\mathcal{T}^- V)(\hat{x}(t), q) e^{-\lambda t} \right\}$$

for all $0 \leq t < \tilde{t}(x, q)$. Then by the same procedure as in the proof in Theorem 4.3, $(\mathcal{T}^- V)|_{\overline{F^c}}$ is also a super-solution of (22) with boundary condition (23). Since we know that

$$\mathcal{T}^- V = \mathcal{M}^- V \geq \mathcal{M}^-(\mathcal{T}^- V) \quad \text{in } \tilde{F},$$

it holds that $\mathcal{T}^- V$ also satisfies conditions (48) and (49), and thus

$$\mathcal{T}^- V \geq \mathcal{T}^-(\mathcal{T}^- V) = (\mathcal{T}^-)^2 V.$$

The above argument shows that if we start from any bounded function $V \in \mathcal{V}^-$ satisfying conditions (48) and (49), we can get a sequence of nonincreasing functions

$$\{V, \mathcal{T}^- V, (\mathcal{T}^-)^2 V, \dots, (\mathcal{T}^-)^n V, \dots\}$$

all of which satisfy conditions (48) and (49). By a similar argument, we can show that for any fixed $q \in Q$, the sequence

$$\{V(\cdot, q), (\mathcal{T}^- V)(\cdot, q), ((\mathcal{T}^-)^2 V)(\cdot, q), \dots, ((\mathcal{T}^-)^n V)(\cdot, q), \dots\}$$

uniformly converges. Let $\widetilde{\mathbf{V}}(\cdot, q) = \lim_{n \to \infty}((\mathcal{T}^-)^n V)(\cdot, q)$.

We now show that $\widetilde{\mathbf{V}} = \mathbf{V}^-$.

Since for any n,

$$\widetilde{\mathbf{V}} \leq (\mathcal{T}^-)^n V = \mathcal{T}^-((\mathcal{T}^-)^{n-1} V) \leq \mathcal{M}^-((\mathcal{T}^-)^{n-1} V),$$

we get

$$\widetilde{\mathbf{V}} \leq \mathcal{M}^- \widetilde{\mathbf{V}}.$$

Assume for some $(x,q) \in F^c$ that

$$\widetilde{\mathbf{V}}(x,q) < (\mathcal{M}^-\widetilde{\mathbf{V}})(x,q).$$

For any n, let $(\mu_n(\cdot), \theta_n)$ be

$$\begin{aligned}((\mathcal{T}^-)^n V)(x,q) + \frac{1}{n} > \int_0^{\theta_n} \int_U \ell(\hat{x}_n(s),q,u) e^{-\lambda s} d\mu_n(s) ds \\ + (\mathcal{M}^-((\mathcal{T}^-)^{n-1}V))(\hat{x}_n(\theta_n),q) e^{-\lambda\theta_n};\end{aligned}$$

then there exists $\hat{t}(x,q) > 0$ such that $\theta_n \geq \hat{t}(x,q)$ for n large enough. Otherwise, assume there exists a convergent subsequence (still denoted as θ_n); then we easily get a contradiction:

$$\widetilde{\mathbf{V}}(x,q) \geq (\mathcal{M}^-\widetilde{\mathbf{V}})(x,q).$$

Hence for every $0 \leq t < \hat{t}(x,q)$, we have

$$((\mathcal{T}^-)^n V)(x,q) + \frac{1}{n} \geq \int_0^t \int_U \ell(\hat{x}_n(s),q,u) e^{-\lambda s} d\mu_n(s) ds + ((\mathcal{T}^-)^n V)(\hat{x}_n(t),q) e^{-\lambda t}$$

for all large n. By the argument of weak convergence, we know that there exits $\mu_{(\cdot)}$ such that $\mu_n(\cdot)$ converges to $\mu_{(\cdot)}$ weakly, and

$$\begin{aligned}\widetilde{\mathbf{V}}(x,q) &\geq \int_0^t \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds + \widetilde{\mathbf{V}}(\hat{x}(t),q) e^{-\lambda t} \\ &\geq \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}} \left\{ \int_0^t \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds + \widetilde{\mathbf{V}}(\hat{x}(t),q) e^{-\lambda t} \right\}.\end{aligned}$$

This is the super-solution property for $\widetilde{V}$; hence we know $\widetilde{\mathbf{V}}|_{\overline{F^c}}$ is also a super-solution of (22) with boundary condition (23). When $(x,q) \in \widetilde{F}$, $(\mathcal{T}^-)^n V$ satisfies

$$((\mathcal{T}^-)^n V)(x,q) \geq (\mathcal{M}^-((\mathcal{T}^-)^n V))(x,q)$$

for every n; then

$$\widetilde{\mathbf{V}}(x,q) \geq (\mathcal{M}^-\widetilde{\mathbf{V}})(x,q) \quad \forall (x,q) \in \widetilde{F}.$$

We have just shown that conditions (48) and (49) are satisfied by the function $\widetilde{\mathbf{V}}$; hence

$$\widetilde{\mathbf{V}} \geq \mathcal{T}^-\widetilde{\mathbf{V}}.$$

On the other hand, let

$$\sigma_n = \sup_{x,q}((\mathcal{T}^-)^n V - \widetilde{\mathbf{V}})(x,q);$$

then $\sigma_n \geq 0$ and $\lim_n \sigma_n = 0$ by the uniform convergence, and hence for any n,

$$\begin{aligned}\widetilde{\mathbf{V}} &\leq (\mathcal{T}^-)^{n+1} V \\ &= \mathcal{T}^-((\mathcal{T}^-)^n V) \\ &\leq \mathcal{T}^-(\widetilde{\mathbf{V}} + \sigma_n) \\ &= \mathcal{T}^-\widetilde{\mathbf{V}} + \sigma_n.\end{aligned}$$

Hence

$$\widetilde{\mathbf{V}} \le \mathcal{T}^- \widetilde{\mathbf{V}},$$

and then

$$\widetilde{\mathbf{V}} = \mathcal{T}^- \widetilde{\mathbf{V}}.$$

Thanks to property 5.4 in Lemma 5.4 and Proposition 5.5, we know that

$$\mathbf{V}^- = \widetilde{\mathbf{V}} \le V.$$

When restricting $\mathbf{V}^-$ and V on $\overline{F^c}$, we have

$$\mathbf{V}^-|_{\overline{F^c}} \le V|_{\overline{F^c}} = \mathbf{v},$$

which shows that $\mathbf{V}^-|_{\overline{F^c}}$ is the minimum super-solution.

From the definition of $\mathbf{V}^-$ (47), it is easy to check that for any $(x, q) \in (X \times Q)$

$$\mathbf{V}^-(x,q) \le (\mathcal{M}^- \mathbf{V}^-)(x,q) \le (\mathcal{M}\mathbf{V}^-)(x,q),$$

and for any $\mu_{(\cdot)} \in \widehat{\mathcal{U}}$ and $0 \le T \le \widehat{T}^F(x,q,\mu_{(\cdot)})$

$$\mathbf{V}^-(x,q) \le \int_0^T \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds + \mathbf{V}^-(\hat{x}(T),q) e^{-\lambda T}.$$

These inequalities are the sub-optimality principles for $\mathbf{V}^-$, and consequently, it can be shown that $(\mathbf{V}^-)^*|_{\overline{F^c}}$ is a sub-solution of (22) with boundary condition (23) by routine.

The proof is complete. □

5.3. Weak sense uniqueness. We have specified two extreme solutions for the SQVI (22) with boundary condition (23); next we show that they share the same lower envelope. This property also holds for the extreme solutions for the equations arising from the optimal stopping problem [4] and the optimal exit problem [5], [10]; however, their approach in proving this point is not readily applicable for our case since multiple (possibly an infinite number of) exits are involved, and it can be seen that we need a few more conditions to guarantee this property.

THEOREM 5.7. *If* $\mathcal{M}^- v_* = (\mathcal{M}^+ v)_*$ *for all* $v \in \mathcal{V}^+$, *then* $\mathbf{V}^- = \mathbf{V}^+_*$.

Proof. By definition of $\mathcal{T}^-$,

$$\begin{aligned}(\mathcal{T}^- \mathbf{V}^+_*)(x,q) &= \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}, 0 \le \theta \le \widehat{T}^F(x,q,\mu_{(\cdot)})} \left\{ \int_0^\theta \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds \right. \\ &\qquad \left. + (\mathcal{M}^- \mathbf{V}^+_*)(\hat{x}(\theta),q) e^{-\lambda\theta} \right\} \\ &= \inf_{\mu_{(\cdot)} \in \widehat{\mathcal{U}}, 0 \le \theta \le \widehat{T}^F(x,q,\mu_{(\cdot)})} \left\{ \int_0^\theta \int_U \ell(\hat{x}(s),q,u) e^{-\lambda s} d\mu_s ds \right. \\ &\qquad \left. + (\mathcal{M}^+ \mathbf{V}^+)_*(\hat{x}(\theta),q) e^{-\lambda\theta} \right\}.\end{aligned}$$

By Proposition A.4 in Appendix A.1 concerning optimal stopping in a restricted area, we know that

$$\begin{aligned}(\mathcal{T}^-\mathbf{V}_*^+)(x,q) &= (\mathcal{T}^+\mathbf{V}^+)_*(x,q)\\ &= \mathbf{V}_*^+(x,q);\end{aligned}$$

hence $\mathbf{V}_*^+$ is the fixed point of $\mathcal{T}^-$, but we know that $\mathbf{V}^-$ is the unique fixed point of $\mathcal{T}^-$, so we conclude that $\mathbf{V}^- = \mathbf{V}_*^+$. □

Let V be any solution of (22); then $\mathbf{V}^- \le V \le \mathbf{V}^+$. Hence $\mathbf{V}^- \le V_* \le \mathbf{V}_*^+ = \mathbf{V}^-$, and we get $V_* = \mathbf{V}^-$. Thus all the solutions have the same lower envelope which is the minimum solution. If we define SQVI (22) as having a unique solution in the weak sense if all its solutions share the same lower envelope, then we have just proved that the uniqueness in the weak sense holds for SQVI (22) with boundary condition (23) under the condition of Theorem 5.7. Another implication of Theorem 5.7 is that SQVI (22) with boundary condition (23) has a unique lower semi-continuous solution $\mathbf{V}^-$.

6. Special cases and optimal hybrid controllers. We have shown that extreme solutions $\mathbf{V}^-$ and $\mathbf{V}^+$ of the dynamic programming equations are the value functions of slightly different hybrid optimal control problems. The original value function we are really interested in resides somewhere between $\mathbf{V}^-$ and $\mathbf{V}^+$. Thus $\mathbf{V}^-$ and $\mathbf{V}^+$ can be regarded as the lower and upper bound of the value function we want to compute. Although these boundary functions themselves are the value functions of optimal hybrid control problems, from the proofs of Theorems 5.3 and 5.6 we know that they are actually the limit functions of corresponding monotone sequences,

$$\{V, \mathcal{T}^+V, (\mathcal{T}^+)^2V, \dots, (\mathcal{T}^+)^nV, \dots\}$$

and

$$\{V, \mathcal{T}^-V, (\mathcal{T}^-)^2V, \dots, (\mathcal{T}^-)^nV, \dots\}.$$

Every element in these two sequences can be computed according to

$$\begin{aligned}((\mathcal{T}^+)^nV)(x,q) = \inf_{u(\cdot)\in\mathcal{U}, 0\le\theta\le T^F(x,q,u(\cdot))} \Big\{ &\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds\\ &+ (\mathcal{M}^+((\mathcal{T}^+)^{n-1}V))(x(\theta),q)e^{-\lambda\theta}\Big\}\end{aligned}$$

and

$$\begin{aligned}((\mathcal{T}^-)^nV)(x,q) = \inf_{\mu_{(\cdot)}\in\widehat{\mathcal{U}}, 0\le\theta\le\widehat{T}^F(x,q,\mu_{(\cdot)})} \Big\{ &\int_0^\theta\int_U \ell(\hat{x}(s),q,u)e^{-\lambda s}d\mu_s ds\\ &+ (\mathcal{M}^-((\mathcal{T}^+)^{n-1}V))(\hat{x}(\theta),q)e^{-\lambda\theta}\Big\},\end{aligned}$$

which essentially are optimization problems over nonhybrid controls. So in this sense, computing $\mathbf{V}^-$ and $\mathbf{V}^+$ is easier than computing $\mathbf{V}$.

But it is still desirable to identify some special cases in which the value function itself can be determined completely from the SQVI, i.e., from $\mathbf{V}^-$ and $\mathbf{V}^+$.

6.1. Lower semi-continuous value function. We first discuss the case of a lower semi-continuous value function. We show under some conditions that the value function is the minimum solution.

THEOREM 6.1. *If the value function* $\mathbf{V}$ *satisfies* $\mathcal{M}^-\mathbf{V}_* = (\mathcal{M}\mathbf{V})_*$, *then* $\mathbf{V}_* = \mathbf{V}^-$; *in particular, if furthermore* $\mathbf{V}$ *is lower semi-continuous, then* $\mathbf{V} = \mathbf{V}^-$.

Proof. First, as in the proof of Proposition 5.2, we can show that

$$\mathbf{V}(x,q) = \inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))} \left\{ \int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds + (\mathcal{M}\mathbf{V})(x(\theta),q)e^{-\lambda\theta} \right\}.$$

Actually, this can be regarded as another version of DPP for $\mathbf{V}$. Next, for any (x,q)

$$\begin{aligned}(\mathcal{T}^-\mathbf{V}_*)(x,q) &= \inf_{\mu_{(\cdot)}\in\widehat{\mathcal{U}},0\le\theta\le\widehat{T}^F(x,q,\mu_{(\cdot)})} \left\{ \int_0^\theta \int_U \ell(\hat{x}(s),q,u)e^{-\lambda s}d\mu_s ds \right.\\ &\qquad \left. + (\mathcal{M}^-\mathbf{V}_*)(\hat{x}(\theta),q)e^{-\lambda\theta} \right\}\\ &= \inf_{\mu_{(\cdot)}\in\widehat{\mathcal{U}},0\le\theta\le\widehat{T}^F(x,q,\mu_{(\cdot)})} \left\{ \int_0^\theta \int_U \ell(\hat{x}(s),q,u)e^{-\lambda s}d\mu_s ds \right.\\ &\qquad \left. + (\mathcal{M}\mathbf{V})_*(\hat{x}(\theta),q)e^{-\lambda\theta} \right\}.\end{aligned}$$

Also, by Proposition A.4, we know that

$$(\mathcal{T}^-\mathbf{V}_*)(x,q) = \mathbf{V}_*(x,q);$$

hence $\mathbf{V}_*$ is the fixed point of $\mathcal{T}^-$, but we know $\mathbf{V}^-$ is the unique fixed point of $\mathcal{T}^-$, and so we conclude that $\mathbf{V}_* = \mathbf{V}^-$. □

6.2. Continuous value function. If we let $O = F^c$ and the jumping cost ρ be bounded and uniformly continuous, then according to the definition of $\mathbf{V}^-$ and $\mathbf{V}^+$, we know that $\mathbf{V}^+ = \mathbf{V}^- = \mathbf{V}$; i.e., the maximum sub-solution coincides with the minimum super-solution, which implies that $\mathbf{V}$ is continuous and is the unique viscosity solution of SQVI (22) with boundary condition (23). In this case, the operators become $\mathcal{T}^+ = \mathcal{T}^- = \mathcal{T}$, where $\mathcal{T}$ is defined as

$$(\mathcal{T}v)(x,q) = \inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))} \left\{ \int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds + (\mathcal{M}v)(x(\theta),q)e^{-\lambda\theta} \right\}.$$

The value function $\mathbf{V}$ is the unique fixed point of $\mathcal{T}$. Then the analysis and the results developed in the previous sections reduce to the case of [9].

Based on the above discussion, we conclude that the following theorem holds.

THEOREM 6.2. *In the special case of* $O = F^c$ *and* $\rho \in C(X\times Q)$, *the value function* $\mathbf{V} \in C(X\times Q)$ *is the unique viscosity solution of SQVI* (22) *with boundary condition* (23).

6.3. Smooth value functions and optimal controller. While in general value functions and viscosity solutions are not smooth, it is helpful to consider the smooth case, since light is shed on the nature of the optimal solution. So let us assume that the value function is smooth and therefore the concept of viscosity solution reduces to the notion of a classical smooth solution. Then consequently the value

function is the unique classic solution of SQVI (22). In this special case, we consider the nature of the optimal controller.

Define the optimal jumping set to be

$$\mathcal{S} = \{(x,q) \in (\mathbf{R}^n \times Q) | \mathbf{V}(x,q) = (\mathcal{M}\mathbf{V})(x,q)\}. \tag{50}$$

It is obvious that $F \subset S$. When the state enters $\mathcal{S}$, the optimal control is to have a discrete transition with an optimal discrete input $w^*(x,q)$ defined by

$$w^*(x,q) = \arg\min\{\rho(x,q,w) + \mathbf{V}(g(x,q,w))\} \quad \forall (x,q) \in S.$$

Outside $\mathcal{S}$, from SQVI (22) it must hold that

$$\lambda \mathbf{V}(x,q) + \mathcal{H}(x,q,\nabla_x \mathbf{V}(x,q)) = 0;$$

from this, we get the optimal continuous controller $u^*(x,q)$ through

$$u^*(x,q) = \arg\max\{-\langle f(x,q,u), \nabla_x \mathbf{V}(x,q)\rangle - \ell(x,q,u)\}.$$

Finally, $(\mathcal{S}, w^*, u^*)$ are the optimal hybrid controllers.

We now give the algorithm to compute the optimal state trajectory initiated at (x,q) under the optimal controllers.

Step 1. **Initialization**, $i = 0$, $j = 0$, $\tau_0^* = 0$, $(x^*((\tau^*)_0^+), q_0^*) = (x,q)$.

Step 2. For $i = 0, 1, 2, \dots$,

if $(x^*((\tau^*)_i^+), q_i^*) \in \mathcal{S}$, then "**JUMP**" and set $\tau_{i+1}^* = \tau_i^*$,

$$(x^*((\tau^*)_{i+1}^+), q_{i+1}^*) = g(x^*((\tau^*)_i^+), q_i^*, w^*(x^*((\tau^*)_i^+), q_i^*)).$$

Otherwise, first do "**CONTINUOUS EVOLUTION**"

$$\begin{cases} x^*(t) = \int_0^t f(x^*(t), q^*(t), u^*(x^*(t), q^*(t)))dt - x^*((\tau^*)_i^+), \\ q^*(t) = q_i^* \end{cases}$$

until $\tau_{i+1} = \tau_i + T^{\mathcal{S}}(x^*((\tau_i^*)^+), q_i^*, u^*(\cdot))$ and set $x^*((\tau^*)_{i+1}^-) = \lim_{t \to \tau_{i+1}^*} x^*(t)$; then "**JUMP**"

$$(x^*((\tau^*)_{i+1}^+), q_{i+1}^*) = g(x^*((\tau^*)_{i+1}^-), q_i^*, w^*(x^*((\tau^*)_{i+1}^-), q_i^*)).$$

7. Optimal stopping problem. In this section, we discuss briefly the optimal stopping control problem for hybrid systems. The aim of this section is to show that the approach developed above can be used to solve the optimal stopping problem. For simplicity, we assume that in the hybrid system (3) $F = \varnothing$.

Let us first define our hybrid stopping problem. Given any initial hybrid state $(x,q) \in X \times Q$, admissible hybrid control $\beta(x,q) \in \mathcal{B}(x,q)$, and a fixed time $\theta \geq 0$, denote the number of discrete transitions that happen before time θ to be

$$N_{(x,q)}^{\beta}(\theta) = \max\{i \geq 1 | \tau_i < \theta\}$$

and the number of discrete transitions that happen at time θ to be

$$M_{(x,q)}^{\beta}(\theta) = |\{i \geq 1 | \tau_i = \theta\}|.$$

$M^{\beta}_{(x,q)}(\theta) = 0$ implies that no jumps happen at time θ and that it is a time of continuous evolution.

To stop an admissible hybrid trajectory, we need to decide the stopping time $\theta \geq 0$ and the number of jumps $0 \leq k(\theta) \leq M^{\beta}_{(x,q)}(\theta)$ at time θ, so we call $(\theta, k(\theta))$ a hybrid stopping time and $(\beta(x,q), \theta, k(\theta))$ an admissible hybrid stopping control with respect to (x,q). Define

$$\theta^{+}_{k(\theta)} = \begin{cases} \theta \mathbf{I}\{M^{\beta}_{(x,q)}(\theta) = 0\} + \tau^{-}_{N^{\beta}_{(x,q)}(\theta)+1} \mathbf{I}\{M^{\beta}_{(x,q)}(\theta) > 0\}, & k(\theta) = 0, \\ \tau^{+}_{N^{\beta}_{(x,q)}(\theta)+k(\theta)}, & 1 \leq k(\theta) \leq M^{\beta}_{(x,q)}(\theta). \end{cases}$$

By the definition, we know that $(x(\theta^{+}_{k(\theta)}), q(\theta^{+}_{k(\theta)}))$ is the state where the trajectory is stopped by $(\theta, k(\theta))$.

We consider the problem of stopping the trajectory in some optimal way. Given any initial state $(x,q) \in (\mathbf{R}^n \times Q)$ and admissible hybrid stopping control $(\beta(x,q), \theta, k(\theta))$, define the total stopping cost,

$$\begin{aligned} \mathbf{J}_{(x,q)}(\beta, \theta, k(\theta)) = & \sum_{i=1}^{N^{\beta}_{(x,q)}(\theta)} \left(\int_{\tau_{i-1}}^{\tau_i} \ell(x(s), q_{i-1}, u(s)) e^{-\lambda s} ds + \rho(x(\tau_i^-), q_{i-1}, w_i) e^{-\lambda \tau_i} \right) \\ & + \int_{\tau_{N^{\beta}_{(x,q)}(\theta)}}^{\theta} \ell(x(s), q_{N^{\beta}_{(x,q)}(\theta)}, u(s)) e^{-\lambda s} ds \\ & + \sum_{j=1}^{k(\theta)} \rho\left(x(\tau^{-}_{(N^{\beta}_{(x,q)}(\theta)+j)}), q_{(N^{\beta}_{(x,q)}(\theta)+j-1)}, w_{(N^{\beta}_{(x,q)}(\theta)+j)} \right) e^{-\lambda \theta} \\ & + \psi\left(x(\theta^{+}_{k(\theta)}), q_{(N^{\beta}_{(x,q)}(\theta)+k(\theta))} \right) e^{-\lambda \theta}, \end{aligned} \tag{51}$$

and the value function,

$$\mathbf{V}(x,q) = \inf_{\beta(x,q) \in \mathcal{B}(x,q), \theta \geq 0, 0 \leq k(\theta) \leq N^{\beta}_{(x,q)}(\theta)} \{\mathbf{J}_{(x,q)}(\beta, \theta, k(\theta))\}, \tag{52}$$

where $\psi : \mathbf{R}^n \times Q \to \mathbf{R}^+$ is the stopping cost subject to the following assumption.

Assumption 7.1. There exists $0 < C^1_{\psi} < C^2_{\psi}$ such that

$$0 < C^1_{\psi} \leq \psi(x,q) \leq C^2_{\psi} \quad \forall (x,q) \in \mathbf{R}^n \times Q \tag{53}$$

and

$$(\psi^*)_* = \psi_*. \tag{54}$$

The *optimal stopping control of the hybrid system* is to find an admissible stopping control $(\beta^*(x,q), \theta^*, k^*(\theta^*))$ for every hybrid state (x,q), such that

$$\mathbf{J}_{(x,q)}(\beta^*, \theta^*, k^*(\theta^*)) \leq \mathbf{J}_{(x,q)}(\beta, \theta, k(\theta))$$

for any $\beta(x,q) \in \mathcal{B}(x,q)$, $\theta \geq 0, 0 \leq k(\theta) \leq M^{\beta}_{(x,q)}(\theta)$.

By a technique similar to that of Theorem 4.3, following the DPP for the value function, we can show that the value function is the discontinuous viscosity solution of the following SQVI:

$$\max\{\mathbf{v}(x,q)-\psi(x,q),\mathbf{v}(x,q)-(\mathcal{M}\mathbf{v})(x,q),\lambda\mathbf{v}(x,q)+\mathcal{H}(x,q,\nabla_x\mathbf{v}(x,q))\}=0, \tag{55}$$

where the operator $\mathcal{M}$ and the Hamiltonian $\mathcal{H}$ are defined the same as in (13) and (24).

Similarly, the proofs of comparison theorems for the sub- and super-solutions of the above SQVI rely on the nonlinear operator $\mathcal{T}^+:\mathcal{V}^+\to\mathcal{V}^+$ defined as

$$(\mathcal{T}^+v)(x,q)=\inf_{u(\cdot)\in\mathcal{U},\theta\geq 0}\left\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds+(\mathcal{L}^+v)(x(\theta),q)e^{-\lambda\theta}\right\}, \tag{56}$$

where $x(t)$ is the solution of

$$\dot{x}(t)=f(x(t),q,u(t))$$

with initial state $x(0)=x$ and where $\mathcal{L}^+:\mathcal{V}^+\to\mathcal{V}^+$ is defined as

$$\mathcal{L}^+v=\min\{\psi^*,\mathcal{M}^+v\}. \tag{57}$$

We can show as in Theorem 5.3 that the function defined below is the unique fixed point of $\mathcal{T}^+$ defined above and the maximum sub-solution and solution of SQVI (55).

For the same hybrid system (3), given any initial state $(x,q)\in(\mathbf{R}^n\times Q)$, define a function

$$\mathbf{V}^+(x,q)=\inf_{\beta(x,q)\in\mathcal{B}(x,q),\theta\geq 0,0\leq k(\theta)\leq N^\beta_{(x,q)}(\theta)}\{\mathbf{J}^+_{(x,q)}(\beta,\theta,k(\theta))\}, \tag{58}$$

where

$$\begin{aligned}\mathbf{J}^+_{(x,q)}(\beta,\theta,k(\theta))=&\sum_{i=1}^{N^\beta_{(x,q)}(\theta)}\left(\int_{\tau_{i-1}}^{\tau_i}\ell(x(s),q_{i-1},u(s))e^{-\lambda s}ds+\kappa^*(x(\tau_i^-),q_{i-1},w_i)e^{-\lambda\tau_i}\right)\\&+\int_{\tau_{N^\beta_{(x,q)}(\theta)}}^{\theta}\ell(x(s),q_{N^\beta_{(x,q)}(\theta)},u(s))e^{-\lambda s}ds\\&+\sum_{j=1}^{k(\theta)}\kappa^*\left(x(\tau^-_{(N^\beta_{(x,q)}(\theta)+j)}),q_{(N^\beta_{(x,q)}(\theta)+j-1)},w_{(N^\beta_{(x,q)}(\theta)+j)}\right)e^{-\lambda\theta}\\&+\psi^*\left(x(\theta^+_{k(\theta)}),q_{(N^\beta_{(x,q)}(\theta)+k(\theta))}\right)e^{-\lambda\theta},\end{aligned} \tag{59}$$

where $(\beta,\theta,k(\theta))$ is any hybrid stopping control.

Proceeding as in Theorem 5.6, the minimum super-solution of SQVI (55) is

$$\mathbf{V}^-(x,q)=\inf_{\hat{\beta}(x,q)\in\widehat{\mathcal{B}}(x,q),\theta\geq 0,0\leq k(\theta)\leq M^{\hat{\beta}}_{(x,q)}(\theta)}\{\mathbf{J}^-_{(x,q)}(\hat{\beta},\theta,k(\theta))\}, \tag{60}$$

where

(61)

$$\begin{aligned}\mathbf{J}^-_{(x,q)}(\hat{\beta},\theta,k(\theta)) = &\sum_{i=1}^{N^{\hat{\beta}}_{(x,q)}(\theta)} \int_{\tau_{i-1}}^{\tau_i}\int_U (\ell(\hat{x}(s),q_{i-1},u)e^{-\lambda s}d\mu_s + \kappa_*(x(\tau_i^-),q_{i-1},w_i)e^{-\lambda\tau_i})\\ &+ \int_{\tau_{N^{\hat{\beta}}_{(x,q)}(\theta)}}^{\theta}\int_U \ell\Big(\hat{x}(s),q_{N^{\hat{\beta}}_{(x,q)}(\theta)},u\Big)e^{-\lambda s}d\mu_s\\ &+ \sum_{j=1}^{k(\theta)}\kappa_*\Big(\hat{x}\Big(\tau^-_{(N^{\hat{\beta}}_{(x,q)}(\theta)+j)}\Big),q_{(N^{\hat{\beta}}_{(x,q)}(\theta)+j-1)},w_{(N^{\hat{\beta}}_{(x,q)}(\theta)+j)}\Big)e^{-\lambda\theta}\\ &+ \psi_*\Big(\hat{x}(\theta^+_{k(\theta)}),q_{(N^{\hat{\beta}}_{(x,q)}(\theta)+k(\theta))}\Big)e^{-\lambda\theta},\end{aligned}$$

which is the unique fixed point of the following nonlinear operator $\mathcal{T}^- : \mathcal{V}^- \to \mathcal{V}^-$:

$$(\mathcal{T}^-v)(x,q) = \inf_{\mu_{(\cdot)}\in\widehat{\mathcal{U}},\theta\ge 0}\left\{\int_0^\theta\int_U \ell(\hat{x}(s),q,u)e^{-\lambda s}d\mu_s ds + (\mathcal{L}^-v)(\hat{x}(\theta),q)e^{-\lambda\theta}\right\}, \tag{62}$$

where $\hat{x}(t)$ is the solution of

$$\dot{\hat{x}}(t) = \int_U f(\hat{x}(t),q,u)d\mu_t$$

with initial state $\hat{x}(0) = x$ and where $\mathcal{L}^- : \mathcal{V}^- \to \mathcal{V}^-$ is defined as

$$\mathcal{L}^-v = \min\{\psi_*, \mathcal{M}^-v\}. \tag{63}$$

Appendix A.

A.1. On the restricted optimal stopping problem. In this appendix, we establish the results concerning the restricted optimal stopping problem (ROSP). Consider a continuous system

$$\dot{x}(t) = f(x(t),u(t)),\ x(0)\in\Omega, \tag{64}$$

where $\Omega \subset \mathbf{R}^n$ is a bounded open set with smooth boundary $\partial\Omega$. The continuous vector fields $f : \overline{\Omega}\times U \to \mathbf{R}^n$ satisfy

$$\begin{cases}|f(x,u)| \le C & \forall x\in\overline{\Omega},\quad \forall u\in U,\\ |f(x,u)-f(y,u)| \le C|x-y| & \forall x,y\in\overline{\Omega},\quad \forall u\in U.\end{cases}$$

Given initial state $x\in\overline{\Omega}$ and control $(u(\cdot),\theta)$, the associated cost functional is defined as

$$\begin{aligned}J(x,u(\cdot),\theta) = &\int_0^{\theta\wedge T(x,u(\cdot))} \ell(x(t),u(t))e^{-\lambda t}dt + \varphi_1(x(\theta))e^{-\lambda\theta}\mathbf{I}\{\theta < T(x,u(\cdot))\}\\ &+ \varphi_2(x(T(x,u(\cdot))))e^{-\lambda T(x,u(\cdot))}\mathbf{I}\{\theta\ge T(x,u(\cdot))\},\end{aligned} \tag{65}$$

where $T(x,u(\cdot)) = \inf\{t\ge 0 | x(t)\in\Omega^c\}$ is the first exit time from Ω, $\ell : \overline{\Omega}\times U\to\mathbf{R}$ is the continuous running cost satisfying

$$\begin{cases}|\ell(x,u)| \le C & \forall x\in\overline{\Omega},\quad \forall u\in U,\\ |\ell(x,u)-\ell(y,u)| \le C|x-y| & \forall x,y\in\overline{\Omega},\quad \forall u\in U,\end{cases}$$

and $\varphi_1 : \Omega \to \mathbf{R}$ and $\varphi_2 : \partial\Omega \to \mathbf{R}$ are the bounded stopping cost and exit cost, respectively, both of which can be discontinuous.

The value function is defined as

$$V(x) = \inf_{u(\cdot)\in\mathcal{U},\theta\geq 0} J(x, u(\cdot), \theta), x \in \overline{\Omega}. \tag{66}$$

From the definition of the cost function (65), we know we can either stop the trajectory with a stopping cost before it exits Ω or exit Ω with an exit cost. The ROSP is to find a way to do it optimally. If we let

$$\psi(x) = \begin{cases} \varphi_1(x), & x \in \Omega, \\ \varphi_2(x), & x \in \partial\Omega, \end{cases}$$

then the value function can be expressed in a more compact form by

$$V(x) = \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T(x,u(\cdot))} \left(\int_0^\theta \ell(x(t), u(t))e^{-\lambda t}dt + \psi(x(\theta))e^{-\lambda\theta} \right). \tag{67}$$

From now on, we take (67) as the definition of the value function.

Since our main purpose is to prove the results needed in the proof of Theorems 5.3 and 5.6, we will not discuss the most general case. Instead, we assume the following transversality condition:

$$f(x, u) \cdot n(x) > \beta > 0 \quad \forall x \in \partial\Omega, \ \forall u \in U, \tag{68}$$

where $n(x)$ is the unit outer normal to Ω at x. Since the ROSP essentially is a combination of the optimal stopping problem (see Barles and Perthame [4]) and optimal exit problem (see Blanc [10]), the transversality condition eliminates the complexities due to the exit problem in the boundary, and hence the results of ROSP are very similar to those of the optimal stopping problem without constraint.

The goal of this section of the appendix is to explore the relation between the value function (67) and the HJB equation

$$\begin{cases} \max\{v(x) - \psi(x), \lambda v(x) + \mathcal{H}(x, \nabla v(x))\} = 0 & \text{in } \Omega, \\ v(x) = \psi(x) & \text{in } \partial\Omega. \end{cases} \tag{69}$$

We especially want to show that it has two extreme solutions under assumption (68) defined by

$$V^+(x) = \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T(x,u(\cdot))} \left(\int_0^\theta \ell(x(t), u(t))e^{-\lambda t}dt + \varphi^*(x(\theta))e^{-\lambda\theta} \right) \tag{70}$$

and minimum super-solution and solution

$$V^-(x) = \inf_{\mu_{(\cdot)}\in\mathcal{U},0\leq\theta\leq T(x,\mu_{(\cdot)})} \left(\int_0^\theta \int_U \ell(\hat{x}(t), u)e^{-\lambda t}d\mu_t dt + \varphi_*(\hat{x}(\theta))e^{-\lambda\theta} \right), \tag{71}$$

where $\hat{x}(\cdot)$ is the relaxed trajectory of $\dot{\hat{x}}(t) = \int_U f(\hat{x}(t), u)d\mu_t$ with relaxed control input $\mu_{(\cdot)}$.

We begin with simple cases.

THEOREM A.1. *Under the transversality condition* (68)*, the value function* (67) *is uniformly continuous in* $\overline{\Omega}$ *if* ψ *is uniformly continuous in* $\overline{\Omega}$*, and it is the unique viscosity solution of HJB equation* (69)*.*

Proof. Introduce $\widetilde{U} = U \cup \{w\}$, dynamics

$$\tilde{f}(x,\tilde{u}) = \begin{cases} f(x,u), & \tilde{u} = u \in U, \\ 0, & \tilde{u} = w, \end{cases}$$

and cost

$$\tilde{\ell}(x,\tilde{u}) = \begin{cases} \ell(x,u), & \tilde{u} = u \in U, \\ \psi(x), & \tilde{u} = w; \end{cases}$$

then ROSP is transformed into an optimal exit problem for new dynamics under new cost, i.e.,

$$V(x) = \inf_{\tilde{u}(\cdot)\in\widetilde{\mathcal{U}}} \left(\int_0^{T(x,\tilde{u}(\cdot))} \tilde{\ell}(\tilde{x}(t),\tilde{u}(t)) e^{-\lambda t} dt + \psi(\tilde{x}(T(x,\tilde{u}(\cdot)))) e^{-\lambda T(x,\tilde{u}(\cdot))} \right).$$

By the known result of the optimal exit problem (see Barles and Perthame [5]), we know that under the tranversality condition, the value function is uniformly continuous (the maximum solution and minimum solution coincide) and is the unique viscosity solution of the HJB equation

$$\begin{cases} \lambda v(x) + \widetilde{\mathcal{H}}(x, \nabla v(x)) = 0 & \text{in } \Omega, \\ v(x) = \psi(x) & \text{in } \partial\Omega, \end{cases}$$

where $\widetilde{\mathcal{H}}(x,p) = \sup_{\tilde{u}\in\widetilde{U}}(-\tilde{f}(x,\tilde{u})\cdot p - \tilde{\ell}(x,\tilde{u}))$. It is easy to check that this equation is equivalent to HJB equation (69). The proof is complete. □

Next, we consider a discontinuous stopping cost. Based on the above theorem, we can proceed as Barles and Perthame [5] to obtain a stability result.

LEMMA A.2. *Let ψ_n be a nonincreasing (resp., nondecreasing) sequence of bounded uniformly continuous functions such that*

$$\inf_n \psi_n = \psi^* \quad \Big(\sup_n \psi_n = \psi_* \Big);$$

then $V^+ = \inf_n V_n$ ($V^- = \sup_n V_n$), where V_n is defined as in (67) *with ψ_n replacing ψ.*

The proof of the lemma is similar to the proof of Theorem II.3 of Barles and Perthame [4] and is omitted here. Then basing our work on Lemma A.2 and following the proof of Theorems II.4 and II.5 in [4], we come to the main theorem.

THEOREM A.3. *Under the transversality conditions* (68), *V^+ (V^-) is the maximum sub-solution and solution (minimum super-solution and solution) of HJB equation* (37).

As for the uniqueness in a weak sense, following the similar proof of Proposition I.4 of [4], we have the following proposition.

PROPOSITION A.4. *Under condition* (68), *we have*

$$V_*^+ = V_* = V^-.$$

A.2. Proof of Lemma 5.1. Item 1. Since for any $(x,q) \in (X \times Q)$,

$$\begin{aligned}(\mathcal{M}^+\mathbf{v}_1)(x,q) &= \min_{w\in W}\{\kappa^*(x,q,w) + \mathbf{v}_1(g(x,q,w))\}\\ &= \kappa^*(x,q,\hat{w}) + \mathbf{v}_1(g(x,q,\hat{w}))\\ &\geq \kappa^*(x,q,\hat{w}) + \mathbf{v}_2(g(x,q,\hat{w}))\\ &\geq \min_{w\in W}\{\kappa^*(x,q,w) + \mathbf{v}_2(g(x,q,w))\}\\ &= (\mathcal{M}^+\mathbf{v}_2)(x,q),\end{aligned}$$

it holds that

$$\begin{aligned}(\mathcal{T}^+\mathbf{v}_1)(x,q) &= \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds\\ &\qquad + (\mathcal{M}^+\mathbf{v}_1)(x(\theta),q)e^{-\lambda\theta}\Big\}\\ &\geq \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds\\ &\qquad + (\mathcal{M}^+\mathbf{v}_2)(x(\theta),q)e^{-\lambda\theta}\Big\}\\ &= (\mathcal{T}^+\mathbf{v}_2)(x,q).\end{aligned}$$

Item 2. Since for any $(x,q) \in (X \times Q)$,

$$\begin{aligned}(\mathcal{M}^+(\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2))(x,q) &= \min_{w\in W}\{\kappa^*(x,q,w) + (\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2)(g(x,q,w))\}\\ &= (\alpha + (1-\alpha))\kappa^*(x,q,\hat{w}) + (\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2)(g(x,q,\hat{w}))\\ &\geq \alpha\Big\{\min_{w\in W}\{\kappa^*(x,q,w) + \mathbf{v}_1(g(x,q,w))\Big\}\\ &\quad + (1-\alpha)\Big\{\min_{w\in W}\{\kappa^*(x,q,w) + \mathbf{v}_2(g(x,q,w))\Big\}\\ &= \alpha(\mathcal{M}^+\mathbf{v}_1)(x,q) + (1-\alpha)(\mathcal{M}^+\mathbf{v}_2)(x,q),\end{aligned}$$

it holds that

$$\begin{aligned}&(\mathcal{T}^+(\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2))(x,q)\\ &= \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds\\ &\qquad + (\mathcal{M}^+(\alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2))(x(\theta),q)e^{-\lambda\theta}\Big\}\\ &\geq \inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds\\ &\qquad + (\alpha\mathcal{M}^+\mathbf{v}_1 + (1-\alpha)\mathcal{M}^+\mathbf{v}_2)(x(\theta),q)e^{-\lambda\theta}\Big\}\\ &\geq \alpha\inf_{u(\cdot)\in\mathcal{U},0\leq\theta\leq T^F(x,q,u(\cdot))}\Big\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds\\ &\qquad + (\mathcal{M}^+\mathbf{v}_1)(x(\theta),q)e^{-\lambda\theta}\Big\}\end{aligned}$$

$$+(1-\alpha)\inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))}\left\{\int_0^\theta \ell(x(s),q,u(s))e^{-\lambda s}ds + (\mathcal{M}^+\mathbf{v}_2)(x(\theta),q)e^{-\lambda\theta}\right\}$$
$$=\alpha(\mathcal{T}^+\mathbf{v}_1)(x,q)+(1-\alpha)(\mathcal{T}^+\mathbf{v}_2)(x,q).$$

Item 3. By Item 1, we know that

$$\mathcal{T}^+((1-\gamma)\mathbf{v}_1)\le\mathcal{T}^+\mathbf{v}_2.$$

Then using Item 2 with $\alpha=1-\gamma$, we have

$$\mathcal{T}^+((1-\gamma)\mathbf{v}_1+\gamma 0)\ge(1-\gamma)\mathcal{T}^+\mathbf{v}_1+\gamma\mathcal{T}^+0$$

and then

$$\mathcal{T}^+\mathbf{v}_2\ge(1-\gamma)\mathcal{T}^+\mathbf{v}_1+\gamma\mathcal{T}^+0.$$

Let

$$\bar{\mathbf{v}}_1(x,q)=\inf_{u(\cdot)\in\mathcal{U}}\left\{\int_0^{T^F(x,q,u(\cdot))}\ell(x(s),q,u(s))e^{-\lambda s}ds + (\mathcal{M}^+\mathbf{v}_1)(x(T^F(x,q,u(\cdot))),q)e^{-\lambda T^F(x,q,u(\cdot))}\right\}.$$

This function is obviously bounded, so there exists C_1 such that $\bar{v}_1\le C_1$ and

$$(\mathcal{M}^+0)(x,q)=\min_{w\in W}\{\kappa^*(x,q,w)\}\ge C_\rho.$$

Let $\alpha\le C_T<(\frac{C_\rho}{C_1}\wedge 1)$; then

$$\begin{aligned}(\mathcal{T}^+0)(x,q)&=\inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))}\left\{\int_0^\theta\ell(x(s),q,u(s))e^{-\lambda s}ds+(\mathcal{M}^+0)(x(\theta),q)e^{-\lambda\theta}\right\}\\&\ge\inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))}\left\{\int_0^\theta\ell(x(s),q,u(s))e^{-\lambda s}ds+C_\rho e^{-\lambda\theta}\right\}\\&\ge\inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))}\left\{\int_0^\theta\alpha\ell(x(s),q,u(s))e^{-\lambda s}ds+\alpha\bar{\mathbf{v}}_1(x(\theta),q)e^{-\lambda\theta}\right\}\\&=\alpha\bar{\mathbf{v}}_1(x,q)\end{aligned}$$

and

$$\begin{aligned}(\mathcal{T}^+\mathbf{v}_1)(x,q)&=\inf_{u(\cdot)\in\mathcal{U},0\le\theta\le T^F(x,q,u(\cdot))}\left\{\int_0^\theta\ell(x(s),q,u(s))e^{-\lambda s}ds+(\mathcal{M}^+\mathbf{v}_1)(x(\theta),q)e^{-\lambda\theta}\right\}\\&\le\inf_{u(\cdot)\in\mathcal{U}}\left\{\int_0^{T^F(x,q,u(\cdot))}\ell(x(s),q,u(s))e^{-\lambda s}ds+(\mathcal{M}^+\mathbf{v}_1)(x(T^F(x,q,u(\cdot))),q)e^{-\lambda T^F(x,q,u(\cdot))}\right\}\\&=\bar{\mathbf{v}}_1(x,q).\end{aligned}$$

Hence

$$\mathcal{T}^+ 0 \geq \alpha \mathcal{T}^+ \mathbf{v}_1;$$

then

$$\mathcal{T}^+ \mathbf{v}_2 \geq (1-\gamma)\mathcal{T}^+ \mathbf{v}_1 + \gamma\alpha \mathcal{T}^+ \mathbf{v}_1.$$

We have

$$\mathcal{T}^+ \mathbf{v}_1 - \mathcal{T}^+ \mathbf{v}_2 \leq \gamma(1-\alpha)\mathcal{T}^+ \mathbf{v}_1.$$

Item 4. Since

$$\mathbf{v}_1 - \mathbf{v}_2 \leq \mathbf{v}_1,$$

using Item 3, with $\gamma = 1$, we know there exists $0 < \alpha < 1$ such that

$$\mathcal{T}^+ \mathbf{v}_1 - \mathcal{T}^+ \mathbf{v}_2 \leq (1-\alpha)\mathcal{T}^+ \mathbf{v}_1,$$

and since $\mathbf{v}_1, \mathbf{v}_2$ are fixed points,

$$\mathbf{v}_1 - \mathbf{v}_2 \leq (1-\alpha)\mathbf{v}_1.$$

Using Item 3 again, with $\gamma = 1-\alpha$, we have

$$\mathcal{T}^+ \mathbf{v}_1 - \mathcal{T}^+ \mathbf{v}_2 \leq (1-\alpha)^2 \mathcal{T}^+ \mathbf{v}_1,$$

which means

$$\mathbf{v}_1 - \mathbf{v}_2 \leq (1-\alpha)^2 \mathbf{v}_1.$$

Continuing this procedure, we have

$$\mathbf{v}_1 \leq \mathbf{v}_2.$$

By the same argument we can show that

$$\mathbf{v}_2 \leq \mathbf{v}_1.$$

Then

$$\mathbf{v}_1 = \mathbf{v}_2.$$

REFERENCES

[1] J. A. Ball, J. Chudoung, and M. V. Day, *Robust optimal switching control for nonlinear systems*, SIAM J. Control Optim., 41 (2002), pp. 900–931.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser Boston, Boston, 1997.

[3] G. Barles, *Deterministic impulse control problems*, SIAM J. Control Optim., 23 (1985), pp. 419–432.

[4] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél Math. Anal. Numer., 21 (1987), pp. 557–579.

[5] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[6] G. Barles, *Discontinuous viscosity solutions of first order Hamilton–Jacobi equations: A guided visit*, Nonlinear Anal., 20 (1993), pp. 1123–1134.
[7] S. A. Belbas and S. M. Lenhart, *Nonlinear PDEs for stochastic optimal control with switchings and impulses*, Appl. Math. Optim., 14 (1986), pp. 215–227.
[8] A. Bensoussan and J.-L. Lions, *Impulse Control and Quasivariational Inequalities*, Gauthier-Villars, Paris, 1984.
[9] A. Bensoussan and J. L. Menaldi, *Hybrid control and dynamic programming.* Dynam. Contin. Discrete and Impuls. Systems, 3 (1997), pp. 395–442.
[10] A.-P. Blanc, *Deterministic exit time control problems with discontinuous exit costs*, SIAM J. Control Optim., 35 (1997), pp. 399–434.
[11] M. S. Branicky, V. S. Borkar, and S. K. Mitter, *A unified famework for hybrid control: Model and optimal control theory*, IEEE Trans. Automat. Control, 43 (1998), pp. 31–45.
[12] I. Capuzzo Dolcetta and L. C. Evans, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143–161.
[13] C. G. Cassandras, D. L. Pepyne, and Y. Wardi, *Optimal control of a class of hybrid systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 398–415.
[14] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.
[15] P. Dupuis and A. Nagurney, *Dynamical systems and variational inequalities*, Ann. Oper. Res., 44 (1993), pp. 9–42.
[16] S. Dharmatti and M. Ramaswamy, *Hybrid control systems and viscosity solutions*, SIAM J. Control Optim., 44 (2005), pp. 1259–1288.
[17] J. P. Hespanha, S. Bohacek, K. Obraczka, and J. Lee, *Hybrid modeling of TCP congestion control*, in Proceedings of the 4th International Workshop on Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2034, 2001, Springer-Verlag, London, pp. 291–304.
[18] W. Heemels, B. D. Schutter, and A. Bemporad, *Equivalence of hybrid dynamical models*, Automatica J. IFAC, 37 (2001), pp. 1085–1091.
[19] H. Ishii, *Perron's method for Hamilton–Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.
[20] J. Lygeros, D. N. Godbole, and S. Sastry, *Verified hybrid controllers for automatied vehicles*, IEEE Trans. Automat. Control, 43 (1998), pp. 522–539.
[21] J. Lygeros, C. Tomlin, and S. Sastry, *Controllers for reachability specifications for hybrid systems*, Automatica J. IFAC, 35 (1999), pp. 349–370.
[22] M. S. Shaikh and P. E. Caines, *On the optimal control of hybrid systems: Optimization of trajectories, switching times, and location schedules*, in Proceedings of the 6th International Workshop on Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2623, 2003, Springer-Verlag, Berlin, pp. 466–481.
[23] H. Sussmann, *A maximum principle for hybrid optimal control problems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 425–430.
[24] C. Tomlin, G. L. Pappas, and S. Sastry, *Conflict resolution for air traffic management: A case study in multiagent hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 509–521.
[25] P. P. Varaiya, *Smart cars on smart roads: Problem of control*, IEEE Trans. Automat. Control, 38 (1993), pp. 195-207.
[26] J. Yong, *Systems governed by ordinary differential equations with continuous, switching and impulse controls*, Appl. Math. Optim., 20 (1989), pp. 223–235.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 2, pp. 762–772

THE CONTROL TRANSMUTATION METHOD AND THE COST OF FAST CONTROLS*

LUC MILLER†

Abstract. In this paper, the null-controllability in any positive time T of the first-order equation (1) $\dot{x}(t) = e^{i\theta}Ax(t) + Bu(t)$ ($|\theta| < \pi/2$ fixed) is deduced from the null-controllability in some positive time L of the second-order equation (2) $\ddot{z}(t) = Az(t) + Bv(t)$. The differential equations (1) and (2) are set in a Banach space, B is an admissible unbounded control operator, and A is a generator of cosine operator function. The control transmutation method makes explicit the input function u of (1) in terms of the input function v of (2): $u(t) = \int_{\mathbb{R}} k(t,s)v(s)\,ds$, where the compactly supported kernel k depends only on T and L. This method proves roughly that the norm of a u steering the system (1) from an initial state $x(0) = x_0$ to the final state $x(T) = 0$ grows at most like $\|x_0\| \exp(\alpha_* L^2/T)$ as the control time T tends to zero. (The rate α_* is characterized independently by a one-dimensional controllability problem.) In applications to the cost of fast controls for the heat equation, L is roughly the length of the longest ray of geometric optics which does not intersect the control region.

Key words. controllability, fast controls, control cost, transmutation, cosine operator function, heat equation, linear Ginzburg–Landau equation

AMS subject classifications. 93B05, 93B17, 47D09

DOI. 10.1137/S0363012904440654

1. Introduction.

This paper concerns the relationship between the null-controllability of the following first- and second-order controllable systems:

$$\dot{x}(t) = e^{i\theta}Ax(t) + Bu(t), \qquad t \in \mathbb{R}_+, \quad x(0) = x_0, \tag{1.1}$$

$$\ddot{z}(t) = Az(t) + Bv(t), \qquad t \in \mathbb{R}, \quad z(0) = z_0,\ \dot{z}(0) = 0, \tag{1.2}$$

where x and z are the systems trajectories in the Banach space X, x_0 and z_0 are initial states, u and v are input functions with values in the Banach space U, A is an unbounded generator, B is an unbounded control operator, θ is a given angle in $]-\pi/2, \pi/2[$, each dot denotes a derivative with respect to the time t, and $\mathbb{R}_+ = [0, \infty)$. (The detailed setting is given in section 2.)

Equation (1.1) with $u = 0$ describes an irreversible system (always smoothing), and we think of it as a parabolic distributed system with infinite propagation speed. Equation (1.2) with $v = 0$ describes a reversible system (e.g., conservative), and we think of it as a hyperbolic distributed system with finite propagation speed.

For example, if A is the negative Laplacian on a Euclidean region and the input function is a locally distributed boundary value set by B, then (1.2) is a boundary controlled scalar wave equation, and (1.1) with $\theta = 0$ is a boundary controlled heat equation (section 6 elaborates on this example). Thus, this paper applies to the controllability of the diffusion equation that models the propagation of heat in a medium at rest when the temperature does not vary too much. This diffusion equation is also a

*Received by the editors February 4, 2004; accepted for publication (in revised form) January 26, 2006; published electronically June 2, 2006.
http://www.siam.org/journals/sicon/45-2/44065.html

†Centre de Mathématiques Laurent Schwartz, UMR CNRS 7640, Ecole Polytechnique, 91128 Palaiseau Cedex, France, and MODAL'X, EA 3454, Bâtiment G, Université de Paris X - Nanterre, 200 Av. de la République, 92001 Nanterre Cedex, France (miller@math.polytechnique.fr). This work was partially supported by the ACI grant "Équation des ondes: Oscillations, dispersion et contrôle."

simplified model for various physical phenomena in neutronics, viscous fluids mechanics, and electromagnetism when A is a more general coercive symmetric second-order elliptic operator (cf. section I.B.1.1 in [11]). If $\theta \neq 0$, then (1.1) interpolates between the heat equation and the Schrödinger equation. Thus, this paper applies to the controllability of the linear part of the complex Ginzburg–Landau equation, which describes a vast variety of physical phenomena from nonlinear waves to second-order phase transitions, from superconductivity, superfluidity, and Bose–Einstein condensation to liquid crystals and strings in field theory (cf. [2]). In particular, the one-dimensional linearized Ginzburg–Landau model considered in [1] enters this setting. If $\pi/2 - \theta$ is a small positive constant, then (1.1) is a viscous perturbation of the Schrödinger equation, which appears naturally in the building of stable numerical schemes (cf. [19]). Although conservative physical systems (e.g., the linear system of elasticity, or Maxwell's equations; cf. section I.B.1.2 in [11]) provide examples of (1.2) for which some controllability results are known, we have not yet found one for which the corresponding first-order system (1.1) is relevant to physics or engineering (besides the scalar wave equation considered in section 6).

This paper presents the *control transmutation method* (cf. [18] for a survey on transmutations in other contexts), which can be seen as a shortcut to Russell's famous harmonic analysis method in [26]. It consists of explicitly constructing controls u in any time T for the heat-like equation (1.1) in terms of controls v in time L for the corresponding wave-like equation (1.2), i.e., $u(t) = \int_{\mathbb{R}} k(t,s)v(s)\,ds$, where the compactly supported kernel k depends on T and L.[1] It proves that the exact controllability of (1.2) in some time L implies the null-controllability of (1.1) in any time with a relevant upper bound on the cost of controls for (1.1) as T tends to 0, in short, *the cost of fast controls* (as in the title of this paper and of [30]). Thanks to the geodesic condition of Bardos, Lebeau, and Rauch (cf. [6]) for the controllability of the wave equation, the application of this method to the boundary controllability of the heat equation (cf. section 6) yields new geometric bounds on the cost of fast controls, extending the results of [22] on internal controllability.[2] This method applies as soon as the wave equation is controllable, no matter how singular the coefficients in A are (e.g., discontinuous coefficients in transmission problems in [12], coefficients of bounded variation in [16]) but it does not seem to adapt to time-dependent coefficients. The companion paper [21] concerns the quite different case $|\theta| = \pi/2$ (in particular, there is no smoothing effect) and applies to the Schrödinger equation, which is the control problem to which a transmutation method was first applied by Phung in [25].

The relationship between first- and second-order controllable systems has been investigated in previous papers, always with $\theta = 0$ and the additional initial data $\dot{z}(0) = z_1$ in X (i.e., considering trajectories of (1.2) in the state space $X \times X$). In [13], it was proved that the approximate controllability of (1.2) with $\dot{z}(0) = z_1$ in some time implies the approximate controllability of (1.1) for any time (the control transmutation method yields an alternative proof) and proves the converse under some assumptions on the spectrum of A. The converse is investigated further in [32]

[1] Although L denotes a time here, it also denotes a length in the construction of k and in the main application to the wave and heat equations; the notation is meant as a reminder of this key fact of the method.

[2] The controllability of the heat equation holds even if the geometry prevents the controllability of the wave equation, but in that case no explicit geometric bound on the cost of fast controls is known. For interior controllability, a bound of the form $\exp(C_\beta/T^\beta)$ for any $\beta > 1$ is proved in [24], but the positive constant C_β is not explicit. A better bound of the form $\exp(C_1/T)$ is proved in [15], at least in the Euclidean case (cf. Remark 6.4), but the positive constant C_1 is not explicit either.

(in Hilbert spaces) and [31]. In a restricted setting, the null-controllability of (1.1) was deduced from the exact controllability of (1.2) with $\dot{z}(0) = z_1$ in [26] and [27] by the indirect method of biorthogonal bases[3] (recently used in [16]).

The study of the cost of fast controls was initiated by Seidman in [28] with a result on the heat equation obtained by Russell's method. Seidman also obtained results on the Schrödinger equation by working directly on the corresponding window problem for series of complex exponentials (see [23] for improvements and references). With collaborators, he later treated the case of finite-dimensional linear systems (cf. [30]) and generalized the window problem to a larger class of complex exponentials (cf. [29]). Fernández-Cara and Zuazua obtained results on the cost of fast interior controls for the heat equation discussed in Remark 6.4 (see [15]). Since their approach is based on global Carleman estimates (cf. [17]) it should extend to boundary controllability and to a very large class of equations with variable coefficients possibly depending on time as well. The control transmutation method generalizes upper bounds on the cost of fast controls from the one-dimensional setting (which reduces to a window problem) to the general setting, which we specify in the next section. Some open problems are stated in Remarks 3.3 and 6.4.

2. The setting. We assume that A is the generator of a strongly continuous cosine operator function Cos; i.e., the second-order Cauchy problem for $\ddot{z}(t) = A(t)z$ is well-posed and Cos is its propagator. (For a textbook presentation of cosine operator functions, we refer to Chapter 2 of [14] or section 3.14 of [3].) The associated sine operator function is $\mathrm{Sin}(t) = \int_0^t \mathrm{Cos}(s)ds$ (with the usual Bochner integral). Cos and Sin are strongly continuous functions on $\mathbb{R}$ of bounded operators on X. Moreover, A generates a holomorphic semigroup T of angle $\pi/2$ (cf. Theorem 3.14.17 of [3]). In particular, $S(t) = T(e^{i\theta}t)$ defines a strongly continuous semigroup $(S(t))_{t\in\mathbb{R}_+}$ of bounded operators on X. In this setting, for any source term $f \in L^1_{\mathrm{loc}}(\mathbb{R}, X)$, for any initial data x_0, z_0, and z_1 in X, the inhomogeneous first- and second-order Cauchy problems

$$\dot{x}(t) = e^{i\theta}Ax(t) + f(t), \qquad t \in \mathbb{R}_+, \quad x(0) = z_0, \tag{2.1}$$

$$\ddot{z}(t) = Az(t) + f(t), \qquad t \in \mathbb{R}, \quad z(0) = z_0,\ \dot{z}(0) = z_1, \tag{2.2}$$

have unique mild solutions $x \in C^0(\mathbb{R}_+, X)$ and $z \in C^0(\mathbb{R}, X)$ defined by

$$x(t) = S(t)x_0 + \int_0^t S(t-s)f(s)dt, \quad z(t) = \mathrm{Cos}(t)z_0 + \mathrm{Sin}(t)z_1 + \int_0^t \mathrm{Sin}(t-s)f(s)dt.$$

Remark 2.1. When A is a negative self-adjoint unbounded operator on a Hilbert space T, Cos and Sin are simply defined by the functional calculus as $T(t) = \exp(tA)$, $\mathrm{Cos}(t) = \cos(t\sqrt{-A})$, and $\mathrm{Sin}(t) = (\sqrt{-A})^{-1}\sin(t\sqrt{-A})$.

Following [33], we now make natural assumptions on B for any initial data in the state space X to define a unique continuous trajectory of each system (1.1) and (1.2). Let X_{-1} be the completion of X with respect to the norm $\|x\|_{-1} = \|(A-\beta)^{-1}x\|$ for some $\beta \in \mathbb{C}$ outside the spectrum of A. X_{-1} is also the dual of the space X_1 defined as

[3]The systematic approach to controllability based on biorthogonal bases (cf. [4]) does not apply in our general setting where A may not have eigenfunctions (e.g., on some unbounded domains). When it does apply, it builds on some knowledge about the eigenvalues and the analysis of some exponential series, whereas the control transmutation method per se does not (cf. Remark 3.3). A more detailed comparison can be found in [22], in the specific case which lead the author to this new method.

the domain $D(A)$ with the norm $\|x\|_1 = \|(A-\beta)x\|$. We assume that $B \in \mathcal{L}(U, X_{-1})$ is *an admissible unbounded control operator* in the following sense:

(2.3)
$$\forall t > 0, \forall u \in L^2([0,t];U),\ \int_0^t S(s)Bu(s)ds \in X \quad \text{and} \quad \int_0^t \mathrm{Sin}(s)Bu(s)ds \in X.$$

In this setting, for any x_0 and z_0 in X, for any u and v in $L^2_{\mathrm{loc}}(\mathbb{R}_+;U)$, the unique solutions x and z in $C^0(\mathbb{R};X)$ of (1.1) and (1.2), respectively, are defined by

(2.4)
$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)dt, \quad z(t) = \mathrm{Cos}(t)z_0 + \int_0^t \mathrm{Sin}(t-s)Bv(s)ds.$$

The natural notions of controllability cost for the linear systems (1.1) and (1.2) are as follows.

DEFINITION 2.2. *The system* (1.1) *is* null-controllable in time T *if for all* x_0 *in* X *there is a* u *in* $L^2(\mathbb{R}_+;U)$ *such that* $u(t) = 0$ *for* $t > T$ *and* $x(T) = 0$. *The* controllability cost *for* (1.1) *in time* T *is the smallest positive constant* $\kappa_{1,T}$ *such that, for all* ϕ_0, *there is such a* u *satisfying* $\int_0^T \|u(t)\|^2 dt \leqslant \kappa_{1,T}\|x_0\|^2$.

The system (1.2) *is* null-controllable in time T *if for all* z_0 *in* X *there is a* v *in* $L^2(\mathbb{R}_+;U)$ *such that* $v(t) = 0$ *for* $t > T$ *and* $z(T) = \dot{z}(T) = 0$. *The* controllability cost *for* (1.2) *in time* T *is the smallest positive constant* $\kappa_{2,T}$ *such that, for all* z_0, *there is such a* v *satisfying* $\int_0^T \|v(t)\|^2 dt \leqslant \kappa_{2,T}\|z_0\|^2$.

Remark 2.3. Equivalently, for all x_T in $S(T)X$, there is a u in $L^2(0,T;U)$ such that $x(0) = 0$ and $x(T) = x_T$, and, for all z_0, z_1, z_2, and z_3 in X, there is a v in $L^2(0,T;U)$ such that $(z(0), \dot{z}(0)) = (z_0, z_1)$ and $(z(T), \dot{z}(T)) = (z_2, z_3)$. This is the usual notion of exact controllability for (1.2) (cf. Lemma 2.4).

Although the notion of null-controllability for (1.2) in Definition 2.2 seems weaker than usual, the following lemma proves that it is equivalent to the usual notion.

LEMMA 2.4. *If* (1.2) *is null-controllable in time* T *in the sense of Definition* 2.2 *then, for all* z_0 *and* z_1 *in* X, *there is an input* v *in* $L^2(\mathbb{R}_+;U)$ *such that* $v(t) = 0$ *for* $t > T$ *and the solution of* $\ddot{z}(t) = Az(t) + Bv(t)$ $(t \in \mathbb{R})$ *with initial condition* $(z(0), \dot{z}(0)) = (z_0, z_1)$ *satisfies the final condition* $(z(T), \dot{z}(T)) = (0,0)$.

Proof. Assume that the system (1.2) satisfies Definition 2.2, and let z_0 and z_1 be in X. We shall build an input w in $L^2(0,T;U)$ such that the solution y of $\ddot{y}(t) = Ay(t) + Bw(t)$ with $(y(0), \dot{y}(0)) = (0, z_1)$ satisfies $(y(T), \dot{y}(T)) = (y_0, 0)$ for some $y_0 \in X$. Let v_1 be the input v in Definition 2.2, and let v_2 be the input v in Definition 2.2, with z_0 replaced by y_0. Due to the linearity of the system, the input $v_1(t) - v_2(T-t) + w(t)$ steers the system from (z_0, z_1) to $(0,0)$, as required to complete the proof of the lemma. The rest of the proof builds w and y.

Applying Definition 2.2 with z_0 replaced by z_1 yields an input v and the solution

$$z(s) = \mathrm{Cos}(s)z_1 + \int_0^s \mathrm{Sin}(r)Bv(s-r)dr, \quad z(T) = \dot{z}(T) = 0.$$

We define w and y by $w(t) = \int_0^t v(s)ds$ and $y(t) = \int_0^t z(s)ds$, for $t \in [0,T]$. The final

condition $\dot{y}(T) = z(T) = 0$ is immediately satisfied. Since $\mathrm{Sin}(t) = \int_0^t \mathrm{Cos}(s)ds$,

$$y(t) - \mathrm{Sin}(t)z_1 = \int_0^t \int_0^s \mathrm{Sin}(r)Bv(s-r)dr\,ds = \int_0^t \int_r^t \mathrm{Sin}(r)Bv(s-r)ds\,dr$$
$$= \int_0^t \mathrm{Sin}(r)B \int_r^t v(s-r)ds\,dr = \int_0^t \mathrm{Sin}(r)Bw(t-r)\,dr.$$

The exchange of integrals in the second step is justified since $(s,r) \mapsto \mathrm{Sin}(r)Bv(s-r)$ is continuous with values in X_{-1}. (Cos and Sin extend by density to strongly continuous functions on $\mathbb{R}$ of bounded operators on X_{-1} since they commute with $A - \beta$.) This integral formula means that y is the mild solution of $\ddot{y}(t) = Ay(t) + Bw(t)$ with initial condition $(y(0), \dot{y}(0)) = (0, z_1)$. □

3. The results and the method. Our estimate of the cost of fast controls for (1.1) builds, through the control transmutation method, on the same estimate for a simple system of type (1.1), i.e., on a segment $[0, L]$ with Dirichlet ($N = 0$) or Neumann ($N = 1$) condition at the left end controlled at the right end through a Dirichlet condition,

$$\partial_t \phi = e^{i\theta}\partial_s^2 \phi \text{ on }]0,T[\times]0,L[\,, \quad \partial_s^N \phi_{\restriction s=0} = 0, \quad \phi_{\restriction s=L} = u, \quad \phi_{\restriction t=0} = \phi_0. \tag{3.1}$$

With the notation of section 2, $x = \phi$, $A = \partial_s^2$ on $X = L^2(0,L)$ with $D(A) = \{f \in H^2(0,L) \mid \partial_s^N f(0) = f(L) = 0\}$, $\|\cdot\|_1$ with $\beta = 0$ is the homogeneous Sobolev $\dot{H}^2(0,L)$ norm, and B on $U = \mathbb{C}$ is the dual of $C \in \mathcal{L}(X_1; U)$ defined by $Cf = \partial_s f(L)$.

It is well known that the controllability of this system reduces by spectral analysis to classical results on nonharmonic Fourier series. The following upper bound for the cost of fast controls, proved in section 4, is an application of a refined result of Seidman, Avdonin, and Ivanov in [29].

THEOREM 3.1. *There are positive constants α and γ such that, for all $N \in \{0,1\}$, $L > 0$, $T \in \left]0, \inf(\pi, L)^2\right]$, the controllability cost $\kappa_{L,T}$ of the system* (3.1) *satisfies* $\kappa_{L,T} \leqslant \gamma \exp(\alpha L^2/T)$.

This theorem leads to the following definition of the optimal fast control cost rate for (3.1).

DEFINITION 3.2. *The rate α_* is the smallest positive constant such that for all $\alpha > \alpha_*$ there exists $\gamma > 0$ satisfying the property stated in Theorem* 3.1.

Remark 3.3. Computing α_* is an interesting open problem, and its solution does not have to rely on the analysis of a series of complex exponentials. The best estimate so far is $\alpha_* \in [1/2, 4(36/37)^2]$ for $\theta = 0$ (cf. [22]).

Our main result is a generalization of Theorem 3.1 to the first-order system (1.1) under some condition on the second-order system (1.2), as follows.

THEOREM 3.4. *If the system* (1.2) *is null-controllable for times greater than L_*, then the system* (1.1) *is null-controllable in any time T. Moreover, the controllability cost $\kappa_{1,T}$ of* (1.1) *satisfies the following upper bound (with α_* defined above):*

$$\limsup_{T \to 0} T \ln \kappa_{1,T} \leqslant \alpha_* L_*^2. \tag{3.2}$$

Remark 3.5. The upper bound (3.2) means that the norm of an input function u steering the system (1.1) from an initial state x_0 to zero grows at most like $\gamma \|x_0\| \exp(\alpha L^2/(2T))$ as the control time T tends to zero (for any α_* and some $\gamma > 0$). The falsity of the converse of the first statement in Theorem 3.4 is well known, e.g., in the more specific setting of section 6.

Remark 3.6. As observed in [9], (3.2) yields a logarithmic modulus of continuity for the minimal time function $T_{\min} : X \to [0, +\infty)$ of (1.1); i.e., $T_{\min}(x_0)$, defined as the infimum of the times $T > 0$ for which there is a u in $L^2(\mathbb{R}; Y)$ such that $\int_0^T \|u(t)\|^2 dt \leqslant 1$, $u(t) = 0$ for $t \notin [0, T]$, and $x(T) = 0$, satisfies the following continuity property: $\alpha > \alpha_*$ there is a $c > 0$ such that, for all x_0 and x_0' in X with $\|x_0 - x_0'\|$ small enough, $|T_{\min}(x_0) - T_{\min}(x_0')| \leqslant \alpha L_*^2 / \ln(c/\|x_0 - x_0'\|)$.

It is well known that the semigroup T can be expressed as an integral over the cosine operator function Cos (cf. the second proof of Theorem 3.14.17 in [3]):

$$\forall x \in X, \forall t \in \mathbb{C} \text{ such that } |\arg t| < \pi/2, \quad T(t)x = \int k(t, s) \operatorname{Cos}(s) x ds, \tag{3.3}$$

where $k(0, s) = \delta(s)$ and $k(t, s) = \exp(-s^2/(4t))/\sqrt{\pi t}$ for $\operatorname{Re} t > 0$. This equation has been referred to as the abstract Poisson or Weierstrass formula. Starting with the observation that k is the fundamental solution of the heat equation on the line, i.e., k is the solution of $\partial_t k = \partial_s^2 k$ with the Dirac measure at the origin as initial condition, the *transmutation control method* consists of replacing the kernel k in (3.3) by some *fundamental controlled solution* on the segment $[-L, L]$, controlled at both ends (cf. (5.8)). The one-dimensional Theorem 3.1 is used to construct this fundamental controlled solution in section 4, and the transmutation is performed in section 5.

4. The fundamental controlled solution. This section begins with an outline of the standard application of [29] to the proof of Theorem 3.1. Following closely section 5 of [22], the rest of the section outlines the construction of a "fundamental controlled solution" k in the following sense, where $\mathcal{D}'(\mathcal{O})$ denotes the space of distributions on the open set $\mathcal{O}$ endowed with the weak topology, $\mathcal{M}(\mathcal{O})$ denotes the subspace of Radon measures on $\mathcal{O}$, and δ denotes the Dirac measure at the origin.

Definition 4.1. *The distribution $k \in C^0([0, T]; \mathcal{M}(]-L, L[))$ is a fundamental controlled solution for* (4.1) *at cost* (γ, α) *if*

$$\partial_t k = e^{-i\theta} \partial_s^2 k \quad \text{in } \mathcal{D}'(]0, T[\times]-L, L[), \tag{4.1}$$

$$k_{\restriction t=0} = \delta \quad \text{and} \quad k_{\restriction t=T} = 0, \tag{4.2}$$

$$\|k\|^2_{L^2(]0,T[\times]-L,L[)} \leqslant \gamma e^{\alpha L^2/T}. \tag{4.3}$$

The operator A defined at the beginning of section 3 is negative self-adjoint on the Hilbert space $L^2(0, L)$. It has a sequence $\{\mu_n\}_{n\in\mathbb{N}^*}$ of negative decreasing eigenvalues and an orthonormal Hilbert basis $\{e_n\}_{n\in\mathbb{N}^*}$ in $L^2(0, L)$ of corresponding eigenfunctions. Explicitly, $\sqrt{-\mu_n} = (n + \nu)\pi/L$ with $\nu = 0$ for $N = 0$ (Dirichlet) and $\nu = 1/2$ for $N = 1$ (Neumann). First note that Theorem 3.1 can be reduced to the case $L = \pi$ by the rescaling $(t, s) \mapsto (\sigma^2 t, \sigma s)$ with $\sigma = L/\pi$. In terms of the coordinates $c = (c_n)_{n\in\mathbb{N}^*}$ of $A^{N/2} f_0$ in the Hilbert basis $(e_n)_{n\in\mathbb{N}^*}$, where f_0 is the initial state of the dual observability problem, Theorem 3.1 with $L = \pi$ reduces by duality to the following window problem: there exist $\alpha > 0$ and $\gamma > 0$ such that, for all $T \in \left]0, \pi^2\right]$,

$$\forall c \in l^2(\mathbb{N}^*), \quad \sum_{n\in\mathbb{N}^*} |c_n|^2 \leqslant \gamma e^{\alpha\pi^2/T} \int_0^T |F(t)|^2 dt, \quad \text{where } F(t) = \sum_{n=1}^{\infty} c_n e^{\exp(i\theta)\mu_n t}.$$

Since this results from Theorem 1 of [29] with $\lambda_n \sim i e^{i\theta} n^2$, as in section 5:2 of [29], the proof of Theorem 3.1 is complete.

Now we consider a system governed by the same equation as (3.1) but on the twofold segment $[-L, L]$ controlled at both ends:

$$\partial_t \phi - e^{-i\theta}\partial_s^2 \phi = 0 \quad \text{in }]0,T[\times]-L,L[, \quad \phi_{\restriction s=\pm L} = u_\pm, \quad \phi_{\restriction t=0} = \phi_0, \tag{4.4}$$

with initial state $\phi_0 \in L^2(0,L)$ and input functions u_- and u_+ in $L^2(0,T)$. As in Proposition 5.1 of [22], applying Theorem 3.1 with $N = 0$ to the odd part of ϕ_0 and with $N = 1$ to the even part of ϕ_0 proves that the controllability cost of (4.4) is not greater than the controllability cost of (3.1) and therefore satisfies the same estimate stated in Theorem 3.1. As in Proposition 5.2 of [22], we may now successively combine the smoothing effect of (4.4) with no input (i.e., $u_+ = u_- = 0$) and this controllability cost estimate (plugged into the integral formula expressing ϕ in terms of ϕ_0 and $u_\pm = \phi_{\restriction s=\pm L}$) to obtain the following result.

PROPOSITION 4.2. *For all $\alpha > \alpha_*$ there exists $\gamma > 0$ such that for all $L > 0$ and $T \in]0, \inf(\pi/2, L)^2]$ there is a fundamental controlled solution for* (4.4) *at cost* (γ, α) *(cf. Definition* 4.1*).*

5. The transmutation of second-order controls into first-order controls. In this section we prove Theorem 3.4.

Let $x_0 \in X$ be an initial state for (1.1), and let $L > L_*$. Let $z \in C^0(\mathbb{R}_+; X)$ and $v \in L^2(\mathbb{R}_+; U)$ be the solution and input function obtained by applying the exact controllability of (1.2) in time L to the initial state $z_0 = x_0$.

We define $\underline{z} \in C^0(\mathbb{R}; X)$ and $\underline{v} \in L^2(\mathbb{R}; U)$ as the extensions of ζ and v by reflection with respect to $s = 0$; i.e., $\underline{z}(s) = \zeta(s) = \underline{z}(-s)$ and $\underline{v}(s) = v(s) = \underline{v}(-s)$ for $s \geqslant 0$. They inherit the following from (2.4):

$$\underline{z}(t) = \mathrm{Cos}(t)x_0 + \int_0^t \mathrm{Sin}(t-s)B\underline{v}(s)ds. \tag{5.1}$$

Definition 2.2 of $\kappa_{2,L}$ implies the following cost estimate for $\underline{v}$:

$$\int \|\underline{v}(s)\|^2 ds = 2\int_0^L \|v(s)\|^2 ds \leqslant 2\kappa_{2,L}\|x_0\|^2. \tag{5.2}$$

Since $D(A)$ is dense in X, there is a sequence $(x_n)_{n\in N^*}$ in $D(A)$ converging to x_0 in X. Since X_1 is dense in X_{-1}, there is a sequence $(f_n)_{n\in\mathbb{N}^*}$ in $C^1(\mathbb{R}; X_1)$ converging to $B\underline{v}$ in $L^2(\mathbb{R}; X_{-1})$. For each $n \in \mathbb{N}^*$, let $\underline{z}_n$ be defined in $C^2(\mathbb{R}; X)$ by

$$\underline{z}_n(t) = \mathrm{Cos}(t)x_n + \int_0^t \mathrm{Sin}(t-s)f_n(s)ds,$$

which converges to $\underline{z}(t)$ in X for all t due to (5.1). Since $\underline{z}_n$ is a genuine solution of $\ddot{\underline{z}}(t) = A\underline{z}_n(t) + f_n(t)$ (cf. Lemma 4.1 of [14]), we have for all φ in $D(A')$

$$s \mapsto \langle \underline{z}_n(s), \varphi\rangle \in H^2(\mathbb{R}) \quad \text{and} \quad \frac{d^2}{ds^2}\langle \underline{z}_n(s), \varphi\rangle = \langle \underline{z}_n(s), A'\varphi\rangle + \langle f_n(s), \varphi\rangle.$$

Hence, $\langle \underline{z}_n(t), \varphi\rangle = \langle x_n, \varphi\rangle + \int_0^t (t-s)\langle \underline{z}_n(s), A'\varphi\rangle + \int_0^t (t-s)\langle f_n(s), \varphi\rangle$. Passing to the limit yields $\langle \underline{z}(t), \varphi\rangle = \langle x_0, \varphi\rangle + \int_0^t (t-s)\langle \underline{z}(s), A'\varphi\rangle + \int_0^t (t-s)\langle B\underline{v}(s), \varphi\rangle$. Therefore

$$s \mapsto \langle \underline{z}(s), \varphi\rangle \in H^2(\mathbb{R}) \quad \text{and} \quad \frac{d^2}{ds^2}\langle \underline{z}(s), \varphi\rangle = \langle \underline{z}(s), A\varphi\rangle + \langle B\underline{v}(s), \varphi\rangle, \tag{5.3}$$

$$\langle \underline{z}(s), \varphi\rangle = 0 \quad \text{and} \quad \frac{d}{ds}\langle \underline{z}(s), \varphi\rangle = 0 \quad \text{for } |s| = L. \tag{5.4}$$

Let $\alpha > \alpha_*$ and $T \in]0, \inf(1, L^2)[$. Let $\gamma > 0$ and $k \in C^0([0,T]; \mathcal{M}(]-L, L[))$ be the corresponding constant and fundamental controlled solution given by Proposition 4.2. We define $\underline{k} \in C^0(\mathbb{R}_+; \mathcal{M}(\mathbb{R}))$ as the extension of k by zero; i.e., $\underline{k}(t,s) = \bar{k}(t,s)$ on $[0,T]\times]-L, L[$ and $\underline{k}$ is zero everywhere else. It inherits from k the following properties:

$$\partial_t \underline{k} = e^{i\theta} \partial_s^2 \underline{k} \quad \text{in } \mathcal{D}'(]0,T[\times]-L,L[), \tag{5.5}$$

$$\underline{k}_{]t=0} = \delta \quad \text{and} \quad \underline{k}_{]t=T} = 0, \tag{5.6}$$

$$\|\underline{k}\|^2_{L^2(\mathbb{R}_+\times\mathbb{R})} \leqslant \gamma e^{\alpha L^2/T}. \tag{5.7}$$

The main idea of the proof is to use $\underline{k}$ as a kernel to transmute $\underline{z}$ and $\underline{v}$ into a solution x and a control u for (1.1). The transmutation formulas

$$x(t) = \int \underline{k}(t,s)\underline{z}(s)\,ds \quad \text{and} \quad \forall t > 0,\ u(t) = \int \underline{k}(t,s)\underline{v}(s)\,ds \tag{5.8}$$

define $x \in C^0(\mathbb{R}_+; X)$ and $u \in L^2(\mathbb{R}_+; U)$ since $\underline{k} \in C^0(\mathbb{R}_+; \mathcal{M}(\mathbb{R})) \cap L^2(\mathbb{R}_+; L^2(\mathbb{R}))$, $\underline{z} \in C^0(\mathbb{R}; X)$, and $\underline{v} \in L^2(\mathbb{R}; U)$. The property (5.6) of $\underline{k}$ implies $x(0) = x_0$ and $x(T) = 0$. Equations (5.3), (5.4), and (5.5) imply, by integrating by parts,

$$\forall \varphi \in D(\mathcal{A}'), \quad t \mapsto \langle x(t), \varphi\rangle \in H^1(\mathbb{R}_+) \quad \text{and} \quad \frac{d}{dt}\langle x(t), \varphi\rangle = \langle x(t), A'\varphi\rangle + \langle Bu(t), \varphi\rangle.$$

This characterizes x as the unique solution of (1.1) in the weak sense (cf. [5]), which implies that x and u satisfy (2.4). Since $\int \|u(t)\|^2 dt \leqslant \iint |\underline{k}(t,s)|^2 ds\, dt \int \|\underline{v}(s)\|^2 ds$, (5.7) and (5.2) imply the following cost estimate, which completes the proof of Theorem 3.4:

$$\int_0^T \|u(t)\|^2 dt \leqslant 2\kappa_{2,L}\gamma e^{\alpha L^2/T}\|x_0\|^2.$$

6. Geometric bounds on the cost of fast boundary controls for the heat equation. When the second-order equation (1.2) has a finite propagation speed and is controllable, the control transmutation method yields geometric upper bounds on the cost of fast controls for the first-order equation (1.1). From this point of view, this method is an adaptation of the kernel estimates method of Cheeger, Gromov, and Taylor in [10]. This was illustrated in [22] and [23] on the internal controllability of heat and Schrödinger equations on Riemannian manifolds, which have the wave equation as corresponding second-order equation. Some similar lower bounds are proved in these papers (without assuming the controllability of the wave equation) which imply that the upper bounds are optimal with respect to time-dependence. In this section, we illustrate the control transmutation method on the analogous boundary control problems for the heat equation.

Let (M, g) be a smooth connected compact n-dimensional Riemannian manifold with metric g and smooth boundary ∂M. When $\partial M \neq \emptyset$, M denotes the interior and $\overline{M} = M \cup \partial M$. Let Δ denote the (negative) Laplacian on (M, g) and ∂_ν denote the exterior Neumann vector field on ∂M. The characteristic function of a set S is denoted by χ_S.

Let $X = L^2(M)$. Let A be defined by $Af = \Delta f$ on $D(A) = H^2(M) \cap H_0^1(M)$. Let C be defined from $D(A)$ to $U = L^2(\partial M)$ by $Cf = \partial_\nu f_{]\Gamma}$, where Γ is an open subset of ∂M, and let B be the dual of C. With this setting, (1.1) with $\theta = 0$ and

(1.2) are the heat and wave equations controlled by the Dirichlet boundary condition on Γ. In particular, (1.2) becomes

$$\begin{aligned} &\partial_t^2 z - \Delta z = 0 \ \text{ on } \ \mathbb{R}_t \times M, \quad z = \chi_\Gamma v \ \text{ on } \ \mathbb{R}_t \times \partial M, \\ &z(0) = z_0 \in L^2(M), \quad \dot z(0) = 0, \quad v \in L^2_{\mathrm{loc}}(\mathbb{R}; L^2(\partial M)), \end{aligned} \tag{6.1}$$

It is well known that B is an admissible observation operator (cf. Corollary 3.9 in [6]). To ensure existence of a null-control for the wave equation we use the geometrical optics condition of Bardos, Lebeau, and Rauch (specifically Example 1 after Corollary 4.10 in [6]):

(6.2) There is a positive constant L_Γ such that every generalized geodesic of length greater than L_Γ passes through Γ at a nondiffractive point.

Generalized geodesics are the rays of geometrical optics (we refer to [20] for a presentation of this condition with a discussion of its significance). We make the additional assumption that they can be uniquely continued at the boundary ∂M. As in [6], to ensure this, we may assume either that ∂M has no contacts of infinite order with its tangents (e.g., $\partial M = \emptyset$), or that g and ∂M are real analytic. For instance, we recall that (6.2) holds when Γ contains a closed hemisphere of a Euclidean ball M of diameter $L_\Gamma/2$, or when $\Gamma = \partial M$ and M is a strictly convex bounded Euclidean set which does not contain any segment of length L_Γ.

THEOREM 6.1 (after [6]). *If* (6.2) *holds, then the wave equation* (6.1) *is null-controllable in any time greater than* L_Γ.

Thanks to this theorem, Theorem 3.4 implies the following claim.

THEOREM 6.2. *If* (6.2) *holds, then the equation*

$$\begin{aligned} &\partial_t y - e^{i\theta}\Delta y = 0 \ \text{ on } \ \mathbb{R}_t \times M, \quad y = \chi_\Gamma u \ \text{ on } \ \mathbb{R}_t \times \partial M, \\ &y(0) = y_0 \in H^{-1}(M), \quad u \in L^2_{loc}(\mathbb{R}; L^2(\partial M)), \end{aligned}$$

is null-controllable in any time T. *Moreover, the controllability cost* $\kappa_{1,T}$ *(cf. Definition* 2.2*) satisfies (with* α_* *as in Definition* 3.2*)* $\limsup_{T\to 0} T \ln \kappa_{1,T} \leqslant \alpha_* L_\Gamma^2$.

Remark 6.3. Note that Theorem 3.2 in [6] almost proves the converse of Theorem 6.1, thus showing that (6.2) is a sharp condition. Moreover [7, 8] prove that the full equivalence does hold when χ_Γ is replaced by a smooth function with the same zero level set $\partial\Omega \setminus \Gamma$. (This slight difference of model should not be perceptible in applications.) Hence the control transmutation method does not allow us to improve L_Γ in Theorem 6.2.

Remark 6.4. For more results and references on the cost of fast controls for the heat equation, we refer to [15, 22]. It is a known fact that the heat equation is null-controllable for any nonempty control region Γ, and hence the converse to the first statement of this theorem is not true. When M is a bounded Euclidean set and Δ has constant coefficients, [15] proves that the lim sup in the second statement is finite for any interior control region, but does not give an explicit bound. Motivated by the nonlinear heat equation, it considers the heat equation with a time-dependent potential and describes how the cost depends on the norm of this potential, whereas the control transmutation method seems to apply only to time-independent potentials (on which the lim sup above does not depend). As explained in section 2.4 of [22], in the general case without condition (6.2), one could try to adapt Carleman inequalities with phases ϕ to obtain a bound in terms of the following distance function d :

$d(x,y) = \sup\{\phi(y) - \phi(x)\}$, for all x and y in M, where the supremum is taken over all Lipschitz functions $\phi : M \to \mathbb{R}$ with $|\nabla\phi| \leqslant 1$ almost everywhere. (d also has a geometric characterization in terms of path of least action.)

Acknowledgments. References [4, 16] were pointed out to me by a referee.

REFERENCES

[1] O. M. Aamo, A. Smyshlyaev, and M. Krstić, *Boundary control of the linearized Ginzburg–Landau model of vortex shedding*, SIAM J. Control Optim., 43 (2005), pp. 1953–1971.

[2] I. S. Aranson and L. Kramer, *The world of the complex Ginzburg–Landau equation*, Rev. Modern Phys., 74 (2002), pp. 99–143.

[3] W. Arendt, C. J. K. Batty, M. Hieber, and F. Neubrander, *Vector-valued Laplace Transforms and Cauchy Problems*, Monogr. Math. 96, Birkhäuser-Verlag, Basel, 2001.

[4] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials. The Method of Moments in Control Lability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[5] J. M. Ball, *Strongly continuous semigroups, weak solutions, and the variation of constants formula*, Proc. Amer. Math. Soc., 63 (1977), pp. 370–373.

[6] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[7] N. Burq, *Contrôlabilité exacte de l'équation des ondes dans des ouverts peu réguliers*, Asymptotic Anal., 14 (1997), pp. 157–191.

[8] N. Burq and P. Gérard, *Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes.*, C. R. Acad. Sci. Paris Sér. I Math., 25 (1997), pp. 749–752.

[9] O. Cârjă, *The minimal time function in infinite dimensions*, SIAM J. Control Optim., 31 (1993), pp. 1103–1114.

[10] J. Cheeger, M. Gromov, and M. Taylor, *Finite propagation speed, kernel estimates for functions of the Laplace operator, and the geometry of complete Riemannian manifolds*, J. Differential Geom., 17 (1982), pp. 15–53.

[11] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 1, Springer-Verlag, Berlin, 1990.

[12] A. Doubova, A. Osses, and J.-P. Puel, *Exact controllability to trajectories for semilinear heat equations with discontinuous diffusion coefficients*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 621–661.

[13] H. O. Fattorini, *Controllability of higher order linear systems*, in Mathematical Theory of Control (Proc. Conf., Los Angeles, CA, 1967), Academic Press, New York, 1967, pp. 301–311.

[14] H. O. Fattorini, *Second order linear differential equations in Banach spaces*, North-Holland Math. Stud. 108, North–Holland, Amsterdam, 1985.

[15] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[16] E. Fernández-Cara and E. Zuazua, *On the null controllability of the one-dimensional heat equation with nonsmooth coefficients*, Comput. Appl. Math., 21 (2002), pp. 167–190.

[17] A. V. Fursikov and O. Y. Imanuvilov, *Controllability of Evolution Equations*, Lecture Notes Ser. 34, Seoul National University, Research Institute of Mathematics, Global Analysis Research Center, Seoul, 1996.

[18] R. Hersh, *The method of transmutations*, in Partial Differential Equations and Related Topics (Program, Tulane University, New Orleans, LA, 1974), Lecture Notes in Math. 446, Springer, Berlin, 1975, pp. 264–282.

[19] L. I. Ignat and E. Zuazua, *A two-grid approximation scheme for nonlinear Schrödinger equations: Dispersive properties and convergence*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 381–386.

[20] L. Miller, *Escape function conditions for the observation, control, and stabilization of the wave equation*, SIAM J. Control Optim., 41 (2003), pp. 1554–1566.

[21] L. Miller, *Controllability cost of conservative systems: Resolvent condition and transmutation*, J. Funct. Anal., 236 (2006), pp. 592–608.

[22] L. Miller, *Geometric bounds on the growth rate of null-controllability cost for the heat equation in small time*, J. Differential Equations, 204 (2004), pp. 202–226.

[23] L. Miller, *How violent are fast controls for Schrödinger and plates vibrations?*, Arch. Ration. Mech. Anal., 172 (2004), pp. 429–456.
[24] L. Miller, *On the controllability of anomalous diffusions generated by the fractional laplacian*, Math. Control Signals Systems, to appear.
[25] K.-D. Phung, *Observability and control of Schrödinger equations*, SIAM J. Control Optim., 40 (2001), pp. 211–230.
[26] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.
[27] T. I. Seidman, *Exact boundary control for some evolution equations*, SIAM J. Control Optim., 16 (1978), pp. 979–999.
[28] T. I. Seidman, *Two results on exact boundary control of parabolic equations*, Appl. Math. Optim., 11 (1984), pp. 145–152.
[29] T. I. Seidman, S. A. Avdonin, and S. A. Ivanov, *The "window problem" for series of complex exponentials*, J. Fourier Anal. Appl., 6 (2000), pp. 233–254.
[30] T. I. Seidman and J. Yong, *How violent are fast controls?* II, Math. Control Signals Systems, 9 (1996), pp. 327–340.
[31] R. Triggiani, *On the relationship between first and second order controllable systems in Banach spaces*, SIAM J. Control Optim., 16 (1978), pp. 847–859.
[32] K. Tsujioka, *Remarks on controllability of second order evolution equations in Hilbert spaces*, SIAM J. Control, 8 (1970), pp. 90–99.
[33] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 773–789

ZERO-SUM ERGODIC STOCHASTIC GAMES WITH FELLER TRANSITION PROBABILITIES*

ANNA JAŚKIEWICZ† AND ANDRZEJ S. NOWAK‡

Abstract. We consider zero-sum stochastic games with Borel state spaces satisfying a generalized geometric ergodicity condition. The main objective of this paper is to establish a minimax theorem for a class of ergodic stochastic games with the Feller transition probability function. All previous results on ergodic stochastic games are based on a strong continuity of the transition probabilities in the actions of the players.

Key words. discrete-time zero-sum stochastic games, Borel state space, average optimal strategies, optimality equation

AMS subject classifications. 91A10, 91A15, 93D05, 93E15

DOI. 10.1137/S0363012904443257

1. Introduction. This paper deals with zero-sum Borel state space stochastic games under the average payoff criterion. We make certain stochastic stability assumptions on the transition structure of the game which imply the so-called V-uniform geometric ergodicity of the Markov chains governed by the stationary strategies of the players. Such conditions are inspired by the works of Kartashov [17] and Meyn and Tweedie [23, 24] and by some recent results concerning Markov control processes; see [9, 10] and the references cited therein. Other related papers on stochastic games are [11, 14, 15, 20, 21, 29, 30, 38]. A common feature of the aforementioned works is the assumption that the transition probabilities are strongly continuous in the actions of the players. Such a restriction is often not satisfied in applications. Weakly continuous transition probabilities are, in some situations (e.g., dynamic programming problems or stochastic games), much more natural; see [8] for a relevant example.

The aim of this paper is to establish a new minimax theorem for a class of ergodic zero-sum stochastic games with the Borel state space and Feller transition (i.e., weakly continuous) probability function. To the best of our knowledge no results of this type are known in the existing literature. Nevertheless, one can refer to [16, 34] for a related result in the context of dynamic programming (one person games). Furthermore, we also study the so-called optimality or Shapley equation for the games under consideration. We show that stationary strategies obtained from this equation are strong average optimal. The proof makes use of the vanishing discount factor approach. In order to overcome some difficulties (compared with earlier papers) we need to use Michael's theorem on continuous selections [25] and a version of Serfozo's extension of Fatou's lemma for varying probability measures.

A general approach to Borel state space stochastic games was developed by Maitra and Sudderth [22]. They have considered a very large class of payoffs (including the average payoff case) and proved the existence of the game value for general universally

*Received by the editors April 23, 2004; accepted for publication (in revised form) March 17, 2005; published electronically June 21, 2006. This work is supported by MEiN grant 1 P03A 010 30.
http://www.siam.org/journals/sicon/45-3/44325.html

†Institute of Mathematics and Computer Science, Wrocław University of Technology, Wyspiańskiego 27, 50-370 Wrocław, Poland (ajaskiew@im.pwr.wroc.pl).

‡Faculty of Mathematics, Computer Science and Econometrics, University of Zielona Góra, Podgórna 50, 65-246 Zielona Góra, Poland (a.nowak@wmie.uz.zgora.pl).

measurable strategies. Let us note, however, that optimal strategies need not exist in the games considered in [22]. Their assumption on the transition probability function is different from weak continuity. In fact, it is assumed to be strongly continuous, but only in the actions of one player. Therefore, our theorem on the existence of the game value does not follow from Maitra and Sudderth's work [22]. The ergodicity assumptions which are made in our paper are crucial and may be regarded as quite restrictive. However, they allow us to prove the existence of optimal (or even strong optimal) strategies for a very large class of stochastic games. It is worth mentioning that many optimization models and dynamic games in engineering and economics are ergodic [5, 6, 13, 19, 23]. From a mathematical point of view, it is straightforward to construct weakly continuous ergodic transition laws.

The paper is organized as follows. In section 2 we describe the model and basic assumptions. Section 3 contains some auxiliary material concerning generalized limits of sequences of functions and Fatou's lemma. In section 4 we state our main results with the proofs.

2. The model and assumptions. Before we present our model and state the results, we recall some terminology and describe some auxiliary lemmas on continuous selections.

Let Y be a Borel space, i.e., a nonempty Borel subset of a complete separable metric space. By $\mathcal{B}(Y)$, we denote the σ-algebra of all Borel subsets of Y. Let $P(Y)$ be the space of all probability measures on $\mathcal{B}(Y)$, endowed with the weak topology (see [3] or [32]). This topology, for any Borel space Y, can be characterized in terms of convergent sequences (Proposition 7.21 in [3]). Namely, a sequence $\{p_n\}$ converges to some $p \in P(Y)$ in the weak topology if and only if

$$\int_Y u(y)p_n(dy) \to \int_Y u(y)p(dy)$$

for every $u \in C(Y)$. Here $C(Y)$ denotes the space of all bounded continuous functions on Y, endowed with the supremum metric.

For the reader's convenience, we recall some basic results which we shall be using later on. If Y is compact, then $P(Y)$ is compact as well (see Proposition 7.22 in [3]). If Y is Borel, then $P(Y)$ is Borel too (see Corollary 7.25.1 in [3]).

Let X and Y be Borel spaces. By a Borel measurable transition probability from X to Y we mean a function $\phi : \mathcal{B}(Y) \times X \mapsto [0,1]$ such that, for each $B \in \mathcal{B}(Y)$, $\phi(B|\cdot)$ is a Borel measurable function on X, and $\phi(\cdot|x) \in P(Y)$ for each $x \in X$. It is well known that every Borel measurable mapping $g : X \mapsto P(Y)$ induces a transition probability by setting $\phi(\cdot|x) := g(x)(\cdot)$ (see Proposition 7.25 in [3]). If g is continuous (with respect to the weak topology on $P(Y)$), then the corresponding transition probability is called *weakly continuous* or *Feller*.

Let $M : X \mapsto \mathcal{B}(Y)$ be a set-valued mapping. For any $D \subset Y$, define

$$M^{-1}(D) := \{x \in X : M(x) \cap D \neq \emptyset\}.$$

If $M^{-1}(D)$ is a closed (resp., an open) subset in X for each closed (resp., open) subset D of Y, then M is said to be upper (resp., lower) semicontinuous. A mapping $M : X \mapsto \mathcal{B}(Y)$ is called continuous if it is both lower and upper semicontinuous. For a further discussion of semicontinuous set-valued mappings consult [18, 33].

Let $\Phi : X \mapsto \mathcal{B}(P(Y))$ be the set-valued mapping defined by $\Phi(x) := P(M(x))$, $x \in X$.

LEMMA 2.1 (Theorem 3 in [12]). *If M is nonempty compact valued and continuous, then Φ is also continuous and compact valued.*

The next lemma concerns continuous selections for lower semicontinuous set-valued mappings and is essentially due to Michael [25].

LEMMA 2.2. *Assume that M is nonempty compact valued and continuous. Let $x_0 \in X$ and $\nu_0 \in \Phi(x_0)$. Then Φ admits a continuous selection whose graph contains the point (x_0, ν_0), that is, there exists a continuous mapping $\phi : X \mapsto P(Y)$ such that $\phi(x) \in \Phi(x)$ for each $x \in X$ and $\phi(x_0) = \nu_0$.*

Note that $\Psi(x) := \Phi(x)$ for $x \neq x_0$ and $\Psi(x_0) := \{\nu_0\}$ is lower semicontinuous and compact convex valued. The proof of Lemma 2.2 relies on showing that Ψ has a continuous selection and follows along the same lines as that of Theorem 1.5^* in [33], which in turn is a modification of Michael's proof (see [25] or Theorem 8.1.8 in [18]) given for lower semicontinuous set-valued mappings with closed convex values in a Banach space.

A *zero-sum stochastic game* is described by the following objects:

(i) S is the *set of states* for the game and is assumed to be a Borel space.

(ii) A and B are the *action spaces* for players 1 and 2, respectively, and are also assumed to be Borel spaces.

(iii) $A(s) \subset A$ and $B(s) \subset B$ are nonempty compact *sets of actions available* to players 1 and 2, respectively, in state $s \in S$. It is assumed that the set-valued mappings $s \mapsto A(s)$ and $s \mapsto B(s)$ are continuous. Let

$$K = \{(s, a, b) : s \in S,\ a \in A(s),\ \text{and } b \in B(s)\}$$

and

$$\bar{K} = \{(s, \nu, \rho) : s \in S,\ \nu \in P(A(s)),\ \text{and } \rho \in P(B(s))\}.$$

It is obvious that K is a closed subset of $S \times A \times B$. By Lemma 2.1, $\bar{K}$ is a closed subset of the Borel space $S \times P(A) \times P(B)$.

(iv) q is a Borel measurable transition probability from K to S, called the *law of motion among states*. If s is a state at some stage of the game and the players select actions $a \in A(s)$ and $b \in B(s)$, then $q(\cdot|s, a, b)$ is the probability distribution of the next state of the game.

(v) $r : K \mapsto R$ is a Borel measurable *(daily) reward function* for player 1 (*cost function* for player 2).

A *strategy* for player 1 is a sequence $\pi = \{\pi_n\}$, where π_n $(n \geq 0)$ is a transition probability such that $\pi_n(\cdot|h_n) \in P(A)$ for any given history $h_n = (s_0, a_0, b_0, \dots, s_n)$ of the game and $\pi_n(A(s_n)|h_n) = 1$. (Clearly, $h_0 = s_0$.) A strategy for player 2 is defined in a similar way. By Π $[\Gamma]$ we denote the *class of all strategies for player* 1 (resp., *player* 2). Let F be the set of all Borel measurable transition probabilities f from S to $P(A)$ such that $f(s) \in P(A(s))$ for each $s \in S$. A *Markov strategy* π for player 1 is of the form $\pi = \{f_0, f_1, \dots\}$, where $f_n \in F$, $n \geq 0$. A *stationary strategy* π for player 1 is of the form $\pi = \{f, f, \dots\}$, where $f \in F$. Thus, every stationary strategy $\pi = \{f, f, \dots\}$ for player 1 can be identified with the mapping $f \in F$. Similarly, we define the set G of all stationary strategies for player 2.

Let $H = K \times K \times K \times \cdots$ be the space of all infinite histories of the game endowed with the Borel σ-algebra. For any $\pi \in \Pi$ and $\gamma \in \Gamma$, every initial state $s_0 = s \in S$ a probability measure $P_s^{\pi\gamma}$ and a stochastic process $\{s_m, a_m, b_m\}$ are defined on H in a canonical way, where the random variables s_m, a_m, and b_m describe the state and the action chosen by players 1 and 2, respectively, at the mth stage of the game.

A formal construction of the stochastic process and $P_s^{\pi\gamma}$ is based on the Ionescu–Tulcea theorem (see Proposition V.1.1 in [26], [4], or Chapter 7 in [3]). Thus, for each initial state $s_0 = s \in S$, any strategies $\pi \in \Pi$, $\gamma \in \Gamma$, and any finite horizon n, the total *expected n-stage reward* to player 1 is

$$J_n(s,\pi,\gamma) = E_s^{\pi\gamma}\left[\sum_{m=0}^{n-1} r(s_m, a_m, b_m)\right],$$

where $E_s^{\pi\gamma}$ means the expectation operator with respect to the probability measure $P_s^{\pi\gamma}$ with $s = s_0$. (The assumptions imposed on r and q below guarantee that all expectations considered in what follows are well defined.) If β is a fixed real number in $(0,1)$, called the *discount factor*, then the *expected discounted reward* to player 1 is

$$J_\beta(s,\pi,\gamma) = E_s^{\pi\gamma}\left[\sum_{m=0}^{\infty} \beta^m r(s_m, a_m, b_m)\right].$$

The *expected average reward per unit time* to player 1 is defined as

$$J(s,\pi,\gamma) = \liminf_{n\to\infty} \frac{J_n(s,\pi,\gamma)}{n}.$$

For any initial state $s \in S$, define

$$L(s) := \sup_{\pi\in\Pi} \inf_{\gamma\in\Gamma} J(s,\pi,\gamma) \quad \text{and} \quad U(s) := \inf_{\gamma\in\Gamma} \sup_{\pi\in\Pi} J(s,\pi,\gamma). \tag{1}$$

Then L (U) is called the *lower (upper) value*, respectively, in the average payoff stochastic game. It is always true that $L(s) \le U(s)$ for $s \in S$. If $L(s) = U(s)$ for all $s \in S$, then this common function is called the *value* of the stochastic game and is denoted by ξ.

Suppose that the average reward stochastic game has a value ξ. A strategy $\pi^* \in \Pi$ is called *optimal for player* 1 in the average payoff stochastic game iff

$$\inf_{\gamma\in\Gamma} J(s,\pi^*,\gamma) = \xi(s)$$

for all $s \in S$, and a strategy $\gamma^* \in \Gamma$ is called *optimal for player* 2 in the average payoff stochastic game iff

$$\sup_{\pi\in\Pi} J(s,\pi,\gamma^*) = \xi(s)$$

for all $s \in S$. Of course, the value and optimal strategies are defined similarly for the β-discounted and n-stage stochastic games.

We impose the following continuity assumptions.

C1: $r : K \mapsto R$ is continuous.

C2: $q : K \mapsto P(S)$ is weakly continuous.

C3: There exists a continuous function $V : S \mapsto [1,\infty)$ such that $|r(s,a,b)| \le V(s)$ for every $(s,a,b) \in K$.

C4: The function

$$(s,a,b) \mapsto \int_S V(y) q(dy|s,a,b)$$

is continuous on K.

The following assumption has been recently made in the theory of Markov control processes and stochastic games [9, 10, 16, 38]. Some special cases were considered in [11, 14, 15, 30].

C5: (1) There exist a Borel function $\delta : K \mapsto [0,1]$ and a probability measure φ on S such that

$$q(B|s,a,b) \geq \delta(s,a,b)\varphi(B)$$

for any Borel set $B \subset S$ and $s \in S$.

(2) $\int_S \inf_{a\in A(s)} \inf_{b\in B(s)} \delta(s,a,b)\varphi(ds) > 0$.

(3) $\varphi(V) := \int_S V(s)\varphi(ds) < \infty$.

(4) For some $\lambda \in (0,1)$ and every $(s,a,b) \in K$, it holds that

$$\int_S V(y)q(dy|s,a,b) \leq \lambda V(s) + \delta(s,a,b)\varphi(V).$$

By Proposition 7.50 in [3] the function $s \mapsto \inf_{a\in A(s)} \inf_{b\in B(s)} \delta(s,a,b)$ is universally measurable. Therefore, the integral in C5(2) is well defined.

Let $C_V(S)$ denote the subset of all continuous functions on S for which the so-called V-norm

$$\|u\|_V := \sup_{s\in S} \frac{|u(s)|}{V(s)}$$

is finite. The space of Borel measurable functions u on S for which $\|u\|_V < \infty$ is denoted by $L_V^\infty(S)$. Let μ be a finite signed measure on $\mathcal{B}(S)$. The V-norm of μ is defined by

$$\|\mu\|_V := \sup_{\|u\|_V\leq 1} \left| \int_S u(s)\mu(ds) \right| = \int_S V(s)|\mu|(ds),$$

where $|\mu| = \mu^+ + \mu^-$ denotes the total variation of μ, and μ^+, μ^- stand for the positive and negative parts of μ, respectively. From C5(3)–(4), it follows that the integral $\int_S |u(y)|q(dy|s,a,b)$ is finite for all $(s,a,b) \in K$ and each $u \in L_V^\infty(S)$.

Let $s \in S$, $\nu \in P(A(s))$, and $\rho \in P(B(s))$. We define

$$r(s,\nu,\rho) = \int_{A(s)} \int_{B(s)} r(s,a,b)\rho(db)\nu(da),$$

and for any Borel set $D \subset S$, we put

$$q(D|s,\nu,\rho) = \int_{A(s)} \int_{B(s)} q(D|s,a,b)\rho(db)\nu(da).$$

Thus, for any $f \in F$ and $g \in G$, $r(s,f(s),g(s))$ and $q(D|s,f(s),g(s))$ ($s \in S$, $D \in \mathcal{B}(S)$) have clear meanings.

An important consequence of C5 is that for any $f \in F$ and $g \in G$, the state process $\{s_n\}$ is a positive recurrent aperiodic Markov chain with the unique invariant probability measure (also called stationary distribution) denoted by π_{fg}. In addition, $\{s_n\}$ is V-uniformly ergodic; that is, there exist $\theta > 0$ and $\alpha \in (0,1)$ such that

$$\left| \int_S u(y)q^n(dy|s,f(s),g(s)) - \int_S u(y)\pi_{fg}(dy) \right| \leq V(s)\|u\|_V\theta\alpha^n \tag{2}$$

for every $s \in S$, $n \geq 1$ and each $u \in L_V^\infty(S)$. Here $q^n(\cdot|s, f(s), g(s))$ denotes the n-step transition probability induced by q, $f \in F$, and $g \in G$. Clearly, for integers $n \geq 2$ and $D \in \mathcal{B}(S)$, we have

$$q^n(D|s, f(s), g(s)) := \int_S q^{(n-1)}(D|s', f(s'), g(s'))q(ds'|s, f(s), g(s))$$

and $q^1(D|s', f(s'), g(s')) := q(D|s', f(s'), g(s'))$. It is obvious that (2) means that

$$\|q^n(\cdot|s, f(s), g(s)) - \pi_{fg}(\cdot)\|_V \leq V(s)\theta\alpha^n. \tag{3}$$

For a detailed discussion of C5 and its consequences consult Proposition 10.2.5 in [10], Theorem 3.6 in [17], and Proposition 4.4 in [38]. If V is bounded, (3) implies the convergence of the n-step transition probabilities to the invariant measure in the usual total variation norm. See section 16.1 in [23] and Theorem 2.3 in [24] for a further discussion of this issue. It is worth mentioning that the aperiodicity of the Markov chains involved is connected with assumption C5(1). Clearly, allowing for unbounded functions V enables us to consider a much larger class of chains induced by the stationary strategies of the players. On the other hand, one can also consider unbounded payoff functions. A special case of C5, where $\delta(s, a, b) = 1_C(s)$ and $C \subset S$ is a nonempty Borel set, is extensively studied in [23, 24]. The set C is called a "small set" and C5(4) is called the "drift inequality." In interesting examples C is a relatively compact set in S. If C5 is satisfied with such a special function δ, one can observe a "drift" towards the set C no matter which strategies of the players are used. The expected time for reaching the set C from a state $s \in S$ is bounded by $V(s)$ up to a multiplicative constant (see Theorem 14.2.2 in [23]). This property clearly shows that models with unbounded function V are more interesting for practical applications even in the case where the payoff functions are bounded.

From (2), we conclude that

$$J(f, g) := J(s, f, g) = \int_S r(s, f(s), g(s))\pi_{fg}(ds) \tag{4}$$

for every $f \in F$ and $g \in G$; that is, the expected average payoff is independent of the initial state.

Let $u \in L_V^\infty(S)$. Define

$$u'(s, a, b) = r(s, a, b) + \beta \int_S u(y)q(dy|s, a, b)$$

and

$$\bar{u}(s, \nu, \rho) = \int_{A(s)} \int_{B(s)} u'(s, a, b)\rho(db)\nu(da),$$

where $(s, a, b) \in K$ and $(s, \nu, \rho) \in \bar{K}$. We shall need the following auxiliary result.

LEMMA 2.3. *Let* C1–C4 *hold. If u is lower (resp., upper) semicontinuous, then u' (resp., $\bar{u}$) is lower (resp., upper) semicontinuous on K (resp., $\bar{K}$).*

Proof. This proof is standard (see Lemma 3.3(a) in [9] or Lemmas 5.4 and 5.5 in [28]). □

To simplify our notation, we shall use the following (lower) value operators. For each function $u \in L_V^\infty(S)$ and $\beta \in (0, 1]$, we put

$$(T_\beta u)(s) = \sup_{\nu \in P(A(s))} \inf_{\rho \in P(B(s))} \left[r(s, \nu, \rho) + \beta \int_S u(y)q(dy|s, \nu, \rho) \right], \tag{5}$$

where $s \in S$, and we set $Tu = T_\beta u$ when $\beta = 1$. We close this section with some auxiliary results on minimax selections which are closely related to Theorem 5.1 in [28]. We note that r is continuous on K and q is Feller. Moreover, by Lemma 2.1 the set-valued mappings $s \mapsto P(A(s))$ and $s \mapsto P(B(s))$ are continuous. If u is upper semicontinuous, then by Lemma 2.3 and Berge's theorems (see pp. 115–116 in [2]) $T_\beta u$ is upper semicontinuous, and by a minimax selection theorem [27], there exists some $f \in F$ such that

$$\begin{aligned}(T_\beta u)(s) &= \max_{\nu \in P(A(s))} \inf_{\rho \in P(B(s))} \left[r(s,\nu,\rho) + \beta \int_S u(y) q(dy|s,\nu,\rho) \right] \\ &= \inf_{\rho \in P(B(s))} \left[r(s,f(s),\rho) + \beta \int_S u(y) q(dy|s,f(s),\rho) \right] \qquad (6)\end{aligned}$$

for each $s \in S$. Similarly, if u is lower semicontinuous, then by Lemma 2.3 and Berge's theorems, $T_\beta u$ is lower semicontinuous, and by Fan's minimax theorem [7] and a measurable selection theorem [27], there exists some $g \in G$ such that

$$\begin{aligned}(T_\beta u)(s) &= \min_{\rho \in P(B(s))} \sup_{\nu \in P(A(s))} \left[r(s,\nu,\rho) + \beta \int_S u(y) q(dy|s,\nu,\rho) \right] \\ &= \sup_{\nu \in P(A(s))} \left[r(s,\nu,g(s)) + \beta \int_S u(y) q(dy|s,\nu,g(s)) \right] \qquad (7)\end{aligned}$$

for each $s \in S$.

Let $u \in C_V(S)$. Then, by Fan's minimax theorem [7], (6), and (7), $T_\beta u \in C_V(S)$. Moreover, there exist $f \in F$ and $g \in G$ such that

$$\begin{aligned}(T_\beta u)(s) &= \max_{\nu \in P(A(s))} \min_{\rho \in P(B(s))} \left[r(s,\nu,\rho) + \beta \int_S u(y) q(dy|s,\nu,\rho) \right] \\ &= \min_{\rho \in P(B(s))} \max_{\nu \in P(A(s))} \left[r(s,\nu,\rho) + \beta \int_S u(y) q(dy|s,\nu,\rho) \right] \\ &= r(s,f(s),g(s)) + \beta \int_S u(y) q(dy|s,f(s),g(s)) \\ &= \max_{\nu \in P(A(s))} \left[r(s,\nu,g(s)) + \beta \int_S u(y) q(dy|s,\nu,g(s)) \right] \qquad (8) \\ &= \min_{\rho \in P(B(s))} \left[r(s,f(s),\rho) + \beta \int_S u(y) q(dy|s,f(s),\rho) \right] \qquad (9)\end{aligned}$$

for each $s \in S$.

3. Fatou's lemma for varying probability measures. Let $\{w_n\}$ be a sequence of functions in $L_V^\infty(S)$. As in [35], we consider the following "generalized liminf (limsup)":

$$w_*(s) := \inf \left\{ \liminf_{n\to\infty} w_n(s_n) : s_n \to s \right\} \qquad (10)$$

and

$$w^*(s) := \sup \left\{ \limsup_{n\to\infty} w_n(s_n) : s_n \to s \right\}. \qquad (11)$$

LEMMA 3.1. *The function w_* (resp., w^*) is lower (resp., upper) semicontinuous.*

Proof. Let ρ be the metric on S defining the topology. For any $s \in S$ and $r > 0$, put $B(s,r) := \{y : \rho(s,y) \le r\}$. Note that $w_*(s)$ can be written in the following form:

$$w_*(s) = \sup_n \inf_{k \ge n} \left(\inf_{y \in B(s,1/n)} w_k(y) \right).$$

Fix $s \in S$ and assume that $w_*(s) > d$ for some real number d. To prove that $w_*(s)$ is lower semicontinuous at s, we need to show that there exists $\delta > 0$ such that $w_*(y) > d$ for every $y \in B(s,\delta)$. Since

$$\inf_{k \ge n} \left(\inf_{y \in B(s,1/n)} w_k(y) \right) \uparrow w_*(s),$$

there exists an N such that

$$\inf_{k \ge l} \left(\inf_{y \in B(s,1/l)} w_k(y) \right) > d \quad \forall l > N.$$

Let $m > 2l > l > N$ and $\delta := \frac{1}{2l}$. Then we get

$$\begin{aligned} w_*(y) &= \sup_n \inf_{k \ge n} \left(\inf_{z \in B(y,1/n)} w_k(z) \right) \ge \inf_{k \ge m} \left(\inf_{z \in B(y,1/m)} w_k(z) \right) \\ &\ge \inf_{k \ge m} \left(\inf_{z \in B(s,1/l)} w_k(z) \right) \ge \inf_{k \ge l} \left(\inf_{z \in B(s,1/l)} w_k(z) \right) > d. \end{aligned}$$

The proof of the upper semicontinuity of the function $w^*(s)$ at s is similar to the above one with starting point

$$w^*(s) = \inf_n \sup_{k \ge n} \left(\sup_{y \in B(s,1/n)} w_k(y) \right). \qquad \square$$

LEMMA 3.2. *Let $\{\mu_n\} \subset P(S)$ be converging weakly to some $\mu_0 \in P(S)$. If $\{v_n\}$ is a sequence of nonnegative Borel measurable functions on S and v_* is defined as in* (10), *then*

$$\int_S v_*(s)\mu_0(ds) \le \liminf_{n \to \infty} \int_S v_n(s)\mu_n(ds), \tag{12}$$

and if the functions $\{v_n\}$ are nonpositive, then

$$\int_S v^*(s)\mu_0(ds) \ge \limsup_{n \to \infty} \int_S v_n(s)\mu_n(ds) \tag{13}$$

with v^* defined as in (11).

Proof. Inequality (12) was established by Serfozo [35] for the vague convergence of (possibly unboundeded) measures and under an additional assumption that S is a locally compact space. One can note, however, that when the weak convergence of μ_n to μ_0 in $P(S)$ is considered, then $K = S$ can be assumed at the beginning of the proof of Lemma 3.2 in [35] and the result follows. Obviously, (13) can be easily concluded from (12) by taking into account the sequence $\{-v_n\}$. $\square$

LEMMA 3.3. *Assume that* $\{\mu_n\}$ *converges weakly to some* $\mu_0 \in P(S)$ *and* $\{w_n\}$ *is a sequence of functions in* $C_V(S)$ *such that* $\|w_n\|_V \le b$ *for all* n *and some constant* $b > 0$. *If* V *is a continuous function and* $\int_S V(s)\mu_m(ds) < \infty$ *for every* $m \ge 0$ *and*

$$\int_S V(s)\mu_m(ds) \to \int_S V(s)\mu_0(ds) \tag{14}$$

as $m \to \infty$, *then*

$$\int_S w_*(s)\mu_0(ds) \le \liminf_{n\to\infty} \int_S w_n(s)\mu_n(ds) \tag{15}$$

and

$$\int_S w^*(s)\mu_0(ds) \ge \limsup_{n\to\infty} \int_S w_n(s)\mu_n(ds). \tag{16}$$

Proof. Define $v_n(s) := w_n(s) + bV(s)$ and note that $v_n \ge 0$. For any $s \in S$ and arbitrary sequence $s_n \to s$ as $n \to \infty$, we have

$$\liminf_{n\to\infty} v_n(s_n) = bV(s) + \liminf_{n\to\infty} w_n(s_n).$$

Hence $v_*(s) = bV(s) + w_*(s)$, $s \in S$, and consequently

$$\int_S v_*(s)\mu_0(ds) = b\int_S V(s)\mu_0(ds) + \int_S w_*(s)\mu_0(ds).$$

Applying (12) to the sequence $\{v_n\}$ and (14), we easily get

$$\liminf_{n\to\infty} \int_S w_n(s)\mu_n(ds) + b\int_S V(s)\mu_0(ds) = \liminf_{n\to\infty} \int_S v_n(s)\mu_n(ds)$$

$$\ge \int_S v_*(s)\mu_0(ds) = \int_S w_*(s)\mu_0(ds) + b\int_S V(s)\mu_0(ds)$$

which immediately gives (15).

Similarly, (16) can be concluded from (13) by taking $v_n(s) := w_n(s) - bV(s) \le 0$. □

4. Main results. Now we are ready to present our main results. First of all, we state some theorems for the finite horizon and discounted stochastic games. Similar results are given for stochastic games with Feller transition probabilities in [4] and Theorem 4.1 in [28], but with some differences in other assumptions.

THEOREM 4.1. *Assume* (i)–(v), C1–C4, *and* C5(4). *Then, the finite horizon discounted stochastic game has a value and both players have optimal Markov strategies. Moreover, if* ξ_k *is the value function for the* k*-stage game, then* $\xi_k \in C_V(S)$ *and*

$$\xi_{n+1}(s) = (T_\beta \xi_n)(s) \quad \textit{for each} \quad s \in S,\ n \ge 1.$$

Proof. The proof proceeds by induction and makes use of standard dynamic programming arguments, Lemma 2.3, and measurable selections results described in section 2. Optimal Markov strategies $\pi^* = \{f_0, f_1, \dots\}$ and $\gamma^* = \{g_0, g_1, \dots\}$ are

constructed such that for $k \geq 0$, $(f_k(s), g_k(s))$ is a pair of optimal (mixed) strategies in the zero-sum game with the payoff function

$$r(s,a,b) + \beta \int_S \xi_k(y) q(dy|s,a,b),$$

where $a \in A(s)$, $b \in B(s)$, and $\xi_0(y) := 0$ for all $y \in S$. □

THEOREM 4.2. *Assume* (i)–(v), C1–C4, *and* C5(4). *Then the discounted stochastic game has a value* ξ_β *and both players have optimal stationary strategies* $f_\beta \in F$ *and* $g_\beta \in G$. *Moreover,* $\xi_\beta \in C_V(S)$ *and*

$$\begin{aligned} \xi_\beta(s) &= (T_\beta \xi_\beta)(s) = r(s, f_\beta(s), g_\beta(s)) + \beta \int_S \xi_\beta(y) q(dy|s, f_\beta(s), g_\beta(s)) \\ &= \max_{\nu \in P(A(s))} \left[r(s, \nu, g_\beta(s)) + \beta \int_S \xi_\beta(y) q(dy|s, \nu, g_\beta(s)) \right] \\ &= \min_{\rho \in P(B(s))} \left[r(s, f_\beta(s), \rho) + \beta \int_S \xi_\beta(y) q(dy|s, f_\beta(s), \rho) \right] \end{aligned} \tag{17}$$

for every $s \in S$.

Proof. First, we show that T_β has a fixed point for each $\beta \in (0,1)$. By Theorem 4.1, for each $n \geq 0$, we have

$$\begin{aligned} \xi_n(s) &= \min_{\gamma \in \Gamma} \max_{\pi \in \Pi} E_s^{\pi\gamma} \left(\sum_{m=0}^{n-1} \beta^m r(s_m, a_m, b_m) \right) \\ &= \max_{\pi \in \Pi} \min_{\gamma \in \Gamma} E_s^{\pi\gamma} \left(\sum_{m=0}^{n-1} \beta^m r(s_m, a_m, b_m) \right) \end{aligned}$$

and

$$\xi_n(s) = (T_\beta^n \xi_0)(s)$$

for every $s \in S$, $n \geq 1$. From assumption C5, one can easily prove by induction that

$$E_s^{\pi\gamma}(V(s_n)) \leq \lambda^n V(s) + \varphi(V) \sum_{k=0}^{n-1} \lambda^k \leq V(s) + \frac{\varphi(V)}{1-\lambda}. \tag{18}$$

Assume that $n > k$. Note that

$$\begin{aligned} |\xi_n(s) - \xi_k(s)| &\leq \sup_{\pi \in \Pi} \sup_{\gamma \in \Gamma} \left| E_s^{\pi\gamma} \left(\sum_{m=0}^{n-1} \beta^m r(s_m, a_m, b_m) \right) \right. \\ &\qquad \left. - E_s^{\pi\gamma} \left(\sum_{m=0}^{k-1} \beta^m r(s_m, a_m, b_m) \right) \right| \\ &\leq \sup_{\pi \in \Pi} \sup_{\gamma \in \Gamma} E_s^{\pi\gamma} \left(\sum_{m=k}^{n-1} \beta^m \left| r(s_m, a_m, b_m) \right| \right) \\ &\leq \sup_{\pi \in \Pi} \sup_{\gamma \in \Gamma} E_s^{\pi\gamma} \left(\sum_{m=k}^{n-1} \beta^m V(s_m) \right). \end{aligned}$$

Hence from (18), it follows that

$$|\xi_n(s) - \xi_k(s)| \le \sum_{m=k}^{n-1} \beta^k V(s)d \le \frac{\beta^m}{(1-\beta)} V(s)d$$

for some constant $d > 0$. This implies that $\{\xi_n\}$ is a Cauchy sequence in $C_V(S)$. Put

$$\xi_\beta(s) = \lim_{n\to\infty} \xi_n(s), \quad s \in S.$$

Note that $\xi_\beta \in C_V(S)$ and

$$\|\xi_n - \xi_\beta\|_V = \sup_{s\in S} \frac{|\xi_n(s) - \xi_\beta(s)|}{V(s)} \longrightarrow 0 \quad \text{when} \quad n \to \infty. \tag{19}$$

For $n \ge 2$, it holds that

$$\xi_n(s) = (T_\beta^n \xi_0)(s) = (T_\beta \xi_{n-1})(s).$$

It is also true that $T_\beta \xi_{n-1} \to T_\beta \xi_\beta$ when $n \to \infty$. For this, note that

$$\begin{aligned} |(T_\beta \xi_n)(s) - (T_\beta \xi_\beta)(s)| &\le \max_{a\in A(s)} \max_{b\in B(s)} \int_S \frac{|\xi_n(y) - \xi_\beta(y)|}{V(y)} V(y) q(dy|s,a,b) \\ &\le \|\xi_n - \xi_\beta\|_V \max_{a\in A(s)} \max_{b\in B(s)} \int_S V(y) q(dy|s,a,b). \end{aligned}$$

Now from C5(4) and (19), we conclude that $T_\beta \xi_n \to T_\beta \xi_\beta$ as $n \to \infty$. Hence ξ_β is a fixed point of T_β. The existence of f_β and g_β satisfying (17) follows from a measurable selection theorem (compare with (8) and (9)). □

Remark 4.1. The contraction condition made in [4] is different from our assumptions. Similar assumptions are made in [9], where some minimax control processes are studied. The proof of Theorem 4.2 is quite standard, but from a formal point of view the result is new.

THEOREM 4.3. *Assume* (i)–(v) *and* C1–C5. *Then the average payoff stochastic game has a value* ξ*, which is independent of the initial state, and both players have optimal stationary strategies. Moreover,* $\xi = \lim_{\beta\to 1}(1-\beta)\xi_\beta(s^*)$ *for any state* $s^* \in S$.

Proof. By Theorem 4.2 the value ξ_β of the β-discounted game and stationary strategies $f_\beta \in F$, $g_\beta \in G$ exists for every $\beta \in (0,1)$. Fix a state $s^* \in S$ and consider a sequence $\{\beta_n\}$ of discount factors converging to one. Define

$$w_n(s) = \xi_{\beta_n}(s) - \xi_{\beta_n}(s^*), \quad \xi_n = (1-\beta_n)\xi_{\beta_n}(s^*).$$

Then, from (17), it follows that

$$\begin{aligned} \xi_n + w_n(s) &= \min_{\rho\in P(B(s))} \max_{\nu\in P(A(s))} \left[r(s,\nu,\rho) + \beta_n \int_S w_n(t) q(dt|s,\nu,\rho) \right] \\ &= \max_{\nu\in P(A(s))} \left[r(s,\nu,g_{\beta_n}(s)) + \beta_n \int_S w_n(t) q(dt|s,\nu,g_{\beta_n}(s)) \right] \quad s \in S. \end{aligned} \tag{20}$$

By Lemma 7 in [30] (also see the argument given on page 135 in [10]), the sequence $\{\xi_n\}$ is bounded and there is no loss of generality when assuming that ξ_n converges to some real number ξ^* as $n \to \infty$. Moreover, by Lemma 8 in [30] (also see Lemma 10.4.2

in [10]), it follows that $\|w_n\|_V \le b$ for some constant b. Now, we fix a state $s_0 \in S$ and consider an arbitrary sequence of states $\{s_n\}$ such that $s_n \to s_0$ as $n \to \infty$. Then from (20), putting $g_n := g_{\beta_n}$, we obtain

$$\xi_n + w_n(s_n) = \max_{\nu \in P(A(s_n))} \left[r(s_n, \nu, g_n(s_n)) + \beta_n \int_S w_n(y) q(dy|s_n, \nu, g_n(s_n)) \right]. \tag{21}$$

Let $\{n_k\}$ be a subsequence of positive integers for which

$$\liminf_{n\to\infty} w_n(s_n) = \lim_{k\to\infty} w_{n_k}(s_{n_k}).$$

Obviously, $\lim_{k\to\infty} \xi_{n_k} = \xi^*$ and

$$\xi^* + \liminf_{n\to\infty} w_n(s_n) = \liminf_{n\to\infty} [\xi_n + w_n(s_n)] = \lim_{k\to\infty} [\xi_{n_k} + w_{n_k}(s_{n_k})].$$

Consequently, from (21), we obtain

$$\begin{aligned} &\xi^* + \liminf_{n\to\infty} w_n(s_n) \\ &\quad = \lim_{k\to\infty} \max_{\nu \in P(A(s_{n_k}))} \left[r(s_{n_k}, \nu, g_{n_k}(s_{n_k})) + \beta_{n_k} \int_S w_{n_k}(y) q(dy|s_{n_k}, \nu, g_{n_k}(s_{n_k})) \right]. \end{aligned}$$

Let F_C be the set of all continuous stationary strategies for player 1. By Lemmas 2.1 and 2.2, F_C is nonempty. Choose any $f \in F_C$. Then, we have

$$\begin{aligned} &\xi^* + \liminf_{n\to\infty} w_n(s_n) \\ &\quad \ge \liminf_{k\to\infty} \left[r(s_{n_k}, f(s_{n_k}), g_{n_k}(s_{n_k})) + \beta_{n_k} \int_S w_{n_k}(y) q(dy|s_{n_k}, f(s_{n_k}), g_{n_k}(s_{n_k})) \right]. \end{aligned} \tag{22}$$

Note that $Z := \{s_0\} \cup \{s_n\}$ is compact in S. We know that the set-valued mapping $s \mapsto P(B(s))$ is continuous and compact valued. These facts together with Berge's theorems (see [2] or Theorem 7.4.2 in [18]) imply that $\bigcup_{z\in Z} P(B(z))$ is compact in $P(B)$. Therefore, $\{g_{n_k}(s_{n_k})\}$ has a subsequence converging to some $\rho_0 \in P(B)$. Without loss of generality, let $g_{n_k}(s_{n_k}) \to \rho_0$, as $k \to \infty$. By the continuity of $s \mapsto P(B(s))$, $\rho_0 \in P(B(s_0))$. On the other hand, $f(s_{n_k}) \to f(s_0) \in P(A(s_0))$, because $f \in F_C$. Clearly, $q(\cdot|s_{n_k}, f(s_{n_k}), g_{n_k}(s_{n_k})) \to q(\cdot|s_0, f(s_0), \rho_0)$ weakly as $k \to \infty$. By (22) and Lemma 3.3, we infer that

$$\begin{aligned} &\xi^* + \liminf_{n\to\infty} w_n(s_n) \\ &\quad \ge \liminf_{k\to\infty} r(s_{n_k}, f(s_{n_k}), g_{n_k}(s_{n_k})) + \liminf_{k\to\infty} \int_S w_{n_k}(y) q(dy|s_{n_k}, f(s_{n_k}), g_{n_k}(s_{n_k})) \\ &\quad \ge r(s_0, f(s_0), \rho_0) + \liminf_{k\to\infty} \int_S w_{n_k}(y) q(dy|s_{n_k}, f(s_{n_k}), g_{n_k}(s_{n_k})) \\ &\quad \ge r(s_0, f(s_0), \rho_0) + \int_S \bar{w}_*(y) q(dy|s_0, f(s_0), \rho_0), \end{aligned} \tag{23}$$

where $\bar{w}_*$ is the generalized liminf of the sequence $\bar{w}_k = w_{n_k}$. Let $w_*(s_0)$ be the generalized liminf of $\{w_n\}$ defined in (10). Then $w_* \le \bar{w}_*$, and applying this fact to (23) we get

$$\xi^* + \liminf_{n\to\infty} w_n(s_n) \ge r(s_0, f(s_0), \rho_0) + \int_S w_*(y) q(dy|s_0, f(s_0), \rho_0).$$

Furthermore, by the fact that a continuous selector $f \in F_C$ can be chosen in such a way that an arbitrary value from $P(A(s))$ is assigned to the point s_0 (see Lemma 2.2), we infer that

$$\xi^* + \liminf_{n\to\infty} w_n(s_n) \geq \sup_{f\in F_C} \left[r(s_0, f(s_0), \rho_0) + \int_S w_*(y) q(dy|s_0, f(s_0), \rho_0) \right]$$
$$= \sup_{\nu\in P(A(s_0))} \left[r(s_0, \nu, \rho) + \int_S w_*(y) q(dy|s_0, \nu, \rho) \right].$$

Since w_* is lower semicontinuous (see Lemma 3.1), we can write

$$\xi^* + \liminf_{n\to\infty} w_n(s_n) \geq \min_{\rho\in P(B(s_0))} \sup_{\nu\in P(A(s_0))} \left[r(s_0, \nu, \rho) + \int_S w_*(y) q(dy|s_0, \nu, \rho) \right].$$

By the definition of w_* (see (10)), we have

$$\xi^* + w_*(s_0) \geq \min_{\rho\in P(B(s_0))} \sup_{\nu\in P(A(s_0))} \left[r(s_0, \nu, \rho) + \int_S w_*(y) q(dy|s_0, \nu, \rho) \right]. \tag{24}$$

Since s_0 was chosen arbitrarily, then (24) holds with s_0 replaced by any $s \in S$.

We already know that the function w_* is lower semicontinuous. By (24) and a minimax measurable selection theorem (see (7)), there exists some $g^* \in G$ such that

$$\xi^* + w_*(s) \geq r(s, \nu, g^*(s)) + \int_S w_*(y) q(dy|s, \nu, g^*(s)) \tag{25}$$

for every $s \in S$ and $\nu \in P(A(s))$. Iterating (25), one can show in a standard manner (see [10, 30]) that

$$\xi^* \geq \sup_{\pi\in\Pi} J(s, \pi, g^*) \geq U(s) \tag{26}$$

for each $s \in S$ (U is defined in (1)).

Now, let $f_{\beta_n} = f_n$ be a stationary optimal strategy to player 1 in the β_n-discounted stochastic game. We need to prove that $\xi^* \leq L(s)$. From (20) and Fan's minimax theorem [7] (see also (9)), it can be easily seen that

$$\xi_n + w_n(s_n) = \min_{\rho\in P(B(s_n))} \left[r(s_n, f_n(s_n), \rho) + \beta_n \int_S w_n(y) q(dy|s_n, f_n(s_n), \rho) \right].$$

Assume that $s_n \to s_0$ and consider the generalized limsup $w^*(s_0)$ (see (11)). Proceeding along similar lines, we take again a subsequence $\{n_k\}$ of positive integers such that

$$\limsup_{n\to\infty} w_n(s_n) = \lim_{k\to\infty} w_{n_k}(s_{n_k}).$$

Then using Lemma 3.3 and (6), we obtain some $f^* \in F$ such that

$$\xi^* + w^*(s) \leq r(s, f^*(s), \rho) + \int_S w^*(y) q(dy|s, f^*(s), \rho) \tag{27}$$

for every $s \in S$ and $\rho \in P(B(s))$. Again standard dynamic programming arguments, based on (27), show that

$$\xi^* \leq \inf_{\gamma\in\Gamma} J(s, f^*, \gamma) \leq L(s) \tag{28}$$

for each $s \in S$ (L is defined in (1)). By (26) and (28), we have

$$\xi^* = \sup_{\pi\in\Pi} J(s,\pi,g^*) = \inf_{\gamma\in\Gamma} J(s,f^*,\gamma) = J(s,f^*,g^*) = J(f^*,g^*); \tag{29}$$

that is, the game has a value $\xi(s) = \xi^*$, and f^*, g^* are stationary optimal strategies for players 1 and 2, respectively. □

Remark 4.2. Theorem 4.3 is a *first* result on ergodic stochastic games with Feller transition probabilities satisfying fairly general assumptions. All related papers [14, 15, 20, 21, 29, 30, 38] are based on the strong continuity assumption on q that the mapping $(s,a,b) \mapsto q(D|s,a,b)$ is continuous in (a,b) for every Borel subset D of S. The payoffs and transitions in the mentioned papers need not be continuous with respect to the state variable. The situation studied in this paper is somewhat more delicate. Note that to overcome some technical difficulties we have to use Michael's theorem on continuous selections [25], which plays no part in the proofs given in the aforementioned papers. Feller transition probabilities are, in some applications, more natural than strongly continuous ones.

Remark 4.3. A general class of Borel stochastic games with "limsup payoffs" was studied by Maitra and Sudderth [22]. The average payoffs belong to this class. They proved the existence of the value in history-dependent strategies in a very general setup. The payoff functions in [22] are bounded. What is more important for us is their assumption on q that $b \mapsto q(D|s,a,b)$ is continuous for any Borel set $D \subset S$ and (s,a). Clearly, the weak continuity of the transition law is something different and Theorem 4.3 is not a corollary to their work.

Our last result concerns the optimality equation (sometimes called the Shapley equation [38]) for the average payoff stochastic games. It is well known that under our assumption C5 the Markov chain induced by any stationary strategies of the players is ϕ-irreducible. The ϕ-irreducibility with a probability (or σ-finite) *irreducibility measure* ϕ means that if $\phi(B) > 0$ for some set $B \in \mathcal{B}(S)$, then the chance that a Markov chain (starting at any $s \in S$) ever enters B is positive. By Remark 2.1 on page 15 in [31], one can take $\phi = \varphi$ in our setup, but the family of irreducibility measures is infinite, even for the Markov chain induced by a fixed pair of stationary strategies of the players. In particular, note that any nontrivial restriction of an irreducibility measure is still an irreducibility measure and such a restriction can give zero weight to any selected part of the state space. To avoid such a situation it is desirable to introduce the definition of the *maximal irreducibility measure* for the Markov chain under consideration. It is usually denoted by ψ and dominates all possible irreducibility measures of the chain and has some other natural properties. For a formal definition of ψ consult pages 88–89 in [23]. The maximal irreducibility measure describes the range of the chain more completely than any other irreducibility measure which one could construct initially. By Proposition 10.1.2 in [23], ψ (which may depend on $f \in F$, $g \in G$) and π_{fg} are equivalent. Below we make a stronger irreducibility assumption.

IA: The Markov chains induced by any stationary strategies of the players have a common maximal irreducibility measure denoted by ψ.

Usually it is assumed that ψ is σ-finite, but there is no restriction if we consider ψ to be a probability measure [23]. An important consequence of assumption IA is that all the invariant probability measures π_{fg} ($f \in F$, $g \in G$) are equivalent (see Proposition 10.1.2 in [23]). Clearly, IA is a generalization of the irreducible case for countable state space games, where it is assumed that any two states are communicating under any stationary pair of strategies. The maximal irreducibility measure in such a special case is the geometric distribution or counting measure.

THEOREM 4.4. *Let* (i)–(v), C1–C4, *and* IA *hold. Then there exist a constant* ξ^*, *a function* $h \in L_V^\infty(S)$, *and strategies* $f^* \in F$ *and* $g^* \in G$ *such that*

$$\begin{aligned}(30)\quad \xi^* + h(s) &= r(s, f^*(s), g^*(s)) + \int_S h(y) q(dy|s, f^*(s), g^*(s)) \\ &= \max_{\nu \in P(A(s))} \left[r(s, \nu, g^*(s)) + \int_S h(y) q(dy|s, \nu, g^*(s)) \right] \\ &= \min_{\rho \in P(B(s))} \left[r(s, f^*(s), \rho) + \int_S h(y) q(dy|s, f^*(s), \rho) \right] \quad \psi - a.e.\end{aligned}$$

Proof. By Theorem 4.3, there exists a pair (f^*, g^*) of stationary optimal strategies and the value $\xi(s) = \xi^* = J(f^*, g^*)$. Put $h(s) := w_*(s)$, $s \in S$. By (24) and (25), and Fan's minimax theorem for each $s \in S$, we have

$$\begin{aligned}\xi^* + h(s) &\geq \sup_{\nu \in P(A(s))} \left[r(s, \nu, g^*(s)) + \int_S h(y) q(dy|s, \nu, g^*(s)) \right] \\ &= \min_{\rho \in P(B(s))} \sup_{\nu \in P(A(s))} \left[r(s, \nu, \rho) + \int_S h(y) q(dy|s, \nu, \rho) \right] \\ &= \sup_{\nu \in P(A(s))} \min_{\rho \in P(B(s))} \left[r(s, \nu, \rho) + \int_S h(y) q(dy|s, \nu, \rho) \right] \\ &\geq \min_{\rho \in P(B(s))} \left[r(s, f^*(s), \rho) + \int_S h(y) q(dy|s, f^*(s), \rho) \right].\end{aligned}$$

Applying a measurable selection theorem [27], we conclude that there exists some $g_0 \in G$ such that

$$(31)\qquad \xi^* + h(s) \geq R(s, f^*, g_0) := r(s, f^*(s), g_0(s)) + \int_S h(y) q(dy|s, f^*(s), g_0(s))$$

for each $s \in S$. Define

$$D := \{s \in S : \quad \xi^* + h(s) > R(s, f^*, g_0)\}.$$

Assume that $\psi(D) > 0$. Since $\psi \ll \pi_{f^* g_0}$ (see Proposition 10.1.2 in [23]), we have $\pi_{f^* g_0}(D) > 0$. Integrating both sides of (31) with respect to the invariant probability measure $\pi_{f^* g_0}$ and using (4), we get

$$J(f^*, g^*) = \xi^* > \int_S r(y, f^*(y), g_0(y)) \pi_{f^* g_0}(dy) = J(f^*, g_0).$$

This contradicts the fact that f^* is an optimal strategy for player 1. Therefore $\psi(D) = 0$, and (30) follows with $\xi^* = J(f^*, g^*)$ and $h = w_*$. □

Remark 4.4. The optimality equation has relevance to studying strong average optimality in games [10]. The question as to whether (30) holds for every $s \in S$ remains open. It has an affirmative answer when q is strongly continuous in actions (a, b) of the players, i.e., if we assume that the mapping $(a, b) \mapsto q(D|s, a, b)$ is continuous on $A(s) \times B(s)$ for every $s \in S$ and any Borel set $D \subset S$. Let us make this assumption. A simple modification of Lemma 2.1 in [28] shows that there exists a sequence of Borel measurable mappings $\{\phi_n\}$ from S to $A \times B$ such that $A(s) \times B(s) = \mathrm{cl}\{\phi_n(s)\}$, $s \in S$. Define the transition probability Q from S to S by

$$Q(\cdot|s) := \sum_{n=1}^{\infty} \frac{1}{2^n} q(\cdot|s, \phi_n(s)).$$

It is easy to see that Q corresponds to a stationary (correlated) strategy of the players $\phi(s) := \sum_{n=1}^{\infty} \frac{1}{2^n} \delta_{\phi_n(s)}$, $s \in S$, where $\delta_{\phi_n(s)}$ is the probability measure concentrated at the point $\phi_n(s)$. Assuming IA and C5, and using Proposition 4.2.3 in [23], we infer that there exists a Borel absorbing set $S_1 \subset S \setminus D$ (D is defined in the proof of Theorem 4.4) for Q, that is, $Q(S_1|s) = 1$ for every $s \in S_1$. This implies that $q(S_1|s, \phi_n(s)) = 1$ for each $n \geq 1$ and $s \in S_1$. In other words, S_1 is absorbing for every $q(\cdot|s, \phi_n(s))$, $n \geq 1$. Since $\{\phi_n(s)\}$ is dense in $A(s) \times B(s)$ and q is strongly continuous in actions, we obtain $q(S_1|s, a, b) = 1$ for each $a \in A(s)$, $b \in B(s)$, and $s \in S_1$. Consequently, $q(S_1|s, f(s), g(s)) = 1$ for $s \in S_1$ and $f \in F$, $g \in G$. The existence of such an absorbing set is a crucial factor used in the proofs of Theorem in [15] and Theorem 2 in [14], where a sequence of functions converging to the solution of the Shapley equation (for every $s \in S$) is defined. In other words, the results given in [14, 15] for games with strongly continuous transition probabilities are valid under assumption IA. The same procedure leading to a solution of the Shapley equation for every $s \in S$ can be used in the framework of this paper whenever q is strongly continuous in actions.

Remark 4.5. In a recent paper [38], Vega-Amaya proposes an approach based on a contraction mapping, but the value ξ^* must be known in advance. His proof applies to a strongly continuous transition probability q and can be modified to cover the weakly continuous case as well if we additionally assume that the function δ in C5 is *continuous* on K. This assumption excludes a very important case, in which $\delta(s, a, b) = 1_C(s)$, $(s, a, b) \in K$, and C is a Borel (or even a closed or an open) set. If δ is not a continuous function, then the mapping, given by Vega-Amaya, maps continuous functions into Borel (or even universally measurable) ones. Therefore the Banach fixed point theorem cannot be applied.

Remark 4.6. The Markov chains induced by stationary strategies of the players are, under our assumptions, aperiodic. As noted by Tijms [37], the results obtained for aperiodic models can also be applied to periodic ones by a suitable transformation. The unichain assumption which we make in this paper is difficult to relax in general state space models, especially when we wish to have optimal strategies for both players. When the state space is countable some weakening of the assumptions is possible; see [1] and its references.

Acknowledgments. The authors wish to thank four anonymous referees for useful comments.

REFERENCES

[1] E. Altman, A. Hordijk, and F. M. Spieksma, *Contraction conditions for average and α-discount optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 22 (1997), pp. 588–618.
[2] E. Berge, *Topological Spaces*, MacMillan, New York, 1963.
[3] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
[4] H. A. M. Couwenbergh, *Stochastic games with metric state space*, Internat. J. Game Theory, 9 (1980), pp. 25–36.
[5] L. O. Curtat, *Markov equilibria of stochastic games with complementarities*, Games and Econ. Behavior, 17 (1996), pp. 177–199.
[6] D. Duffie, J. Geanakoplos, A. Mas-Colell, and A. McLennan, *Stationary Markov equilibria*, Econometrica, 62 (1994), pp. 745–782.
[7] K. Fan, *Minimax theorems*, Proc. Nat. Acad. Sci. USA, 39 (1953), pp. 42–47.
[8] E. A. Feinberg and M. E. Lewis, *Optimality of four-threshold policies in inventory systems with customer returns and borrowing/storage options*, Probab. Engrg. Inform. Sci., 19 (2005), pp. 45–71.

[9] J. I. González-Trejo, O. Hernández-Lerma, and L. F. Hoyos-Reyes, *Minimax control of discrete-time stochastic systems*, SIAM J. Control Optim., 41 (2003), pp. 1626–1659.
[10] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
[11] O. Hernández-Lerma and J. B. Lasserre, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.
[12] C. J. Himmelberg and F. S. Van Vleck, *Multifunctions with values in a space of probability measures*, J. Math. Anal. Appl., 50 (1975), pp. 108–112.
[13] H. A. Hopenhayn and E. C. Prescott, *Stochastic monotonicity and stationary distributions for dynamic economies*, Econometrica, 60 (1992), pp. 1387–1406.
[14] A. Jaśkiewicz, *Zero-sum semi-Markov games*, SIAM J. Control Optim., 41 (2002), pp. 723–739.
[15] A. Jaśkiewicz and A. S. Nowak, *On the optimality equation for zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 54 (2001), pp. 291–301.
[16] A. Jaśkiewicz and A. S. Nowak, *On the average cost optimality equation for Markov control processes with Feller transition probabilities*, J. Math. Anal. Appl., 316 (2006), pp. 495–509.
[17] N. V. Kartashov, *Strong Stable Markov Chains*, VSP, Utrecht, The Netherlands, 1996.
[18] E. Klein and A. C. Thompson, *Theory of Correspondences*, Wiley, New York, 1984.
[19] H. J. Kushner, *Numerical approximations for stochastic differential games: The ergodic case*, SIAM J. Control Optim., 42 (2004), pp. 1911–1933.
[20] H.-U. Küenle, *Markov games under a geometric drift condition*, in Advances in Dynamic Games, Ann. Internat. Soc. Dynam. Games 7, Birkhäuser Boston, Boston, MA, 2005, pp. 21–38.
[21] H.-U. Küenle and R. Schurath, *The optimality equation and ε-optimal strategies in Markov games with average reward criterion*, Math. Methods Oper. Res., 56 (2003), pp. 439–449.
[22] A. Maitra and W. Sudderth, *Borel stochastic games with limsup payoffs*, Ann. Probab., 21 (1993), pp. 861–885.
[23] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1993.
[24] S. P. Meyn and R. L. Tweedie, *Computable bounds for geometric convergence rates of Markov chains*, Ann. Appl. Probab., 4 (1994), pp. 981–1011.
[25] E. Michael, *Continuous selections* I, Ann. Math., 63 (1956), pp. 361–382.
[26] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.
[27] A. S. Nowak, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466–476.
[28] A. S. Nowak, *Semicontinuous nonstationary stochastic games*, J. Math. Anal. Appl., 117 (1986), pp. 84–99.
[29] A. S. Nowak, *Zero-sum average payoff stochastic games with general state space*, Games and Econ. Behavior, 7 (1994), pp. 221–232.
[30] A. S. Nowak, *Optimal strategies in a class of zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 50 (1999), pp. 399–420.
[31] E. Nummelin, *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press, Cambridge, UK, 1983.
[32] K. R. Parthasarathy, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
[33] D. Repovš and P. V. Semenov, *Continuous Selections of Multivalued Mappings*, Kluwer, Dordrecht, 1998.
[34] M. Schäl, *Average optimality in dynamic programming with general state space*, Math. Oper. Res., 18 (1993), pp. 163–172.
[35] R. Serfozo, *Convergence of Lebesgue integrals with varying measures*, Sankhyã, Ser. A, 44 (1982), pp. 380–402.
[36] L. S. Shapley, *Stochastic games*, Proc. Nat. Acad. Sci. USA, 39 (1953), pp. 1095–1100.
[37] H. C. Tijms, *Average reward optimality equation in Markov decision processes*, in Probability, Statistics and Optimization: A Tribute to Peter Whittle, F. P. Kelly, ed., Wiley, New York, 1994, pp. 485–495.
[38] O. Vega-Amaya, *Zero-sum average semi-Markov games: Fixed-point solutions of the Shapley equation*, SIAM J. Control Optim., 42 (2003), pp. 1876–1894.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 790–828

FEEDBACK BOUNDARY STABILIZATION OF THE TWO-DIMENSIONAL NAVIER–STOKES EQUATIONS*

JEAN-PIERRE RAYMOND†

Abstract. We study the exponential stabilization of the linearized Navier–Stokes equations around an unstable stationary solution, by means of a feedback boundary control, in dimension 2 or 3. The feedback law is determined by solving a linear-quadratic control problem. We do not assume that the normal component of the control is equal to zero. In the nonzero case the state equation, satisfied by the velocity field $\mathbf{y}$, is decoupled into an evolution equation, satisfied by $P\mathbf{y}$, where P is the so-called Helmholtz projection operator, and a quasi-stationary elliptic equation, satisfied by $(I-P)\mathbf{y}$. Using this decomposition, we show that the feedback law can be expressed as a function only of $P\mathbf{y}$. In the two-dimensional case we show that the linear feedback law provides a local exponential stabilization of the Navier–Stokes equations.

Key words. Dirichlet control, feedback control, stabilization, Navier–Stokes equations, Oseen equations, Riccati equation

AMS subject classifications. 93B52, 93C20, 93D15, 35Q30, 76D55, 76D05, 76D07

DOI. 10.1137/050628726

1. Setting of the problem.

Let Ω be a bounded and connected domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ with a regular boundary Γ, $\nu > 0$, and consider a pair $(\mathbf{w}, \chi)$—a velocity field and a pressure—solution to the stationary Navier–Stokes equations in Ω:

$$-\nu\Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{w} + \nabla\chi = \mathbf{f} \quad \text{and} \quad \operatorname{div}\mathbf{w} = 0 \quad \text{in } \Omega, \qquad \mathbf{w} = \mathbf{u}_s^\infty \quad \text{on } \Gamma. \tag{1.1}$$

We assume that $\mathbf{w}$ is regular and is an unstable solution of the instationary Navier–Stokes equations.

The purpose of this paper is to determine a Dirichlet boundary control $\mathbf{u}$, in feedback form, localized in a part of the boundary Γ, so that the corresponding controlled system

$$\begin{aligned} &\frac{\partial\mathbf{y}}{\partial t} - \nu\Delta\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{y} + \nabla p = 0, \quad \operatorname{div}\mathbf{y} = 0 \quad \text{in } Q_\infty, \\ &\mathbf{y} = M\mathbf{u} \quad \text{on } \Sigma_\infty, \quad \mathbf{y}(0) = \mathbf{y}_0 \quad \text{in } \Omega, \end{aligned} \tag{1.2}$$

is stable for initial values $\mathbf{y}_0$ small enough in an appropriate space $\mathbf{X}(\Omega)$. In this setting, $Q_\infty = \Omega\times(0,\infty)$, $\Sigma_\infty = \Gamma\times(0,\infty)$, $\mathbf{X}(\Omega)$ is a subspace of $\mathbf{V}_n^0(\Omega) = \{\mathbf{y}\in \mathbf{L}^2(\Omega) \mid \operatorname{div}\mathbf{y} = 0 \text{ in } \Omega,\ \mathbf{y}\cdot\mathbf{n} = 0 \text{ on } \Gamma\}$, and the operator M is a restriction operator precisely defined in section 2. If we set $(\mathbf{z}, q) = (\mathbf{w}+\mathbf{y}, \chi+p)$ and if $\mathbf{u} = 0$, we see that $(\mathbf{z}, q)$ is the solution to the Navier–Stokes equations

$$\begin{aligned} &\frac{\partial\mathbf{z}}{\partial t} - \nu\Delta\mathbf{z} + (\mathbf{z}\cdot\nabla)\mathbf{z} + \nabla q = \mathbf{f}, \quad \operatorname{div}\mathbf{z} = 0 \quad \text{in } Q_\infty, \\ &\mathbf{z} = \mathbf{u}_s^\infty \ \text{ on } \Sigma_\infty, \quad \mathbf{z}(0) = \mathbf{w}+\mathbf{y}_0 \quad \text{in } \Omega. \end{aligned}$$

*Received by the editors April 7, 2005; accepted for publication (in revised form) January 15, 2006; published electronically June 21, 2006.
http://www.siam.org/journals/sicon/45-3/62872.html

†Laboratoire Mathematics for Industry and Physics (MIP), UMR CNRS 5640, Université Paul Sabatier, 31062 Toulouse Cedex 9, France (raymond@mip.ups-tlse.fr).

Thus $\mathbf{y}_0$ is a perturbation of the stationary solution $\mathbf{w}$.

To study the local feedback stabilization of system (1.2), we first study the feedback stabilization of the corresponding linearized system

$$\begin{aligned} &\frac{\partial \mathbf{y}}{\partial t} - \nu\Delta\mathbf{y} + (\mathbf{w}\cdot\nabla)\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{w} + \nabla p = 0 \quad \text{in } Q_\infty, \\ &\text{div } \mathbf{y} = 0 \quad \text{in } Q_\infty, \quad \mathbf{y} = M\mathbf{u} \quad \text{on } \Sigma_\infty, \quad \mathbf{y}(0) = \mathbf{y}_0 \quad \text{in } \Omega. \end{aligned} \tag{1.3}$$

To stabilize this system we can look for a control $\mathbf{u}$ belonging either to $L^2(0,\infty;\mathbf{V}^0(\Gamma))$ or to $L^2(0,\infty;\mathbf{V}_n^0(\Gamma))$, where

$$\begin{aligned} \mathbf{V}^0(\Gamma) &= \Big\{\mathbf{y}\in\mathbf{L}^2(\Gamma) \mid \langle \mathbf{y}\cdot\mathbf{n}, 1\rangle_{H^{-1/2}(\Gamma),H^{1/2}(\Gamma)} = 0\Big\} \quad \text{and} \\ \mathbf{V}_n^0(\Gamma) &= \Big\{\mathbf{y}\in\mathbf{L}^2(\Gamma) \mid \mathbf{y}\cdot\mathbf{n} = 0 \text{ on } \Gamma\Big\}. \end{aligned}$$

A linear feedback law stabilizing (1.3) in $\mathbf{X}(\Omega) \subset \mathbf{V}_n^0(\Omega)$, with a control $\mathbf{u} \in L^2(0,\infty;\mathbf{V}^0(\Gamma))$, is determined by a bounded operator $K \in \mathcal{L}(\mathbf{X}(\Omega),\mathbf{V}^0(\Gamma))$ such that the solution of the closed loop system

$$\begin{aligned} &\frac{\partial \mathbf{y}}{\partial t} - \nu\Delta\mathbf{y} + (\mathbf{w}\cdot\nabla)\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{w} + \nabla p = 0 \quad \text{in } \Omega\times(0,T), \\ &\text{div } \mathbf{y} = 0 \quad \text{in } \Omega\times(0,T), \quad \mathbf{y} = MK\mathbf{y} \quad \text{on } \Gamma\times(0,T), \quad \mathbf{y}(0) = \mathbf{y}_0 \quad \text{in } \Omega, \end{aligned}$$

obeys

$$|\mathbf{y}(t)|_{\mathbf{X}(\Omega)} \le C|\mathbf{y}_0|_{\mathbf{X}(\Omega)} \quad \text{for all } t\in(0,\infty). \tag{1.4}$$

The feedback control law is given by

$$\mathbf{u}(t) = K\mathbf{y}(t) \quad \text{for all } t\in(0,\infty). \tag{1.5}$$

In place of (1.4) we can look for an exponential decay

$$|\mathbf{y}(t)|_{\mathbf{X}(\Omega)} \le Ce^{-\sigma t}|\mathbf{y}_0|_{\mathbf{X}(\Omega)} \quad \text{for all } t\in(0,\infty), \quad \text{with } \sigma>0. \tag{1.6}$$

Let us underline that a feedback law of the form (1.5) is a pointwise (in time) feedback law. A feedback law may be of a different form, for example of the form

$$\mathbf{u} = L_0\mathbf{y}_0, \qquad L_0 \in \mathcal{L}(\mathbf{X}(\Omega), L^2(0,\infty;\mathbf{V}^0(\Gamma))).$$

In engineering applications, pointwise feedback laws are needed because they are more robust with respect to perturbations in the models. This paper is focused on the characterization of pointwise feedback laws for the Oseen and the Navier–Stokes equations.

Several important questions must be addressed when we look for a pointwise feedback law able to stabilize system (1.3) or (1.2):

(Q_1) Does there exist a control $\mathbf{u}\in L^2(0,\infty;\mathbf{V}^0(\Gamma))$ such that the solution of (1.3) obeys (1.6) or (1.4)? In other words, is the system (1.3) stabilizable?

(Q_2) Assume that K is a pointwise feedback law able to stabilize system (1.3) in $\mathbf{X}(\Omega)$. Does K also stabilize the nonlinear system (1.2) for $|\mathbf{y}_0|_{\mathbf{X}(\Omega)}$ small enough?

(Q_3) Assume that we have proved the existence of a feedback law stabilizing system (1.3). Can we find an equation characterizing K which can be numerically solved by classical methods?

We stop the list here, but lot of other questions are very important in applications, such as the robustness of feedback laws and the numerical accuracy of approximations.

In this paper, we do not address the stabilizability of system (1.3), and we shall use the existing results in the literature. This paper is mainly devoted to (Q_2) and (Q_3). But first we mention some results addressing (Q_1).

When $\mathbf{w} \in \mathbf{L}^\infty(\Omega)$ and $\mathbf{X}(\Omega) = \mathbf{V}_n^0(\Omega) \cap \mathbf{L}^4(\Omega)$, the existence of a boundary control $\mathbf{u} \in L^2(0,\infty;\mathbf{V}^0(\Gamma))$ such that the solution of system (1.3) obeys (1.6) may be deduced, by using an extension method, from exact controllability results with internal controls stated in [8].

In the three-dimensional case, and when $\mathbf{X}(\Omega) = \{\mathbf{y} \in \mathbf{H}^1(\Omega) \mid \text{ div } \mathbf{y} = 0 \text{ in } \Omega\}$, the existence of a feedback control law exponentially stabilizing (1.2) or (1.3) is proved in [12]. However, the feedback operator constructed in [12] is of the form L_0. (It is not a pointwise feedback operator, and it requires the knowledge of the eigenfunctions and the eigenvalues of the Oseen operator of (1.3).)

In the three-dimensional cases, when Ω is simply connected, Barbu, Lasiecka, and Triggiani [4] have proved the stabilizability of system (1.3) in $\mathbf{X}(\Omega) = \{\mathbf{y} \in \mathbf{H}^{1/2+\varepsilon}(\Omega) \mid \text{div } \mathbf{y} = 0 \text{ in } \Omega,\ \mathbf{y}\cdot\mathbf{n} = 0 \text{ on } \Gamma\}$ with controls $\mathbf{u} \in L^2(0,\infty;\mathbf{V}_n^0(\Gamma))$ acting everywhere on the boundary Γ. Some additional results in the two-dimensional case are also stated in [4] (see also section 7.1).

Let us now focus on (Q_2) and (Q_3). One way to address these questions is to use optimal control theory, where the pointwise feedback law is obtained by solving an infinite time horizon control problem of the form

$$(\mathcal{Q}) \qquad \inf\Big\{ J(\mathbf{y},\mathbf{u}) \ \mid \ (\mathbf{y},\mathbf{u}) \text{ satisfies (1.3)},\ \mathbf{u} \in L^2(0,T;\mathbf{U}) \Big\}.$$

More precisely, one has to prove that the value function of problem ($\mathcal{Q}$) obeys

$$\inf(\mathcal{Q}) = \Big(\Pi \mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}_n^0(\Omega)},$$

where $\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega))$, $\Pi = \Pi^* \geq 0$. Next the feedback law is defined thanks to the operator Π. In this framework, question (Q_3) can be reformulated as follows: Does the operator Π satisfy an algebraic Riccati equation in the domain of the Oseen operator?

The answer depends on the choice of J and $\mathbf{U}$ and is not necessarily obvious in the case of a boundary control problem.

According to the stabilizability results mentioned above, we have two possible choices for $\mathbf{U}$:

$$\mathbf{U} = \mathbf{V}_n^0(\Gamma) \quad \text{or} \quad \mathbf{U} = \mathbf{V}^0(\Gamma).$$

Both choices are interesting for applications [9, 10, 11, 13]. Even if $\mathbf{U} = \mathbf{V}^0(\Gamma)$ leads to a slightly complicated analysis, this case is interesting, since the normal velocity is used as a control in many applications [14, 15, 16, 20, 25]. These papers are devoted to the suppression of vortex shedding past a cylinder. That corresponds to the case where $\Omega = \Omega_e \setminus \overline{\Omega_i}$, Ω_i is a bounded domain in $\mathbb{R}^2$ with a regular boundary Γ_i, and Ω_e is another bounded domain in $\mathbb{R}^2$ with a regular boundary Γ_e such that $\overline{\Omega_i} \subset \Omega_e$ (Ω_i is a disk in the case of a circular cylinder). Given a stationary velocity $\mathbf{u}_s^\infty$, we can take $(\mathbf{w},\chi)$ as the solution to the stationary Navier–Stokes equation (1.1) with $\mathbf{f} = 0$. In that case the control has to be localized in a part of the boundary Γ_i, and Γ_e must be far enough from the cylinder so that the solution to the stationary or instationary

Navier–Stokes equations gives a good approximation of the corresponding equation in the exterior domain $\mathbb{R}^2 \setminus \overline{\Omega_i}$. Some heuristic feedback laws have been successfully tested in numerical experiments [25, 20], but in general these feedback laws, which are designed for a very specific value of $\mathbf{u}_s^\infty$, are not robust with respect to perturbations of $\mathbf{u}_s^\infty$, in contrast to feedback laws obtained by solving Riccati equations.

This paper is written in the case when $\mathbf{U} = \mathbf{V}^0(\Gamma)$, and the adaptation to the case when $\mathbf{U} = \mathbf{V}_n^0(\Gamma)$ is given in section 7.1. We would like to underline that the choice of J is critical. Indeed once $\mathbf{X}(\Omega)$, $\mathbf{U}$, and J are chosen, K is generally uniquely determined. Thus we have to choose J so that the corresponding feedback law also stabilizes system (1.2). The functional J is usually of the form

$$J(\mathbf{y}, \mathbf{u}) = \frac{1}{2} \int_0^\infty |C\mathbf{y}(t)|^2_{\mathbf{V}^0(\Omega)}\, dt + \frac{1}{2} \int_0^\infty |\mathbf{u}(t)|^2_{\mathbf{U}}\, dt,$$

where the observation operator C may be a bounded or an unbounded operator in $\mathbf{V}^0(\Omega) = \{\mathbf{y} \in \mathbf{L}^2(\Omega) \mid \operatorname{div} \mathbf{y} = 0\}$ when $\mathbf{U} = \mathbf{V}^0(\Gamma)$ (or in $\mathbf{V}_n^0(\Omega)$ when $\mathbf{U} = \mathbf{V}_n^0(\Gamma)$). If $D(C)$ denotes the domain of C in $\mathbf{V}^0(\Omega)$ (or in $\mathbf{V}_n^0(\Omega)$ when $\mathbf{U} = \mathbf{V}_n^0(\Gamma)$), the solution $\mathbf{y}$ of (1.3) is searched in $L^2(0, \infty; D(C))$. In [4, 2], C is chosen so that $|C\mathbf{y}|_{\mathbf{V}^0(\Omega)}$ is a norm in $\mathbf{V}_n^0(\Omega)$ equivalent to the usual norm of the space $\mathbf{H}^{3/2+\varepsilon}(\Omega)$, for some $\varepsilon > 0$. The idea in [4, 2] is to choose the operator C so that the norm $|C\mathbf{y}|_{\mathbf{V}^0(\Omega)}$ is strong enough to dominate the nonlinearity of the Navier–Stokes equations. In that way a Lyapunov function for the nonlinear closed loop system can be defined thanks to Π. The price to pay is that the operator Π, corresponding to this problem, does not satisfy a Riccati equation in the domain of the Oseen operator. An algebraic Riccati equation is stated only in $D((A_\Pi)^2)$ [4, section 4.5], where A_Π is the infinitesimal generator of the associated closed loop system. Thus the domain in which the Riccati equation is stated depends on the unknown Π of the equation. This is a serious drawback.

Here we follow a different approach. We choose J as follows:

$$J(\mathbf{y}, \mathbf{u}) = \frac{1}{2} \int_0^T \int_\Omega |\mathbf{y}|^2\, dxdt + \frac{1}{2} \int_0^T \int_\Gamma |\mathbf{u}|^2\, dxdt. \tag{1.7}$$

One difficulty comes from the fact that, choosing $\mathbf{U} = \mathbf{V}^0(\Gamma)$, the system (1.3) cannot be written in the form of a classical evolution equation. We have shown in [26] that (1.3) can be rewritten in the form

$$\begin{aligned} &P\mathbf{y}' = AP\mathbf{y} + BM\mathbf{u}, \qquad \mathbf{y}(0) = \mathbf{y}_0, \\ &(I - P)\mathbf{y} = (I - P)D_A\gamma_n M\mathbf{u}, \end{aligned} \tag{1.8}$$

where A is the corresponding Oseen operator, B is a boundary control operator, P is the so-called Helmholtz or Leray projector, and D_A the Dirichlet operator associated with A. For the precise definitions and properties of these operators we refer to section 2. Writing system (1.3) in the form (1.8), we show that the control problem $(\mathcal{Q})$ can be rewritten in the form of another control problem in which the state variable is $P\mathbf{y}$ and not $\mathbf{y}$. This transformation is essential in our approach. It leads to a Riccati equation which is the natural one for the new control problem, but which is not the expected one if we consider only problem $(\mathcal{Q})$. Moreover, the Riccati equation is satisfied in the domain of A.

Now the challenge is to prove that the linear feedback law determined in this way also stabilizes the nonlinear equation (1.2). This is not obvious, because the norm

of $\mathbf{y}$ involved in J is too weak to dominate the nonlinearity of the Navier–Stokes equations. Thus we cannot follow the Lyapunov function approach as in [2] and [4]; we develop a completely different approach. We study the regularizing properties of the operator Π. Thanks to these results—which, to the best of our knowledge, are new even in the case of other parabolic equations like the heat equation—we are able to establish a local feedback stabilization result for the Navier–Stokes equation for initial data $\mathbf{y}_0$ small enough in the space $\mathbf{V}_n^{1/2-\varepsilon}(\Omega) = \mathbf{V}_n^0(\Omega) \cap \mathbf{H}^{1/2-\varepsilon}(\Omega)$, for all $0 < \varepsilon < 1/4$. We show that the solution to the closed loop nonlinear system obeys an exponential decay in the corresponding space $\mathbf{V}^{1/2-\varepsilon}(\Omega) = \mathbf{V}^0(\Omega) \cap \mathbf{H}^{1/2-\varepsilon}(\Omega)$.

Finally, we note that in the three-dimensional case, the stabilization of the nonlinear problem cannot be treated with the same tools and requires a more delicate analysis [27, 28].

The paper is organized as follows. The Oseen operator and some associated boundary control operators are studied in section 2. Optimality conditions for the finite horizon control problem of the Oseen equations are established in section 3. We study the corresponding infinite time horizon control problem in section 4. We show that the optimal solution of this problem is characterized by an optimality system. This kind of characterization, which is known in the case of bounded control operators [1, 3] is to the best of our knowledge new in the case of unbounded operators. Thanks to this optimality system, we are able to study the regularity properties of the feedback operator. In order to study the nonlinear problem, we first study in section 5 a nonhomogeneous linear-quadratic control problem. The nonhomogeneous term will play the role of the nonlinearity in the next section. The local stabilization of the Navier–Stokes equations is studied in section 6. Some additional results are stated in section 7. We have collected some regularity results needed throughout the paper in an appendix.

2. Functional framework.

2.1. Notation and assumptions. Let us introduce the following spaces: $H^s(\Omega;\mathbb{R}^N) = \mathbf{H}^s(\Omega)$, $L^2(\Omega;\mathbb{R}^N) = \mathbf{L}^2(\Omega)$; the same notation conventions will be used for trace spaces and for the spaces $H_0^s(\Omega;\mathbb{R}^N)$. We also introduce different spaces of free divergence functions and some corresponding trace spaces:

$$\mathbf{V}^s(\Omega) = \left\{\mathbf{y} \in \mathbf{H}^s(\Omega) \mid \text{ div } \mathbf{y} = 0 \text{ in } \Omega, \ \langle \mathbf{y}\cdot\mathbf{n}, 1\rangle_{H^{-1/2}(\Gamma),H^{1/2}(\Gamma)} = 0\right\} \quad \text{for } s \geq 0,$$

$$\mathbf{V}_n^s(\Omega) = \left\{\mathbf{y} \in \mathbf{H}^s(\Omega) \mid \text{ div } \mathbf{y} = 0 \text{ in } \Omega, \ \mathbf{y}\cdot\mathbf{n} = 0 \text{ on } \Gamma\right\} \quad \text{for } s \geq 0,$$

$$\mathbf{V}_0^s(\Omega) = \left\{\mathbf{y} \in \mathbf{H}^s(\Omega) \mid \text{ div } \mathbf{y} = 0 \text{ in } \Omega, \ \mathbf{y} = 0 \text{ on } \Gamma\right\} \quad \text{for } s > 1/2,$$

$$\mathbf{V}^s(\Gamma) = \left\{\mathbf{y} \in \mathbf{H}^s(\Gamma) \mid \langle \mathbf{y}\cdot\mathbf{n}, 1\rangle_{H^{-1/2}(\Gamma),H^{1/2}(\Gamma)} = 0\right\} \quad \text{for } s \geq -1/2\,.$$

In the above setting, $\mathbf{n}$ denotes the unit normal to Γ outward Ω. We shall use the notation $Q_T = \Omega \times (0,T)$, $\Sigma_T = \Gamma \times (0,T)$, $Q_{\bar{t},T} = \Omega \times (\bar{t},T)$, and $\Sigma_{\bar{t},T} = \Gamma \times (\bar{t},T)$ for $\bar{t} > 0$, and $0 < T \leq \infty$. For spaces of time dependent functions we set

$$\mathbf{V}^{s,\sigma}(Q_T) = H^\sigma(0,T;\mathbf{V}^0(\Omega)) \cap L^2(0,T;\mathbf{V}^s(\Omega))$$

and

$$\mathbf{V}^{s,\sigma}(\Sigma_T) = H^\sigma(0,T;\mathbf{V}^0(\Gamma)) \cap L^2(0,T;\mathbf{V}^s(\Gamma)).$$

We assume that Ω is of class C^4 and $\mathbf{w} \in \mathbf{V}^3(\Omega)$.

In order to find a control $\mathbf{u}$, supported in an open subset Γ_c of Γ, we introduce a weight function $m \in C^2(\Gamma)$ with values in $[0,1]$, with support in Γ_c, equal to 1 in Γ_0, where Γ_0 is an open subset in Γ_c. Associated with this function m we introduce the operator $M \in \mathcal{L}(\mathbf{V}^0(\Gamma))$ defined by

$$M\mathbf{u}(x) = m(x)\mathbf{u}(x) - \frac{m}{\int_\Gamma m}\left(\int_\Gamma m\mathbf{u}\cdot\mathbf{n}\right)\mathbf{n}(x),$$

where $|\Gamma|$ is the $(N-1)$-dimensional Lebesgue measure of Γ. In this way, we can replace the condition $\mathrm{supp}(\mathbf{u}) \subset \Gamma_c$ by considering a boundary condition of the form

$$\mathbf{z} - \mathbf{w} = M\mathbf{u} \quad \text{on} \quad \Sigma_\infty.$$

The main interest of this operator M is that if $\mathbf{u} \in L^2(0,\infty;H^s(\Gamma_c;\mathbb{R}^N)) \cap H^{s/2}(0,\infty;L^2(\Gamma_c;\mathbb{R}^N))$ for some $0 < s \le 2$, and if $\tilde{\mathbf{u}}$ denotes the extension of $\mathbf{u}$ by zero to $\Sigma_\infty \backslash \big(\Gamma_c \times (0,\infty)\big)$, then $M\tilde{\mathbf{u}}$ belongs to $L^2(0,\infty;H^s(\Gamma;\mathbb{R}^N)) \cap H^{s/2}(0,\infty;L^2(\Gamma;\mathbb{R}^N))$, which is not true for $\tilde{\mathbf{u}}$.

For all $\psi \in H^{1/2+\varepsilon'}(\Omega)$, with $\varepsilon' > 0$, we denote by $c(\psi)$ and $c(m\psi)$ the constants defined by

$$c(\psi) = \frac{1}{|\Gamma|}\int_\Gamma \psi \quad \text{and} \quad c(m\psi) = \frac{1}{|\Gamma|}\int_\Gamma m\psi. \tag{2.1}$$

2.2. Properties of some operators. In the following we consider the linearized Navier–Stokes equation

$$\begin{aligned} &\frac{\partial \mathbf{y}}{\partial t} - \nu\Delta\mathbf{y} + (\mathbf{w}\cdot\nabla)\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{w} + \nabla p = 0 \quad \text{in } Q_T, \\ &\operatorname{div}\mathbf{y} = 0 \quad \text{in } Q_T, \quad \mathbf{y} = M\mathbf{u} \quad \text{on } \Sigma_T, \quad \mathbf{y}(0) = \mathbf{y}_0 \quad \text{in } \Omega, \end{aligned} \tag{2.2}$$

and the adjoint equation

$$\begin{aligned} &-\frac{\partial \mathbf{\Phi}}{\partial t} - \nu\Delta\mathbf{\Phi} - (\mathbf{w}\cdot\nabla)\mathbf{\Phi} + (\nabla\mathbf{w})^T\mathbf{\Phi} + \nabla\psi = \mathbf{y} \quad \text{in } Q_T, \\ &\operatorname{div}\mathbf{\Phi} = 0 \quad \text{in } Q_T, \quad \mathbf{\Phi} = 0 \quad \text{on } \Sigma_T, \quad \mathbf{\Phi}(T) = 0 \quad \text{in } \Omega, \end{aligned} \tag{2.3}$$

where T is finite or infinite. To study these equations, we introduce the Stokes and the Oseen operators associated with (2.2) and (2.3). Let P be the orthogonal projector in $\mathbf{L}^2(\Omega)$ onto $\mathbf{V}_n^0(\Omega)$, and denote by $(A_0, D(A_0))$, $(A, D(A))$, and $(A^*, D(A^*))$ the unbounded operators in $\mathbf{V}_n^0(\Omega)$ defined by

$$\begin{aligned} &D(A_0) = \mathbf{H}^2(\Omega) \cap \mathbf{V}_0^1(\Omega), \quad A_0\mathbf{y} = P\Delta\mathbf{y} \quad \text{for all } \mathbf{y} \in D(A_0), \\ &D(A) = \mathbf{H}^2(\Omega) \cap \mathbf{V}_0^1(\Omega), \quad A\mathbf{y} = \nu P\Delta\mathbf{y} - P((\mathbf{w}\cdot\nabla)\mathbf{y}) - P((\mathbf{y}\cdot\nabla)\mathbf{w}), \\ &D(A^*) = \mathbf{H}^2(\Omega) \cap \mathbf{V}_0^1(\Omega), \quad A^*\mathbf{y} = \nu P\Delta\mathbf{y} + P((\mathbf{w}\cdot\nabla)\mathbf{y}) - P((\nabla\mathbf{w})^T\mathbf{y}). \end{aligned}$$

Throughout the following we denote by $\lambda_0 > 0$ an element in the resolvent set of A satisfying

$$\begin{aligned} &\big((\lambda_0 I - A)\mathbf{y}, \mathbf{y}\big)_{\mathbf{V}_n^0(\Omega)} \ge \omega_0 |\mathbf{y}|^2_{\mathbf{V}_0^1(\Omega)} \quad \text{for all } \mathbf{y} \in D(A) \\ &\text{and} \\ &\big((\lambda_0 I - A^*)\mathbf{y}, \mathbf{y}\big)_{\mathbf{V}_n^0(\Omega)} \ge \omega_0 |\mathbf{y}|^2_{\mathbf{V}_0^1(\Omega)} \quad \text{for all } \mathbf{y} \in D(A^*) \end{aligned} \tag{2.4}$$

for some $0 < \omega_0 < 1$.

THEOREM 2.1. *The unbounded operator $(A-\lambda_0 I)$ (respectively, $(A^*-\lambda_0 I)$) with domain $D(A-\lambda_0 I)=D(A)$ (respectively, $D(A^*-\lambda_0 I)$) is the infinitesimal generator of a bounded analytic semigroup on $\mathbf{V}_n^0(\Omega)$. Moreover, for all $0\le\alpha\le 1$, we have*

$$D((\lambda_0 I-A)^\alpha)=D((\lambda_0 I-A^*)^\alpha)=D((\lambda_0 I-\nu A_0)^\alpha)=D((-A_0)^\alpha)\,.$$

Proof. Under condition (2.4) the analyticity of the semigroup generated by $(A-\lambda_0 I)$ is well known (see [6, Chapter 1, Theorem 2.1.2]). The characterization of the domains of fractional powers of $(\lambda_0 I-A)$ and $(\lambda_0 I-A^*)$ may be deduced from [24]. □

Observe that the semigroups $(e^{t(A-\lambda_0 I)})_{t\ge 0}$ and $(e^{t(A^*-\lambda_0 I)})_{t\ge 0}$ are exponentially stable on $\mathbf{V}_n^0(\Omega)$ and that

$$\|e^{t(A-\lambda_0 I)}\|_{\mathcal{L}(\mathbf{V}_n^0(\Omega))}\le e^{-\omega t}\quad\text{and}\quad\|e^{t(A^*-\lambda_0 I)}\|_{\mathcal{L}(\mathbf{V}_n^0(\Omega))}\le e^{-\omega t}$$

for all $\omega<\omega_0$ (see [6, Chapter 1, Theorem 2.12]).

LEMMA 2.2. *If $\mathbf{\Phi}$ belongs to $\mathbf{V}^{3/2+\varepsilon}(\Omega)\cap\mathbf{V}_0^1(\Omega)$ for some $\varepsilon>0$, then $\frac{\partial\mathbf{\Phi}}{\partial\mathbf{n}}\cdot\mathbf{n}=0$ on Γ.*

Proof. This result is proved in [4, Lemma 3.3.1] for functions $\mathbf{\Phi}$ belonging to $\mathbf{V}_0^1(\Omega)\cap(C^1(\overline{\Omega}))^N$. Since $\mathbf{V}_0^1(\Omega)\cap(C^2(\overline{\Omega}))^N$ is dense in $\mathbf{V}^{3/2+\varepsilon}(\Omega)\cap\mathbf{V}_0^1(\Omega)$, the result is still true for functions in $\mathbf{V}^{3/2+\varepsilon}(\Omega)\cap\mathbf{V}_0^1(\Omega)$ when $\varepsilon>0$. □

Let us introduce D_A and D_p, two Dirichlet operators associated with A, defined as follows. For $\mathbf{u}\in\mathbf{V}^0(\Gamma)$, set $D_A\mathbf{u}=\mathbf{y}$ and $D_p\mathbf{u}=q$, where $(\mathbf{y},q)$ is the unique solution in $\mathbf{V}^{1/2}(\Omega)\times(H^{1/2}(\Omega)/\mathbb{R})'$ to the equation

$$\begin{aligned}&\lambda_0\mathbf{y}-\nu\Delta\mathbf{y}+(\mathbf{w}\cdot\nabla)\mathbf{y}+(\mathbf{y}\cdot\nabla)\mathbf{w}+\nabla q=0\quad\text{in }\Omega,\\&\operatorname{div}\mathbf{y}=0\quad\text{in }\Omega,\quad\mathbf{y}=\mathbf{u}\quad\text{on }\Gamma.\end{aligned}$$

LEMMA 2.3. (i) *The operator D_A is a bounded operator from $\mathbf{V}^0(\Gamma)$ into $\mathbf{V}^0(\Omega)$; moreover, it satisfies*

$$\|D_A\mathbf{u}\|_{\mathbf{V}^{s+1/2}(\Omega)}\le C(s)\|\mathbf{u}\|_{\mathbf{V}^s(\Omega)}\qquad\textit{for all } 0\le s\le 2\,.$$

(ii) *The operator $D_A^*\in\mathcal{L}(\mathbf{V}^0(\Omega),\mathbf{V}^0(\Gamma))$, the adjoint operator of $D_A\in\mathcal{L}(\mathbf{V}^0(\Gamma),\mathbf{V}^0(\Omega))$, is defined by*

$$D_A^*\mathbf{g}=-\nu\frac{\partial\mathbf{z}}{\partial\mathbf{n}}+\pi\mathbf{n}-c(\pi)\mathbf{n},\tag{2.5}$$

where $(\mathbf{z},\pi)$ is the solution of

$$\lambda_0\mathbf{z}-\nu\Delta\mathbf{z}-(\mathbf{w}\cdot\nabla)\mathbf{z}+(\nabla\mathbf{w})^T\mathbf{z}+\nabla\pi=\mathbf{g}\ \textit{ and }\ \operatorname{div}\mathbf{z}=0\ \textit{ in }\Omega,\ \ \mathbf{z}=0\ \textit{ on }\Gamma\tag{2.6}$$

and $c(\pi)$ is defined by (2.1).

Proof. Part (i) is well known when $\mathbf{w}=0$ (see, e.g., [31]). Its adaptation in the case when $\mathbf{w}\in\mathbf{V}^3(\Omega)$ is given in [26, Corollary 7.1]. Part (ii) is stated in [26, Lemma 7.4]. □

Let us define the operators $\gamma_\tau\in\mathcal{L}(\mathbf{V}^0(\Gamma))$ and $\gamma_n\in\mathcal{L}(\mathbf{V}^0(\Gamma))$ by

$$\gamma_\tau\mathbf{u}=\mathbf{u}-\big(\mathbf{u}\cdot\mathbf{n}\big)\mathbf{n}\quad\text{and}\quad\gamma_n\mathbf{u}=\big(\mathbf{u}\cdot\mathbf{n}\big)\mathbf{n}=\mathbf{u}-\gamma_\tau\mathbf{u}\qquad\text{for all }\mathbf{u}\in\mathbf{V}^0(\Gamma)\,.$$

Let us also denote by P_Γ the projector in $\mathbf{L}^2(\Gamma)$ onto $\mathbf{V}^0(\Gamma)$ defined by $P_\Gamma \mathbf{u} = \mathbf{u} - \frac{m}{\int_\Gamma m}\big(\int_\Gamma \mathbf{u}\cdot\mathbf{n}\big)\mathbf{n}$. Observe that $M = P_\Gamma m$, where m denotes the multiplication operator by the function m. Introducing the spaces

$$\mathbf{V}^0_\tau(\Gamma) = \Big\{\mathbf{u} \in \mathbf{V}^0(\Gamma) \mid \gamma_\tau \mathbf{u} = 0\Big\} \qquad \text{and} \qquad \mathbf{V}^0_n(\Gamma) = \Big\{\mathbf{u} \in \mathbf{V}^0(\Gamma) \mid \gamma_n \mathbf{u} = 0\Big\},$$

we have $\mathbf{V}^0(\Gamma) = \mathbf{V}^0_\tau(\Gamma) \oplus \mathbf{V}^0_n(\Gamma)$.

LEMMA 2.4. *The operator M obeys the following properties:*

$$M = M^*, \qquad M\gamma_\tau = \gamma_\tau M = m\gamma_\tau, \quad \textit{and} \quad M\gamma_n = \gamma_n M\,.$$

The operators γ_τ and γ_n satisfy

$$\gamma_\tau = \gamma_\tau^*, \qquad \gamma_n = \gamma_n^*, \quad \textit{and} \quad (I-P)D_A = (I-P)D_A\gamma_n.$$

Proof. From the definition of M it follows that

$$\Big(M\mathbf{u}, \mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)} = \int_\Gamma m\mathbf{u}\cdot\mathbf{v} = \int_\Gamma \mathbf{u}\cdot P_\Gamma m\mathbf{v} = \Big(\mathbf{u}, M\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)},$$

for all $\mathbf{u} \in \mathbf{V}^0(\Gamma)$ and all $\mathbf{v} \in \mathbf{V}^0(\Gamma)$. Thus $M = M^*$. Observe that $P_\Gamma\gamma_\tau = \gamma_\tau = \gamma_\tau P_\Gamma$ and $m\gamma_\tau = \gamma_\tau m$. Thus we have

$$M\gamma_\tau = P_\Gamma m\gamma_\tau = m\gamma_\tau = \gamma_\tau m = \gamma_\tau M.$$

Thus the second identity is established. Since $\gamma_\tau + \gamma_n = I$, where I is the identity operator in $\mathbf{V}^0(\Gamma)$, we have $M(\gamma_\tau + \gamma_n) = (\gamma_\tau + \gamma_n)M$, and the identity $M\gamma_n = \gamma_n M$ follows from the equality $M\gamma_\tau = \gamma_\tau M$. The first two identities satisfied by γ_τ and γ_n are obvious. Moreover, if $\gamma_n\mathbf{u} = 0$, then $D_A\mathbf{u} \in \mathbf{V}^0_n(\Omega)$ and $(I-P)D_A\mathbf{u} = 0$. This proves the last identity. ∎

In the next lemma we study the properties of an operator R_A, which plays a crucial role in optimality conditions of control problems that we consider.

LEMMA 2.5. *The operator*

$$R_A = MD_A^*(I-P)D_AM + I$$

is an isomorphism from $\mathbf{V}^0(\Gamma)$ into itself. Moreover, for all $0 \le s \le 3/2$, its restriction to $\mathbf{V}^s(\Gamma)$ is an isomorphism from $\mathbf{V}^s(\Gamma)$ into itself. In addition, R_A satisfies

$$R_A\gamma_n = \gamma_n R_A\gamma_n \quad \textit{and} \quad R_A\gamma_\tau = \gamma_\tau R_A\gamma_\tau = \gamma_\tau.$$

The restriction of R_A to $\mathbf{V}^0_\tau(\Gamma)$ is an isomorphism from $\mathbf{V}^0_\tau(\Gamma)$ into itself, and we have

$$R_A^{-1}\mathbf{u} = (\gamma_n R_A\gamma_n)^{-1}\mathbf{u} = \gamma_n R_A^{-1}\gamma_n\mathbf{u} = \gamma_n R_A^{-1}\mathbf{u} \quad \textit{for all } \mathbf{u} \in \mathbf{V}^0_\tau(\Gamma)\,.$$

Proof. Let s belong to $[0, 3/2]$. The operator D_A is continuous from $\mathbf{V}^s(\Gamma)$ into $\mathbf{V}^{s+1/2}(\Omega)$, $(I-P)$ is bounded from $\mathbf{V}^{s+1/2}(\Omega)$ into itself, and D_A^* is bounded from $\mathbf{V}^{s+1/2}(\Omega)$ into $\mathbf{V}^{s+1}(\Gamma)$. Thus $MD_A^*(I-P)D_AM$ is bounded from $\mathbf{V}^s(\Gamma)$ into $\mathbf{V}^{s+1}(\Gamma)$, and compact in $\mathbf{V}^s(\Gamma)$. Let us show that the kernel of R_A is reduced to 0. If $\mathbf{u} \in \mathbf{V}^0(\Gamma)$ obeys $R_A\mathbf{u} = 0$, then we have

$$\begin{aligned} 0 &= \Big(R_A\mathbf{u}, \mathbf{u}\Big)_{\mathbf{V}^0(\Gamma)} = \Big((I-P)D_AM\mathbf{u}, (I-P)D_AM\mathbf{u}\Big)_{\mathbf{V}^0(\Omega)} + \Big(\mathbf{u}, \mathbf{u}\Big)_{\mathbf{V}^0(\Gamma)} \\ &\ge \Big(\mathbf{u}, \mathbf{u}\Big)_{\mathbf{V}^0(\Gamma)}. \end{aligned}$$

Thus $\mathbf{u} = 0$, which shows that R_A is injective. From the Fredholm alternative it follows that R_A is an isomorphism from $\mathbf{V}^s(\Gamma)$ into itself.

With the identities stated in Lemma 2.4, we can write

$$\Big(R_A\gamma_n\mathbf{u},\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)} = \Big((I-P)D_AM\gamma_n\mathbf{u},(I-P)D_A\gamma_nM\mathbf{v}\Big)_{\mathbf{V}^0(\Omega)} + \Big(\gamma_n\mathbf{u},\gamma_n\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)}$$
$$= \Big((I-P)D_AM\gamma_n\mathbf{u},(I-P)D_AM\gamma_n\mathbf{v}\Big)_{\mathbf{V}^0(\Omega)} + \Big(\gamma_n\mathbf{u},\gamma_n\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)} = \Big(R_A\gamma_n\mathbf{u},\gamma_n\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)}$$

for all $\mathbf{u}\in\mathbf{V}^0(\Gamma)$ and all $\mathbf{v}\in\mathbf{V}^0(\Gamma)$. Thus $R_A\gamma_n = \gamma_nR_A\gamma_n$.

Similarly, we have

$$\Big(R_A\gamma_\tau\mathbf{u},\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)} = \Big(\gamma_\tau\mathbf{u},\mathbf{v}\Big)_{\mathbf{V}^0(\Gamma)}$$

for all $\mathbf{u}\in\mathbf{V}^0(\Gamma)$ and all $\mathbf{v}\in\mathbf{V}^0(\Gamma)$, because $(I-P)D_AM\gamma_\tau\mathbf{u} = 0$. Thus $R_A\gamma_\tau = \gamma_\tau$, from which we deduce $\gamma_\tau R_A\gamma_\tau = \gamma_\tau$.

From the above identities it follows that

$$R_A\mathbf{u} = R_A\gamma_n\mathbf{u} = \gamma_nR_A\gamma_n\mathbf{u}\in\mathbf{V}^0_\tau(\Gamma) \qquad \text{for all } \mathbf{u}\in\mathbf{V}^0_\tau(\Gamma).$$

Thus the restriction of R_A to $\mathbf{V}^0_\tau(\Gamma)$ is an isomorphism from $\mathbf{V}^0_\tau(\Gamma)$ into itself. Also observe that the restriction of $\gamma_nR_A\gamma_n$ to $\mathbf{V}^0_\tau(\Gamma)$ enjoys the same property. The last identities in the lemma follow from these properties. □

We introduce the operators

$$B_n = (\lambda_0I - A)PD_A\gamma_n, \qquad B_\tau = (\lambda_0I-A)D_A\gamma_\tau\,, \qquad B = B_n + B_\tau.$$

Let us set

$$B_{n,\alpha} = (\lambda_0I-A)^{\alpha-1}B_n = (\lambda_0I-A)^\alpha PD_A\gamma_n,$$
$$B_{\tau,\alpha} = (\lambda_0I-A)^{\alpha-1}B_\tau = (\lambda_0I-A)^\alpha D_A\gamma_\tau,$$

and $B_\alpha = B_{n,\alpha} + B_{\tau,\alpha}$.

THEOREM 2.6. *For all $\alpha\in\,]0,\frac14[$, $B_{n,\alpha}$, and $B_{\tau,\alpha}$ belong to $\mathcal{L}(\mathbf{V}^0(\Gamma),\mathbf{V}^0_n(\Omega))$.*

Proof. Let us prove the result for $B_{n,\alpha}$; the other one can be shown in a similar way. Let $\mathbf{u}\in\mathbf{V}^0(\Gamma)$, and set $\mathbf{y} = D_A\gamma_n(\mathbf{u})$. Then $P\mathbf{y} = PD_A\gamma_n(\mathbf{u})$ belongs to $\mathbf{V}^{1/2}_n(\Omega)$. We have

(2.7)

$$\Big\langle(\lambda_0I-A)P\mathbf{y},\mathbf{z}\Big\rangle_{(D(A^*))',D(A^*)}$$
$$= \Big\langle(\lambda_0I-A)^\alpha P\mathbf{y},(\lambda_0I-A^*)^{1-\alpha}\mathbf{z}\Big\rangle_{(D((-A^*)^{1-\alpha}))',D((-A^*)^{1-\alpha})}$$
$$= \int_\Omega \mathbf{y}P(\lambda_0\mathbf{z} - \nu\Delta\mathbf{z} - (\mathbf{w}\cdot\nabla)\mathbf{z} + (\nabla\mathbf{w})^T\mathbf{z}) \quad \text{for all } \mathbf{z}\in D(A^*) = \mathbf{V}^2(\Omega)\cap\mathbf{V}^1_0(\Omega).$$

Every $\mathbf{z}\in D(A^*)$ is the solution to the equation

$$\lambda_0\mathbf{z} - \nu\Delta\mathbf{z} - (\mathbf{w}\cdot\nabla)\mathbf{z} + (\nabla\mathbf{w})^T\mathbf{z} + \nabla q = \mathbf{f}, \quad \text{div }\mathbf{z} = 0 \quad \text{in }\Omega, \quad \mathbf{z} = 0 \quad \text{on }\Gamma,$$

where $\mathbf{f} = P(\lambda_0\mathbf{z} - \nu\Delta\mathbf{z} - (\mathbf{w}\cdot\nabla)\mathbf{z} + (\nabla\mathbf{w})^T\mathbf{z})$. It is well known that $\nu\frac{\partial\mathbf{z}}{\partial\mathbf{n}} - q\,\mathbf{n}$ satisfies the estimate

$$\Big|\nu\frac{\partial\mathbf{z}}{\partial\mathbf{n}} - q\,\mathbf{n}\Big|_{\mathbf{V}^0(\Gamma)} \le C|\mathbf{f}|_{D((-A^*)^{-\alpha})} \le C|\mathbf{z}|_{D((-A^*)^{1-\alpha})},$$

provided that $0 < \alpha < 1/4$. Thus we have

$$\left|\int_\Omega \mathbf{y}\mathbf{f}\right| = \left|\int_\Omega \mathbf{y}P\Big(\lambda_0\mathbf{z} - \nu\Delta\mathbf{z} - (\mathbf{w}\cdot\nabla)\mathbf{z} + (\nabla\mathbf{w})^T\mathbf{z}\Big)\right|$$
$$= \left|\int_\Omega \mathbf{y}\Big(\lambda_0\mathbf{z} - \nu\Delta\mathbf{z} - (\mathbf{w}\cdot\nabla)\mathbf{z} + (\nabla\mathbf{w})^T\mathbf{z} + \nabla q\Big)\right| = \left|\int_\Gamma \left(-\nu\frac{\partial\mathbf{z}}{\partial\mathbf{n}} + q\,\mathbf{n}\right)\cdot\gamma_n(\mathbf{u})\right|$$
$$\le |\mathbf{u}|_{\mathbf{V}^0(\Gamma)}\left|\nu\frac{\partial\mathbf{z}}{\partial\mathbf{n}} - q\,\mathbf{n}\right|_{\mathbf{V}^0(\Gamma)} \le C|\mathbf{u}|_{\mathbf{V}^0(\Gamma)}|\mathbf{z}|_{D((-A^*)^{1-\alpha})}.$$

Combining this inequality with (2.7), we obtain

$$|(\lambda_0 I - A)^\alpha P\mathbf{y}|_{\mathbf{V}_n^0(\Omega)}$$
$$\le \sup_{|\mathbf{z}|_{D((-A^*)^{1-\alpha})}=1}\left|\int_\Omega \mathbf{y}P\Big(\lambda_0\mathbf{z} - \nu\Delta\mathbf{z} - (\mathbf{w}\cdot\nabla)\mathbf{z} + (\nabla\mathbf{w})^T\mathbf{z}\Big)\right|$$
$$\le C|\mathbf{u}|_{\mathbf{V}^0(\Gamma)}.$$

Thus the proof is complete. □

PROPOSITION 2.7. *For all $\mathbf{\Phi} \in D(A^*)$, $B^*\mathbf{\Phi}$ belongs to $\mathbf{V}^{1/2}(\Gamma)$, and we have*

$$B^*\mathbf{\Phi} = D_A^*(\lambda_0 I - A^*)\mathbf{\Phi}, \quad B_\tau^*\mathbf{\Phi} = \gamma_\tau D_A^*(\lambda_0 I - A^*)\mathbf{\Phi}, \quad B_n^*\mathbf{\Phi} = \gamma_n D_A^*(\lambda_0 I - A^*)\mathbf{\Phi},$$

and

$$B^*\mathbf{\Phi} = -\nu\frac{\partial\mathbf{\Phi}}{\partial\mathbf{n}} + \psi\mathbf{n} - c(\psi)\mathbf{n}, \qquad B_\tau^*\mathbf{\Phi} = \gamma_\tau B^*\mathbf{\Phi} = -\nu\frac{\partial\mathbf{\Phi}}{\partial\mathbf{n}},$$
$$\textit{and} \quad B_n^*\mathbf{\Phi} = \gamma_n B^*\mathbf{\Phi} = \psi\mathbf{n} - c(\psi)\mathbf{n},$$

with

$$\nabla\psi = (I - P)\Big[\nu\Delta\mathbf{\Phi} + (\mathbf{w}\cdot\nabla)\mathbf{\Phi} - (\nabla\mathbf{w})^T\mathbf{\Phi}\Big],$$

and $c(\psi)$ is defined by (2.1). *In particular, if $\mathbf{\Phi} \in \mathbf{V}^s(\Omega) \cap \mathbf{V}_0^1(\Omega)$ with $s \ge 2$, then the following estimate holds:*

$$|B^*\mathbf{\Phi}|_{\mathbf{V}^{s-3/2}(\Gamma)} \le C|\mathbf{\Phi}|_{\mathbf{V}^s(\Omega)\cap\mathbf{V}_0^1(\Omega)}.$$

Proof. The first result is a direct consequence of the definition of B and of the fact that B is a bounded operator from $\mathbf{V}^0(\Gamma)$ into $(D(A^*))'$. The identities for $B_\tau^*\mathbf{\Phi}$ and $B_n^*\mathbf{\Phi}$ follow from Lemma 2.4. From the definitions of B^* and D_A^* it follows that

$$B^*\mathbf{\Phi} = -\nu\frac{\partial\hat{\mathbf{\Phi}}}{\partial\mathbf{n}} + \hat{\psi}\mathbf{n} - c(\hat{\psi})\mathbf{n},$$

where $(\hat{\mathbf{\Phi}}, \hat{\psi})$ is defined by

$$\lambda_0\hat{\mathbf{\Phi}} - \nu\Delta\hat{\mathbf{\Phi}} - (\mathbf{w}\cdot\nabla)\hat{\mathbf{\Phi}} + (\nabla\mathbf{w})^T\hat{\mathbf{\Phi}} + \nabla\hat{\psi} = P(\lambda_0 I - A^*)\mathbf{\Phi} = (\lambda_0 I - A^*)\mathbf{\Phi} \quad \text{in } \Omega,$$
$$\text{div } \hat{\mathbf{\Phi}} = 0 \quad \text{in } \Omega, \quad \hat{\mathbf{\Phi}} = 0 \quad \text{on } \Gamma,$$

and $c(\hat{\psi})$ is defined by (2.1). From the first equation we deduce $(\lambda_0 I - A^*)\mathbf{\Phi} = (\lambda_0 I - A^*)\hat{\mathbf{\Phi}}$. Thus $\mathbf{\Phi} = \hat{\mathbf{\Phi}}$. Next we can see that $\hat{\psi}$ is the unique element in $H^1(\Omega)/\mathbb{R}$ defined by

$$\begin{aligned}\nabla\hat{\psi} &= (I-P)\Big[-\lambda_0\hat{\mathbf{\Phi}} + \nu\Delta\hat{\mathbf{\Phi}} + (\mathbf{w}\cdot\nabla)\hat{\mathbf{\Phi}} - (\nabla\mathbf{w})^T\hat{\mathbf{\Phi}}\Big]\\ &= (I-P)\Big[\nu\Delta\mathbf{\Phi} + (\mathbf{w}\cdot\nabla)\mathbf{\Phi} - (\nabla\mathbf{w})^T\mathbf{\Phi}\Big].\end{aligned}$$

The estimate of $B^*\mathbf{\Phi}$ follows directly from its definition. The proof is complete. ∎

3. A finite time horizon control problem. To deal with the stabilization problem formulated in section 1, we first study the following optimal control problem:

$$(\mathcal{Q}^T_{s,\zeta}) \qquad \inf\Big\{J_T(s,\mathbf{y},\mathbf{u}) \;\Big|\; (\mathbf{y},\mathbf{u}) \text{ satisfies (3.1)},\ \mathbf{u}\in\mathbf{V}^{0,0}(\Sigma_{s,T})\Big\},$$

where

$$J_T(s,\mathbf{y},\mathbf{u}) = \frac{1}{2}\int_s^T\int_\Omega |\mathbf{y}|^2\,dxdt + \frac{1}{2}\int_s^T |\mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)}\,dt$$

and

$$\begin{aligned}&\frac{\partial\mathbf{y}}{\partial t} - \nu\Delta\mathbf{y} + (\mathbf{w}\cdot\nabla)\mathbf{y} + (\mathbf{y}\cdot\nabla)\mathbf{w} + \nabla p = 0 \quad \text{in } Q_{s,T},\\ &\text{div } \mathbf{y} = 0 \quad \text{in } Q_{s,T}, \quad \mathbf{y} = M\mathbf{u} \quad \text{on } \Sigma_{s,T}, \quad \mathbf{y}(s) = \zeta \quad \text{in } \Omega.\end{aligned} \tag{3.1}$$

The main objective of this section is to establish optimality conditions for $(\mathcal{Q}^T_{s,\zeta})$ and to show that the optimal solution $(\bar{\mathbf{y}},\bar{\mathbf{u}})$ obeys a pointwise feedback formulation defined thanks to the solution to a differential Riccati equation. To look for the solution $\mathbf{u}$ to problem $(\mathcal{Q}^T_{s,\zeta})$ in feedback form, we rewrite (3.1) in the form

$$\begin{aligned}&P\mathbf{y}' = AP\mathbf{y} + BM\mathbf{u} \quad \text{in } (s,T), \qquad \mathbf{y}(s) = \zeta,\\ &(I-P)\mathbf{y} = (I-P)D_A\gamma_n M\mathbf{u} \quad \text{in } (s,T).\end{aligned}$$

As we see that the situation is more complicated than in the case when we have a single equation of the form $\mathbf{y}' = A\mathbf{y} + B\mathbf{u}$ and when the feedback law is of the form $\bar{\mathbf{u}}(t) = -B^*\Pi(t)\bar{\mathbf{y}}(t)$ (Π being the solution of some differential Riccati equation). To overcome the difficulty arising from the presence of the second equation $(I-P)\mathbf{y} = (I-P)D_A\gamma_n M\mathbf{u}$, which is not an evolution equation, we also transform the functional J_T in the following way:

$$\begin{aligned}J_T(s,\mathbf{y},\mathbf{u}) &= \frac{1}{2}\int_s^T\int_\Omega |\mathbf{y}|^2 + \frac{1}{2}\int_s^T\int_\Gamma |\mathbf{u}|^2\\ &= \frac{1}{2}\int_s^T\int_\Omega |P\mathbf{y}|^2 + \frac{1}{2}\int_s^T\int_\Omega |(I-P)\mathbf{y}|^2 + \frac{1}{2}\int_s^T\int_\Gamma |\mathbf{u}|^2\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{2}\int_s^T\int_\Omega |(I-P)\mathbf{y}|^2 + \frac{1}{2}\int_s^T\int_\Gamma |\mathbf{u}|^2\\ &= \frac{1}{2}\int_s^T\int_\Omega |(I-P)D_A M\gamma_n\mathbf{u}|^2 + \frac{1}{2}\int_s^T\int_\Gamma (|\gamma_\tau\mathbf{u}|^2 + |\gamma_n\mathbf{u}|^2)\\ &= \frac{1}{2}\int_s^T\int_\Gamma \Big(|(MD_A^*(I-P)D_AM)^{1/2}\gamma_n\mathbf{u}|^2 + |\gamma_n\mathbf{u}|^2 + |\gamma_\tau\mathbf{u}|^2\Big).\end{aligned}$$

Thus, if $\mathbf{y}$ is the solution of (3.1), we have

$$J_T(s,\mathbf{y},\mathbf{u}) = I_T(s,\mathbf{y},\mathbf{u}) = \frac{1}{2}\int_s^T\int_\Omega |P\mathbf{y}|^2 + \frac{1}{2}\int_s^T\int_\Gamma |R_A^{1/2}\gamma_n\mathbf{u}|^2 + \frac{1}{2}\int_s^T\int_\Gamma |\gamma_\tau\mathbf{u}|^2,$$

where R_A is the operator introduced in Lemma 2.5.

The control problem $(\mathcal{Q}^T_{s,\zeta})$ is equivalent to

$$(\mathcal{P}^T_{s,\zeta}) \qquad \inf\Big\{ I_T(s,\mathbf{y},\mathbf{u}) \;\mid\; (\mathbf{y},\mathbf{u}) \text{ satisfies (3.2)},\ \mathbf{u}\in \mathbf{V}^{0,0}(\Sigma_{s,T})\Big\},$$

where

$$P\mathbf{y}' = AP\mathbf{y} + BM\mathbf{u} \quad \text{in } (s,T), \qquad \mathbf{y}(s) = \zeta. \tag{3.2}$$

Even if problems $(\mathcal{Q}^T_{s,\zeta})$ and $(\mathcal{P}^T_{s,\zeta})$ are equivalent, we are going to see that the optimality conditions for $(\mathcal{P}^T_{s,\zeta})$ allow us to prove regularity results for the optimal solution $(\bar{\mathbf{y}},\bar{\mathbf{u}})$, which we cannot obtain with the optimality system of problem $(\mathcal{Q}^T_{s,\zeta})$.

THEOREM 3.1. *For all $s\in[0,T]$ and all $\zeta\in\mathbf{V}^0_n(\Omega)$, problem $(\mathcal{P}^T_{s,\zeta})$ admits a unique solution $(\mathbf{y}^s_\zeta,\mathbf{u}^s_\zeta)$ and*

$$\mathbf{u}^s_\zeta = -MB^*_\tau\mathbf{\Phi}^s_\zeta - R_A^{-1}MB^*_n\mathbf{\Phi}^s_\zeta \quad \textit{in } (s,T), \tag{3.3}$$

where $\mathbf{\Phi}^s_\zeta$ is solution to the equation

$$-\mathbf{\Phi}' = A^*\mathbf{\Phi} + P\mathbf{y}^s_\zeta \quad \textit{in } (s,T), \qquad \mathbf{\Phi}(T) = 0\,. \tag{3.4}$$

Conversely the system

$$\begin{aligned} &P\mathbf{y}' = AP\mathbf{y} - B_\tau M^2 B^*_\tau\mathbf{\Phi} - B_n M R_A^{-1} M B^*_n\mathbf{\Phi} \quad \textit{in } (s,T), \qquad P\mathbf{y}(s) = \zeta\,,\\ &-\mathbf{\Phi}' = A^*\mathbf{\Phi} + P\mathbf{y} \quad \textit{in } (s,T), \qquad \mathbf{\Phi}(T) = 0\,, \end{aligned} \tag{3.5}$$

admits a unique solution $(P\mathbf{y}^s_\zeta,\mathbf{\Phi}^s_\zeta)$ in $L^2(s,T;\mathbf{V}^0_n(\Omega))\times(\mathbf{V}^{2,1}(Q_{s,T})\cap L^2(s,T;\mathbf{V}^1_0(\Omega))$, and $(P\mathbf{y}^s_\zeta - (I-P)D_AMR_A^{-1}MB^*_n\mathbf{\Phi}^s_\zeta, -MB^*_\tau\mathbf{\Phi}^s_\zeta - R_A^{-1}MB^*_n\mathbf{\Phi}^s_\zeta)$ is the optimal solution to $(\mathcal{P}^T_{s,\zeta})$.

Remark 3.2. Due to Lemma 2.5, we have

$$MB^*_\tau\mathbf{\Phi}^s_\zeta + R_A^{-1}MB^*_n\mathbf{\Phi}^s_\zeta = R_A^{-1}MB^*\mathbf{\Phi}^s_\zeta,$$

but we prefer to write it using the decomposition of B^* in the form $B^*_\tau + B^*_n$ to give the respective expressions of the tangential and the normal components.

Proof. Step 1. The existence of a unique solution $(\mathbf{y}^s_\zeta,\mathbf{u}^s_\zeta)$ to problem $(\mathcal{P}^T_{s,\zeta})$ is obvious. Let $\mathbf{u}$ be in $L^2(s,T;\mathbf{V}^0(\Gamma))$ and $\mathbf{v}\in L^2(s,T;\mathbf{V}^0(\Gamma))$. Denote by $(\mathbf{y}_\mathbf{u},p_\mathbf{u})$ the solution to (3.2) corresponding to $\mathbf{u}$, and set

$$I_T(s,\mathbf{y}_\mathbf{u},\mathbf{u}) = \mathbf{I}_T(\mathbf{u}).$$

We have

$$\mathbf{I}'_T(\mathbf{u})\mathbf{v} = \int_s^T\int_\Omega P\mathbf{y}_\mathbf{u}\cdot P\mathbf{z} + \int_s^T\int_\Gamma R_A\gamma_n\mathbf{u}\cdot\gamma_n\mathbf{v} + \int_s^T\int_\Gamma \gamma_\tau\mathbf{u}\cdot\gamma_\tau\mathbf{v},$$

where $\mathbf{z}$ is the solution to

$$P\mathbf{z}' = AP\mathbf{z} + BM\mathbf{v} \quad \text{in } (s,T), \qquad P\mathbf{z}(s) = 0, \qquad (I-P)\mathbf{z} = (I-P)D_A M\mathbf{v}\,.$$

Let $\mathbf{\Phi}$ be the solution to the equation

$$-\mathbf{\Phi}' = A^*\mathbf{\Phi} + P\mathbf{y}_\mathbf{u}, \qquad \mathbf{\Phi}(T) = 0\,.$$

Due to Lemma 8.6, $\mathbf{\Phi}$ belongs to $\mathbf{V}^{2,1}(Q_{s,T})$. Thus $B^*\mathbf{\Phi}$ belongs to $L^2(s,T;\mathbf{V}^0(\Gamma))$. The functions $\mathbf{z}$ and $\mathbf{\Phi}$ obey the following identity:

$$\int_s^T \int_\Omega P\mathbf{y}_\mathbf{u} \cdot P\mathbf{z} = \int_s^T \Big\langle BM\mathbf{v}(\tau), \mathbf{\Phi}(\tau)\Big\rangle_{(D(A^*))',D(A^*)} d\tau = \int_s^T \int_\Gamma \mathbf{v} \cdot MB^*\mathbf{\Phi}\,.$$

Thus

$$\mathbf{I}_T'(\mathbf{u})\mathbf{v} = \int_s^T \int_\Gamma \mathbf{v} \cdot MB^*\mathbf{\Phi} + \int_s^T \int_\Gamma R_A\gamma_n\mathbf{u} \cdot \gamma_n\mathbf{v} + \int_s^T \int_\Gamma \gamma_\tau\mathbf{u} \cdot \gamma_\tau\mathbf{v}. \tag{3.6}$$

If $(\mathbf{y}_\zeta^s, \mathbf{u}_\zeta^s)$ is the solution to problem $(\mathcal{P}_{s,\zeta}^T)$, we have $\mathbf{I}_T'(\mathbf{u}_\zeta^s) = 0$, which gives

$$\gamma_n\mathbf{u}_\zeta^s = -\gamma_n R_A^{-1} MB^*\mathbf{\Phi}_\zeta^s = -R_A^{-1} MB_n^*\mathbf{\Phi}_\zeta^s \quad \text{and} \quad \gamma_\tau\mathbf{u}_\zeta^s = -\gamma_\tau MB^*\mathbf{\Phi}_\zeta^s = -MB_\tau^*\mathbf{\Phi}_\zeta^s\,.$$

Step 2. From Step 1 of the proof it follows that $(\mathbf{y}_\zeta^s, \mathbf{\Phi}_\zeta^s)$ is a solution of system (3.5). ($\mathbf{\Phi}_\zeta^s$ is the solution to (3.4).) If $(\bar{\mathbf{y}}, \bar{\mathbf{\Phi}})$ is a solution of system (3.5), and if we set $\bar{\mathbf{u}} = -MB_\tau^*\bar{\mathbf{\Phi}} - R_A^{-1}MB_n^*\bar{\mathbf{\Phi}}$, with (3.6) we can verify that $\mathbf{I}_T'(\bar{\mathbf{u}}) = 0$, which implies that $\bar{\mathbf{u}} = \mathbf{u}_\zeta^s$. Thus $\bar{\mathbf{y}} = \mathbf{y}_\zeta^s$ and $\bar{\mathbf{\Phi}} = \mathbf{\Phi}_\zeta^s$. The proof is complete. □

There is another way to characterize the optimal solution of $(\mathcal{P}_{s,\zeta}^T)$; it consists of writing the optimality conditions for $(\mathcal{Q}_{s,\zeta}^T)$. More precisely, we can state the following theorem.

THEOREM 3.3. *For all $s \in [0,T]$ and all $\zeta \in \mathbf{V}_n^0(\Omega)$, the unique solution $(\mathbf{y}_\zeta^s, \mathbf{u}_\zeta^s)$ of problem $(Q_{s,\zeta}^T)$ is characterized by*

$$\mathbf{u}_\zeta^s = m\left(\nu\frac{\partial\mathbf{\Phi}_\zeta^s}{\partial\mathbf{n}} - \psi_\zeta^s\mathbf{n}\right) + c(m\psi_\zeta^s)\mathbf{n}\,, \tag{3.7}$$

where $(\mathbf{\Phi}_\zeta^s, \psi_\zeta^s)$ is the solution to

$$\begin{aligned} &-\frac{\partial\mathbf{\Phi}}{\partial t} - \nu\Delta\mathbf{\Phi} - (\mathbf{w}\cdot\nabla)\mathbf{\Phi} + (\nabla\mathbf{w})^T\mathbf{\Phi} + \nabla\psi = \mathbf{y}_\zeta^s \quad \textit{in } Q_{s,T},\\ &\operatorname{div}\mathbf{\Phi} = 0 \quad \textit{in } Q_{s,T}, \quad \mathbf{\Phi} = 0 \quad \textit{on } \Sigma_{s,T}, \quad \mathbf{\Phi}(T) = 0 \quad \textit{in } \Omega\,, \end{aligned} \tag{3.8}$$

and $c(m\psi_\zeta^s)$ is the constant defined by (2.1) *corresponding to $m\psi_\zeta^s$.*

Remark 3.4. It is clear that the solution $\mathbf{\Phi}_\zeta^s$ of (3.8) is equal to the solution of (3.4). This is the reason why they are denoted in the same way.

Proof. Let $\mathbf{u}$ be in $L^2(s,T;\mathbf{V}^0(\Gamma))$ and $\mathbf{v} \in L^2(s,T;\mathbf{V}^0(\Gamma))$. Denote by $\mathbf{y}_\mathbf{u}$ the solution to (3.1), and set

$$J_T(s, \mathbf{y}_\mathbf{u}, \mathbf{u}) = \mathbf{J}_T(\mathbf{u}).$$

We have

$$\mathbf{J}_T'(\mathbf{u})\mathbf{v} = \int_s^T \int_\Omega \mathbf{y}_\mathbf{u} \cdot \mathbf{z} + \int_s^T \int_\Gamma \mathbf{u} \cdot \mathbf{v},$$

where $(\mathbf{z}, q)$ is the solution to

$$\frac{\partial \mathbf{z}}{\partial t} - \nu\Delta\mathbf{z} + (\mathbf{w}\cdot\nabla)\mathbf{z} + (\mathbf{z}\cdot\nabla)\mathbf{w} + \nabla q = 0 \quad \text{in } Q_{s,T},$$
$$\text{div } \mathbf{z} = 0 \quad \text{in } Q_{s,T}, \quad \mathbf{y} = M\mathbf{v} \quad \text{on } \Sigma_{s,T}, \quad \mathbf{z}(s) = 0 \quad \text{in } \Omega.$$

Let $(\mathbf{\Phi}, \psi)$ be the solution to the equation

$$-\frac{\partial \mathbf{\Phi}}{\partial t} - \nu\Delta\mathbf{\Phi} - (\mathbf{w}\cdot\nabla)\mathbf{\Phi} + (\nabla\mathbf{w})^T\mathbf{\Phi} + \nabla\psi = \mathbf{y_u} \quad \text{in } Q_{s,T},$$
$$\text{div } \mathbf{\Phi} = 0 \quad \text{in } Q_{s,T}, \quad \mathbf{\Phi} = 0 \quad \text{on } \Sigma_{s,T}, \quad \mathbf{\Phi}(T) = 0 \quad \text{in } \Omega\,.$$

We can verify that

$$\int_s^T \int_\Omega \mathbf{y_u}\cdot\mathbf{z} = \int_s^T \int_\Gamma \left(-\nu\frac{\partial \mathbf{\Phi}}{\partial \mathbf{n}} + \psi\mathbf{n}\right)\cdot M\mathbf{v} = \int_s^T \int_\Gamma m\left(-\nu\frac{\partial \mathbf{\Phi}}{\partial \mathbf{n}} + \psi\mathbf{n}\right)\cdot\mathbf{v}.$$

Thus $\mathbf{u}_\zeta^s$ is characterized by

$$\int_s^T \int_\Gamma \left(-m\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}} + m\psi_\zeta^s\mathbf{n} + \mathbf{u}_\zeta^s\right)\cdot\mathbf{v} = 0$$

for all $\mathbf{v} \in \mathbf{V}^0(\Gamma)$; that is,

$$\gamma_\tau\mathbf{u}_\zeta^s = m\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}} \quad \text{and} \quad \gamma_n\mathbf{u}_\zeta^s = -m\psi_\zeta^s\mathbf{n} + c(m\psi_\zeta^s)\mathbf{n}.$$

The proof is complete. □

Since $(\mathcal{P}_{s,\zeta}^T)$ and $(\mathcal{Q}_{s,\zeta}^T)$ admit the same solution, we have

$$-MB_\tau^*\mathbf{\Phi}_\zeta^s - R_A^{-1}MB_n^*\mathbf{\Phi}_\zeta^s = m\left(\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}} - \psi_\zeta^s\mathbf{n}\right) + c(m\psi_\zeta^s)\mathbf{n}\,.$$

The characterization given in (3.7) seems to be easier to use than the one in (3.3). However, we are going to see that we are able to characterize the regularity of the solution $(\mathbf{y}_\zeta^s, \mathbf{u}_\zeta^s)$ in an optimal way by using (3.3), which is not the case with (3.7). To understand the relationship between (3.3) and (3.7), let us rewrite the second equation in (3.5) in the form

$$-\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial t} - \nu\Delta\mathbf{\Phi}_\zeta^s - (\mathbf{w}\cdot\nabla)\mathbf{\Phi}_\zeta^s + (\nabla\mathbf{w})^T\mathbf{\Phi}_\zeta^s + \nabla\tilde{\psi}_\zeta^s = P\mathbf{y}_\zeta^s \quad \text{in } Q_{s,T},$$
$$\text{div } \mathbf{\Phi}_\zeta^s = 0 \quad \text{in } Q_{s,T}, \quad \mathbf{\Phi}_\zeta^s = 0 \quad \text{on } \Sigma_{s,T}, \quad \mathbf{\Phi}(T) = 0 \quad \text{in } \Omega. \tag{3.9}$$

Observe that $\nabla\tilde{\psi}_\zeta^s = \nabla\psi_\zeta^s - (I-P)\mathbf{y}_\zeta^s$, where ψ_ζ^s is the pressure appearing in (3.8).

LEMMA 3.5. *If $(\mathbf{\Phi}_\zeta^s, \tilde{\psi}_\zeta^s)$ is the solution to* (3.9)*, then*

$$B^*\mathbf{\Phi}_\zeta^s(t) = -\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}}(t) + \tilde{\psi}_\zeta^s(t)\mathbf{n} - c(\tilde{\psi}_\zeta^s(t))\mathbf{n} \quad \textit{for almost all } t \in (s,T),$$

and the optimal control $\mathbf{u}_\zeta^s$ is defined by

$$\mathbf{u}_\zeta^s = m\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}} - R_A^{-1}M\tilde{\psi}_\zeta^s\mathbf{n}. \tag{3.10}$$

Proof. From Proposition 2.7 it follows that

$$B^*\mathbf{\Phi}^s_\zeta(t) = -\nu\frac{\partial \mathbf{\Phi}^s_\zeta}{\partial \mathbf{n}}(t) + \hat\psi^s_\zeta(t)\mathbf{n} - c(\hat\psi^s_\zeta(t))\mathbf{n}\,,$$

where $\nabla\hat\psi^s_\zeta$ is defined by

$$\nabla\hat\psi^s_\zeta = (I-P)\Big[\nu\Delta\mathbf{\Phi}^s_\zeta + (\mathbf{w}\cdot\nabla)\mathbf{\Phi}^s_\zeta - (\nabla\mathbf{w})^T\mathbf{\Phi}^s_\zeta\Big]\,.$$

Thus $\nabla\hat\psi^s_\zeta = \nabla\tilde\psi^s_\zeta$ because $(I-P)\big(\frac{\partial\mathbf{\Phi}^s_\zeta}{\partial t}\big) = 0$. Therefore $\mathbf{u}^s_\zeta$ is defined by

$$\mathbf{u}^s_\zeta = -MB^*_\tau\mathbf{\Phi}^s_\zeta - R_A^{-1}MB^*_n\mathbf{\Phi}^s_\zeta = \Big(-M\gamma_\tau - R_A^{-1}M\gamma_n\Big)\left(\nu\frac{\partial\mathbf{\Phi}^s_\zeta}{\partial\mathbf{n}} - \tilde\psi^s_\zeta\mathbf{n}\right),$$

and due to Lemma 2.5, the proof is complete. □

In the following theorem we improve the regularity result of the optimal solution.

THEOREM 3.6. *The solution* $(\mathbf{y}^s_\zeta, \mathbf{\Phi}^s_\zeta)$ *to system* (3.5) *belongs to* $\mathbf{V}^{1,1/2}(Q_{s,T}) \times L^2(s,T;\mathbf{V}^3(\Omega)\cap\mathbf{V}^1_0(\Omega)) \cap H^{3/2}(s,T;\mathbf{V}^0_n(\Omega))$. *In particular,* $\mathbf{\Phi}^s_\zeta$ *belongs to* $C([s,T];\mathbf{V}^2(\Omega)\cap\mathbf{V}^1_0(\Omega))$.

Proof. Since $B^*\mathbf{\Phi}^s_\zeta \in L^2(s,T;\mathbf{V}^0(\Gamma))$, due to Lemmas 8.1 and 8.3, $\mathbf{y}^s_\zeta$ belongs to $\mathbf{V}^{1/2-\varepsilon,1/4-\varepsilon}(Q_{s,T})$ for all $\varepsilon>0$. From Lemmas 8.6 and 8.8 it follows that

$$\frac{\partial\mathbf{\Phi}^s_\zeta}{\partial\mathbf{n}} \in \mathbf{V}^{1-\varepsilon,1/2-\varepsilon/2}(\Sigma_{s,T}) \qquad \text{and} \qquad \nabla\tilde\psi^s_\zeta \in \mathbf{H}^{1/2-\varepsilon,1/4-\varepsilon/2}(Q_{s,T})$$

for all $\varepsilon>0$, where $\tilde\psi^s_\zeta$ is the pressure appearing in (3.9). From this regularity result and from Lemma 3.5, we deduce that $B^*\mathbf{\Phi}^s_\zeta \in \mathbf{V}^{1/2-\varepsilon,1/4-\varepsilon/2}(\Sigma_{s,T})$ and $\mathbf{u}^s_\zeta \in \mathbf{V}^{1/2-\varepsilon,1/4-\varepsilon/2}(\Sigma_{s,T})$ for all $\varepsilon>0$, where $\mathbf{u}^s_\zeta = -MB^*_\tau\mathbf{\Phi}^s_\zeta - R_A^{-1}MB^*_n\mathbf{\Phi}^s_\zeta$ is the solution of problems $(Q^T_{s,\zeta})$ and $(P^T_{s,\zeta})$. Still applying Lemma 8.3, we obtain $\mathbf{y}^s_\zeta \in \mathbf{V}^{1-\varepsilon,1/2-\varepsilon}(Q_{s,T})$ for all $\varepsilon>0$, and repeating the above analysis for $B^*\mathbf{\Phi}^s_\zeta$, we can show that $\mathbf{u}^s_\zeta \in \mathbf{V}^{1-\varepsilon,1/2-\varepsilon/2}(\Sigma_{s,T})$ for all $\varepsilon>0$. Still with Lemmas 8.1 and 8.3, we prove that $\mathbf{y}^s_\zeta$ belongs to $\mathbf{V}^{1,1/2}(Q_{s,T})$. Due to Lemma 8.6, $\mathbf{\Phi}^s_\zeta$ belongs to $\mathbf{V}^{3,3/2}(Q_{s,T})$. In particular, $\mathbf{\Phi}^s_\zeta$ belongs to $C([s,T];\mathbf{V}^2(\Omega)\cap\mathbf{V}^1_0(\Omega))$. The proof is now complete. □

Remark 3.7. Observe that the regularity result stated in Theorem 3.6 has been obtained by using the expression of $\mathbf{u}^s_\zeta$ given in (3.10). In (3.10), only the pressure $\tilde\psi^s_\zeta$ appears, and $\tilde\psi^s_\zeta$ depends only on $\mathbf{\Phi}^s_\zeta$. Thus by improving the regularity of $\mathbf{\Phi}^s_\zeta$ we are able to improve the regularity of $\tilde\psi^s_\zeta$, and this is what is done in Lemma 8.8. If we use the expression of $\mathbf{u}^s_\zeta$ given in (3.7), the pressure ψ^s_ζ obeys $\nabla\psi^s_\zeta = \nabla\tilde\psi^s_\zeta + (I-P)\mathbf{y}^s_\zeta$. Since the regularity with respect to the time variable of $(I-P)\mathbf{y}^s_\zeta$ cannot be improved, we are not able to improve the regularity of $\mathbf{u}^s_\zeta$ by using (3.7).

COROLLARY 3.8. *For all* $s\in[0,T]$ *and all* $\zeta\in\mathbf{V}^0_n(\Omega)$, *the unique solution* $(\mathbf{y}^s_\zeta,\mathbf{u}^s_\zeta)$ *to problem* $(\mathcal{P}^T_{s,\zeta})$ *and the corresponding solution* $(\mathbf{y}^s_\zeta,\mathbf{\Phi}^s_\zeta)$ *to system* (3.5) *obey*

$$I_T(s,\mathbf{y}^s_\zeta,\mathbf{u}^s_\zeta) = \frac{1}{2}\int_\Omega \mathbf{\Phi}^s_\zeta(s)\cdot\zeta\,.$$

Proof. By multiplying by $\mathbf{\Phi}_\zeta^s$ the equation (3.1) satisfied by $(\mathbf{y}_\zeta^s, \mathbf{u}_\zeta^s)$, integrating in space and time, making integration by parts, and using (3.9) satisfied by $\mathbf{\Phi}_\zeta^s$, we can prove that

$$\int_\Omega \mathbf{\Phi}_\zeta^s(s)\cdot\zeta = \int_s^T\int_\Omega |P\mathbf{y}_\zeta^s|^2 + \int_s^T\int_\Gamma \left(\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}} - \tilde{\psi}_\zeta^s\mathbf{n} + c(\tilde{\psi}_\zeta^s)\mathbf{n}\right)\cdot M\mathbf{u}_\zeta^s.$$

Next using (3.10), we can show that

$$\int_s^T\int_\Gamma \left(\nu\frac{\partial \mathbf{\Phi}_\zeta^s}{\partial \mathbf{n}} - \tilde{\psi}_\zeta^s\mathbf{n} + c(\tilde{\psi}_\zeta^s)\mathbf{n}\right)\cdot M\mathbf{u}_\zeta^s = \int_s^T\int_\Gamma |R_A^{1/2}\gamma_n\mathbf{u}_\zeta^s|^2 + \int_s^T\int_\Gamma |\gamma_\tau\mathbf{u}_\zeta^s|^2,$$

and the proof is complete. ∎

Let $\Pi(s)$ be the operator defined by

$$\Pi(s)\ :\ \zeta \longmapsto \mathbf{\Phi}_\zeta^s(s)\,, \tag{3.11}$$

where $(P\mathbf{y}_\zeta^s, \mathbf{\Phi}_\zeta^s)$ is the unique solution to system (3.5). From Theorem 3.6 it follows that $\Pi(s) \in \mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^2(\Omega)\cap\mathbf{V}_0^1(\Omega))$. Using Theorem 3.6, we can prove that the family of operators $(\Pi(s))_{s\in[0,T]}$ defined by (3.11) belongs to $C_s([0,T];\mathcal{L}(\mathbf{V}_n^0(\Omega)))$ (the space of functions Π from $[0,T]$ into $\mathcal{L}(\mathbf{V}_n^0(\Omega))$ such that, for all $\mathbf{y}\in\mathbf{V}_n^0(\Omega)$, $\Pi(\cdot)\mathbf{y}$ is continuous from $[0,T]$ into $\mathbf{V}_n^0(\Omega)$). Next, using the optimality system (3.5), we can show that Π is the unique solution in $C_s([0,T];\mathcal{L}(\mathbf{V}_n^0(\Omega)))$ to the Riccati equation

$$\begin{gathered}
\Pi^*(t) = \Pi(t) \quad\text{and}\quad \Pi(t)\ge 0;\\
\text{for all } \mathbf{y}\in\mathbf{V}_n^0(\Omega),\ t\in[0,T],\\
\Pi(t)\mathbf{y}\in\mathbf{V}^2(\Omega)\cap\mathbf{V}_0^1(\Omega) \quad\text{and}\quad |\Pi(t)\mathbf{y}|_{\mathbf{V}^2(\Omega)} \le C|\mathbf{y}|_{\mathbf{V}_n^0(\Omega)};\\
-\Pi'(t) = A^*\Pi(t) + \Pi(t)A - \Pi(t)B_\tau M^2 B_\tau^*\Pi(t) - \Pi(t)B_n M R_A^{-1} M B_n^*\Pi(t) + I,\\
\Pi(T) = 0.
\end{gathered} \tag{3.12}$$

From the definition of Π, from Theorem 3.1 and Corollary 3.8, we deduce the following theorem. We also refer to [23, Theorem 1.2.2.1], where the existence of a unique solution to (3.12) is established.

THEOREM 3.9. *The solution* $(\mathbf{y},\mathbf{u})$ *to problem* $(\mathcal{P}_{0,\mathbf{y}_0}^T)$ *belongs to* $C([0,T];\mathbf{V}^0(\Omega))\times C([0,T];\mathbf{V}^0(\Gamma))$*; it obeys the feedback formula*

$$\mathbf{u}(t) = -\Big(MB_\tau^* + R_A^{-1}MB_n^*\Big)\Pi(t)P\mathbf{y}(t),$$

and the optimal cost is given by

$$J(\mathbf{y},\mathbf{u}) = \frac{1}{2}\Big(\Pi(0)\mathbf{y}_0,\mathbf{y}_0\Big)_{\mathbf{V}_n^0(\Omega)}.$$

If we set $\widehat{\Pi}(t) = \Pi(T-t)$, then $\widehat{\Pi}$ is the unique solution in $C_s([0,T];\mathcal{L}(\mathbf{V}_n^0(\Omega)))$ to the Riccati equation

$$\begin{gathered}
\widehat{\Pi}^*(t) = \widehat{\Pi}(t) \quad\text{and}\quad \widehat{\Pi}(t)\ge 0;\\
\text{for all } \mathbf{y}\in\mathbf{V}_n^0(\Omega),\ t\in[0,T],\\
\widehat{\Pi}(t)\mathbf{y}\in\mathbf{V}^2(\Omega)\cap\mathbf{V}_0^1(\Omega) \quad\text{and}\quad |\widehat{\Pi}(t)\mathbf{y}|_{\mathbf{V}^2(\Omega)} \le C|\mathbf{y}|_{\mathbf{V}_n^0(\Omega)};\\
\widehat{\Pi}'(t) = A^*\widehat{\Pi}(t) + \widehat{\Pi}(t)A - \widehat{\Pi}(t)B_\tau M^2 B_\tau^*\widehat{\Pi}(t) - \widehat{\Pi}(t)B_n M R_A^{-1} M B_n^*\widehat{\Pi}(t) + I,\\
\widehat{\Pi}(0) = 0.
\end{gathered} \tag{3.13}$$

From the definition of $\widehat{\Pi}$ it follows that $\Pi(0) = \widehat{\Pi}(T)$.

4. An infinite time horizon problem. In this section we want to study problem $(\mathcal{P}^{\infty}_{0,\mathbf{y}_0})$, and we want to study the regularity of its solution according to the regularity of $\mathbf{y}_0$. For notational simplicity, problem $(\mathcal{P}^{\infty}_{0,\mathbf{y}_0})$ will now be denoted by $(\mathcal{P}_{0,\mathbf{y}_0})$, and the state variable by $\mathbf{y}$ and not $P\mathbf{y}$ as in the previous section. With this notation the problem we consider is

$$(\mathcal{P}_{0,\mathbf{y}_0}) \qquad \inf\Big\{ I(\mathbf{y},\mathbf{u}) \ \mid \ (\mathbf{y},\mathbf{u}) \text{ satisfies (4.1)}, \ \mathbf{u} \in \mathbf{V}^{0,0}(\Sigma_\infty)\Big\},$$

where

$$I(\mathbf{y},\mathbf{u}) = \frac{1}{2}\int_0^\infty \int_\Omega |\mathbf{y}|^2 \, dxdt + \frac{1}{2}\int_0^\infty \Big(|\gamma_\tau \mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n \mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)}\Big)\, dt$$

and

$$\mathbf{y}' = A\mathbf{y} + BM\mathbf{u} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = \mathbf{y}_0. \tag{4.1}$$

Accordingly, for $0 \le s < k < \infty$, problem $(\mathcal{P}^k_{s,\zeta})$ will now be defined by

$$(\mathcal{P}^k_{s,\zeta}) \qquad \inf\Big\{ I_k(s,\mathbf{y},\mathbf{u}) \ \mid \ (\mathbf{y},\mathbf{u}) \text{ satisfies (4.2)}, \ \mathbf{u} \in \mathbf{V}^{0,0}(\Sigma_{s,k})\Big\},$$

where

$$I_k(s,\mathbf{y},\mathbf{u}) = \frac{1}{2}\int_s^k \int_\Omega |\mathbf{y}|^2 \, dxdt + \frac{1}{2}\int_s^k \Big(|\gamma_\tau \mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n \mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)}\Big)\, dt$$

and

$$\mathbf{y}' = A\mathbf{y} + BM\mathbf{u} \quad \text{in } (s,k), \qquad \mathbf{y}(s) = \zeta. \tag{4.2}$$

THEOREM 4.1. *For all $\mathbf{y}_0 \in \mathbf{V}^0_n(\Omega)$, problem $(\mathcal{P}_{0,\mathbf{y}_0})$ admits a unique solution $(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0})$. There exists $\Pi \in \mathcal{L}(\mathbf{V}^0_n(\Omega))$, obeying $\Pi = \Pi^*$, such that the optimal cost is given by*

$$I(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0}) = \frac{1}{2}\Big(\Pi \mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}^0_n(\Omega)}.$$

Proof. From the null controllability results stated in [8], we can deduce that there exist controls $\mathbf{u} \in L^2(0,\infty;\mathbf{V}^0(\Gamma))$ such that $I(\mathbf{y}_\mathbf{u},\mathbf{u}) < \infty$, where $\mathbf{y}_\mathbf{u}$ is the solution of (4.1) corresponding to $\mathbf{u}$. The null controllability results in [8] are stated for a distributed control. Using an extension of the domain, this result also provides a null controllability result for a control localized on the boundary. The existence of a unique solution $(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0})$ to $(\mathcal{P}_{0,\mathbf{y}_0})$ follows from classical arguments.

From the dynamic programming principle, it follows that the mapping

$$T \longmapsto \Big(\widehat{\Pi}(T)\mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}^0_n(\Omega)}$$

is nondecreasing, and we have

$$\frac{1}{2}\Big(\widehat{\Pi}(T)\mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}^0_n(\Omega)} \le I(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0}) < \infty\,.$$

As in [7] or in [23], we can show that there exists an operator $\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega))$ satisfying $\Pi = \Pi^* \geq 0$ and

$$\Pi \mathbf{y}_0 = \lim_{T\to\infty} \widehat{\Pi}(T)\mathbf{y}_0 \qquad \text{for all } \mathbf{y}_0 \in \mathbf{V}_n^0(\Omega).$$

Let us show that $I(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0}) = \frac{1}{2}\big(\Pi \mathbf{y}_0, \mathbf{y}_0\big)_{\mathbf{V}_n^0(\Omega)}$. Problem $(\mathcal{P}^k_{0,\mathbf{y}_0})$ admits a unique solution $(\mathbf{y}_k, \mathbf{u}_k)$ characterized by

$$\begin{aligned} &\mathbf{y}_k' = A\mathbf{y}_k + BM\mathbf{u}_k \quad \text{in } (0,k), \qquad \mathbf{y}_k(0) = \mathbf{y}_0, \\ &-\mathbf{\Phi}_k' = A^*\mathbf{\Phi}_k + \mathbf{y}_k \quad \text{in } (0,k), \qquad \mathbf{\Phi}_k(k) = 0, \\ &\gamma_\tau \mathbf{u}_k = -MB_\tau^* \mathbf{\Phi}_k, \qquad \gamma_n \mathbf{u}_k = -R_A^{-1} MB_n^* \mathbf{\Phi}_k. \end{aligned} \tag{4.3}$$

Convergence of $\mathbf{y}_k$ *and* $\mathbf{u}_k$. Denote by $\tilde{\mathbf{u}}_k$ the extension by zero of $\mathbf{u}_k$ to (k,∞), and by $\tilde{\mathbf{y}}_k$ the extension by zero of $\mathbf{y}_k$ to (k,∞). Since we have

$$\begin{aligned} &\int_0^k \int_\Omega |\mathbf{y}_k|^2 \, dxdt + \int_0^k \Big(|\gamma_\tau \mathbf{u}_k(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n \mathbf{u}_k(t)|^2_{\mathbf{V}^0(\Gamma)} \Big)\, dt \\ &\leq \int_0^\infty \int_\Omega |\mathbf{y}_{\mathbf{y}_0}|^2 \, dxdt + \int_0^\infty \Big(|\gamma_\tau \mathbf{u}_{\mathbf{y}_0}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n \mathbf{u}_{\mathbf{y}_0}(t)|^2_{\mathbf{V}^0(\Gamma)} \Big)\, dt\,, \end{aligned}$$

the sequences $(\tilde{\mathbf{y}}_k)_k$ and $(\tilde{\mathbf{u}}_k)_k$ are respectively bounded in $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ and $L^2(0,\infty;\mathbf{V}^0(\Gamma))$. Thus there exist $\mathbf{y}_\infty \in L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ and $\mathbf{u}_\infty \in L^2(0,\infty;\mathbf{V}^0(\Gamma))$ such that

$$\begin{aligned} &\tilde{\mathbf{u}}_k \rightharpoonup \mathbf{u}_\infty \quad \text{weakly in } L^2(0,\infty;\mathbf{V}^0(\Gamma)), \\ &\tilde{\mathbf{y}}_k \rightharpoonup \mathbf{y}_\infty \quad \text{weakly in } L^2(0,\infty;\mathbf{V}_n^0(\Omega)). \end{aligned}$$

By passing to the limit in the above inequality, we obtain

$$\begin{aligned} &\int_0^\infty \int_\Omega |\mathbf{y}_\infty|^2 \, dxdt + \int_0^\infty \Big(|\gamma_\tau \mathbf{u}_\infty(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n \mathbf{u}_\infty(t)|^2_{\mathbf{V}^0(\Gamma)} \Big)\, dt \\ &\leq \int_0^\infty \int_\Omega |\mathbf{y}_{\mathbf{y}_0}|^2 \, dxdt + \int_0^\infty \Big(|\gamma_\tau \mathbf{u}_{\mathbf{y}_0}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n \mathbf{u}_{\mathbf{y}_0}(t)|^2_{\mathbf{V}^0(\Gamma)} \Big)\, dt\,. \end{aligned}$$

And by passing to the limit in the equation satisfied by $(\mathbf{y}_k, \mathbf{u}_k)$, we have

$$\mathbf{y}_\infty' = A\mathbf{y}_\infty + BM\mathbf{u}_\infty \quad \text{in } (0,\infty), \qquad \mathbf{y}_\infty(0) = \mathbf{y}_0.$$

Thus the pair $(\mathbf{y}_\infty, \mathbf{u}_\infty)$ is admissible, and we have

$$(\mathbf{y}_\infty, \mathbf{u}_\infty) = (\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0}),$$

because $I(\mathbf{y}_\infty, \mathbf{u}_\infty) \leq I(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0})$. Therefore we can claim that

$$\tilde{\mathbf{u}}_k \to \mathbf{y}_{\mathbf{y}_0} \quad \text{in } L^2(0,\infty;\mathbf{V}^0(\Gamma)) \quad \text{and} \quad \tilde{\mathbf{y}}_k \to \mathbf{u}_{\mathbf{y}_0} \quad \text{in } L^2(0,\infty;\mathbf{V}_n^0(\Omega)).$$

Since

$$I_k(0, \mathbf{y}_k, \mathbf{u}_k) = \frac{1}{2}\Big(\widehat{\Pi}(k)\mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}_n^0(\Omega)},$$

by passing to the limit when k tends to infinity, we obtain

$$I(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0}) = \frac{1}{2}\Big(\Pi \mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}_n^0(\Omega)}. \qquad \square$$

We denote by $\varphi(\mathbf{y}_0)$ the value function of problem $(\mathcal{P}_{0,\mathbf{y}_0})$, that is,

$$\varphi(\mathbf{y}_0) = I(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0}).$$

LEMMA 4.2. *For every $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$, the system*

$$\begin{aligned} &\mathbf{y}' = A\mathbf{y} - B_\tau M^2 B_\tau^* \mathbf{\Phi} - B_n M R_A^{-1} M B_n^* \mathbf{\Phi} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = \mathbf{y}_0, \\ &-\mathbf{\Phi}' = A^* \mathbf{\Phi} + \mathbf{y} \quad \text{in } (0,\infty), \qquad \mathbf{\Phi}(\infty) = 0, \\ &\mathbf{\Phi}(t) = \Pi \mathbf{y}(t) \quad \text{for all } t \in (0,\infty), \end{aligned} \tag{4.4}$$

admits a unique solution in $L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \times \mathbf{V}^{2,1}(Q_\infty)$. This solution belongs to $C_b(\mathbb{R}^+;\mathbf{V}_n^0(\Omega)) \cap \mathbf{V}^{1,1/2}(Q_\infty) \times \mathbf{V}^{3,3/2}(Q_\infty)$, and it satisfies

$$\|\mathbf{y}\|_{C_b(\mathbb{R}^+;\mathbf{V}_n^0(\Omega))} + \|\mathbf{y}\|_{\mathbf{V}^{1,1/2}(Q_\infty)} + \|\mathbf{\Phi}\|_{\mathbf{V}^{3,3/2}(Q_\infty)} \le C|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)}.$$

The pair $(\mathbf{y}, -MB_\tau^*\mathbf{\Phi} - R_A^{-1}MB_n^*\mathbf{\Phi})$ is the solution of $(\mathcal{P}_{0,\mathbf{y}_0})$.

Proof. For notational simplicity the solution to $(\mathcal{P}_{0,\mathbf{y}_0})$ will now be denoted by $(\hat{\mathbf{y}}, \hat{\mathbf{u}})$. We denote by $\varphi_k(0,\mathbf{y}_0)$ the value function of problem $(\mathcal{P}^k_{0,\mathbf{y}_0})$ and by $\varphi_k(\bar{t},\zeta)$ the value function of problem $(\mathcal{P}^k_{\bar{t},\zeta})$.

Let $(\mathbf{y}_k^{\bar{t}}, \mathbf{u}_k^{\bar{t}})$ be the solution of $(\mathcal{P}^k_{\bar{t},\mathbf{y}_k(\bar{t})})$, and let $(\mathbf{y}_k, \mathbf{u}_k)$ be the solution of $(\mathcal{P}^k_{0,\mathbf{y}_0})$ characterized by (4.3). Denote by $\mathbf{\Phi}_k^{\bar{t}}$ the adjoint state corresponding to $(\mathbf{y}_k^{\bar{t}}, \mathbf{u}_k^{\bar{t}})$, and by $\mathbf{\Phi}_k$ the adjoint state corresponding to $(\mathbf{y}_k, \mathbf{u}_k)$. From the dynamic programming principle it follows that $(\mathbf{y}_k^{\bar{t}}, \mathbf{u}_k^{\bar{t}}, \mathbf{\Phi}_k^{\bar{t}})(t) = (\mathbf{y}_k, \mathbf{u}_k, \mathbf{\Phi}_k)(t)$ for all $t \in (\bar{t}, k)$. Therefore we have $\mathbf{\Phi}_k^{\bar{t}}(\bar{t}) = \mathbf{\Phi}_k(\bar{t}) \in \partial_y \varphi_k(\bar{t}, \mathbf{y}_k(\bar{t}))$; that is, $\mathbf{\Phi}_k(\bar{t}) = \widehat{\Pi}(k-\bar{t})\mathbf{y}_k(\bar{t})$. Thus

$$|\mathbf{\Phi}_k(\bar{t})|_{\mathbf{V}_n^0(\Omega)} \le \|\widehat{\Pi}(k-\bar{t})\| |\mathbf{y}_k(\bar{t})|_{\mathbf{V}_n^0(\Omega)} \le C|\mathbf{y}_k(\bar{t})|_{\mathbf{V}_n^0(\Omega)}$$

and

$$\|\tilde{\mathbf{\Phi}}_k\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \le C\|\tilde{\mathbf{y}}_k\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))},$$

where $\tilde{\mathbf{\Phi}}_k$ (respectively, $\tilde{\mathbf{y}}_k$) is the extension by zero of $\mathbf{\Phi}_k$ (respectively, $\mathbf{y}_k$) to (k,∞). In the proof of Theorem 4.1 we have shown that $(\tilde{\mathbf{y}}_k)_k$ is bounded in $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ and that it converges to $\hat{\mathbf{y}}$ in $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$. Therefore $(\tilde{\mathbf{\Phi}}_k)_k$ is also bounded in $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$. Observe that $\tilde{\mathbf{\Phi}}_k$ is also the solution of the equation

$$-\tilde{\mathbf{\Phi}}_k' = (A^* - \lambda_0 I)\tilde{\mathbf{\Phi}}_k + \tilde{\mathbf{y}}_k + \lambda_0 \tilde{\mathbf{\Phi}}_k, \qquad \tilde{\mathbf{\Phi}}_k(\infty) = 0.$$

Thus

$$\tilde{\mathbf{\Phi}}_k(t) = \int_t^\infty e^{(A^* - \lambda_0 I)(\tau - t)}(\tilde{\mathbf{y}}_k(\tau) + \lambda_0 \tilde{\mathbf{\Phi}}_k(\tau))\, d\tau \qquad \text{for all } t \ge 0.$$

From Young's inequality for convolutions it follows that $(\tilde{\mathbf{\Phi}}_k)_k$ is also bounded in $L^\infty(0,\infty;\mathbf{V}_n^0(\Omega))$. There then exists $\hat{\mathbf{\Phi}} \in L^\infty(0,\infty;\mathbf{V}_n^0(\Omega)) \cap L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ such that, after extraction of a subsequence, we have

$$\tilde{\mathbf{\Phi}}_k \rightharpoonup \hat{\mathbf{\Phi}} \quad \text{weakly in } L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \quad \text{and} \quad \text{weak-star in } L^\infty(0,\infty;\mathbf{V}_n^0(\Omega)),$$

and $\hat{\mathbf{\Phi}}$ obeys the equation

$$\hat{\mathbf{\Phi}}(t) = \int_t^\infty e^{(A^*-\lambda_0 I)(\tau-t)}(\hat{\mathbf{y}}(\tau) + \lambda_0\hat{\mathbf{\Phi}}(\tau))\,d\tau \qquad \text{for all } t \geq 0.$$

Step 1. *Regularity of* $\hat{\mathbf{\Phi}}$. We have

$$-\hat{\mathbf{\Phi}}' = (A^* - \lambda_0 I)\hat{\mathbf{\Phi}} + \lambda_0\hat{\mathbf{\Phi}} + \hat{\mathbf{y}} \quad \text{in } (0,\infty), \qquad \hat{\mathbf{\Phi}}(\infty) = 0\,.$$

Since $\hat{\mathbf{\Phi}}$ belongs to $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$, we deduce that $\hat{\mathbf{\Phi}}$ belongs to $L^2(0,\infty;\mathbf{V}_0^1(\Omega)) \cap \mathbf{V}^{2,1}(Q_\infty)$ (see Lemma 8.5). Moreover, due to Lemma 8.9, the sequence $(\tilde{\mathbf{u}}_k)_k = (-MB_\tau^*\tilde{\mathbf{\Phi}}_k - R_A^{-1}MB_n^*\tilde{\mathbf{\Phi}}_k)_k$ converges weakly in $L^2(0,\infty;\mathbf{V}^0(\Gamma))$ to $-MB_\tau^*\hat{\mathbf{\Phi}} - R_A^{-1}MB_n^*\hat{\mathbf{\Phi}}$. Thus $\hat{\mathbf{u}} = -MB_\tau^*\hat{\mathbf{\Phi}} - R_A^{-1}MB_n^*\hat{\mathbf{\Phi}}$, and $(\hat{\mathbf{y}},\hat{\mathbf{\Phi}})$ obeys the first two equations in (4.4).

Step 2. Let us show that if $(\mathbf{y},\mathbf{\Phi}) \in L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \times \mathbf{V}^{2,1}(Q_\infty)$ is a solution to system (4.4), then $\mathbf{y}$ belongs $C_b([0,\infty);\mathbf{V}_n^0(\Omega) \cap \mathbf{V}^{1,1/2}(Q_\infty)$, $\mathbf{\Phi} \in \mathbf{V}^{3,3/2}(Q_\infty)$, and

$$\begin{aligned} &\|\mathbf{y}\|_{C_b([0,\infty);\mathbf{V}_n^0(\Omega))} + \|\mathbf{y}\|_{\mathbf{V}^{1,1/2}(Q_\infty)} + \|\mathbf{\Phi}\|_{\mathbf{V}^{3,3/2}(Q_\infty)} \\ &\leq C(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)} + \|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))})\,. \end{aligned} \tag{4.5}$$

To establish this result we rewrite system (4.4) as follows:

$$\begin{aligned} &\mathbf{y}' = (A - \lambda_0 I)\mathbf{y} - B_\tau M^2 B_\tau^*\mathbf{\Phi} - B_n M R_A^{-1} M B_n^*\mathbf{\Phi} + \lambda_0\mathbf{y} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = \mathbf{y}_0, \\ &-\mathbf{\Phi}' = (A^* - \lambda_0 I)\mathbf{\Phi} + \mathbf{y} + \lambda_0\mathbf{\Phi} \quad \text{in } (0,\infty), \qquad \mathbf{\Phi}(\infty) = 0. \end{aligned} \tag{4.6}$$

Due to Lemma 8.9 we know that $B^*\mathbf{\Phi} \in L^2(0,\infty;\mathbf{V}^0(\Gamma))$. Applying Lemmas 8.1 and 8.3, we obtain

$$\|\mathbf{y}\|_{\mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(Q_\infty)} \leq C(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)} + \|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))}) \quad \text{for all } \varepsilon' > 0.$$

Still from Lemma 8.9, we deduce that $B^*\mathbf{\Phi} \in \mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(\Sigma_\infty)$ for all $\varepsilon' > 0$. Applying successively Lemmas 8.1, 8.3, and 8.9, we can prove that $\mathbf{y}$ belongs to $\mathbf{V}^{1-\varepsilon',1/2-\varepsilon'/2}(Q_\infty)$ and $B^*\mathbf{\Phi}$ belongs to $\mathbf{V}^{1-\varepsilon',1/2-\varepsilon'/2}(\Sigma_\infty)$ for all $\varepsilon' > 0$. Another iteration gives $\mathbf{y} \in \mathbf{V}^{1,1/2}(Q_\infty) \cap C_b([0,\infty);\mathbf{V}_n^0(\Omega))$ (because $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$). From Lemma 8.5 we deduce that $\mathbf{\Phi} \in \mathbf{V}^{3,3/2}(Q_\infty)$, and the estimate (4.5) holds true.

Step 3. *We show that the pair* $(\hat{\mathbf{y}},\hat{\mathbf{\Phi}})$ *obeys the third equation in* (4.4). With Lemma 8.5 we can show that

$$\tilde{\mathbf{\Phi}}_k(t) \rightharpoonup \hat{\mathbf{\Phi}}(t) \quad \text{weakly in } \mathbf{V}_n^0(\Omega) \qquad \text{for all } t \geq 0.$$

Since

$$\mathbf{\Phi}_k(t) \in \partial_{\mathbf{y}}\varphi_k(t,\mathbf{y}_k(t)), \qquad \mathbf{\Phi}_k(t) \rightharpoonup \hat{\mathbf{\Phi}}(t) \quad \text{weakly in } \mathbf{V}_n^0(\Omega),$$

and

$$\varphi_k(t,\mathbf{y}_k(t)) \to \varphi(\hat{\mathbf{y}}(t)) \quad \text{as } k \to \infty,$$

we deduce that

$$\hat{\mathbf{\Phi}}(t) \in \partial\varphi(\hat{\mathbf{y}}(t)), \quad \text{i.e.,} \quad \hat{\mathbf{\Phi}}(t) = \Pi\hat{\mathbf{y}}(t).$$

Thus we have shown that $(\hat{\mathbf{y}},\hat{\mathbf{\Phi}})$ obeys the third equation in (4.4).

Step 4. *Uniqueness.* If a solution $(\mathbf{y}, \mathbf{\Phi}) \in L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \times \mathbf{V}^{2,1}(Q_\infty)$ is a solution to system (4.4), due to Step 2 it obeys (4.5), and we can show that

$$\int_0^k |\mathbf{y}(t)|^2_{\mathbf{V}_n^0(\Omega)}dt + \int_0^k |MB_\tau^*\mathbf{\Phi}(t)|^2_{\mathbf{V}^0(\Gamma)}dt + \int_0^k |R_A^{-1}MB_n^*\mathbf{\Phi}(t)|^2_{\mathbf{V}^0(\Gamma)}dt$$
$$= \int_\Omega \mathbf{y}_0\mathbf{\Phi}(0) - \int_\Omega \mathbf{y}(k)\mathbf{\Phi}(k).$$

Passing to the limit when k tends to infinity, we obtain

$$\int_0^\infty |\mathbf{y}(t)|^2_{\mathbf{V}_n^0(\Omega)}dt + \int_0^\infty |MB_\tau^*\mathbf{\Phi}(t)|^2_{\mathbf{V}^0(\Gamma)}dt + \int_0^\infty |R_A^{-1}MB_n^*\mathbf{\Phi}(t)|^2_{\mathbf{V}^0(\Gamma)}dt = \int_\Omega \mathbf{y}_0\mathbf{\Phi}(0).$$

Thus if $\mathbf{y}_0 = 0$, we have $\mathbf{y} = 0$. From the relation $\mathbf{\Phi} = \Pi\mathbf{y}$ we deduce that $\mathbf{\Phi} = 0$, and the uniqueness is established.

Step 5. *Final estimate.* From the previous steps it follows that $(\hat{\mathbf{y}}, \hat{\mathbf{\Phi}})$ is the unique solution to system (4.4). Since $\|\hat{\mathbf{y}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \le C|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)}$, we have

$$\|\hat{\mathbf{\Phi}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \le C|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)},$$

and the estimate of the lemma follows from (4.5). ∎

COROLLARY 4.3. *If* $\mathbf{y}_0 \in \mathbf{V}_n^{1/2-\varepsilon}(\Omega)$ *for some* $0 < \varepsilon \le 1/2$, *then the solution* $(\mathbf{y}, \mathbf{\Phi})$ *of system* (4.4) *belongs to* $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty) \times \mathbf{V}^{7/2-\varepsilon,7/4-\varepsilon/2}(Q_\infty)$, *and we have*

$$\|\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} + \|\mathbf{\Phi}\|_{\mathbf{V}^{7/2-\varepsilon,7/4-\varepsilon/2}(Q_\infty)} + \|B^*\mathbf{\Phi}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}$$
$$\le C|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)}.$$

Proof. With Lemmas 4.2 and 8.9, we first prove that $\mathbf{u} \in \mathbf{V}^{1,1/2}(\Sigma_\infty)$. Applying Lemmas 8.1 and 8.3, it follows that $\mathbf{y}$ belongs to $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$. The estimate for $\mathbf{\Phi}$ follows from Lemma 8.5, and we deduce that $B^*\mathbf{\Phi} \in \mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$ with Lemma 8.9. ∎

Remark 4.4. Due to Lemma 4.2, if $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$, then $\mathbf{\Phi}$ belongs to $C([0,T];\mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$. In particular, $\mathbf{\Phi}(0)$ belongs to $\mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega)$. From Corollary 4.3 we deduce that if $\mathbf{y}_0 \in \mathbf{V}_n^{1/2-\varepsilon}(\Omega)$, then $\mathbf{\Phi} \in C([0,T];\mathbf{V}^{5/2-\varepsilon}(\Omega) \cap \mathbf{V}_0^1(\Omega))$ and $\mathbf{\Phi}(0)$ belongs to $\mathbf{V}^{5/2-\varepsilon}(\Omega) \cap \mathbf{V}_0^1(\Omega)$. This means that Π belongs to $\mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$, and Π also belongs to $\mathcal{L}(\mathbf{V}_n^{1/2-\varepsilon}(\Omega), \mathbf{V}^{5/2-\varepsilon}(\Omega) \cap \mathbf{V}_0^1(\Omega))$.

Since B^* is continuous from $\mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega)$ into $\mathbf{V}^{1/2}(\Gamma)$ (see Proposition 2.7), the operator $B^*\Pi$ belongs to $\mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^{1/2}(\Gamma))$, and $B^*\Pi$ also belongs to $\mathcal{L}(\mathbf{V}_n^{1/2-\varepsilon}(\Omega), \mathbf{V}^{1-\varepsilon}(\Gamma))$ for all $\varepsilon > 0$.

From Lemma 4.2 it follows that, for all $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$, the evolution equation

$$\mathbf{y}' = A\mathbf{y} - B_\tau M^2 B_\tau^*\Pi\mathbf{y} - B_n M R_A^{-1} M B_n^*\Pi\mathbf{y} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = \mathbf{y}_0, \tag{4.7}$$

admits at least one weak solution belonging to $L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \cap C_b(\mathbb{R};\mathbf{V}_n^0(\Omega))$. It is easy to show that this solution is unique. Due to Lemma 4.2, it is equal to $\mathbf{y}_{\mathbf{y}_0}$, where $(\mathbf{y}_{\mathbf{y}_0}, \mathbf{u}_{\mathbf{y}_0})$ is the solution of $(\mathcal{P}_{0,\mathbf{y}_0})$. Still from Lemma 4.2, we deduce that the family of operators $(S(t))_{t\ge0}$ defined by

$$S(t) \ : \ \mathbf{y}_0 \longmapsto \mathbf{y}_{\mathbf{y}_0}(t)$$

is a strongly continuous exponentially stable semigroup on $\mathbf{V}_n^0(\Omega)$. Let $(A_\Pi, D(A_\Pi))$ be its infinitesimal generator, and let us denote by $(e^{tA_\Pi})_{t\geq 0}$ the semigroup $(S(t))_{t\geq 0}$. Since the semigroup $(e^{tA_\Pi})_{t\geq 0}$ is exponentially stable on $\mathbf{V}_n^0(\Omega)$, we can define $D(A_\Pi)$ by

$$D(A_\Pi) = \left\{ \mathbf{y} \in \mathbf{V}_n^0(\Omega) \mid \mathbf{y} = -\int_0^\infty e^{\tau A_\Pi}\mathbf{f}\, d\tau \text{ for some } \mathbf{f} \in \mathbf{V}_n^0(\Omega) \right\},$$

and $A_\Pi \mathbf{y} = \mathbf{f}$ if $\mathbf{y} = -\int_0^\infty e^{\tau A_\Pi}\mathbf{f}\, d\tau \in D(A_\Pi)$.

We want to find a more explicit characterization of $D(A_\Pi)$ and A_Π. From Remark 4.4, it follows that the operator $A - BMR_A^{-1}MB^*\Pi$ may be considered as a bounded operator from $\mathbf{V}_n^0(\Omega)$ into $(D(A_0))'$, the dual of $D(A_0)$ with respect to the pivot space $\mathbf{V}_n^0(\Omega)$. We want to prove that

$$D(A_\Pi) = \left\{ \mathbf{y} \in \mathbf{V}_n^0(\Omega) \mid A - BMR_A^{-1}MB^*\Pi \in \mathbf{V}_n^0(\Omega) \right\}. \tag{4.8}$$

For that we are going to use the characterization of $(A_\Pi^*, D(A_\Pi^*))$, the adjoint of $(A_\Pi, D(A_\Pi))$. Let us define the unbounded operator $(A_\Pi^\sharp, D(A_\Pi^\sharp))$ in $\mathbf{V}_n^0(\Omega)$ by

$$D(A_\Pi^\sharp) = D(A_0) \quad \text{and} \quad A_\Pi^\sharp \mathbf{\Phi} = A^*\mathbf{\Phi} - (B^*\Pi)^* MR_A^{-1}MB^*\mathbf{\Phi} \quad \text{for all } \mathbf{\Phi} \in D(A_\Pi^\sharp),$$

where $(B^*\Pi)^* \in \mathcal{L}(\mathbf{V}^{-1/2}(\Gamma), \mathbf{V}_n^0(\Omega))$ is the adjoint of $B^*\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^{1/2}(\Gamma))$. We know that $(A^* - \lambda_0 I, D(A^*))$ is the infinitesimal generator of a bounded analytic semigroup on $\mathbf{V}_n^0(\Omega)$. Let $(L, D(L))$ be the unbounded operator in $\mathbf{V}_n^0(\Omega)$ defined by $D(L) = \mathbf{V}^{3/2+\varepsilon}(\Omega) \cap \mathbf{V}_0^1(\Omega)$ with $0 < \varepsilon < 1/2$, and $L\mathbf{\Phi} = (B^*\Pi)^* MR_A^{-1}MB^*\mathbf{\Phi}$. We can easily verify that L is $(A^* - \lambda_0 I)$-bounded with relative bound zero. From [21, Corollary 2.5, p. 500] it follows that $(A_\Pi^\sharp - \lambda_0 I, D(A_0))$ is the infinitesimal generator of a quasi-bounded analytic semigroup on $\mathbf{V}_n^0(\Omega)$. Thus $(A_\Pi^\sharp, D(A_0))$ is also the infinitesimal generator of an analytic semigroup on $\mathbf{V}_n^0(\Omega)$. Consider the equations

$$\mathbf{\Phi}' = A_\Pi^\sharp \mathbf{\Phi}, \qquad \mathbf{\Phi}(0) = \mathbf{\Phi}_0, \tag{4.9}$$

and

$$-\hat{\mathbf{\Phi}}' = A_\Pi^\sharp \hat{\mathbf{\Phi}}, \qquad \hat{\mathbf{\Phi}}(T) = \mathbf{\Phi}_0. \tag{4.10}$$

We clearly have $\mathbf{\Phi}(T-t) = \hat{\mathbf{\Phi}}(t)$ for all $t \in [0,T]$. If $\mathbf{\Phi}_0 \in D(A_0)$, then the solution $\mathbf{\Phi}$ to (4.9) belongs to $C([0,T]; D(A_0)) \cap C^1([0,T]; D(A_0))$. With (4.7) and (4.10), we can show that

$$\left(\mathbf{y}_0, e^{TA_\Pi^\sharp}\mathbf{\Phi}_0\right)_{\mathbf{V}_n^0(\Omega)} = \left(e^{TA_\Pi}\mathbf{y}_0, \mathbf{\Phi}_0\right)_{\mathbf{V}_n^0(\Omega)} = \left(\mathbf{y}_0, e^{TA_\Pi^*}\mathbf{\Phi}_0\right)_{\mathbf{V}_n^0(\Omega)}$$

for all $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$, $\mathbf{\Phi}_0 \in \mathbf{V}_n^0(\Omega)$, and $T > 0$. Thus the two semigroups $\left(e^{TA_\Pi^\sharp}\right)_{t\geq 0}$ and $\left(e^{TA_\Pi^*}\right)_{t\geq 0}$ on $\mathbf{V}_n^0(\Omega)$ are identical, and we have $(A_\Pi^\sharp, D(A_0)) = (A_\Pi^*, D(A_\Pi^*))$. For all $\mathbf{g} \in \mathbf{V}_n^0(\Omega)$, the equation $A_\Pi^*\mathbf{\Phi} = \mathbf{g}$ admits $-\int_0^\infty e^{\tau A_\Pi^*}\mathbf{g}\, d\tau$ as unique solution in $D(A_\Pi^*)$. Thus, for all $\mathbf{g} \in \mathbf{V}_n^0(\Omega)$, the solution $-\int_0^\infty e^{\tau A_\Pi^*}\mathbf{g}\, d\tau$ to the equation $A_\Pi^*\mathbf{\Phi} = A_\Pi^\sharp\mathbf{\Phi} = \mathbf{g}$ belongs to $D(A_0)$.

Let us come back to the characterization of $D(A_\Pi)$. Let $\mathbf{y}$ be in $D(A_\Pi)$, and set $A_\Pi\mathbf{y} = \mathbf{f}$. Setting $\mathbf{z}(t) = -\int_0^t e^{\tau A_\Pi}\mathbf{f}\, d\tau = -\int_0^t e^{(t-s)A_\Pi}\mathbf{f}\, ds$, using the definition of

weak solution for $\mathbf{z}$, by passing to the limit when t tends to infinity, we can show that $\mathbf{y}$ satisfies the equation

$$A\mathbf{y} - BMR_A^{-1}MB^*\Pi\mathbf{y} = \mathbf{f}, \tag{4.11}$$

where $\mathbf{f} = A_\Pi\mathbf{y}$. Thus

$$D(A_\Pi) \subset \Big\{\mathbf{y} \in \mathbf{V}_n^0(\Omega) \mid A - BMR_A^{-1}MB^*\Pi \in \mathbf{V}_n^0(\Omega)\Big\}.$$

In (4.11), $A\mathbf{y} - BMR_A^{-1}MB^*\Pi\mathbf{y}$ is considered as an element in $(D(A_0))'$. This means that $\mathbf{y}$ is a solution to (4.11) if and only if

$$\int_\Omega \mathbf{y}A^*\mathbf{\Phi} - \int_\Gamma MR_A^{-1}MB^*\Pi\,\mathbf{y}\,B^*\mathbf{\Phi} = \int_\Omega \mathbf{y}A_\Pi^*\mathbf{\Phi} = \int_\Omega \mathbf{f}\,\mathbf{\Phi} \quad \text{for all } \mathbf{\Phi} \in D(A_0). \tag{4.12}$$

To prove (4.8), it is now sufficient to show that the solution to (4.11) or (4.12) is unique. Assume that $\mathbf{y} \in \mathbf{V}_n^0(\Omega)$ obeys (4.12) with $\mathbf{f} = 0$. If we take the solution $\mathbf{\Phi}$ of $A_\Pi^*\mathbf{\Phi} = \mathbf{y}$ in (4.12), we obtain $\int_\Omega |\mathbf{y}|^2 = 0$. Thus $\mathbf{y} = 0$, and the equality (4.8) is established.

THEOREM 4.5. *The unbounded operator $(A_\Pi, D(A_\Pi))$ defined by*

$$D(A_\Pi) = \Big\{\mathbf{y} \in \mathbf{V}_n^0(\Omega) \mid A - BMR_A^{-1}MB^*\Pi \in \mathbf{V}_n^0(\Omega)\Big\},$$
$$A_\Pi\mathbf{y} = A\mathbf{y} - BMR_A^{-1}MB^*\Pi\mathbf{y} \quad \textit{for all } \mathbf{y} \in D(A_\Pi),$$

is the infinitesimal generator of an analytic exponentially stable semigroup on $\mathbf{V}_n^0(\Omega)$.

The operator Π is the unique weak solution to the algebraic Riccati equation

$$\Pi^* = \Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega)) \quad \textit{and} \quad \Pi \geq 0,$$
$$\textit{for all } \mathbf{y} \in \mathbf{V}_n^0(\Omega), \qquad \Pi\mathbf{y} \in \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega) \quad \textit{and} \quad |\Pi\mathbf{y}|_{\mathbf{V}^2(\Omega)} \leq C|\mathbf{y}|_{\mathbf{V}_n^0(\Omega)},$$
$$A^*\Pi + \Pi A - \Pi B_\tau M^2 B_\tau^*\Pi - \Pi B_n MR_A^{-1}MB_n^*\Pi + I = 0.$$

Moreover, $\Pi \in \mathcal{L}(\mathbf{V}_n^{1/2-\varepsilon}(\Omega); \mathbf{V}^{5/2-\varepsilon}(\Omega) \cap \mathbf{V}_0^1(\Omega))$.

Proof. The first part of the theorem is already proved. We have already shown that $\Pi^* = \Pi \geq 0$ and that $\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$. Since $\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$, it is relatively easy to prove that Π is the unique solution to the algebraic Riccati equation stated in the theorem. The uniqueness is shown in [23, Theorem 2.2.1] under a weaker condition than $\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$. □

5. Problems with a nonhomogeneous source term. For all $\bar{t} \geq 0$ and all $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$ we consider the problem

$$(\mathcal{P}_{\bar{t},\mathbf{y}_0,\mathbf{f}}) \qquad \inf\Big\{I(\bar{t},\mathbf{y},\mathbf{u}) \;\mid\; (\mathbf{y},\mathbf{u}) \text{ satisfies (5.1)},\ \mathbf{u} \in \mathbf{V}^{0,0}(\Sigma_{\bar{t},\infty})\Big\},$$

where

$$I(\bar{t},\mathbf{y},\mathbf{u}) = \frac{1}{2}\int_{\bar{t}}^\infty \int_\Omega |\mathbf{y}|^2\,dxdt + \frac{1}{2}\int_{\bar{t}}^\infty \Big(|\gamma_\tau\mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n\mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)}\Big)\,dt$$

and

$$\mathbf{y}' = A\mathbf{y} + BM\mathbf{u} + \mathbf{f} \quad \text{in } (\bar{t},\infty), \qquad \mathbf{y}(\bar{t}) = \mathbf{y}_0. \tag{5.1}$$

The function $\mathbf{f}$ belongs to $L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))\cap L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')$, with $0<\varepsilon<1/2$. In this section we want to study the regularity of solutions to the control problem $(\mathcal{P}_{0,\mathbf{y}_0,\mathbf{f}})$ according to the regularity of $\mathbf{y}_0$. This result will be used in the next section to study the local stabilization of the two-dimensional Navier–Stokes equations.

THEOREM 5.1. *For all $\bar t\ge 0$, $\mathbf{y}_0\in\mathbf{V}_n^0(\Omega)$, and $\mathbf{f}\in L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))$, problem $(\mathcal{P}_{\bar t,\mathbf{y}_0,\mathbf{f}})$ admits a unique solution $(\mathbf{y}_{\bar t,\mathbf{y}_0,\mathbf{f}},\mathbf{u}_{\bar t,\mathbf{y}_0,\mathbf{f}})$, and the optimal cost $\varphi(\bar t,\mathbf{y}_0,\mathbf{f})=J(\bar t,\mathbf{y}_{\bar t,\mathbf{y}_0,\mathbf{f}},\mathbf{u}_{\bar t,\mathbf{y}_0,\mathbf{f}})$ obeys*

$$\varphi(\bar t,\mathbf{y}_0,\mathbf{f})\le C\Big(|\mathbf{y}_0|^2_{\mathbf{V}_n^0(\Omega)}+\|\mathbf{f}\|^2_{L^{2/(1+2\varepsilon)}(\bar t,\infty;\mathbf{V}_n^0(\Omega))}\Big),\tag{5.2}$$

where the constant C is independent of $\bar t$.

Proof. The semigroup generated by A_Π on $\mathbf{V}_n^0(\Omega)$ is analytic and exponentially stable. Thus the solution to the equation

$$\mathbf{z}'=A_\Pi\mathbf{z}+\mathbf{f},\qquad \mathbf{z}(\bar t)=\mathbf{y}_0,$$

is defined by

$$\mathbf{z}(t)=e^{(t-\bar t)A_\Pi}\mathbf{y}_0+\int_{\bar t}^t e^{(t-\tau)A_\Pi}\mathbf{f}(\tau)\,d\tau,$$

and it obeys

$$|\mathbf{z}(t)|_{\mathbf{V}_n^0(\Omega)}\le Ce^{-\omega(t-\bar t)}|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)}+C\int_{\bar t}^t e^{-\omega(t-\tau)}|\mathbf{f}(\tau)|_{\mathbf{V}_n^0(\Omega)}\,d\tau,$$

for some $C>0$ and some $\omega>0$. It follows that

$$\|\mathbf{z}\|_{L^2(\bar t,\infty;\mathbf{V}_n^0(\Omega))}\le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)}+\|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(\bar t,\infty;\mathbf{V}_n^0(\Omega))}\big),$$

where the constant C is independent of $\bar t$. Since $MB_\tau^*\Pi$ and $R_A^{-1}MB_n^*\Pi$ are bounded operators from $\mathbf{V}_n^0(\Omega)$ into $\mathbf{V}^0(\Gamma)$ (see Remark 4.4 and Lemma 2.5), we also have

$$\begin{aligned}&\|MB_\tau^*\Pi\mathbf{z}\|_{L^2(\bar t,\infty;\mathbf{V}^0(\Gamma))}+\|R_A^{-1}MB_n^*\Pi\mathbf{z}\|_{L^2(\bar t,\infty;\mathbf{V}^0(\Gamma))}\\&\quad\le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)}+\|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(\bar t,\infty;\mathbf{V}_n^0(\Omega))}\big).\end{aligned}$$

The pair $(\mathbf{z},-MB_\tau^*\Pi\mathbf{z}-R_A^{-1}MB_n^*\Pi\mathbf{z})$ is admissible for $(\mathcal{P}_{\bar t,\mathbf{y}_0,\mathbf{f}})$, and we have

$$I(\bar t,\mathbf{z},-MB_\tau^*\Pi\mathbf{z}-R_A^{-1}MB_n^*\Pi\mathbf{z})\le C\big(|\mathbf{y}_0|^2_{\mathbf{V}_n^0(\Omega)}+\|\mathbf{f}\|^2_{L^{2/(1+2\varepsilon)}(\bar t,\infty;\mathbf{V}_n^0(\Omega))}\big).$$

Therefore by classical arguments we can prove that $(\mathcal{P}_{\bar t,\mathbf{y}_0,\mathbf{f}})$ admits a unique solution $(\mathbf{y}_{\bar t,\mathbf{y}_0,\mathbf{f}},\mathbf{u}_{\bar t,\mathbf{y}_0,\mathbf{f}})$ and that the optimal cost $\varphi(\bar t,\mathbf{y}_0,\mathbf{f})=J(\bar t,\mathbf{y}_{\bar t,\mathbf{y}_0,\mathbf{f}},\mathbf{u}_{\bar t,\mathbf{y}_0,\mathbf{f}})$ obeys (5.2). ∎

LEMMA 5.2. *For all $\mathbf{u}\in L^2(0,\infty;\mathbf{V}^0(\Gamma))$, $\mathbf{y}\in L^2(0,\infty;\mathbf{V}_n^0(\Omega))$, the equation*

$$-\mathbf{\Phi}'=A_\Pi^*\mathbf{\Phi}-(MB_\tau^*\Pi)^*\gamma_\tau\mathbf{u}-(R_A^{-1}MB_n^*\Pi)^*\gamma_n\mathbf{u}+\mathbf{y}\quad\text{in }(0,\infty),\qquad\mathbf{\Phi}(\infty)=0,\tag{5.3}$$

admits a unique solution in $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ and

$$\|\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))}+\|\mathbf{\Phi}\|_{L^\infty(0,\infty;\mathbf{V}_n^0(\Omega))}\le C(\|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))}+\|\mathbf{u}\|_{L^2(0,\infty;\mathbf{V}^0(\Gamma))}).$$

Proof. We already know that $(MB_\tau^*\Pi)^* \in \mathcal{L}(\mathbf{V}^0(\Gamma), \mathbf{V}_n^0(\Omega))$ and $(R_A^{-1}MB_n^*\Pi)^* \in \mathcal{L}(\mathbf{V}^0(\Gamma), \mathbf{V}_n^0(\Omega))$. Thus from the exponential stability of the semigroup $(e^{tA_\Pi^*})_{t\geq 0}$ we find

$$|\mathbf{\Phi}(t)|_{\mathbf{V}_n^0(\Omega)} \leq C\int_t^\infty e^{-\omega(\tau-t)}(|\mathbf{y}(\tau)|_{\mathbf{V}_n^0(\Omega)} + |\mathbf{u}(\tau)|_{\mathbf{V}^0(\Gamma)})d\tau.$$

The estimates of the lemma follow from Young's inequality for convolutions. The uniqueness of the solution is obvious. ∎

LEMMA 5.3. *Let* $\mathbf{f}$ *be in* $L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))$, $\mathbf{y}_0$ *be in* $\mathbf{V}_n^0(\Omega)$, *denote by* $(\hat{\mathbf{y}},\hat{\mathbf{u}})$ *the solution to problem* $(\mathcal{P}_{0,\mathbf{y}_0,\mathbf{f}})$, *and let* $\hat{\mathbf{\Phi}}$ *be the solution to* (5.3) *corresponding to* $(\hat{\mathbf{y}},\hat{\mathbf{u}})$. *Then* $(\hat{\mathbf{y}},\hat{\mathbf{\Phi}})$ *is also a solution in* $L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \times L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ *to the system*

$$\begin{aligned} &\mathbf{y}' = A\mathbf{y} - B_\tau M^2 B_\tau^*\mathbf{\Phi} - B_n M R_A^{-1} M B_n^*\mathbf{\Phi} + \mathbf{f} \quad \textit{in } (0,\infty), \qquad \mathbf{y}(0) = \mathbf{y}_0,\\ &-\mathbf{\Phi}' = A^*\mathbf{\Phi} + \mathbf{y} \quad \textit{in } (0,\infty), \qquad \mathbf{\Phi}(\infty) = 0. \end{aligned} \tag{5.4}$$

The following estimate holds:

$$\|\hat{\mathbf{y}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\hat{\mathbf{\Phi}}\|_{\mathbf{V}^{2,1}(Q_\infty)} \leq C\big(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}\big).$$

Proof. Step 1. Due to estimate (5.2), we have

$$\|\hat{\mathbf{y}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\hat{\mathbf{u}}\|_{L^2(0,\infty;\mathbf{V}^0(\Gamma))} \leq C\Big(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}\Big).$$

Consider the problem

$$(\mathcal{P}_{0,\mathbf{y}_0,\mathbf{f}}^k) \qquad \inf\Big\{I_k(0,\mathbf{y},\mathbf{u}) \;\mid\; (\mathbf{y},\mathbf{u}) \text{ satisfies (5.5)},\ \mathbf{u}\in\mathbf{V}^{0,0}(\Sigma_{0,k})\Big\},$$

where

$$I_k(0,\mathbf{y},\mathbf{u}) = \frac{1}{2}\int_0^k\int_\Omega |\mathbf{y}|^2\,dxdt + \frac{1}{2}\int_0^k \Big(|\gamma_\tau\mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n\mathbf{u}(t)|^2_{\mathbf{V}^0(\Gamma)}\Big)\,dt$$

and

$$\mathbf{y}' = A\mathbf{y} + BM\mathbf{u} + \mathbf{f} \quad \text{in } (0,k), \qquad \mathbf{y}(0) = \mathbf{y}_0. \tag{5.5}$$

Problem $(\mathcal{P}_{0,\mathbf{y}_0,\mathbf{f}}^k)$ admits a unique solution $(\mathbf{y}_k,\mathbf{u}_k)$ characterized by

$$\begin{aligned} &\mathbf{y}_k' = A\mathbf{y}_k + BM\mathbf{u}_k + \mathbf{f} \quad \text{in } (0,k), \qquad \mathbf{y}_k(0) = \mathbf{y}_0,\\ &-\mathbf{\Phi}_k' = A^*\mathbf{\Phi}_k + \mathbf{y}_k \quad \text{in } (0,k), \qquad \mathbf{\Phi}_k(k) = 0,\\ &\gamma_\tau\mathbf{u}_k = -MB_\tau^*\mathbf{\Phi}_k, \qquad \gamma_n\mathbf{u}_k = -R_A^{-1}MB_n^*\mathbf{\Phi}_k. \end{aligned} \tag{5.6}$$

Since we have

$$\begin{aligned} &\int_0^k\int_\Omega |\mathbf{y}_k|^2\,dxdt + \int_0^k \Big(|\gamma_\tau\mathbf{u}_k(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n\mathbf{u}_k(t)|^2_{\mathbf{V}^0(\Gamma)}\Big)\,dt\\ &\leq \int_0^\infty\int_\Omega |\hat{\mathbf{y}}|^2\,dxdt + \int_0^\infty \Big(|\gamma_\tau\hat{\mathbf{u}}(t)|^2_{\mathbf{V}^0(\Gamma)} + |R_A^{1/2}\gamma_n\hat{\mathbf{u}}(t)|^2_{\mathbf{V}^0(\Gamma)}\Big)\,dt, \end{aligned}$$

as in the proof of Theorem 4.1, we can show that

$$\tilde{\mathbf{u}}_k \to \hat{\mathbf{u}} \quad \text{in } L^2(0,\infty;\mathbf{V}^0(\Gamma)) \quad \text{and} \quad \tilde{\mathbf{y}}_k \to \hat{\mathbf{y}} \quad \text{in } L^2(0,\infty;\mathbf{V}^0(\Omega)), \tag{5.7}$$

where $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{y}}_k$ denote the extensions by zero of $\mathbf{u}_k$ and $\mathbf{y}_k$ to (k,∞).

Step 2. *Passage to the limit for* $\mathbf{\Phi}_k$. Let $\tilde{\mathbf{\Phi}}_k$ be the extension by zero of $\mathbf{\Phi}_k$ to (k,∞). We have

$$-\tilde{\mathbf{\Phi}}_k' = A^*\tilde{\mathbf{\Phi}}_k + \tilde{\mathbf{y}}_k \quad \text{in } (0,\infty), \qquad \tilde{\mathbf{\Phi}}_k(\infty) = 0\,.$$

We already know that $(MB_\tau^*\Pi)^* \in \mathcal{L}(\mathbf{V}^0(\Gamma),\mathbf{V}_n^0(\Omega))$ and $(R_A^{-1}MB_n^*\Pi)^* \in \mathcal{L}(\mathbf{V}^0(\Gamma),\mathbf{V}_n^0(\Omega))$. We can rewrite the above equation in the form

$$-\tilde{\mathbf{\Phi}}_k' = A_\Pi^*\tilde{\mathbf{\Phi}}_k - (MB_\tau^*\Pi)^*\gamma_\tau\tilde{\mathbf{u}}_k - (R_A^{-1}MB_n^*\Pi)^*\gamma_n\tilde{\mathbf{u}}_k + \tilde{\mathbf{y}}_k \quad \text{in } (0,\infty), \qquad \tilde{\mathbf{\Phi}}_k(\infty) = 0. \tag{5.8}$$

Due to (5.7) and to Lemma 5.2, we can claim that

$$\|\tilde{\mathbf{\Phi}}_k - \hat{\mathbf{\Phi}}\|_{L^\infty(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\tilde{\mathbf{\Phi}}_k - \hat{\mathbf{\Phi}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \longrightarrow 0 \quad \text{when } k \longrightarrow \infty,$$

where $\hat{\mathbf{\Phi}}$ is the solution of (5.3) corresponding to $(\hat{\mathbf{y}},\hat{\mathbf{u}})$. Notice that

$$\begin{aligned} &\|\hat{\mathbf{\Phi}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \le C\big(\|\hat{\mathbf{u}}\|_{L^2(0,\infty;\mathbf{V}^0(\Gamma))} + \|\hat{\mathbf{y}}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))}\big) \\ &\le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}\big). \end{aligned} \tag{5.9}$$

By passing to the limit in the equation

$$\tilde{\mathbf{\Phi}}_k(t) = \int_t^\infty e^{(A^*-\lambda_0 I)(\tau-t)}(\tilde{\mathbf{y}}_k(\tau) + \lambda_0\tilde{\mathbf{\Phi}}_k(\tau))\,d\tau,$$

with Lemmas 8.6 and 8.9, we can show that $(\tilde{\mathbf{\Phi}}_k)_k$ converges to $\hat{\mathbf{\Phi}}$ in $\mathbf{V}^{2,1}(Q_\infty)$, $(B^*\tilde{\mathbf{\Phi}}_k)_k$ converges to $B^*\hat{\mathbf{\Phi}}$ in $L^2(0,\infty;\mathbf{V}^0(\Gamma))$, and $\hat{\mathbf{\Phi}}$ satisfies

$$\hat{\mathbf{\Phi}}(t) = \int_t^\infty e^{(A^*-\lambda_0 I)(\tau-t)}(\hat{\mathbf{y}}(\tau) + \lambda_0\hat{\mathbf{\Phi}}(\tau))\,d\tau.$$

Thus $\hat{\mathbf{\Phi}}$ satisfies the second equation in (5.4) corresponding to $\hat{\mathbf{y}}$. Since $(\tilde{\mathbf{u}}_k)_k = (-MB_\tau^*\tilde{\mathbf{\Phi}}_k - R_A^{-1}MB_n^*\tilde{\mathbf{\Phi}}_k)_k$ converges to $\hat{\mathbf{u}}$, we have $\hat{\mathbf{u}} = -MB_\tau^*\hat{\mathbf{\Phi}} - R_A^{-1}MB_n^*\hat{\mathbf{\Phi}}$, and $(\hat{\mathbf{y}},\hat{\mathbf{\Phi}})$ obeys the system (5.4). The estimate for $\hat{\mathbf{\Phi}}$ follows from Lemma 8.5 and from (5.9). □

THEOREM 5.4. *Assume that* $\mathbf{f} \in L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega)) \cap L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')$. *If* $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$, *then the solution* $(\mathbf{y},\mathbf{\Phi})$ *of system* (5.4) *belongs to* $\mathbf{V}^{1,1/2}(Q)\times\mathbf{V}^{3,3/2}(Q)$ *and we have*

$$\begin{aligned} &\|\mathbf{y}\|_{C_b([0,\infty);\mathbf{V}_n^0(\Omega))} + \|\mathbf{y}\|_{\mathbf{V}^{1,1/2}(Q_\infty)} + \|\mathbf{\Phi}\|_{\mathbf{V}^{3,3/2}(Q_\infty)} \\ &\le C(|\mathbf{y}_0|_{\mathbf{V}_n^0(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')})\,. \end{aligned}$$

If $\mathbf{y}_0 \in \mathbf{V}_n^{1/2-\varepsilon}(\Omega)$ *for some* $0<\varepsilon<1/2$, *then the solution* $(\mathbf{y},\mathbf{\Phi})$ *of system* (5.4) *belongs to* $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q)\times\mathbf{V}^{7/2-\varepsilon,7/4-\varepsilon/2}(Q)$, *and we have*

$$\begin{aligned} &\|\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q)} + \|\mathbf{\Phi}\|_{\mathbf{V}^{7/2-\varepsilon,7/4-\varepsilon/2}(Q)} \\ &\le C(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}) \end{aligned}$$

and

$$\|B^*\mathbf{\Phi}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q)} \le C(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)}+\|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}+\|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}).$$

Proof. Assume that $\mathbf{y}_0 \in \mathbf{V}_n^0(\Omega)$. Applying Lemmas 8.1, 8.2, and 8.3, we first obtain that $\mathbf{y}$ belongs to $\mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(Q_\infty)$ for all $\varepsilon' > 0$. From Lemma 8.9, we deduce that $B^*\mathbf{\Phi} \in \mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(\Sigma_\infty)$ for all $\varepsilon' > 0$. As in the proof of Lemma 4.2, we can use a bootstrap argument to show that $\mathbf{y} \in \mathbf{V}^{1,1/2}(Q_\infty) \cap C_b([0,\infty);\mathbf{V}_n^0(\Omega))$ and $\mathbf{\Phi} \in \mathbf{V}^{3,3/2}(Q_\infty)$, with the corresponding estimates for $\mathbf{y}$ and $\mathbf{\Phi}$. The same procedure can be used to prove the other estimates in the case when $\mathbf{y}_0 \in \mathbf{V}_n^{1/2-\varepsilon}(\Omega)$. ∎

6. Stabilization of the two-dimensional Navier–Stokes equations. Throughout this section, we assume that $N = 2$.

6.1. First stabilization result. Consider the Navier–Stokes equations with the linear feedback law determined in section 4:

$$\begin{aligned} &P\mathbf{y}' = A_\Pi P\mathbf{y} + PF(\mathbf{y}) \quad \text{in } (0,\infty), \qquad P\mathbf{y}(0) = \mathbf{y}_0, \\ &(I-P)\mathbf{y} = -(I-P)D_A M R_A^{-1} M B_n^* \Pi P\mathbf{y} \quad \text{in } (0,\infty), \end{aligned} \tag{6.1}$$

where $F(\mathbf{y}) = -(\mathbf{y}\cdot\nabla)\mathbf{y}$.

THEOREM 6.1. *For all $0 < \varepsilon < 1/4$, there exist $\mu_0 > 0$ and a nondecreasing function η from $\mathbb{R}^+$ into itself such that if $\mu \in (0,\mu_0)$ and $|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} \le \eta(\mu)$, then* (6.1) *admits a unique solution in the set*

$$D_\mu = \Big\{\mathbf{y} \in \mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty) \mid \|\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \le \mu\Big\}.$$

Moreover, $(I-P)\mathbf{y}$ belongs to $H^{3/4-\varepsilon/2}(0,\infty;\mathbf{V}^{1/2}(\Omega)) \cap L^2(0,\infty;\mathbf{V}^{3/2-\varepsilon}(\Omega))$.

LEMMA 6.2. *For all $\mathbf{y}_0 \in \mathbf{V}_n^{1/2-\varepsilon}(\Omega)$ and all $\mathbf{f} \in L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega)) \cap L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')$ the solution to the equation*

$$\mathbf{y}' = A_\Pi \mathbf{y} + \mathbf{f}, \qquad \mathbf{y}(0) = \mathbf{y}_0, \tag{6.2}$$

obeys

$$\|\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \le C_1\big(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)}+\|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}+\|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big).$$

Proof. Since the semigroup $(e^{tA_\Pi})_{t\ge 0}$ is exponentially stable on $\mathbf{V}_n^0(\Omega)$, equation (6.2) admits a unique solution $\mathbf{y}$ belonging to $C_b([0,\infty);\mathbf{V}_n^0(\Omega)) \cap L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ which satisfies

$$\|\mathbf{y}\|_{L^\infty(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}\big).$$

We denote by $(\hat{\mathbf{y}},\hat{\mathbf{\Phi}})$ the solution to system (5.4). We set

$$\mathbf{r}(t) = \hat{\mathbf{\Phi}}(t) - \Pi\mathbf{y}(t).$$

We denote by $\tilde{\mathbf{y}}$ the solution to the equation

$$\tilde{\mathbf{y}}' = A\tilde{\mathbf{y}} + B_\tau M^2 B_\tau^*\mathbf{r} + B_n M R_A^{-1} M B_n^*\mathbf{r}, \qquad \tilde{\mathbf{y}}(0) = 0.$$

From the definition of $\mathbf{r}$ it follows that

$$\begin{aligned}&-B_\tau M^2 B_\tau^* \hat{\mathbf{\Phi}} - B_n M R_A^{-1} M B_n^* \hat{\mathbf{\Phi}}\\ &= -B_\tau M^2 B_\tau^* \Pi \mathbf{y} - B_n M R_A^{-1} M B_n^* \Pi \mathbf{y} - B_\tau M^2 B_\tau^* \mathbf{r} - B_n M R_A^{-1} M B_n^* \mathbf{r}\,.\end{aligned}$$

Thanks to this identity, we can rewrite (6.2) in the form

$$\mathbf{y}' = A\mathbf{y} - B_\tau M^2 B_\tau^* \hat{\mathbf{\Phi}} - B_n M R_A^{-1} M B_n^* \hat{\mathbf{\Phi}} + B_\tau M^2 B_\tau^* \mathbf{r} + B_n M R_A^{-1} M B_n^* \mathbf{r} + \mathbf{f}, \quad \mathbf{y}(0) = \mathbf{y}_0\,.$$

Thus $\mathbf{y} = \hat{\mathbf{y}} + \tilde{\mathbf{y}}$. Due to Theorem 5.4, we notice that

$$\begin{aligned}&\|B^* \hat{\mathbf{\Phi}}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(\Sigma_\infty)}\\ &\quad \le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big).\end{aligned}$$

Moreover, $\mathbf{y}$ belongs to $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ and $B^*\Pi \in \mathcal{L}(\mathbf{V}_n^0(\Omega), \mathbf{V}^{1/2}(\Gamma))$ (see Remark 4.4). Thus $B^*\mathbf{r} = B^*\hat{\mathbf{\Phi}} - B^*\Pi\mathbf{y}$ belongs to $\mathbf{V}^{0,0}(\Sigma_\infty)$ and

$$\|B^*\mathbf{r}\|_{\mathbf{V}^{0,0}(\Sigma_\infty)} \le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big).$$

From Lemma 8.3 it follows that

$$\|\tilde{\mathbf{y}}\|_{\mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'}(Q_\infty)} \le C\|B^*\mathbf{r}\|_{\mathbf{V}^{0,0}(\Sigma_\infty)} \qquad \text{for all } \varepsilon' > 0.$$

Thus $\mathbf{y} = \hat{\mathbf{y}} + \tilde{\mathbf{y}}$ belongs to $\mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(Q_\infty)$ for all $\varepsilon' > 0$. Therefore $B^*\mathbf{r}$ belongs to $\mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(\Sigma_\infty)$ because $B^*\Pi \in \mathcal{L}(\mathbf{V}_n^{1/2-\varepsilon'}(\Omega), \mathbf{V}^{1-\varepsilon'}(\Gamma))$ for all $\varepsilon' > 0$ (see Remark 4.4), and we have

$$\begin{aligned}&\|B^*\mathbf{r}\|_{\mathbf{V}^{1/2-\varepsilon',1/4-\varepsilon'/2}(\Sigma_\infty)}\\ &\quad \le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big) \text{ for all } \varepsilon' > 0.\end{aligned}$$

With another iteration we can prove an estimate of $B^*\mathbf{r}$ in $\mathbf{V}^{1-\varepsilon',1/2-\varepsilon'/2}(\Sigma_\infty)$. Still with Lemma 8.3 we obtain

$$\|\tilde{\mathbf{y}}\|_{\mathbf{V}^{3/2-\varepsilon'',3/4-\varepsilon''}(Q_\infty)} \le C\|B^*\mathbf{r}\|_{\mathbf{V}^{1-\varepsilon',1/2-\varepsilon'/2}(\Sigma)}$$

for all $\varepsilon'' > \varepsilon' > 0$. With Theorem 5.4, we have

$$\|\hat{\mathbf{y}}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \le C\big(|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} + \|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big),$$

and the proof is complete. □

LEMMA 6.3. *If* $\mathbf{z}$ *belongs to* $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$*, then*

$$\begin{aligned}&\|\mathbf{z}\|_{L^{1/\varepsilon}(0,\infty;\mathbf{V}^{1/2+\varepsilon}(\Omega))}\\ &\quad \le C\|\mathbf{z}\|^{(1-2\varepsilon)}_{L^\infty(0,\infty;\mathbf{V}^{1/2-\varepsilon}(\Omega))}\|\mathbf{z}\|^{2\varepsilon}_{L^2(0,\infty;\mathbf{V}^{3/2-\varepsilon}(\Omega))} \le C\|\mathbf{z}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.\end{aligned}$$

Proof. We first observe that

$$|\mathbf{z}(t)|_{\mathbf{V}^{1/2+\varepsilon}(\Omega)} \le C|\mathbf{z}(t)|^{(1-2\varepsilon)}_{\mathbf{V}^{1/2-\varepsilon}(\Omega)}|\mathbf{z}(t)|^{2\varepsilon}_{\mathbf{V}^{3/2-\varepsilon}(\Omega)}.$$

Therefore we have

$$\int_0^\infty |\mathbf{z}(t)|^{1/\varepsilon}_{\mathbf{V}^{1/2+\varepsilon}(\Omega)} dt \le C\|\mathbf{z}\|^{(1-2\varepsilon)/\varepsilon}_{L^\infty(0,\infty;\mathbf{V}^{1/2-\varepsilon}(\Omega))} \int_0^\infty |\mathbf{z}(t)|^2_{\mathbf{V}^{3/2-\varepsilon}(\Omega)} dt.$$

The proof is complete. □

LEMMA 6.4. *Let ε be in $(0,1/4)$. If $\mathbf{z}$ belongs to $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$, then*

$$\|(\mathbf{z}\cdot\nabla)\mathbf{z}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{L}^2(\Omega))}+\|P\mathrm{div}(\mathbf{z}\otimes\mathbf{z})\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\le C_2\|\mathbf{z}\|^2_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

Proof. We have

$$\|(\mathbf{z}\cdot\nabla)\mathbf{z}\|_{\mathbf{L}^2(\Omega)}\le\|\mathbf{z}\|_{\mathbf{L}^{4/(1-2\varepsilon)}(\Omega)}\|\nabla\mathbf{z}\|_{\mathbf{L}^{4/(1+2\varepsilon)}(\Omega)}\le C\|\mathbf{z}\|_{\mathbf{V}^{1/2+\varepsilon}(\Omega)}\|\mathbf{z}\|_{\mathbf{V}^{3/2-\varepsilon}(\Omega)}.$$

Therefore with Lemma 6.3, it follows that

$$\|(\mathbf{z}\cdot\nabla)\mathbf{z}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{L}^2(\Omega))}\le C\|\mathbf{z}\|_{L^{1/\varepsilon}(0,\infty;\mathbf{V}^{1/2+\varepsilon}(\Omega))}\|\mathbf{z}\|_{L^2(0,\infty;\mathbf{V}^{3/2-\varepsilon}(\Omega))}$$

$$\le C\|\mathbf{z}\|^2_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

From [19, Theorem B.3] we deduce that

$$\|\mathbf{z}\otimes\mathbf{z}\|_{\mathbf{H}^{1-2\varepsilon,1/2-\varepsilon}(Q_\infty)}\le C\|\mathbf{z}\|^2_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

The divergence operator is continuous from $\mathbf{L}^2(\Omega)$ into $\mathbf{H}^{-1}(\Omega)$, and from $\mathbf{H}^1(\Omega)$ to $\mathbf{L}^2(\Omega)$. Thus it is also continuous from $\mathbf{H}^{1-2\varepsilon}(\Omega)$ into $(\mathbf{H}^{2\varepsilon}(\Omega))'$. The projector P is continuous from $\mathbf{L}^2(\Omega)$ into $\mathbf{V}^0_n(\Omega)$, and from $\mathbf{H}^{-1}(\Omega)$ into $\mathbf{V}^{-1}(\Omega)$ (see, e.g., [32]). Thus it is also continuous from $(\mathbf{H}^{2\varepsilon}(\Omega))'$ into $(\mathbf{V}^{2\varepsilon}_n(\Omega))'$. Therefore we have

$$\|P\mathrm{div}(\mathbf{z}\otimes\mathbf{z})\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}_n(\Omega))')}\le C\|\mathbf{z}\otimes\mathbf{z}\|_{L^2(0,\infty;\mathbf{H}^{1-2\varepsilon}(\Omega))}\le C\|\mathbf{z}\|^2_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

Observe that the condition $0<\varepsilon<1/4$ is needed to have $(\mathbf{H}^{2\varepsilon}(\Omega))'\hookrightarrow\mathbf{H}^{-1}(\Omega)$. ∎

LEMMA 6.5. *Let ε be in $(0,1/4)$. The mapping PF is locally Lipshitz continuous from $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$ into $L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}^0_n(\Omega))\cap L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')$. More precisely we have*

$$\|PF(\mathbf{z}_1)-PF(\mathbf{z}_2)\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}^0_n(\Omega))}+\|PF(\mathbf{z}_1)-PF(\mathbf{z}_2)\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}$$
$$\le C_2\big(\|\mathbf{z}_1\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}+\|\mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\big)\|\mathbf{z}_1-\mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

Proof. From Lemma 6.4 it follows that PF is a mapping from $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$ into $L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}^0_n(\Omega))\cap L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')$. By calculations similar to those in Lemma 6.4, we obtain

$$\|PF(\mathbf{z}_1)-PF(\mathbf{z}_2)\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}^0_n(\Omega))}+\|PF(\mathbf{z}_1)-PF(\mathbf{z}_2)\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}$$
$$\le\|P((\mathbf{z}_1\cdot\nabla)(\mathbf{z}_1-\mathbf{z}_2))\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}^0_n(\Omega))}+\|P(((\mathbf{z}_1-\mathbf{z}_2)\cdot\nabla)\mathbf{z}_2)\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}^0_n(\Omega))}$$
$$+\|P\mathrm{div}(\mathbf{z}_1\otimes(\mathbf{z}_1-\mathbf{z}_2))\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}_n(\Omega))')}+\|P\mathrm{div}((\mathbf{z}_1-\mathbf{z}_2)\otimes\mathbf{z}_2)\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}_n(\Omega))')}$$
$$\le C_2\big(\|\mathbf{z}_1\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}+\|\mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\big)\|\mathbf{z}_1-\mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

The proof is complete. ∎

LEMMA 6.6. *If $P\mathbf{y}$ belongs to $\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)$ for some $0<\varepsilon<1/2$, then*

$$\|(I-P)D_AMR_A^{-1}MB_n^*\Pi P\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\le C_3\|P\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

Proof. The operator $B^*\Pi$ belongs to $\mathcal{L}(\mathbf{V}^{1/2-\varepsilon}_n(\Omega),\mathbf{V}^{1-\varepsilon}(\Gamma))$ (see Remark 4.4). Thus $MR_A^{-1}MB_n^*\Pi$ belongs to $\mathcal{L}(\mathbf{V}^{1/2-\varepsilon}_n(\Omega),\mathbf{V}^{1-\varepsilon}(\Gamma))$, and $D_AMR_A^{-1}MB_n^*\Pi$ belongs

to $\mathcal{L}(\mathbf{V}_n^{1/2-\varepsilon}(\Omega), \mathbf{V}^{3/2-\varepsilon}(\Omega))$. The lemma follows from the continuity properties of the operator $(I-P)$. ∎

Proof of Theorem 6.1. We set $\mu_0 = \frac{1}{4C_1C_2(1+C_3)}$ and $\eta(\mu) = \frac{3\mu}{4C_1(1+C_3)}$. For $\mathbf{z} \in \mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(\Omega)$, we denote by $\mathbf{y_z}$ the solution to the equation

$$\begin{aligned} &P\mathbf{y}' = A_\Pi P\mathbf{y} + PF(\mathbf{z}), \qquad \mathbf{y}(0) = \mathbf{y}_0, \\ &(I-P)\mathbf{y} = -(I-P)D_A M R_A^{-1} M B_n^* \,\Pi P\mathbf{y}\,. \end{aligned} \tag{6.3}$$

We are going to prove that the mapping $\mathbf{M} : \mathbf{z} \mapsto \mathbf{y_z}$ is a contraction in D_μ.

Step 1. From Lemmas 6.2, 6.4, and 6.6 it follows that

$$\begin{aligned} &\|\mathbf{y_z}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \le \|P\mathbf{y_z}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} + \|(I-P)\mathbf{y_z}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \\ &\le C_1(1+C_3)\big(|\mathbf{y}_0|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)} + \|PF(\mathbf{z})\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|PF(\mathbf{z})\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big) \\ &\le C_1(1+C_3)\left(\frac{3\mu}{4C_1(1+C_3)} + C_2\|\mathbf{z}\|^2_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\right) \le \frac{3\mu}{4} + C_1(1+C_3)C_2\mu^2 \le \mu \end{aligned}$$

if $\mu < \mu_0$. Thus $\mathbf{M}$ is a mapping from D_μ into itself.

Step 2. From Lemmas 6.2 and 6.5 it follows that

$$\begin{aligned} &\|P\mathbf{y}_{\mathbf{z}_1} - P\mathbf{y}_{\mathbf{z}_2}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \\ &\le C_1\big(\|PF(\mathbf{z}_1) - PF(\mathbf{z}_2)\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} + \|PF(\mathbf{z}_1) - PF(\mathbf{z}_2)\|_{L^2(0,\infty;(\mathbf{V}^{2\varepsilon}(\Omega))')}\big) \\ &\le C_2C_1\big(\|\mathbf{z}_1\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} + \|\mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\big)\|\mathbf{z}_1 - \mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \\ &\le 2C_2C_1\mu\|\mathbf{z}_1 - \mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}. \end{aligned}$$

Thus with Lemma 6.6 and the previous estimate we obtain

$$\|\mathbf{y}_{\mathbf{z}_1} - \mathbf{y}_{\mathbf{z}_2}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \le 2C_2C_1\mu(1+C_3)\|\mathbf{z}_1 - \mathbf{z}_2\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}.$$

Thus if $\mu < \mu_0$, the mapping $\mathbf{M}$ is a contraction in D_μ, and the system (6.1) admits a unique solution in D_μ. ∎

6.2. Second stabilization result. To obtain a feedback law providing an exponential stabilization of the Navier–Stokes equation, we are going to use the linear feedback law determined thanks to an auxiliary problem. For that, we set

$$\hat{\mathbf{y}} = e^{\omega t}\mathbf{y}, \qquad \hat{\mathbf{u}} = e^{\omega t}\mathbf{u}.$$

If

$$\begin{aligned} &P\mathbf{y}' = AP\mathbf{y} + PF(\mathbf{y}) + BM\mathbf{u}, \qquad P\mathbf{y}(0) = \mathbf{y}_0, \\ &(I-P)\mathbf{y} = (I-P)D_A\gamma_n M\mathbf{u}, \end{aligned}$$

then $\hat{\mathbf{y}}$ is the solution to the system

$$\begin{aligned} &P\hat{\mathbf{y}}' = AP\hat{\mathbf{y}} + \omega P\hat{\mathbf{y}} + e^{-\omega t}PF(\hat{\mathbf{y}}) + BM\hat{\mathbf{u}}, \qquad P\hat{\mathbf{y}}(0) = \mathbf{y}_0, \\ &(I-P)\hat{\mathbf{y}} = (I-P)D_A\gamma_n M\hat{\mathbf{u}}. \end{aligned}$$

Set $A_\omega = A + \omega I$, and let $\Pi_\omega \in \mathcal{L}(\mathbf{V}_n^0(\Omega))$ be the solution to the algebraic Riccati equation

$$\begin{aligned} &\Pi_\omega = \Pi_\omega^* \ge 0, \\ &\Pi_\omega A_\omega + A_\omega^*\Pi_\omega - \Pi_\omega B_\tau M^2 B_\tau^*\Pi_\omega - \Pi_\omega B_n M\, R_A^{-1}\, MB_n^*\Pi_\omega + I = 0. \end{aligned} \tag{6.4}$$

The existence of a unique solution in $\mathcal{L}(\mathbf{V}_n^0(\Omega);\mathbf{V}^2(\Omega)\cap\mathbf{V}_0^1(\Omega))$ to this equation may be proved as in section 4. As in section 4, it can be shown that

$$I(\mathbf{z}_{\mathbf{y}_0},\mathbf{v}_{\mathbf{y}_0})=\frac{1}{2}\Big(\Pi_\omega\mathbf{y}_0,\mathbf{y}_0\Big)_{\mathbf{V}_n^0(\Omega)},$$

where $(\mathbf{z}_{\mathbf{y}_0},\mathbf{v}_{\mathbf{y}_0})$ is the solution of the control problem

$$(\mathcal{P}_{0,\mathbf{y}_0})\qquad \inf\Big\{I(\mathbf{z},\mathbf{v})\ \mid\ (\mathbf{z},\mathbf{v})\text{ satisfies (6.5)},\ \mathbf{v}\in\mathbf{V}^{0,0}(\Sigma_\infty)\Big\},$$

where I is the functional of section 4, and

$$\mathbf{z}'=A_\omega\mathbf{z}+BM\mathbf{v}\quad\text{in }(0,\infty),\qquad \mathbf{z}(0)=\mathbf{y}_0. \tag{6.5}$$

Consider the Navier–Stokes equations with the linear feedback law

$$\begin{aligned}&P\hat{\mathbf{y}}'=A_{\omega,\Pi_\omega}P\hat{\mathbf{y}}+e^{-\omega t}PF(\hat{\mathbf{y}}),\qquad P\hat{\mathbf{y}}(0)=\mathbf{y}_0,\\&(I-P)\hat{\mathbf{y}}=-(I-P)D_AMR_A^{-1}MB_n^*\,\Pi_\omega P\hat{\mathbf{y}},\end{aligned} \tag{6.6}$$

where

$$D(A_{\omega,\Pi_\omega})=D(A_{\Pi_\omega})\quad\text{and}\quad A_{\omega,\Pi_\omega}\mathbf{y}=A_{\Pi_\omega}\mathbf{y}+\omega\mathbf{y}\quad\text{for all }\mathbf{y}\in D(A_{\Pi_\omega}).$$

As previously, if $\hat{\mathbf{y}}$ is a solution to (6.6), then $\mathbf{y}=e^{-\omega t}\hat{\mathbf{y}}$ is the solution of

$$\begin{aligned}&P\mathbf{y}'=AP\mathbf{y}-B_\tau M^2B_\tau^*\Pi_\omega P\mathbf{y}-B_n\,MR_A^{-1}M\,B_n^*\Pi_\omega P\mathbf{y}+PF(\mathbf{y}),\qquad P\mathbf{y}(0)=\mathbf{y}_0,\\&(I-P)\mathbf{y}=-(I-P)D_AR_A^{-1}MB_n^*\Pi_\omega P\mathbf{y}.\end{aligned} \tag{6.7}$$

THEOREM 6.7. *For all $0<\varepsilon<1/4$, there exist $\mu_0>0$ and a nondecreasing function η_0 from $\mathbb{R}^+$ into itself such that if $\mu\in(0,\mu_0)$ and $\|\mathbf{y}_0\|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)}\le\eta_0(\mu)$, then (6.7) admits a unique solution in the set*

$$D_\mu=\Big\{\mathbf{y}\in\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)\mid\|e^{\omega(\cdot)}\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\le\mu\Big\}.$$

Moreover, $\mathbf{y}$, which belongs to $C_b([0,\infty);\mathbf{V}^{1/2-\varepsilon}(\Omega))$, satisfies

$$|\mathbf{y}(t)|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)}\le C(\mathbf{w},\mu,\omega)\ e^{-\omega t}.$$

Proof. Substituting $F(\mathbf{y})$ by $e^{-\omega t}F(\hat{\mathbf{y}})$ in the proof of Theorem 6.1, we can show that there exists $\mu_0>0$ and a nondecreasing function η_0 from $\mathbb{R}^+$ into itself such that if $\mu\in(0,\mu_0)$ and $\|\mathbf{y}_0\|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)}\le\eta_0(\mu)$ for some $0<\varepsilon<1/4$, the equation (6.6) admits a unique solution $\hat{\mathbf{y}}$ in $\widehat{D}_\mu=\big\{\mathbf{y}\in\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)\mid\|\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)}\le\mu\big\}$, and $|\hat{\mathbf{y}}(t)|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)}\le C(\mathbf{w},\mu,\omega)$. Thus $\mathbf{y}=e^{-\omega t}\hat{\mathbf{y}}$ is the solution of (6.7), belongs to D_μ, and satisfies

$$|\mathbf{y}(t)|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)}\le C(\mathbf{w},\mu,\omega)\ e^{-\omega t}.$$

The proof is complete. ∎

7. Additional results.

7.1. Control by a tangential velocity. Assume that Ω and $\mathbf{w}$ are such that (1.3) is stabilizable in $\mathbf{V}_n^{1/2-\varepsilon}(\Omega)$ with controls $\mathbf{u} \in L^2(0,\infty;\mathbf{V}_n^0(\Gamma))$ satisfying $\mathrm{supp}(\mathbf{u}) \subset \Gamma_0 \times (0,\infty)$, where Γ_0 is an open set in Γ in which the function m is equal to 1 (see the assumptions on the function m at the beginning of section 2). Such a stabilizability result is stated in [4, Theorem B.1.1]. With such a result, as in section 6.2 (and section 4), we can show that the algebraic Riccati equation

$$\Pi_\omega \in \mathcal{L}(\mathbf{V}_n^0(\Omega)), \quad \Pi_\omega = \Pi_\omega^* \geq 0, \quad \Pi_\omega A_\omega + A_\omega^* \Pi_\omega - \Pi_\omega B_\tau M^2 B_\tau^* \Pi_\omega + I = 0, \tag{7.1}$$

admits a unique solution satisfying $\Pi_\omega \in \mathcal{L}(\mathbf{V}_n^0(\Omega); \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$, which enjoys the same regularizing properties as in Theorem 4.5. Since we deal with controls belonging to $\mathbf{V}_n^0(\Gamma)$, the operator R_A is now the identity. As in Theorem 6.7, we can prove that, for all $0 < \varepsilon < 1/4$, there exist $\mu_0 > 0$ and a nondecreasing function η_0 from $\mathbb{R}^+$ into itself such that if $\mu \in (0,\mu_0)$ and $\|\mathbf{y}_0\|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} \leq \eta_0(\mu)$, the equation

$$\mathbf{y}' = A\mathbf{y} - B_\tau M^2 B_\tau^* \Pi_\omega \mathbf{y} + PF(\mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

admits a unique solution in the set

$$D_\mu = \Big\{ \mathbf{y} \in \mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty) \cap L^2(0,\infty;\mathbf{V}_n^0(\Omega)) \mid \|e^{\omega\,(\cdot)}\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \leq \mu \Big\}.$$

Moreover, $\mathbf{y}$, which belongs to $C_b([0,\infty);\mathbf{V}_n^{1/2-\varepsilon}(\Omega))$, satisfies

$$|\mathbf{y}(t)|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)} \leq C(\mathbf{w},\mu,\omega)\; e^{-\omega t}.$$

Notice that the local feedback stabilization of the two-dimensional Navier–Stokes equations is not studied in [4] (only the stabilization of the linearized Navier–Stokes equations is studied in two dimensions in [4, Appendix B]). If we transpose the results obtained in [4] in the three-dimensional case to the two-dimensional case, the results are completely different from the ones we obtain. Indeed we here prove a local stabilization result for the Navier–Stokes equations thanks to the solution of a classical Riccati equation (equation (7.1)), while the Riccati equation obtained in [4, section 4] is defined only in the domain of the square of the feedback operator, which is unknown.

7.2. Dependence of solutions with respect to ν for the stabilization of the Stokes equation. The dependence of solutions to our control problems with respect to ν is very complicated because the stationary solution $\mathbf{w}$ involved in the Oseen operator also depends on ν in a complicated way. There is one particular situation for which we can clarify this dependence; it is the one corresponding to $\mathbf{w} = 0$. Assume that the semigroup generated by $(\nu P\Delta + \omega I,\ \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$ on $\mathbf{V}_n^0(\Omega)$ is unstable for a given $\omega > 0$. What has been done in section 6.2 is still valid if $\mathbf{w} \equiv 0$. Let Π_ω be the solution to (6.4) corresponding to $A = \nu A_0$ and $A_\omega = \nu A_0 + \omega I$. Consider the solution $\hat{\mathbf{y}}$ to the closed loop system

$$\begin{aligned} &P\hat{\mathbf{y}}' = AP\hat{\mathbf{y}} + \omega P\hat{\mathbf{y}} - B\,MR_A^{-1}M\,B^*\Pi_\omega P\hat{\mathbf{y}}, \qquad P\hat{\mathbf{y}}(0) = \mathbf{y}_0, \\ &(I-P)\hat{\mathbf{y}} = -(I-P)D_A R_A^{-1} M B_n^* \Pi_\omega P\hat{\mathbf{y}}. \end{aligned}$$

From the previous sections we know that

$$I(\hat{\mathbf{y}}, -MR_A^{-1}M\,B^*\Pi_\omega P\hat{\mathbf{y}}) = \frac{1}{2}\Big(\Pi_\omega \mathbf{y}_0, \mathbf{y}_0\Big)_{\mathbf{V}_n^0(\Omega)}$$

and

$$|\hat{\mathbf{y}}(t)|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)} \le C(\varepsilon,\omega)e^{-\sigma t}|\mathbf{y}_0|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)}$$

for all $0 < \varepsilon \le 1/2$, where $C(\varepsilon,\omega) > 0$ depends on ε and ω, and σ is positive but not precisely known. Setting $\mathbf{y} = e^{-\omega t}\hat{\mathbf{y}}$, we can check that $\mathbf{y}$ is the solution to the closed loop system

$$\begin{aligned} &P\mathbf{y}' = AP\mathbf{y} - B\,MR_A^{-1}M\,B^*\Pi_\omega P\mathbf{y}, \qquad P\mathbf{y}(0) = \mathbf{y}_0, \\ &(I-P)\mathbf{y} = -(I-P)D_AR_A^{-1}MB_n^*\Pi_\omega P\mathbf{y}, \end{aligned} \tag{7.2}$$

and $\mathbf{y}$ obeys the decay rate

$$|\mathbf{y}(t)|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)} \le C(\varepsilon)e^{-(\omega+\sigma)t}|\mathbf{y}_0|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)} \le C(\varepsilon)e^{-\omega t}|\mathbf{y}_0|_{\mathbf{V}^{1/2-\varepsilon}(\Omega)}.$$

Thus we have solved the boundary feedback stabilization problem of the Stokes equations with a prescribed exponential decay rate ω. Now we would like to clarify the dependence of the solution $\mathbf{y}$ of (7.2) with respect to ν.

We have to clarify the dependence of Π_ω with respect to ν. For that we first have to clarify the dependence of R_A, B, and B^* with respect to ν. We can choose $\lambda_0 = 0$ in the definition of D_A. We denote by $D_{A_0}\mathbf{u}$ the solution $\mathbf{y}$ to the equation

$$-\Delta\mathbf{y} + \nabla q = 0 \quad \text{and} \quad \text{div } \mathbf{y} = 0 \quad \text{in } \Omega, \qquad \mathbf{y} = \mathbf{u} \quad \text{on } \Gamma,$$

and we have $D_A = \frac{1}{\nu}D_{A_0}$ because $\lambda_0 = 0$. Thus $D_A^* = \frac{1}{\nu}D_{A_0}^*$, and $R_A = MD_A^*(I-P)D_AM + I = \frac{1}{\nu^2}MD_{A_0}^*(I-P)D_{A_0}M + I$. Since $B = (-A)PD_A = (-A_0)PD_{A_0}$, we observe that B and B^* do not depend on ν.

Let ν_1 and ν_2 be two viscosity coefficients such that $\nu_1 > \nu_2$. Assume that the semigroup generated by $(\nu_1 P\Delta + \omega I,\ \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega))$ on $\mathbf{V}_n^0(\Omega)$ is unstable for a given $\omega > 0$.

For $i = 1, 2$, let $\Pi_{\nu_i,\omega}$ be the solution to (6.4) corresponding to $A_\omega = A_{\nu_i,\omega} = \nu_i A_0 + \omega I$.

With this notation we can observe that $\Pi_{\nu_1,\omega}$ is also the solution to the equation

$$\Pi_{\nu_1,\omega}A_{\nu_2,\omega} + A_{\nu_2,\omega}^*\Pi_{\nu_1,\omega}$$
$$-\frac{\nu_2}{\nu_1}\Pi_{\nu_1,\omega}BM\left(\frac{1}{\nu_1^2}MD_{A_0}^*(I-P)D_{A_0}M + I\right)^{-1}MB^*\Pi_{\nu_1,\omega} + \frac{\nu_2}{\nu_1}I + \omega I\left(\frac{\nu_2}{\nu_1} - 1\right) = 0.$$

Since $\nu_1 > \nu_2$, we can verify that

$$\begin{aligned} &\frac{\nu_2}{\nu_1}BM\left(\frac{1}{\nu_1^2}MD_{A_0}^*(I-P)D_{A_0}M + I\right)^{-1}MB^* \\ &= BM\left(\frac{1}{\nu_1\nu_2}MD_{A_0}^*(I-P)D_{A_0}M + I\right)^{-1}MB^* \\ &\ge BM\left(\frac{1}{\nu_2^2}MD_{A_0}^*(I-P)D_{A_0}M + I\right)^{-1}MB^* \end{aligned}$$

(where the inequality $\le$ is the inequality between quadratic forms), and

$$\frac{\nu_2}{\nu_1}I + \omega I\left(\frac{\nu_2}{\nu_1} - 1\right) \le I.$$

Thus, as in [7], we can claim that

$$\Pi_{\nu_1,\omega} \le \Pi_{\nu_2,\omega}.$$

If $\mathbf{y}_{\nu_i}$ is the solution to the closed loop system (7.2) for $A = \nu_i A_0$, we have

$$I(e^{-\omega t}\mathbf{y}_{\nu_1}, -e^{-\omega t} M R^{-1}_{\nu_1 A_0} M\, B^* \Pi_{\nu_1,\omega} P \mathbf{y}_{\nu_1}) = \tfrac{1}{2}\Big(\Pi_{\nu_1,\omega}\mathbf{y}_0, \mathbf{y}_0\Big)$$
$$\le I(e^{-\omega t}\mathbf{y}_{\nu_2}, -e^{-\omega t} M R^{-1}_{\nu_2 A_0} M\, B^* \Pi_{\nu_2,\omega} P \mathbf{y}_{\nu_2}) = \tfrac{1}{2}\Big(\Pi_{\nu_2,\omega}\mathbf{y}_0, \mathbf{y}_0\Big),$$

which provides a comparison between the L^2 of the solution to the stabilization problems according to the value of ν.

8. Appendix. In this section we prove some regularity results for the state and the adjoint equations.

LEMMA 8.1. *If* $\mathbf{y}_0 \in \mathbf{V}_n^{1/2-\varepsilon}(\Omega)$ *with* $0 < \varepsilon \le 1/2$, *the weak solution to the equation*

$$\mathbf{y}' = (A - \lambda_0)\mathbf{y} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

obeys

$$\|\mathbf{y}\|_{C_b([0,\infty);\mathbf{V}_n^{1/2-\varepsilon}(\Omega))} + \|\mathbf{y}\|_{\mathbf{V}^{3/2-\varepsilon,3/4-\varepsilon/2}(Q_\infty)} \le C|\mathbf{y}_0|_{\mathbf{V}_n^{1/2-\varepsilon}(\Omega)}.$$

Proof. The exponential stability of the semigroup $(e^{t(A-\lambda_0 I)})_{t\ge 0}$ gives a bound for $\mathbf{y}$ in $L^2(0,\infty;\mathbf{V}_n^{1/2-\varepsilon}(\Omega))$. Using the estimate in $L^2(0,\infty;\mathbf{V}_n^{1/2-\varepsilon}(\Omega))$, the other estimates may be obtained by using the proofs in [5] and interpolation results. ∎

LEMMA 8.2. *If* $\mathbf{f} \in L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))$, *the solution* $\mathbf{y}$ *to the equation*

$$\mathbf{y}' = (A - \lambda_0 I)\mathbf{y} + \mathbf{f} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = 0,$$

obeys

$$\|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}^{2-2\varepsilon}(\Omega))} \le C\|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))}.$$

If $0 < \varepsilon < 1/4$ *and* $\mathbf{f} \in L^2(0,\infty;(\mathbf{V}_n^{2\varepsilon}(\Omega))')$, *then*

$$\|\mathbf{y}\|_{\mathbf{V}^{2-2\varepsilon,1-\varepsilon}(Q_\infty)} \le C\|\mathbf{f}\|_{L^2(0,\infty;(\mathbf{V}_n^{2\varepsilon}(\Omega))')}.$$

Proof. From the identity

$$(-A + \lambda_0 I)^{1-\varepsilon}\mathbf{y}(t) = \int_0^t (-A + \lambda_0 I)^{1-\varepsilon} e^{(t-\tau)(A-\lambda_0 I)}\mathbf{f}(\tau)\, d\tau,$$

it follows that

$$\|\mathbf{y}(t)\|_{\mathbf{V}^{2-2\varepsilon}(\Omega)} \le C\int_0^t e^{-\omega(t-\tau)}(t-\tau)^{-1+\varepsilon}|\mathbf{f}(\tau)|_{\mathbf{V}_n^0(\Omega)}\, d\tau.$$

Thus, with the generalized Young inequality for convolutions (see [30, p. 32]), we have

$$\|\mathbf{y}\|_{L^r(0,\infty;\mathbf{V}^{2-2\varepsilon}(\Omega))} \le C\|\mathbf{f}\|_{L^{2/(1+2\varepsilon)}(0,\infty;\mathbf{V}_n^0(\Omega))} \tag{8.1}$$

for all $2/(1+2\varepsilon) \le r \le 2$.

(ii) Now we assume that $\mathbf{f}$ belongs to $L^2(0,\infty;(\mathbf{V}_n^{2\varepsilon}(\Omega))')$. To prove the estimate in that case we proceed by interpolation. We know that

$$\|\mathbf{y}\|_{\mathbf{V}^{2,1}(Q_\infty)} \le C\|\mathbf{f}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \qquad \text{and} \qquad \|\mathbf{y}\|_{\mathbf{V}^{1,1/2}(Q_\infty)} \le C\|\mathbf{f}\|_{L^2(0,\infty;\mathbf{V}^{-1}(\Omega))}.$$

Since we have

$$[L^2(0,\infty;\mathbf{V}_n^0(\Omega)), L^2(0,\infty;\mathbf{V}^{-1}(\Omega))]_{2\varepsilon} = ([L^2(0,\infty;\mathbf{V}_n^0(\Omega)), L^2(0,\infty;\mathbf{V}_0^1(\Omega))]_{2\varepsilon})' = L^2(0,\infty;(\mathbf{V}_n^{2\varepsilon}(\Omega))'),$$

we obtain the desired result by interpolation. (Notice that $[\mathbf{V}_n^0(\Omega),\mathbf{V}_0^1(\Omega)]_{2\varepsilon} = \mathbf{V}_n^{2\varepsilon}(\Omega)$ because $0<\varepsilon<1/4$.) □

LEMMA 8.3. *If* $\mathbf{u}$ *belongs to* $\mathbf{V}^{s,s/2}(\Sigma_\infty)$ *with* $0 \le s < 1$, *then the weak solution to the equation*

$$\mathbf{y}' = (A-\lambda_0)\mathbf{y} + BM\mathbf{u} \quad \text{in } (0,\infty), \qquad \mathbf{y}(0) = 0,$$

obeys

$$\|\mathbf{y}\|_{\mathbf{V}^{1/2+s-\varepsilon,1/4+s/2-\varepsilon/2}(Q_\infty)} \le C\|\mathbf{u}\|_{\mathbf{V}^{s,s/2}(\Sigma_\infty)} \quad \text{for all } \varepsilon > 0.$$

If $\mathbf{u}$ *belongs to* $\mathbf{V}^{s,s/2}(\Sigma_\infty)$ *with* $1 < s \le 2$, *and if* $\mathbf{u}(0) = 0$, *then*

$$\|\mathbf{y}\|_{\mathbf{V}^{1/2+s-\varepsilon,1/4+s/2-\varepsilon/2}(Q_\infty)} \le C\|\mathbf{u}\|_{\mathbf{V}^{s,s/2}(\Sigma_\infty)} \quad \text{for all } \varepsilon > 0.$$

Proof. Similar estimates are proved in [26, Theorem 2.3] for a finite time interval. It is sufficient to rewrite the proof and to use the exponential stability of the semigroup $(e^{t(A-\lambda_0 I)})_{t\ge 0}$ to obtain the desired results. □

LEMMA 8.4. *For all* $\mathbf{y} \in \mathbf{V}_n^2(\Omega)$, *the solution* $\mathbf{\Phi} \in \mathbf{V}^2(\Omega) \cap \mathbf{V}_0^1(\Omega)$ *to the stationary equation* $\lambda_0\mathbf{\Phi} - A^*\mathbf{\Phi} = \mathbf{y}$ *obeys*

$$|\mathbf{\Phi}|_{\mathbf{V}^4(\Omega)} \le C|\mathbf{y}|_{\mathbf{V}_n^2(\Omega)}.$$

Proof. We rewrite the equation in the form

$$\lambda_0\mathbf{\Phi} - \nu P\Delta\mathbf{\Phi} = \mathbf{y} - P((\mathbf{w}\cdot\nabla)\mathbf{\Phi}) + P((\nabla\mathbf{w})^T\mathbf{\Phi}).$$

Since $\mathbf{w} \in \mathbf{V}^3(\Omega)$ and $\mathbf{\Phi} \in \mathbf{V}^2(\Omega)$, then $P((\mathbf{w}\cdot\nabla)\mathbf{\Phi})$ and $P((\nabla\mathbf{w})^T\mathbf{\Phi})$ belong to $\mathbf{V}^1(\Omega)$, which gives an estimate of $\mathbf{\Phi}$ in $\mathbf{V}^3(\Omega)$. Knowing that $\mathbf{\Phi} \in \mathbf{V}^3(\Omega)$, $P((\mathbf{w}\cdot\nabla)\mathbf{\Phi})$, and $P((\nabla\mathbf{w})^T\mathbf{\Phi})$ belong to $\mathbf{V}^2(\Omega)$, the proof is complete. □

LEMMA 8.5. *If the function* $\mathbf{y}$ *belongs to* $\mathbf{V}^{s,s/2}(Q_\infty) \cap L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ *with* $0 \le s \le 2$, *then the solution* $\mathbf{\Phi}$ *to the equation*

$$-\mathbf{\Phi}' = (A^*-\lambda_0 I)\mathbf{\Phi} + \mathbf{y} \quad \text{in } (0,\infty), \qquad \mathbf{\Phi}(\infty) = 0, \tag{8.2}$$

satisfies

$$\|\mathbf{\Phi}\|_{\mathbf{V}^{s+2,s/2+1}(Q_\infty)} \le C\|\mathbf{y}\|_{\mathbf{V}^{s,s/2}(Q_\infty)}. \tag{8.3}$$

Proof. Step 1. Let us first prove estimate (8.3) for $s = 0$. Assume that $\mathbf{y} \in L^2(0,\infty;\mathbf{V}_n^0(\Omega))$. The existence of a unique solution $\mathbf{\Phi}$ to (8.2) in $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$ follows from the exponential stability of the semigroup $(e^{t(A^*-\lambda_0 I)})_{t\ge 0}$ on $\mathbf{V}_n^0(\Omega)$. Moreover, we have

$$\|\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))} \le C\|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))}. \tag{8.4}$$

It is sufficient to prove estimate (8.3) for $s = 0$ and $\mathbf{y} \in C_c(0,\infty;\mathbf{V}_n^0(\Omega))$. In that case, it can be deduced from (8.4) with the method of proof in [5, Theorem 1.1].

Step 2. Now let us assume that $\mathbf{y}$ belongs to $\mathbf{V}^{2,1}(Q_\infty)$, and let us prove estimate (8.3) for $s = 2$. We have

$$(-A^* + \lambda_0 I)\mathbf{\Phi}(t) = \int_t^\infty (-A^* + \lambda_0 I)e^{(\tau-t)(A^*-\lambda_0 I)}\mathbf{y}(\tau)\,d\tau.$$

Integrating by parts, we obtain

$$(-A^* + \lambda_0 I)\mathbf{\Phi}(t) = \mathbf{y}(t) + \int_t^\infty e^{(\tau-t)(A^*-\lambda_0 I)}\mathbf{y}'(\tau)\,d\tau.$$

Since $\mathbf{y}'$ belongs to $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$, with estimate (8.3) for $s = 0$ we have

$$\left\| \int_t^\infty e^{(\tau-t)(A^*-\lambda_0 I)}\mathbf{y}'(\tau)\,d\tau \right\|_{\mathbf{V}^{2,1}(Q_\infty)} \le C\|\mathbf{y}'\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))}.$$

Thus

$$\begin{aligned}\|(-A^* + \lambda_0 I)\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}^2(\Omega))} &\le C(\|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}^2(\Omega))} + \|\mathbf{y}'\|_{L^2(0,\infty;\mathbf{V}_n^0(\Omega))})\\ &\le C\|\mathbf{y}\|_{\mathbf{V}^{2,1}(Q_\infty)}.\end{aligned}$$

From regularity results for the stationary Oseen equation (Lemma 8.4), it follows that

$$\|\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}^4(\Omega))} \le C\|(-A^* + \lambda_0 I)\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}^2(\Omega))}.$$

Thus

$$\|\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}^4(\Omega))} \le C\|\mathbf{y}\|_{\mathbf{V}^{2,1}(Q_\infty)}.$$

From (8.2) we deduce

$$\begin{aligned}\|\mathbf{\Phi}'\|_{L^2(0,\infty;\mathbf{V}^2(\Omega))} &\le C(\|(-A^* + \lambda_0 I)\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}^2(\Omega))} + \|\mathbf{y}\|_{L^2(0,\infty;\mathbf{V}^2(\Omega))})\\ &\le C\|\mathbf{y}\|_{\mathbf{V}^{2,1}(Q_\infty)}.\end{aligned}$$

The estimate (8.3) is obtained by interpolation between the estimates obtained for $s = 0$ and $s = 2$. □

LEMMA 8.6. *For all* $\mathbf{y} \in L^2(0,T;\mathbf{V}_n^0(\Omega))$, *the solution to the equation*

$$-\mathbf{\Phi}' = A^*\mathbf{\Phi} + \mathbf{y} \quad \text{in } (0,T), \qquad \mathbf{\Phi}(T) = 0, \tag{8.5}$$

satisfies

$$\|\mathbf{\Phi}\|_{\mathbf{V}^{2,1}(Q_T)} \le C\|\mathbf{y}\|_{L^2(0,T;\mathbf{V}_n^0(\Omega))}. \tag{8.6}$$

If the function $\mathbf{y}$ *belongs to* $\mathbf{V}^{s,s/2}(Q_T)$ *with* $0 \le s < 3/2$, *then the function* $\mathbf{\Phi}$ *belongs to* $\mathbf{V}^{s+2,s/2+1}(Q_T)$, *and the following estimate holds:*

$$\|\mathbf{\Phi}\|_{\mathbf{V}^{s+2,s/2+1}(Q_T)} \le C\|\mathbf{y}\|_{\mathbf{V}^{s,s/2}(Q_T)}. \tag{8.7}$$

Proof. The estimate (8.6) follows from [5, Theorem 1.1]. If $\mathbf{y}$ belongs to $\mathbf{V}^{2,1}(Q_T)$ and if $\mathbf{y}(T) \in \mathbf{V}_0^1(\Omega)$, we have

$$\|\mathbf{\Phi}'\|_{L^2(0,T;\mathbf{V}^2(\Omega))} + \|\mathbf{\Phi}\|_{L^2(0,T;\mathbf{V}^4(\Omega))} \le C\|\mathbf{y}\|_{\mathbf{V}^{2,1}(Q_T)};$$

see, e.g., [29]. The estimate for $0 \le s < 3/2$ is obtained by interpolation between the estimates obtained for $s = 0$ and $s = 2$. For $s < 3/2$ the regularity condition $y(T) \in \mathbf{V}_0^1(\Omega)$ is not needed. ∎

LEMMA 8.7. *Let $\mathbf{\Phi}$ be the solution to* (8.2)*, and let ψ be the pressure associated with $\mathbf{\Phi}$, that is, the function satisfying*

$$
\begin{aligned}
&-\frac{\partial \mathbf{\Phi}}{\partial t} - \nu \Delta \mathbf{\Phi} - (\mathbf{w}\cdot\nabla)\mathbf{\Phi} + (\nabla \mathbf{\Phi})^T \mathbf{w} + \lambda_0 \mathbf{\Phi} + \nabla \psi = \mathbf{y} \quad \text{in } Q_\infty,\\
&\operatorname{div} \mathbf{\Phi} = 0 \quad \text{in } Q_\infty, \quad \mathbf{\Phi} = 0 \quad \text{on } \Sigma_\infty, \quad \mathbf{\Phi}(\infty) = 0 \quad \text{in } \Omega.
\end{aligned} \tag{8.8}
$$

If, in (8.2)*, $\mathbf{y}$ belongs to $L^2(0,\infty;\mathbf{V}_n^0(\Omega))$, then the function ψ belongs to $L^2(0,\infty;H^1(\Omega))$. If, in addition, $\mathbf{y}$ belongs to $\mathbf{V}^{s,s/2}(Q_\infty)$ with $0 \le s \le 2$, then the function ψ belongs to $L^2(0,\infty;H^{s+1}(\Omega)) \cap H^{s/2}(0,\infty;H^1(\Omega))$.*

Proof. Assume that $\mathbf{y}$ belongs to $L^2(0,\infty;\mathbf{V}_n^s(\Omega)) \cap H^{s/2}(0,\infty;\mathbf{V}_n^0(\Omega))$ with $0 \le s \le 2$. From Lemma 8.5 it follows that

$$
\left(\mathbf{y} + \frac{\partial \mathbf{\Phi}}{\partial t} + \nu \Delta \mathbf{\Phi} + (\mathbf{w}\cdot\nabla)\mathbf{\Phi} - (\nabla \mathbf{\Phi})^T \mathbf{w} \right) \in \mathbf{H}^{s,s/2}(Q_\infty).
$$

Thus $\nabla\psi$ belongs to $\mathbf{H}^{s,s/2}(Q_\infty)$, and the proof is complete. ∎

LEMMA 8.8. *If, in* (8.5)*, $\mathbf{y}$ belongs to $\mathbf{V}^{s,s/2}(Q_T) \cap L^2(0,T;\mathbf{V}_n^s(\Omega))$ with $0 \le s < 3/2$, then the function ψ, the pressure associated with $\mathbf{\Phi}$, belongs to $L^2(0,T;H^{s+1}(\Omega)) \cap H^{s/2}(0,T;H^1(\Omega))$.*

Proof. The proof is similar to the previous one. ∎

LEMMA 8.9. *Let $\mathbf{\Phi} \in \mathbf{V}^{2,1}(Q_\infty)$ be the solution to* (8.2)*, and set*

$$
\mathbf{u} = -(M\gamma_\tau + R_A^{-1} M\gamma_n) B^* \mathbf{\Phi}\,.
$$

If, in (8.2)*, the function $\mathbf{y}$ belongs to $L^2(0,\infty;\mathbf{V}_n^s(\Omega)) \cap H^{s/2}(0,\infty;\mathbf{V}_n^0(\Omega))$ with $0 \le s \le 2$, then*

$$
\begin{aligned}
&\|B^*\mathbf{\Phi}\|_{L^2(0,\infty;\mathbf{V}^{s+1/2}(\Gamma)) \cap H^{s/2}(0,\infty;\mathbf{V}^{1/2}(\Gamma))} + \|\mathbf{u}\|_{L^2(0,\infty;\mathbf{V}^{s+1/2}(\Gamma)) \cap H^{s/2}(0,\infty;\mathbf{V}^{1/2}(\Gamma))}\\
&\le C \|\mathbf{y}\|_{\mathbf{V}^{s,s/2}(Q_\infty)}.
\end{aligned} \tag{8.9}
$$

Proof. As in Lemma 3.5, we can show that

$$
\mathbf{u} = -(M\gamma_\tau + R_A^{-1} M\gamma_n) \left(\nu \frac{\partial \mathbf{\Phi}}{\partial \mathbf{n}} - \psi\, \mathbf{n} + c(\psi)\mathbf{n} \right),
$$

where ψ is the pressure associated with $\mathbf{\Phi}$ and $c(\psi)$ is the constant defined in (2.1).

Since $\mathbf{y}$ belongs to $L^2(0,\infty;\mathbf{V}_n^s(\Omega)) \cap H^{s/2}(0,\infty;\mathbf{V}_n^0(\Omega))$, from Lemma 8.5 we deduce that $\mathbf{\Phi}$ belongs to $\mathbf{V}^{s+2,s/2+1}(Q_\infty)$, and from Lemma 8.7 it follows that ψ belongs to $L^2(0,\infty;H^{s+1}(\Omega)) \cap H^{s/2}(0,\infty;H^1(\Omega))$. Thus $\mathbf{u}$ belongs to $L^2(0,\infty;H^{s+1/2}(\Gamma)) \cap H^{s/2}(0,\infty;H^{1/2}(\Gamma))$. ∎

Acknowledgments. The author thanks the referees for their valuable questions and suggestions, which have contributed to improving the first version of the paper. After this paper was submitted, I. Lasiecka informed the author of her joint work with V. Barbu and R. Triggiani [4]. The author thanks I. Lasiecka for providing him a copy of this joint paper and for discussions in Oberwolfach.

REFERENCES

[1] V. Barbu, *Mathematical Methods in Optimization of Differential Systems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.

[2] V. Barbu, *Feedback stabilization of the Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 197–206.

[3] V. Barbu and S. S. Sritharan, *H^∞-control theory of fluid dynamics*, Proc. R. Soc. Lond. Ser. A, 454 (1998), pp. 3009–3033.

[4] V. Barbu, I. Lasiecka, and R. Triggiani, *Boundary stabilization of Navier-Stokes equations*, Mem. Amer. Math. Soc., to appear.

[5] C. Bardos, *A regularity theorem for parabolic equations*, J. Funct. Anal., 7 (1971), pp. 311–322.

[6] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. 1, Birkhäuser Boston, Cambridge, MA, 1992.

[7] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. 2, Birkhäuser Boston, Cambridge, MA, 1993.

[8] E. Fernandez-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl., 83 (2004), pp. 1501–1542.

[9] A. V. Fursikov, M. D. Gunzburger, and L. S. Hou, *Inhomogeneous boundary value problems for the three-dimensional evolutionary Navier–Stokes equations*, J. Math. Fluid Mech., 4 (2002), pp. 45–75.

[10] A. V. Fursikov, M. D. Gunzburger, and L. S. Hou, *Boundary value problems and optimal boundary control for the Navier–Stokes system: The two-dimensional case*, SIAM J. Control Optim., 36 (1998), pp. 852–894.

[11] A. V. Fursikov, *Stabilizability of two-dimensional Navier-Stokes equations with help of a boundary feedback control*, J. Math. Fluid Mech., 3 (2001), pp. 259–301.

[12] A. V. Fursikov, *Stabilization for the 3D Navier-Stokes system by feedback boundary control*, Discrete Contin. Dyn. Syst., 10 (2004), pp. 289–314.

[13] B. B. King and Y.-R. Ou, *Nonlinear dynamic compensator design for flow control in a driven cavity*, in Proceedings of the 34th IEEE Conference on Decision and Control, 1995, pp. 3741–3746.

[14] Y.-R. Ou, *Design of feedback compensators for viscous flow*, in Optimal Control of Viscous Flow, S. S. Sritharan, ed., SIAM, Philadelphia, 1998, pp. 151–180.

[15] A. V. Fursikov, M. Gunzburger, L. S. Hou, and S. Manservisi, *Optimal control for the Navier-Stokes equations*, in Lectures on Applied Mathematics, H.-J. Bungartz, R. H. W. Hoppe, and C. Zenger, eds., Springer, New York, 2000, pp. 143–155.

[16] J.-W. He, R. Glowinski, R. Metcalfe, A. Nordlander, and J. Periaux, *Active control and drag optimization for flow past a circular cylinder*, J. Comput. Phys., 163 (2000), pp. 83–117.

[17] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vol. 1, Springer-Verlag, New York, Berlin, 1994.

[18] G. Grubb, *Nonhomogeneous Dirichlet Navier-Stokes problems in low regularity L_p Sobolev spaces*, J. Math. Fluid Mech., 3 (2001), pp. 57–81.

[19] G. Grubb and V. A. Solonnikov, *Boundary value problems for the nonstationary Navier-Stokes equations treated by pseudo-differential methods*, Math. Scand., 69 (1991), pp. 217–290.

[20] M. D. Gunzburger and H. C. Lee, *Feedback control of vortex shedding*, Trans. ASME J. Appl. Mech., 63 (1996), pp. 828–835.

[21] T. Kato, *Perturbation Theory for Linear Operators*, reprint of the 1980 ed., Springer-Verlag, New York, 1995.

[22] I. Lasiecka and R. Triggiani, *The regulator problem for parabolic equations with Dirichlet boundary control, Part* 1, Appl. Math. Optim., 16 (1987), pp. 147–168.

[23] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations*, Vol. 1, Cambridge University Press, Cambridge, UK, 2000.

[24] J.-L. Lions, *Espaces d'interpolation et domaines de puissances fractionnaires d'opérateurs*, J. Math. Soc. Japan, 14 (1962), pp. 233–241.

[25] D. S. Park, D. M. Ladd, and E. W. Hendricks, *Feedback control of von Karman vortex shedding behind a circular cylinder at low Reynolds numbers*, Phys. Fluids, 6 (1994), pp. 2390–2405.

[26] J.-P. Raymond, *Stokes and Navier-Stokes equations with nonhomogeneous boundary conditions*, submitted.

[27] J.-P. Raymond, *Boundary Feedback Stabilization of the Three Dimensional Navier-Stokes Equations*, in preparation.
[28] J.-P. Raymond, *Local boundary feedback stabilization of the Navier-Stokes equations*, in Control Systems: Theory, Numerics and Applications, available online at http://pos.sissa.it.
[29] V. A. Solonnikov, *Estimates of the solutions of a nonstationary linearized system of Navier-Stokes equations*, in Amer. Math. Soc. Transl. Ser. 2, Vol. 75, AMS, Providence, RI, 1968, pp. 1–116.
[30] M. Reed and B. Simon, *Fourier Analysis, Self-Adjointness*, Academic Press, San Diego, CA, 1975.
[31] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1984.
[32] W. von Wahl, *The Equations of Navier-Stokes and Abstract Parabolic Equations*, Vieweg and Sohn, Braunschweig, Wiesbaden, 1985.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 829–855

A NEW APPROACH TO ADAPTIVE NONLINEAR REGULATION*

F. DELLI PRISCOLI†, L. MARCONI‡, AND A. ISIDORI§

Abstract. This paper shows how the theory of adaptive observers can be effectively used in the design of internal models for nonlinear output regulation. The main result obtained in this way is a new method for the synthesis of adaptive internal models, which substantially enhances the existing theory of adaptive output regulation by allowing nonlinear internal models and more general classes of controlled plants.

Key words. adaptive observers, internal model, regulation, tracking, nonlinear control

AMS subject classifications. 93C40, 93C10, 93B52, 93C15

DOI. 10.1137/S0363012904442069

1. Introduction.

The problem of controlling the output of a system so as to achieve asymptotic tracking of prescribed trajectories and/or asymptotic rejection of disturbances is a central problem in control theory. There are essentially three different possible approaches to the problem: tracking by dynamic inversion, adaptive tracking, and tracking via internal models.

Tracking by dynamic inversion consists of computing a precise initial state and a precise control input (or equivalently a reference trajectory of the state) such that, if the system is accordingly initialized and driven, its output exactly reproduces the reference signal. The computation of such a control input, though, requires "perfect knowledge" of the entire trajectory to be tracked as well as "perfect knowledge" of the model of the controlled plant. Thus, this type of approach is not suited to the presence of large uncertainties on plant parameters or on the reference signal.

Adaptive tracking can successfully handle parameter uncertainties, but it still presupposes the knowledge of the entire trajectory to be tracked (to be used in the design of the adaptation algorithm), and therefore this approach is not suited to the problem of tracking unknown trajectories.

Internal model–based tracking, on the other hand, is able to handle simultaneously uncertainties in plant parameters as well as in the trajectory to be tracked. It has been proven that, if the trajectory to be tracked belongs to the set of all trajectories generated by some fixed dynamical system, a controller which incorporates an internal model of such a system is able to secure asymptotic decay to zero of the tracking error for every possible trajectory in this set and does it robustly with respect to parameter uncertainties. This is in sharp contrast with the two approaches mentioned above, where, in lieu of the assumption that a signal is within a class of signals generated by an exogenous system, one instead needs to assume complete knowledge of the past, present, and future time history of the trajectory to be tracked. It is for this

*Received by the editors March 15, 2004; accepted for publication (in revised form) December 9, 2005; published electronically June 30, 2006. This work was partially supported by NSF grant ECS-0314004, by ONR grant N00014-03-1-0314, and by MIUR.

http://www.siam.org/journals/sicon/45-3/44206.html

†Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza," 00184 Rome, Italy (dellipriscoli@dis.uniroma1.it).

‡Corresponding author. Dipartimento di Elettronica, Informatica e Sistemistica, University of Bologna, 40136 Bologna, Italy (marconi@deis.unibo.it).

§Department of Electrical and Systems Engineering, Washington University, St. Louis, MO 63130 (isidori@ese.wustl.edu).

reason that the internal model–based approach seems to be best suited to problems of tracking unknown reference trajectories or rejecting unknown disturbances.

A generalized problem of tracking and asymptotic disturbance rejection is usually cast as follows. A nonlinear system is given, modeled by equations of the form

$$\begin{aligned} \dot{x} &= f(x, u, v), \\ y &= k(x, v), \\ e &= h(x, v), \end{aligned}$$

with state x, control input u, measured output y, and regulated output e. In this system, v is an exogenous input, which represents actual *disturbances* as well as *commands* to be followed, and it is assumed that, as a function of time, $v(t)$ can be seen as generated by a separate autonomous dynamical system, called the *exosystem*. Generally speaking, the problem of tracking and asymptotic disturbance rejection (sometimes also referred to as the generalized *servomechanism problem* or the *output regulation problem*) is to design a controller so as to obtain a closed-loop system in which

- all trajectories are bounded, and
- the regulated output $e(t)$ asymptotically decays to 0 as $t \to \infty$.

The peculiar aspect of this design problem is the characterization of the class of all possible exogenous inputs (disturbances as well as commands) as the set of all possible solutions of a fixed (finite-dimensional) differential equation. This can be seen as an intermediate choice between two extremes: the (pessimistic) case in which the design is required to obtain certain goals in the presence of the *worst possible* exogenous input, or the (optimistic) case in which the controller is assumed *to have access* to the exogenous input v. In this design problem, the controller does not have access to the exogenous input in real time, but the latter is restricted to range over a "finite-dimensional" set of functions (such as the set of solutions of a fixed differential equation). The vector v may include constant uncertain parameters, which have a trivial dynamics, and hence can be viewed as solutions of a (trivial) differential equation. In other words, in this setting, *any source of uncertainty* (about an actual disturbance affecting the system, an actual trajectory to be tracked, any unknown constant parameter in the plant, or any unknown constant parameter in the exosystem itself) is treated as *uncertainty in the initial condition* of a fixed autonomous finite-dimensional dynamical system, which is then seen as a source of all possible constant, as well time-varying, uncertainties.

For linear multivariable systems this problem was addressed in very elegant geometric terms by Davison, Francis, Wonham [7, 9, 8], and others. A nonlinear enhancement of this theory, which uses a combination of geometry and nonlinear dynamical systems theory, was presented in [16, 12, 11, 5] in the context of solving the problem near an equilibrium, in the presence of exogenous signals which were produced by a Poisson stable system. In particular, Huang showed how, by appropriately designing the internal model, the controlled output could be steered toward zero in spite of plant parameter uncertainties, thus extending to the nonlinear setting one of the most remarkable features of internal model–based design for linear systems. Under suitable hypotheses, the (local) design methods presented in these works have been extended [18, 15, 24] to the case of arbitrarily large (but compact) sets of initial data. A substantial limitation of classical internal model–based control (for linear as well as nonlinear systems) is the sensitivity to parameter uncertainties in the exosystem. This limitation, though, was later removed, under convenient hypotheses, in the paper [25], where the possibility of using techniques of adaptive control to cope with

unknown parameters in the exosystem was successfully demonstrated. The adaptive design methodology of [25] was soon recognized as a useful way to cope with parameter uncertainties in the exosystem. It was, for instance, used in [29] for the solution of a global decentralized regulation problem for a class of nonlinear systems.

In the recent paper [2], the problem in question has been posed in more general terms, not tied, as all previous contributions were, to the existence of a privileged equilibrium point about which the (local as well semiglobal) analysis was conducted. The more general foundations laid in this way make it possible to overcome certain restrictions of the earlier theory, notably the assumption that the controlled plant has an asymptotically stable zero dynamics, which is replaced by the substantially weaker hypothesis that the latter (the zero dynamics) possess a compact attractor. Another major enhancement of this newer approach is a systematic method for the design of nonlinear internal models (see [3]). The presence of parametric uncertainties in the exosystem, however, is not explicitly addressed in [3, 2].

The purpose of the present paper is to show how the problem of handling *parametric uncertainties in the exosystem* can be successfully addressed by means of a new approach which relies, on one hand, on the general nonequilibrium theory developed in [2] and, on the other hand, on the theory of adaptive observers for nonlinear systems pioneered in [1] and [21]. The result obtained in this way is a totally new method for the synthesis of adaptive internal models and substantially extends the adaptive regulation theory presented in [25] by allowing nonlinear internal models and more general classes of controlled plants.

2. Output regulation and limit sets. The purpose of output regulation is to obtain a closed-loop system in which all trajectories with initial conditions in a fixed (but otherwise arbitrary) compact set are bounded and the regulated output converges to zero as time tends to infinity. As shown in [2], intimately associated with this problem is the notion of a *limit set* of a given bounded set of initial conditions. For the reader's convenience, the notion in question is summarized as follows.

Consider an *autonomous* ordinary differential equation

$$\dot{x} = f(x) \tag{1}$$

in which $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, with $f(x)$ a locally Lipschitz function. Let

$$\phi : \quad (t, x) \mapsto \phi(t, x)$$

define the flow of (1). Suppose the flow is forward complete. The *ω-limit set* of a subset $B \subset \mathbb{R}^n$, written $\omega(B)$, is the totality of all points $x \in \mathbb{R}^n$ for which there exists a sequence of pairs (x_k, t_k), with $x_k \in B$ and $t_k \to \infty$ as $k \to \infty$, such that

$$\lim_{k \to \infty} \phi(t_k, x_k) = x \,.$$

In case $B = \{x_0\}$, the set thus defined, $\omega(x_0)$, is precisely the ω-limit set, as defined by Birkhoff, of the point x_0. Note that, in general,

$$\bigcup_{x_0 \in B} \omega(x_0) \subset \omega(B) \,,$$

but the equality may not hold.

It is well known that if $\phi(t, x_0)$ is bounded in positive time, then the set $\omega(x_0)$ is nonempty, compact, invariant, and

$$\lim_{t \to \infty} \operatorname{dist}(\phi(t, x_0), \omega(x_0)) = 0 \,.$$

If B is not just the singleton $\{x_0\}$, the following more general property holds. Recall that a set A is said to *uniformly attract* a set B under the flow of (1) if for every $\varepsilon > 0$ there exists a time $\bar{t}$ such that

$$\operatorname{dist}(\phi(t,x),A) \leq \varepsilon \qquad \text{for all } t \geq \bar{t} \text{ and for all } x \in B.$$

Then the following holds (see [10, page 8]).

LEMMA 1. *If B is a nonempty bounded set for which there is a compact set J which uniformly attracts B (and thus, in particular, if B is any nonempty bounded set whose positive orbit has a bounded closure), then $\omega(B)$ is nonempty, compact, invariant, and uniformly attracts B. Moreover, if $\omega(B) \in \operatorname{int}(B)$, then $\omega(B)$ is stable in the sense of Lyapunov.*

3. Class of systems and main assumptions. In this paper we discuss the design of output regulators for nonlinear systems modeled by equations of the form

$$\begin{aligned} \dot{z} &= f_0(\varrho,w,z) + f_1(\varrho,w,z,e_1)e_1, \\ \dot{e}_1 &= e_2 \\ &\vdots \\ \dot{e}_{r-1} &= e_r, \\ \dot{e}_r &= q(\varrho,w,z,e_1,\dots,e_r) + u, \\ e &= e_1, \\ y &= \operatorname{col}(e_1,\dots,e_r), \end{aligned} \tag{2}$$

with state $(z,e_1,\dots,e_r) \in \mathbb{R}^n \times \mathbb{R}^r$, control input $u \in \mathbb{R}$, regulated output $e \in \mathbb{R}$, measured output $y \in \mathbb{R}^r$, and in which the exogenous (disturbance) input $w \in \mathbb{R}^s$ is generated by an exosystem

$$\dot{w} = s(\varrho,w)\,. \tag{3}$$

In this model, $\varrho \in \mathbb{R}^p$ is a vector of *constant* uncertain parameters, ranging over a fixed compact set P. The vector ϱ is the aggregate of a finite set of uncertain parameters affecting the controlled plant and another, possibly different, set of uncertain parameters affecting the exosystem. These parameters may be regarded as "trivial components" of an "augmented" exogenous input, but for the sake of clarity, and also consistency with some of the earlier literature, their role will be kept separate. Occasionally, throughout the paper, the "augmented" exosystem

$$\begin{aligned} \dot{\varrho} &= 0, \\ \dot{w} &= s(\varrho,w) \end{aligned} \tag{4}$$

will be rewritten in more compact form as

$$\dot{\mathbf{w}} = \mathbf{s}(\mathbf{w}), \tag{5}$$

where $\mathbf{w} = \operatorname{col}(\varrho,w)$.

The functions $f_0(\cdot), f_1(\cdot), q(\cdot), s(\cdot)$ in (2) and (4) are assumed to be at least continuously differentiable. The initial conditions of (2) range over a set $Z \times E$, in which Z is a fixed *compact* subset of $\mathbb{R}^n$ and $E = \{(e_1,\dots,e_r) \in \mathbb{R}^r : |e_i| \leq c\}$, with c a fixed number. The initial conditions of the exosystem (5) range over a *compact* subset $\mathbf{W}$ of $\mathbb{R}^p \times \mathbb{R}^s$. In this framework the problem of output regulation is to design an output feedback regulator of the form

$$\begin{aligned} \dot{\zeta} &= \varphi(\zeta,y), \\ u &= \gamma(\zeta,y) \end{aligned}$$

such that *for all initial conditions* $\mathbf{w}(0) \in \mathbf{W}$ *and* $(z(0), e_1(0), \dots, e_r(0)) \in Z \times E$ *the trajectories of the closed-loop system are bounded and* $\lim_{t\to\infty} e(t) = 0$.

Augmenting (2) with (4) yields a system which, viewing u as input and e as output, has relative degree r. The associated "augmented" zero dynamics, which is forced by the control

$$u = -q(\varrho, w, z, 0, \dots, 0)\,, \tag{6}$$

is given by

$$\begin{aligned} \dot{\varrho} &= 0, \\ \dot{w} &= s(\varrho, w), \\ \dot{z} &= f_0(\varrho, w, z)\,. \end{aligned} \tag{7}$$

The ability to design feedback controllers which solve the problem of output regulation very much depends, among other things, on certain properties of the autonomous system (7), viewed as a system with state (ϱ, w, z) and "output" provided by (6). For notational convenience, throughout the paper, we will sometime rewrite (7) and (6) in more compact form as

$$\dot{\mathbf{z}} = \mathbf{f}_0(\mathbf{z}) \tag{8}$$

and

$$u = \mathbf{q}_0(\mathbf{z}),$$

having set $\mathbf{z} = \mathrm{col}(\varrho, w, z)$. Accordingly, we set $\mathbf{Z} = \mathbf{W} \times Z$.

In what follows, we retain three of the basic assumptions that were introduced in [2] and express certain properties of the augmented zero dynamics (7). The assumptions in question are the following:

Assumption (i): The set $\mathbf{W}$ is a compact differential submanifold (with boundary) of $\mathbb{R}^p \times \mathbb{R}^s$, and $\mathbf{W}$ is invariant for (5).

Assumption (ii): There exists a compact subset $\mathcal{Z}$ of $\mathbf{W} \times \mathbb{R}^n$ which contains the positive orbit of the set $\mathbf{Z}$ under the flow of (8), and $\omega(\mathbf{Z})$ is a differential submanifold (with boundary) of $\mathbf{W} \times \mathbb{R}^n$. Moreover, there exists a number $d_1 > 0$ such that

$$\mathbf{z} \in \mathbf{W} \times \mathbb{R}^n\,, \quad \mathrm{dist}(\mathbf{z}, \omega(\mathbf{Z})) \le d_1 \Rightarrow \mathbf{z} \in \mathbf{Z}.$$

We remark on the above hypotheses that, since the positive orbit of the set $\mathbf{Z}$ under the flow of (8) is bounded, the set $\omega(\mathbf{Z})$, namely the ω-limit set of $\mathbf{Z}$ under the flow of (8), is a nonempty, compact, and invariant subset of $\mathbf{W} \times \mathbb{R}^n$ which uniformly attracts all trajectories of (8) with initial conditions in $\mathbf{Z}$. It can also be shown (as in [2]) that for every $\mathbf{w} \in \mathbf{W}$ there is $z \in \mathbb{R}^n$ such that $(\mathbf{w}, z) \in \omega(\mathbf{Z})$. In what follows, for convenience, the set $\omega(\mathbf{Z})$ will be simply denoted as $\mathcal{A}_0$.

The last condition in Assumption (ii) implies that for every $\varepsilon > 0$ there is $\delta > 0$ such that

$$\mathbf{z}_0 \in \mathbf{W} \times \mathbb{R}^n\,, \quad \mathrm{dist}(\mathbf{z}_0, \mathcal{A}_0) \le \delta \Rightarrow \mathrm{dist}(\mathbf{z}(t, \mathbf{z}_0),\, \mathcal{A}_0) \le \varepsilon \qquad \text{for all } t \ge 0,$$

in which $\mathbf{z}(t, \mathbf{z}_0)$ denotes the solution of (8) passing through $\mathbf{z}_0$ at time $t = 0$.

In the next hypothesis, which will be used in the last part of the paper, we strengthen this property as follows.

Assumption (iii): There exist $M \geq 1$, $a > 0$, and $d_2 \leq d_1$ such that

$$\mathbf{z}_0 \in \mathbf{W} \times \mathbb{R}^n, \quad \operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0) \leq d_2 \Rightarrow \operatorname{dist}(\mathbf{z}(t, \mathbf{z}_0), \mathcal{A}_0) \leq M e^{-at} \operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0).$$

The results presented in [2], like essentially most previous results on output regulation, relied upon the hypothesis that the set of all "feed-forward inputs capable of securing perfect tracking" (that is, the set of inputs of the form $u(t) = \mathbf{q}_0(\mathbf{z}(t))$, with $\mathbf{z}(t)$ a trajectory of the restriction of (8) to $\mathcal{A}_0$) could be seen as a subset of the set of outputs of a suitable linear system. The system in question was used to construct a (linear, as a matter of fact) internal model. This assumption was weakened in the paper [6], where it was assumed that the controls in question can be put in the form $u(t) = \beta(\xi(t))$, with $\beta : \mathbb{R}^d \to \mathbb{R}$ a smooth function and $\xi(t)$ obeying a linear differential equation. A further extension was proposed in [3], where a general method for the construction of fully nonlinear internal models was presented, but the method in question did not allow for the presence of uncertain parameters in the exosystem. In this paper we introduce a different kind of hypothesis, leading to a somewhat more restricted class of internal models, but which—in return—allows for uncertain parameters in the exosystem.

Assumption (iv): There exist positive integers d and q, a C^d map

$$\begin{aligned} \tau \quad : \quad & \mathcal{Z} \to \mathbb{R}^d, \\ & \mathbf{z} \mapsto \tau(\mathbf{z}), \end{aligned}$$

a C^0 map

$$\begin{aligned} \theta \quad : \quad & P \to \mathbb{R}^q, \\ & \varrho \mapsto \theta(\varrho), \end{aligned}$$

an observable pair $(A, C) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{1 \times d}$, a C^1 map $\phi : \mathbb{R} \to \mathbb{R}^d$, and a C^d map $\Omega : \mathbb{R} \to \mathbb{R}^{d \times q}$ such that the following identities (which we call *immersion property*)

$$\frac{\partial \tau}{\partial \mathbf{z}} \mathbf{f}_0(\mathbf{z}) = A\,\tau(\mathbf{z}) + \phi(C\tau(\mathbf{z})) + \Omega(C\tau(\mathbf{z}))\,\theta(\varrho), \tag{9}$$

$$\mathbf{q}_0(\mathbf{z}) = C\,\tau(\mathbf{z}) \tag{10}$$

hold for all $\mathbf{z} \in \mathcal{A}_0$, $\varrho \in P$.

Remark. Note that Assumption (iv) can be rephrased to say that for each initial condition $\mathbf{z}(0) \in \mathcal{A}_0$ of (8), there is a pair $\xi(0), \theta$ such that the control input $u(t) = \mathbf{q}_0(\mathbf{z}(\mathbf{t}))$ (which is the unique input capable of keeping $e(t)$ identically at zero) can be seen as the output of a system of the form

$$\begin{aligned} \dot{\xi} &= A\xi + \phi(u) + \Omega(u)\theta, \\ \dot{\theta} &= 0, \\ u &= C\xi. \end{aligned} \tag{11}$$

Remark. If the vector ϱ of constant uncertain parameters enters linearly in the vector field $\mathbf{f}_0(\mathbf{z})$ and the map $\mathbf{q}_0(\mathbf{z})$ is not affected by ϱ, necessary and sufficient conditions for the existence of a ϱ-independent *diffeomorphism* $\tau(\cdot)$, which satisfies (9)–(10) with $\theta(\varrho) = \varrho$, can be found in [21]. It must be stressed, though, that Assumption (iv) can be fulfilled under much weaker conditions because (a) the immersion map τ is

not necessarily required to be a diffeomorphism (as the dimension d of the final state space is not restricted to be equal to the dimension $p+s+n$ of the original state space); (b) the mapping $\tau(\cdot)$ can be ϱ-dependent; and (c) "overparametrization" is allowed in the definition of the map $\theta(\cdot)$, which in this way is not necessarily the identity map.

Without loss of generality (see [20, page 208]), we will assume throughout that the matrices A and C in (11) have the form

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Furthermore, since the maps $\Omega(\cdot)$ and $\phi(\cdot)$ are continuously differentiable and relations (9)–(10) are supposed to hold over the compact set $\mathcal{A}_0$, it can be assumed without loss of generality that functions $\phi(\cdot)$ and $\Omega(\cdot)$ have compact support. This being the case, the functions in question can be assumed *globally Lipschitz*; i.e., there exist L_ϕ and L_Ω such that

$$|\phi(s_1)-\phi(s_2)| \le L_\phi |s_1-s_2|, \qquad |\Omega(s_1)-\Omega(s_2)| \le L_\Omega |s_1-s_2|$$

for all s_1, s_2.

The immersion property specified in Assumption (iv) enables us to use, for the design of controllers, the theory of adaptive observers pioneered in [1] and developed in [21]. As shown in that theory, convergence of certain estimates occurs under the assumption that certain signals are "persistently exciting." In the present context, a similar role is played (as we will see later in the proof of Lemma 6) by an assumption which can be expressed as follows. For each $\mathbf{z}_0 \in \mathcal{A}_0$, consider the map

$$\begin{aligned} \varphi_{\mathbf{z}_0} \quad : \quad & \mathbb{R} \to \mathbb{R}^{1\times q}, \\ & t \mapsto \sum_{i=1}^{d} C A^{i-1} \Omega^{(d-i)}(C\tau(\mathbf{z}(t,\mathbf{z}_0))) \end{aligned}$$

in which $\Omega^{(k)}(C\tau(\mathbf{z}(t,\mathbf{z}_0)))$ denotes the kth derivative of $\Omega(C\tau(\mathbf{z}(t,\mathbf{z}_0)))$ with respect to the argument t.

Assumption (v): For each $\mathbf{z}_0 \in \mathcal{A}_0$, the q entries of $\varphi_{\mathbf{z}_0}(\cdot)$ are linearly independent functions, namely

$$\varphi_{\mathbf{z}_0}(t)\theta = 0 \qquad \text{for all } t \in \mathbb{R} \Rightarrow \theta = 0.$$

Remark. While the hypothesis in question plays, in the present context, a role equivalent to the role of the assumption of *persistence of excitation* in classical adaptive control (i.e., it implies convergence of certain parameter estimates to their true values) it must be stressed that Assumption (v) here is different, at least for two reasons. First, this assumption does not require (as the assumption of persistency of excitation does) that certain signals have "finite energy" over any finite time interval of the form $[t, t+\delta]$, but simply that certain signals are just linearly independent over the entire time axis. Second, this assumption is only a property of the solutions of $\dot{\mathbf{z}} = \mathbf{f}_0(\mathbf{z})$ with initial conditions in $\mathcal{A}_0$.

In the remaining part of this section we show that *there is no loss of generality* in addressing the simpler case in which the relative degree of (2) is $r = 1$. As a matter of fact, consider the change of variable

$$e_r \;\mapsto\; \tilde{e} := e_r + g^{r-1}a_0e_1 + g^{r-2}a_1e_2 + \cdots + ga_{r-2}e_{r-1},$$

where g is a *positive* design parameter and a_i, $i = 0, \ldots, r-2$, are such that all roots of the polynomial $\lambda^{r-1} + a_{r-2}\lambda^{r-1} + \cdots + a_1\lambda + a_0 = 0$ have negative real part. This changes system (2) into a system of the form

$$\begin{aligned} \dot{\tilde{z}} &= \tilde{f}_0(\varrho, w, \tilde{z}) + \tilde{f}_1(\varrho, w, \tilde{z}, \tilde{e})\tilde{e}, \\ \dot{\tilde{e}} &= \tilde{q}(\varrho, w, \tilde{z}, \tilde{e}, g) + u \end{aligned} \tag{12}$$

in which

$$\tilde{z} = \operatorname{col}(z, e_1, \ldots, e_{r-1}),$$

$$\tilde{f}_0(\varrho, w, \tilde{z}) = \begin{pmatrix} f_0(\varrho, w, z) + f_1(\varrho, w, z, e_1)e_1 \\ e_2 \\ \cdots \\ e_{r-1} \\ -g^{r-1}a_0e_1 - g^{r-2}a_1e_2 - \cdots - ga_{r-2}e_{r-1} \end{pmatrix}, \qquad \tilde{f}_1(\varrho, w, \tilde{z}, \tilde{e}) = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 1 \end{pmatrix},$$

and

$$\begin{aligned} \tilde{q}(\varrho, w, \tilde{z}, \tilde{e}, g) = {} & q(\varrho, w, z, e_1, \ldots, e_r) - g^{r-1}a_0e_2 - \cdots - g^2a_{r-3}e_{r-1} \\ & - ga_{r-2}[\tilde{e} - g^{r-1}a_0e_1 - g^{r-2}a_1e_2 - \cdots - ga_{r-2}e_{r-1}]. \end{aligned} \tag{13}$$

Let the initial conditions of (12) range over a set of the form $Z \times Z_e \times \tilde{E}$, in which $Z_e = \{(e_1, \ldots, e_{r-1} : |e_i| \le c\}$ and $\tilde{E} = \{\tilde{e} : |\tilde{e}| \le \tilde{c}\}$ with

$$\tilde{c} \ge (1 + g^{r-1}a_0 + g^{r-2}a_1 + \cdots + ga_{r-2})c$$

(note the dependence on the choice of the a_i's and of g).

Let system (12) be augmented with (4) and consider a regulation problem with *regulated* output $\tilde{e}$ and *measured* output $\tilde{y} = \tilde{e}$. The system, viewed as a system with input u and output $\tilde{e}$, has relative degree 1 and its zero dynamics, forced by the control

$$u = -\tilde{q}(\varrho, w, \tilde{z}, 0, g), \tag{14}$$

is given by

$$\begin{aligned} \dot{\varrho} &= 0, \\ \dot{w} &= s(\varrho, w), \\ \dot{\tilde{z}} &= \tilde{f}_0(\varrho, w, \tilde{z})\,. \end{aligned} \tag{15}$$

Consistent with the notation used for (7) and (6), the autonomous system (15) and the map (14) can be rewritten in more succinct form as

$$\dot{\tilde{\mathbf{z}}} = \tilde{\mathbf{f}}_0(\tilde{\mathbf{z}}) \tag{16}$$

and

$$u = \tilde{\mathbf{q}}_0(\tilde{\mathbf{z}})\,.$$

Suppose that a controller of the form

$$\begin{aligned} \dot{\zeta} &= \varphi(\zeta, \tilde{y}), \\ u &= \gamma(\zeta, \tilde{y}) \end{aligned} \tag{17}$$

has been found which solves the problem of output regulation thus defined. Then, we immediately realize that the controller

$$\begin{aligned} \dot{\zeta} &= \varphi(\zeta, e_r + g^{\,r-1}\, a_0\, e_1 + g^{\,r-2}\, a_1\, e_2 + \cdots + g\, a_{r-2}\, e_{r-1}), \\ u &= \gamma(\zeta, e_r + g^{\,r-1}\, a_0\, e_1 + g^{\,r-2}\, a_1\, e_2 + \cdots + g\, a_{r-2}\, e_{r-1}) \end{aligned} \tag{18}$$

solves the problem of output regulation for the original plant (2). To this end, note first of all that (18) is an admissible controller for (2) because it is driven only by the components $e_1, \ldots, e_r$ of the measured output y of (2). Trivially, the composition of (2) with (18) differs from the composition of (12) with (17) only by a linear change of coordinates, and for any initial state of (2) in $Z \times E$, the corresponding initial state of (12) is in $Z \times Z_e \times \tilde{E}$. Thus, all trajectories of (2), controlled by (18), with initial conditions in $Z \times E$, are bounded. The trajectories in question are such that $\lim_{t\to\infty} \tilde{e}(t) = 0$. But since

$$\begin{aligned} \dot{e}_1 &= e_2 \\ &\vdots \\ \dot{e}_{r-1} &= -(g^{r-1}a_0e_1 + g^{r-2}a_1e_2 + \cdots + ga_{r-2}e_{r-1}) + \tilde{e} \end{aligned}$$

and the a_i's are coefficients of a Hurwitz polynomial, it is readily concluded that also $\lim_{t\to\infty} e_1(t) = 0$. Therefore (18) solves the problem of output regulation for system (2).

In light of these considerations, what is left to show in order to prove the desired claim (namely, the fact that there is no loss of generality in addressing the problem for systems having relative degree 1) is that the zero dynamics (15) and the associated map (14) inherit, from (7) and (6), the appropriate properties which make the solution of the problem of output regulation possible. Specifically, we will prove that if (7) and (6) satisfy Assumptions (i)–(iv) above, then (15) and (14) satisfy an *identical* set of assumptions, *provided that* the parameter g is chosen sufficiently large. This is formalized in the next lemma.

LEMMA 2. *Suppose that Assumptions* (i)–(v) *hold for* (7) *and* (6). *Set* $\tilde{\mathbf{Z}} = \mathbf{W} \times Z \times Z_e$. *Then there exists* $g^\star > 0$ *such that for all fixed* $g \geq g^\star$ *the following hold:*

(ii)′ *There exists a compact subset* $\tilde{\mathcal{Z}}$ *of* $\mathbf{W} \times \mathbb{R}^n \times \mathbb{R}^{r-1}$ *which contains the positive orbit of the set* $\tilde{\mathbf{Z}}$ *under the flow of* (16), *and* $\tilde{\mathcal{A}}_0 := \omega(\tilde{\mathbf{Z}})$ *is a differential submanifold (with boundary) of* $\mathbf{W} \times \mathbb{R}^n \times \mathbb{R}^{r-1}$. *Moreover, there exists a number* $\tilde{d}_1 > 0$ *such that*

$$\tilde{\mathbf{z}} \in \mathbf{W} \times \mathbb{R}^n \times \mathbb{R}^{r-1}\,, \quad \operatorname{dist}(\tilde{\mathbf{z}}, \tilde{\mathcal{A}}_0) \leq \tilde{d}_1 \Rightarrow \tilde{\mathbf{z}} \in \tilde{\mathbf{Z}}.$$

(iii)′ *There exist* $\tilde{M} \geq 1$, $\tilde{a} > 0$, *and* $\tilde{d}_2 \leq \tilde{d}_1$ *such that*

$$\tilde{\mathbf{z}}_0 \in \mathbf{W} \times \mathbb{R}^n \times \mathbb{R}^{r-1}, \operatorname{dist}(\tilde{\mathbf{z}}_0, \tilde{\mathcal{A}}_0) \leq \tilde{d}_2 \Rightarrow \operatorname{dist}(\tilde{\mathbf{z}}(t, \tilde{\mathbf{z}}_0), \tilde{\mathcal{A}}_0) \leq \tilde{M} e^{-\tilde{a}t} \operatorname{dist}(\tilde{\mathbf{z}}_0, \tilde{\mathcal{A}}_0)$$

in which $\tilde{\mathbf{z}}(t, \tilde{\mathbf{z}}_0)$ *denotes the solution of* (16) *passing through* $\tilde{\mathbf{z}}_0$ *at time* $t = 0$.

(iv)′ *There exists a* C^d *map*

$$\begin{aligned} \tilde{\tau} : \tilde{\mathcal{Z}} &\to \mathbb{R}^d, \\ \tilde{\mathbf{z}} &\mapsto \tilde{\tau}(\tilde{\mathbf{z}}) \end{aligned}$$

such that the immersion property

$$\frac{\partial \tilde{\tau}}{\partial \tilde{\mathbf{z}}} \tilde{\mathbf{f}}_0(\tilde{\mathbf{z}}) = A\tilde{\tau}(\tilde{\mathbf{z}}) + \phi(C\tilde{\tau}(\tilde{\mathbf{z}})) + \Omega(C\tilde{\tau}(\tilde{\mathbf{z}}))\,\theta(\varrho),$$

$$\tilde{\mathbf{q}}_0(\tilde{\mathbf{z}}) = C\tilde{\tau}(\tilde{\mathbf{z}})$$

holds for all $\tilde{\mathbf{z}} \in \tilde{\mathcal{A}}_0$ *and* $\varrho \in P$.

(v)′ *For each* $\tilde{\mathbf{z}}_0 \in \tilde{\mathcal{A}}_0$ *the* q *entries of the map*

$$\begin{aligned} \tilde{\varphi}_{\tilde{\mathbf{z}}_0} : \mathbb{R} &\to \mathbb{R}^{1\times q}, \\ t &\mapsto \sum_{i=1}^{d} CA^{i-1}\Omega^{(d-i)}(C\tilde{\tau}(\tilde{\mathbf{z}}(t, \tilde{\mathbf{z}}_0))) \end{aligned}$$

are linearly independent functions.

Proof. Consider the change of variable

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{r-1} \end{pmatrix} := D_g^{-1} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{r-1} \end{pmatrix} \qquad \text{with} \qquad D_g = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & g & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g^{r-2} \end{pmatrix},$$

which transforms system (15) into

$$\begin{aligned} \dot{z} &= f_0(\varrho, w, z) + f_1(\varrho, w, z, x_1)x_1, \\ \dot{x} &= gAx, \end{aligned} \tag{19}$$

where A is a Hurwitz matrix. Note that if $g > 1$, which we can assume without loss of generality, then $(e_1(0), \dots, e_r(0)) \in E$ implies $x(0) \in Z_e$. System (19) augmented with (4) can be regarded as a particular case of system (64) of the appendix, to which Lemma 7 applies. In particular by property (b) of the latter, there is a number $g^\star > 0$ such that for all $g \geq g^\star$ the positive orbit of $\mathbf{Z} \times Z_e$ under the flow of (4)–(19) is bounded. As a consequence, the ω-limit set $\tilde{\mathcal{A}}_0$ of $\mathbf{Z} \times Z_e$ is a nonempty, compact, invariant set which uniformly attracts $\mathbf{Z} \times Z_e$. We prove now that $\tilde{\mathcal{A}}_0 = \mathcal{A}_0 \times \{0\}$. To this end, note first of all that $\mathcal{A}_0 \times \{0\}$ by construction is contained in $\tilde{\mathcal{A}}_0$. Moreover, x is necessarily 0 at any point of $\mathcal{A}_0$. In fact suppose, by contradiction, that there is a point $(\mathbf{z}, x)$ of $\tilde{\mathcal{A}}_0$ with $x \neq 0$. As $g > 0$ and A is Hurwitz, it follows that the trajectory $x(t)$ of (4)–(19) originating from $(\mathbf{z}, x)$ is unbounded in backward time, which contradicts the fact that $\tilde{\mathcal{A}}_0$ is a compact invariant (in particular in backward time) set. Finally, since $\mathcal{A}_0$ is the ω-limit set of $\mathbf{Z}$ under the flow of (7), we can conclude that necessarily $\tilde{\mathcal{A}}_0 = \mathcal{A}_0 \times \{0\}$. This in particular proves claim (ii)′. Claim (iii)′, namely exponential attractivity of $\tilde{\mathcal{A}}_0$, is an easy consequence of property (a) of Lemma 7 and of the fact that the lower subsystem of (19) is exponentially stable. To prove claim (iv)′, note that (6), (13), (14) imply $\tilde{\mathbf{q}}_0(\tilde{\mathbf{z}})|_{\mathcal{Z}\times\{0\}} = \mathbf{q}_0(\mathbf{z})$. From this,

claims (iv)′ and (v)′ immediately follow from Assumptions (iv) and (v), taking as $\tilde{\tau}(\tilde{\mathbf{z}})$ any differentiable function such that $\tilde{\tau}(\tilde{\mathbf{z}})|_{\mathcal{A}_0\times\{0\}} = \tau(\mathbf{z})|_{\mathcal{A}_0}$. This completes the proof. □

Motivated by the previous considerations and result, in what follows we focus our attention on the case in which $r = 1$, i.e., on the special case in which system (2) is a system of the form

$$\begin{aligned} \dot{z} &= f_0(\varrho, w, z) + f_1(\varrho, w, z, e_1)e_1, \\ \dot{e}_1 &= q(\varrho, w, z, e_1) + u, \\ e &= e_1, \\ y &= e_1, \end{aligned} \tag{20}$$

and we assume that Assumptions (i)–(iv) hold.

4. The adaptive internal model.

4.1. The structure of the regulator. Inspired by various important results on the design of adaptive observers design (see [1, 17, 20, 22]), we propose in what follows to consider a regulator modeled by equations of the form

$$\begin{aligned} u &= \xi_1 + v, \\ \dot{\xi} &= A\xi + \phi(\xi_1) + \Omega(\xi_1)\hat{\theta} + H(X, \xi_1)v - M(X)\mathrm{dzv}_\ell(\hat{\theta}), \\ \dot{\hat{\theta}} &= \beta(X, \xi_1)v - \mathrm{dzv}_\ell(\hat{\theta}), \\ \dot{X} &= FX + G\Omega(\xi_1) \end{aligned} \tag{21}$$

in which ξ_1 denotes the first component of ξ, the matrix X is a $(d-1)\times q$ matrix, and $M(X)$ is a $d \times q$ matrix defined as

$$M(X) = \begin{pmatrix} 0 \\ X \end{pmatrix},$$

while the vectors $H(X, \xi_1), \beta(X, \xi_1)$ and the matrices F, G have the form described below. The function $\mathrm{dzv}_\ell(\cdot)$, whose role is to introduce the so-called switching σ modification in the adaptation law (see [17, Theorem 8.5.2]), is defined as[1]

$$\mathrm{dzv}_\ell(\mathrm{col}(s_1, \ldots, s_q)) = \mathrm{col}(\mathrm{dz}_\ell(s_1), \ldots, \mathrm{dz}_\ell(s_q))$$

in which $\mathrm{dz}_\ell(\cdot)$ is any continuously differentiable function satisfying

$$\mathrm{dz}_\ell(x) = \begin{cases} 0 & \text{if } |x| \le \ell, \\ x & \text{if } |x| \ge \ell + 1, \end{cases} \tag{22}$$

and the amplitude ℓ of the *dead-zone* is chosen so that

$$\ell > \max_{\varrho\in P} |\theta(\varrho)|\,.$$

[1]It must be stressed that the presence of the dead-zone function in (21) is not, strictly speaking, essential, as the same theoretical results presented in the section could be established also by dropping the terms $M(X)\mathrm{dzv}_\ell(\hat{\theta})$ and $\mathrm{dzv}_\ell(\hat{\theta})$ in the second and third lines of (21). Its presence, though, is appropriate if robustness with respect to bounded additive disturbances is sought. Moreover, it sensibly simplifies the proof of Lemma 4.

This controller can be viewed as a "copy" of (11), corrected by an "innovation term" and augmented with an "adaptation law" for $\hat{\theta}$ and with a "filter" which generates the "auxiliary state" X. The additional input v, which is a "stabilizing control," will eventually be taken as $v = -ky$.

Following the theory of adaptive observers of [1] and [21] (see also [22]), the functions $H(X, \xi_1), \beta(X, \xi_1)$ and the matrices F, G of (21) are chosen as follows. Define new variables

$$\begin{aligned} \tilde{\theta} &= \hat{\theta} - \theta(\varrho), \\ \eta &= \xi - M(X)\tilde{\theta} \end{aligned} \tag{23}$$

(note that $\eta_1 = \xi_1$) and observe that, in the new variables, the second equation of (21) reads as follows (for convenience, we omit the arguments (X, ξ_1) in H and β and the argument X in M):

$$\begin{aligned} \dot{\eta} &= A(\eta + M\tilde{\theta}) + \phi(\xi_1) + \Omega(\xi_1)(\theta(\varrho) + \tilde{\theta}) + Hv - \dot{M}\tilde{\theta} - M\beta v \\ &= A\eta + [AM + \Omega(\xi_1) - \dot{M}]\tilde{\theta} + [H - M\beta]v + \phi(\xi_1) + \Omega(\xi_1)\theta(\varrho)\,. \end{aligned} \tag{24}$$

The third equation, instead, becomes trivially

$$\dot{\tilde{\theta}} = \beta v - \mathrm{dzv}_{\ell}\,(\tilde{\theta} + \theta(\varrho))\,.$$

The choices of $H(X, \xi_1), \beta(X, \xi_1)$ and of F, G are meant to simplify the terms

$$[AM + \Omega(\xi_1) - \dot{M}]\tilde{\theta} + [H - M\beta]v$$

in expression (24). First of all, note that choosing

$$H = M\beta + K,$$

with K a constant vector (whose expression will be determined later), causes the second term to become equal to Kv. As for the first term, the idea is to impose that

$$[AM + \Omega(\xi_1) - \dot{M}]\tilde{\theta} = b\beta^{\mathrm{T}}\tilde{\theta},$$

in which b is a $d \times 1$ fixed vector. The identity in question holds if M satisfies

$$\dot{M} = (A - bCA)M + (I - bC)\Omega(\xi_1)$$

and β is taken as

$$\beta^{\mathrm{T}} = CAM + C\Omega(\xi_1)\,.$$

In this way, the second equation of (21) takes the simplified form

$$\dot{\eta} = A\eta + b\beta^{\mathrm{T}}\tilde{\theta} + Kv + \phi(\eta_1) + \Omega(\eta_1)\theta(\varrho), \tag{25}$$

to which we will return later. To show that the required differential equation for M can be enforced, pick a column vector $b = \mathrm{col}(1, b_2, \ldots, b_d)$. Then, bearing in mind the definition of M, it is easily realized that the required differential equation holds if the matrices F and G in the differential equation for X have the form (see [21])

$$F = \begin{pmatrix} -b_2 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ -b_{d-1} & 0 & \cdots & 0 & 1 \\ -b_d & 0 & \cdots & 0 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} -b_2 & 1 & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ -b_{d-1} & 0 & \cdots & 0 & 1 & 0 \\ -b_d & 0 & \cdots & 0 & 0 & 1 \end{pmatrix}. \tag{26}$$

In summary, the quantities $H(X,\xi_1), \beta(X,\xi_1), F, G$ which appear in the controller (21) are determined as follows: F and G are the matrices in (26), $\beta(X,\xi_1)$ is chosen as

$$\beta(X,\xi_1) = \left[CA\begin{pmatrix}0\\X\end{pmatrix} + C\Omega(\xi_1)\right]^{\mathrm{T}}, \tag{27}$$

and $H(X,\xi_1)$ is chosen as

$$H(X,\xi_1) = \begin{pmatrix}0\\X\end{pmatrix}\left[CA\begin{pmatrix}0\\X\end{pmatrix} + C\Omega(\xi_1)\right]^{\mathrm{T}} + K\,. \tag{28}$$

The vector b, whose entries determine the choice of F and G and the parameter K, which appears in the expression of $H(X,\xi_1)$, will be chosen later.

The controller thus defined determines a closed-loop system which, in the coordinates indicated above, can be written as (recall that $e_1 = e$)

$$\begin{aligned}
\dot{\varrho} &= 0,\\
\dot{w} &= s(\varrho, w),\\
\dot{z} &= f_0(\varrho, w, z) + f_1(\varrho, w, z, e)e,\\
\dot{e} &= q(\varrho, w, z, e) + \eta_1 + v,\\
\dot{\eta} &= A\eta + b\beta^{\mathrm{T}}\tilde{\theta} + Kv + \phi(\eta_1) + \Omega(\eta_1)\theta(\varrho),\\
\dot{\tilde{\theta}} &= \beta v - \mathrm{dzv}_\ell\,(\tilde{\theta} + \theta(\varrho)),\\
\dot{X} &= FX + G\Omega(\eta_1),
\end{aligned} \tag{29}$$

where β is a function of X and η_1. This system, viewed as a system with input v and output e, has relative degree 1, and its zero dynamics is that of

$$\begin{aligned}
\dot{\varrho} &= 0,\\
\dot{w} &= s(\varrho, w),\\
\dot{z} &= f_0(\varrho, w, z),\\
\dot{\eta} &= A\eta - K[q(\varrho, w, z, 0) + \eta_1] + b\beta^{\mathrm{T}}\tilde{\theta} + \phi(\eta_1) + \Omega(\eta_1)\theta(\varrho),\\
\dot{\tilde{\theta}} &= -\beta[q(\varrho, w, z, 0) + \eta_1] - \mathrm{dzv}_\ell\,(\tilde{\theta} + \theta(\varrho)),\\
\dot{X} &= FX + G\Omega(\eta_1).
\end{aligned} \tag{30}$$

Intuition suggests that if the latter have convenient asymptotic properties, and in particular possess a locally exponentially stable compact attractor, an additional control of the form $v = -ke$ (with large $k > 0$) should be able to solve the problem of output regulation. This intuition relies upon the extension of a consolidated design paradigm for semiglobal stabilization of equilibrium points of minimum-phase nonlinear systems, namely the fact that a high-gain feedback $v = -ke$ is sufficient to guarantee asymptotic stability (with an arbitrarily large domain of attraction) for a system of the form

$$\dot{x} = f(x) + g(x)e, \qquad \dot{e} = h(x) + v$$

if $x = 0$ is exponentially stable for $\dot{x} = f(x)$ and $h(0) = 0$ (see, e.g., [14]). This intuition will be rendered rigorous in section 5 (see also the appendix). For the time being we are interested in studying the asymptotic properties of (30). This, indeed, is the main objective of the next three subsections.

4.2. Trajectories of (30) are bounded. In studying the asymptotic properties of this system, it is convenient to take advantage of the "immersion" assumption, Assumption (iii), introduced above. Specifically, suppose that the initial conditions for ϱ, w, z are taken in the set $\mathbf{Z}$, a subset of a set $\mathcal{Z}$ which by hypothesis is positively invariant for the subsystem formed by the first three equations of (30). Thus, for any such initial conditions and for any $t \geq 0$, the function $\tau(\varrho, w(t), z(t))$ is well defined, and it is legitimate to consider the change of variables

$$\chi = \eta - \tau(\varrho, w, z)\,.$$

This transforms system (30) into a system of the form (use here (9) and (10), which hold on $\mathcal{A}_0 \subset \mathcal{Z}$)

$$\begin{aligned}
\dot{\varrho} &= 0,\\
\dot{w} &= s(\varrho, w),\\
\dot{z} &= f_0(\varrho, w, z),\\
\dot{\chi} &= (A - KC)\chi + b\beta^{\mathrm{T}}\tilde{\theta} + \Delta(\chi_1, \tau_1, \theta) + \varsigma_1(\varrho, w, z),\\
\dot{\tilde{\theta}} &= -\beta\chi_1 - \mathrm{dzv}_\ell\,(\tilde{\theta} + \theta(\varrho)) + \beta\varsigma_2(\varrho, w, z),\\
\dot{X} &= FX + G\Omega(\chi_1 + \tau_1)\,,
\end{aligned} \tag{31}$$

with $\chi_1 = C\chi$ and $\tau_1 = C\tau$, in which

$$\Delta(\chi_1, \tau_1, \theta) = \phi(\chi_1 + \tau_1) - \phi(\tau_1) + [\Omega(\chi_1 + \tau_1) - \Omega(\tau_1)]\theta(\varrho)$$

is a term which vanishes at $\chi_1 = 0$ and

$$\begin{aligned}
\varsigma_1(\varrho, w, z) &= -K(q(\varrho, w, z, 0) + \tau_1(\varrho, w, z)) + A\tau(\varrho, w, z) + \phi(\tau_1(\varrho, w, z))\\
&\quad + \Omega(\tau_1(\varrho, w, z))\theta(\varrho) - \frac{\partial\tau}{\partial z} f_0(\varrho, w, z) - \frac{\partial\tau}{\partial w} s(\varrho, w),\\
\varsigma_2(\varrho, w, z) &= \tau_1(\varrho, w, z) + q(\varrho, w, z, 0)
\end{aligned} \tag{32}$$

are terms vanishing on $\mathcal{A}_0$. In particular note that, since $\phi(y)$ and $\Omega(y)$ can be taken to be globally Lipschitz and θ ranges over a compact set, there exists a number L such that

$$|\Delta(\chi_1, \tau_1, \theta)| \leq L_\phi\,|\chi_1| + L_\Omega|\chi_1|\,|\theta| \leq L|\chi_1|$$

for all $\chi_1 \in \mathbb{R}$, all $\mathbf{z} \in \mathcal{Z}$, and all $\theta \in P$.

The idea is now to choose the b_i's and K so that system (31) has certain desirable asymptotic properties. To this end, let b_i be such that the polynomial

$$p(\lambda) = \lambda^{d-1} + b_2\lambda^{d-2} + \cdots + b_{d-1}\lambda + b_d \tag{33}$$

has $d - 1$ roots with negative real part. As a consequence the matrix F in the last equation of (31) is Hurwitz. This, in view of the assumptions on the first three equations, suggests that the asymptotic properties of (31) are entirely determined by those of the fourth and fifth equations.

As indicated in [21, Theorem 2.1], the appropriate choice for K in (25) is

$$K = Ab + \lambda b, \tag{34}$$

in which $\lambda > 0$. To see why this is the case, note first of all that, using a little algebra, it is not difficult to prove the following.

LEMMA 3. *Choose K as in* (34) *and set*

$$T = \begin{pmatrix} 1 & 0 \\ \hat{b} & I \end{pmatrix}, \qquad \hat{b} = -\begin{pmatrix} b_2 \\ \cdot \\ b_{d-1} \\ b_d \end{pmatrix}.$$

Then

$$T(A - KC)T^{-1} = \begin{pmatrix} -\lambda & \hat{c} \\ 0 & F \end{pmatrix}, \quad Tb = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad CT^{-1} = C\,,$$

in which $\hat{c} = (\,1 \quad 0 \quad \cdots \quad 0\,)$ and F is the matrix defined in (26).

From this fact, standard arguments can be invoked to claim boundedness of the trajectories of (31). In fact, the following result holds.

LEMMA 4. *Suppose Assumptions* (i), (ii), (iv) *hold. There is a number λ^* such that, if $\lambda \geq \lambda^*$, all trajectories of* (31) *are bounded.*

Proof. First of all, recall that, by Assumption (ii), $(\varrho, w(t), z(t)) \in \mathcal{Z}$ for all $t \geq 0$, where $\mathcal{Z}$ is a compact set. Thus, looking at expression $\varsigma_1(\varrho, w, z)$ of (32), it is seen that there exists a number $\bar{\varsigma}_1$ (depending on the design parameter λ because the latter appears in K) such that

$$|\varsigma_1(\varrho, w(t), z(t))| \leq \bar{\varsigma}_1 \qquad \text{for all } t \geq 0\,. \tag{35}$$

Observe also that, so long as trajectories of (31) exist on some interval $[0, T]$, $|X(t)|$ is bounded by a number which depends only on $|X(0)|$ (because $|\Omega(\cdot)|$, having compact support, is bounded by some fixed number). As a consequence, $|\beta|$ also is bounded, again by a number which depends only on $|X(0)|$. Thus, system (31) cannot have finite escape times. Moreover, looking at the expression of $\varsigma_2(\varrho, w, z)$ in (32), it is possible to conclude the existence of a number $\bar{\varsigma}_2$ such that

$$|\beta\varsigma_2(\varrho, w(t), z(t))| \leq \bar{\varsigma}_2 \qquad \text{for all } t \geq 0\,. \tag{36}$$

With this in mind, to prove the lemma it remains to show that also χ and $\tilde{\theta}$ are bounded. To this end, let χ be partitioned as $\chi = \mathrm{col}(\chi_1, \chi_2)$, in which χ_2 is a $(d-1) \times 1$ vector, and change χ_2 into

$$\zeta = \hat{b}\chi_1 + \chi_2\,.$$

In this way, the fourth and fifth equations of system (31) are changed into

$$\begin{aligned} \dot{\chi}_1 &= -\lambda\chi_1 + \hat{c}\zeta + \beta^{\mathrm{T}}\tilde{\theta} + C\Delta(\chi_1, \tau_1, \theta) + C\varsigma_1(\varrho, w, z), \\ \dot{\zeta} &= F\zeta + (\,\hat{b} \quad I\,)\,\Delta(\chi_1, \tau_1, \theta) + (\,\hat{b} \quad I\,)\,C\varsigma_1(\varrho, w, z), \\ \dot{\tilde{\theta}} &= -\beta\chi_1 - \mathrm{dzv}_\ell(\tilde{\theta} + \theta(\varrho)) + \beta\varsigma_2(\varrho, w, z,)\,. \end{aligned} \tag{37}$$

With this in mind, choose for (37) the Lyapunov function

$$V(\chi_1, \zeta, \tilde{\theta}) = \chi_1^2 + \zeta^{\mathrm{T}}P\zeta + \tilde{\theta}^{\mathrm{T}}\tilde{\theta}\,, \tag{38}$$

in which P is the positive definite solution of $PF + F^{\mathrm{T}}P = -I$. This yields

$$\begin{aligned}\dot V =& -2\lambda\chi_1^2 + 2\chi_1\hat c\zeta + 2\chi_1\beta^{\mathrm{T}}\tilde\theta + 2\chi_1 C\Delta(\chi_1,\tau_1,\theta) + 2\chi_1 C\varsigma_1(\varrho,w,z)\\ &-|\zeta|^2 + 2\zeta^{\mathrm{T}}P\begin{pmatrix}\hat b & I\end{pmatrix}\Delta(\chi_1,\tau_1,\theta) + 2\zeta^{\mathrm{T}}P\begin{pmatrix}\hat b & I\end{pmatrix}C\varsigma_1(\varrho,w,z)\\ &-2\tilde\theta^{\mathrm{T}}\beta\chi_1 - 2\tilde\theta^{\mathrm{T}}\mathrm{dzv}_\ell(\tilde\theta+\theta(\varrho)) + 2\tilde\theta^{\mathrm{T}}\beta\varsigma_2(\varrho,w,z)\\ \le& -2\lambda\chi_1^2 - |\zeta|^2 - 2\tilde\theta^{\mathrm{T}}\mathrm{dzv}_\ell(\tilde\theta+\theta(\varrho)) + L_1|\chi_1|^2 + L_2|\chi_1|\,|\zeta|\\ &+L_3|\chi_1||\varsigma_1(\varrho,w,z)| + L_4|\zeta||\varsigma_1(\varrho,w,z)| + 2|\tilde\theta||\beta\varsigma_2(\varrho,w,z)|,\end{aligned} \tag{39}$$

in which L_i, $i = 1,\dots,4$, are suitable positive constants. By completing the squares and using (35)–(36), we obtain

$$\dot V \le -\left(2\lambda - L_1 + \frac{1}{2}L_2^2\right)\chi_1^2 - \frac{1}{2}|\zeta|^2 - 2\tilde\theta^{\mathrm{T}}\mathrm{dzv}_\ell(\tilde\theta+\theta(\varrho)) + L_3|\chi_1|\bar\varsigma_1 + L_4|\zeta|\bar\varsigma_1 + 2|\tilde\theta|\bar\varsigma_2. \tag{40}$$

Bearing in mind the definition (22) and the choice of ℓ, observe that

$$\tilde\theta^{\mathrm{T}}\mathrm{dzv}_\ell(\tilde\theta+\theta(\varrho)) \ge 0 \qquad \text{for all } \tilde\theta\in\mathbb{R}^q \quad \text{and} \quad \varrho\in P\,. \tag{41}$$

It is also easy to check that for any $\delta > \sqrt{q}(2\ell+1)$ there is a positive number c_1 such that

$$|\tilde\theta| \ge \delta \Rightarrow 2\tilde\theta^{\mathrm{T}}\mathrm{dzv}_\ell(\tilde\theta+\theta(\varrho)) \ge c_1|\tilde\theta|^2 \qquad \text{for all } \tilde\theta\in\mathbb{R}^q \quad \text{and} \quad \varrho\in P\,. \tag{42}$$

Now pick a value of λ large enough so that $\bar\lambda := 2\lambda - L_1 - L_2^2/2 > 0$ and note that inequality (40), by adding and subtracting the term $c_1|\tilde\theta|^2$ in the right-hand side of (40) and using property (42), yields

$$\dot V \le -c_2|(\chi_1,\zeta,\tilde\theta)|^2 + c_3|(\chi_1,\zeta,\tilde\theta)| + c_1\delta^2,$$

in which $c_2 = \min\{\bar\lambda, \frac{1}{2}, c_1\}$ and $c_3 = 3\max\{L_3\bar\varsigma_1, L_4\bar\varsigma_1, 2\bar\varsigma_2\}$. From this, boundedness of $\chi(t)$ and $\tilde\theta(t)$ follows by standard arguments. ∎

We can therefore draw the following conclusion about system (30). Let the initial conditions $\eta(0), \tilde\theta(0), X(0)$ be taken in fixed compact sets $\mathbf{H}$, Θ, $\mathbf{X}$.[2] Then, the positive orbit of the set

$$\mathbf{B} = \mathbf{Z}\times\mathbf{H}\times\Theta\times\mathbf{X}$$

under the flow of (30) is bounded. As a consequence $\omega(\mathbf{B})$, the ω-limit set of $\mathbf{B}$ under the flow of (30), is a nonempty, compact, and invariant set, which uniformly attracts all trajectories of (30) with initial conditions in $\mathbf{B}$.

4.3. The limit set of (30). We proceed now to investigate the structure of the set $\omega(\mathbf{B})$. To this end, we look at the equivalent system (31); we note that the first three equations are independent of the last three and we rewrite them in compact form as in (8) (and consistently we rewrite the terms $\varsigma_i(\varrho,w,z)$, $i = 1,2$, as $\varsigma_i(\mathbf{z})$ and

[2] Recall that $\tilde\theta(t) = \hat\theta(t) - \theta(\varrho)$ and $\eta(t) = \xi(t) - \beta(X(t),\xi_1(t))\tilde\theta(t)$. Thus, to establish boundedness of trajectories when $\xi(0)$, $\hat\theta(0)$, and $X(0)$ are taken in fixed compact sets, it suffices to consider the case in which $\tilde\theta(0)$, $\eta(0)$, and $X(0)$ are taken in fixed compact sets.

$\tau(z, w, \varrho)$ as $\tau(\mathbf{z})$). In particular, because of the special triangular structure of (31), we note that if $(\mathbf{z}, \chi, \tilde{\theta}, X)$ is a point of $\omega(\mathbf{B})$, necessarily $\mathbf{z}$ is a point in the ω-limit set of $\mathbf{Z}$ under the flow of (8); that is, $\mathbf{z}$ is a point of $\mathcal{A}_0$. This implies that on $\omega(\mathbf{B})$ we have $\varsigma_1(\mathbf{z}) = \varsigma_2(\mathbf{z}) = 0$, and thus system (31) simplifies to

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{f}_0(\mathbf{z}), \\ \dot{\chi} &= (A - KC)\chi + b\beta^{\mathrm{T}}\tilde{\theta} + \Delta(\chi_1, \tau_1, \theta), \\ \dot{\tilde{\theta}} &= -\beta\chi_1 - \mathrm{dzv}_\ell(\tilde{\theta} + \theta(\varrho)), \\ \dot{X} &= FX + G\Omega(\chi_1 + \tau_1(\mathbf{z})). \end{aligned} \tag{43}$$

In the following we will be able to prove that on points of $\omega(\mathbf{B})$, necessarily $\chi = 0$, $\tilde{\theta} = 0$, and the value of X is entirely determined by the properties of the system

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{f}_0(\mathbf{z}), \\ \dot{X} &= FX + G\Omega(\tau_1(\mathbf{z})), \end{aligned} \tag{44}$$

in which $\tau_1(\mathbf{z})$ is the obvious abbreviated notation for $\tau_1(z, w, \varrho)$. To this end, we describe first an interesting feature of the latter system.

LEMMA 5. *The graph of the C^1 map*

$$\begin{aligned} \sigma : \mathcal{A}_0 &\to \mathbb{R}^{(d-1)\times q}, \\ \mathbf{z} &\mapsto \int_{-\infty}^{0} e^{-Fs} G\Omega(\tau_1(\mathbf{z}(s, \mathbf{z})))ds \end{aligned}$$

is invariant for (44). *Moreover, if Assumption* (v) *holds, then for any $\mathbf{z}_0 \in \mathcal{A}_0$, the identity*

$$\theta^{\mathrm{T}}\beta(\sigma(\mathbf{z}(t, \mathbf{z}_0)), \tau_1(\mathbf{z}(t, \mathbf{z}_0))) = 0 \qquad \textit{for all } t \in \mathbb{R} \tag{45}$$

implies $\theta = 0$.

Proof. Let $\mathbf{z}(t, \mathbf{z}_0)$ denote the solution of (8) passing through $\mathbf{z}_0$ at time $t = 0$ and note that, if $\mathbf{z}_0 \in \mathcal{A}_0$, then $\mathbf{z}(t, \mathbf{z}_0) \in \mathcal{A}_0$ for all t (thus, in particular, since $\mathcal{A}_0$ is compact, $|\mathbf{z}(t, \mathbf{z}_0)|$ is bounded by a number which depends only on $\mathcal{A}_0$). Then, since F is a Hurwitz matrix, the map $\sigma(\cdot)$ is well defined. A simple calculation shows that

$$\sigma(\mathbf{z}(t, \mathbf{z}_0)) = e^{Ft}\sigma(\mathbf{z}_0) + \int_0^t e^{F(t-s)} G\Omega(\tau_1(\mathbf{z}(s, \mathbf{z}_0)))ds\,.$$

This shows that

$$\mathrm{graph}(\sigma) = \{(\mathbf{z}, X) : \mathbf{z} \in \mathcal{A}_0, X = \sigma(\mathbf{z})\}$$

is invariant for (44).

To prove the second property in the lemma, let $\bar{C}$ be a $(d-1)$-dimensional row vector in which all entries are 0, except the first one, which is equal to 1. Bearing in

mind the definitions of $\beta(\cdot,\cdot)$ and $\sigma(\cdot)$, it is seen that

$$\begin{aligned}\beta^{\mathrm{T}}(\sigma(\mathbf{z}(t,\mathbf{z}_0)),\tau_1(\mathbf{z}(t,\mathbf{z}_0))) &= \bar{C}\int_{-\infty}^{0} e^{-Fs}G\Omega(\tau_1(\mathbf{z}(s,\mathbf{z}(t,\mathbf{z}_0))))ds + C\Omega(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\\ &= \bar{C}\int_{-\infty}^{0} e^{-Fs}G\Omega(\tau_1(\mathbf{z}(s+t,\mathbf{z}_0)))ds + C\Omega(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\\ &= \bar{C}\int_{-\infty}^{t} e^{F(t-s')}G\Omega(\tau_1(\mathbf{z}(s',\mathbf{z}_0)))ds' + C\Omega(\tau_1(\mathbf{z}(t,\mathbf{z}_0))).\end{aligned}$$

If (45) holds for some $\theta\in\mathbb{R}^q$, then

$$\bar{C}\int_{-\infty}^{t} e^{F(t-s)}G\Omega(\tau_1(\mathbf{z}(s,\mathbf{z}_0)))\theta ds + C\Omega(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\theta = 0 \qquad \text{for all } t\in\mathbb{R}. \tag{46}$$

Now let

$$x(t) := \int_{-\infty}^{t} e^{F(t-s)}G\Omega(\tau_1(\mathbf{z}(s,\mathbf{z}_0)))\theta ds$$

and note that this satisfies

$$\dot{x}(t) = Fx(t) + G\Omega(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\theta\,. \tag{47}$$

Bearing in mind the definitions of F and G, letting x_i denote the i-component of x and $\Omega_i(\cdot)$ the ith row of $\Omega(\cdot)$, this yields

$$\dot{x}_i(t) = b_{i+1}(x_1(t) + \Omega_1(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\theta) + x_{i+1}(t) + \Omega_{i+1}(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\theta, \qquad i=1,\dots,d-1, \tag{48}$$

(with the understanding that $x_d = 0$). Now note that, by definition of $\bar{C}$ and C, condition (46) implies

$$x_1(t) + \Omega_1(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\theta = 0 \qquad \text{for all } t\in\mathbb{R}$$

from which, using (48) recursively, it is easy to conclude that

$$\sum_{i=1}^{d}\Omega_i^{(d-i)}(\tau_1(\mathbf{z}(t,\mathbf{z}_0)))\theta = 0 \qquad \text{for all } t\in\mathbb{R}.$$

Observing that $\Omega_i(\cdot) = CA^{i-1}\Omega(\cdot)$, and bearing in mind the definition of map $\varphi_{\mathbf{z}_0}(\cdot)$, this proves the claim. □

Remark. Consider the restriction of (44) to $\mathcal{A}_0\times\mathbb{R}^{(d-1)\times q}$. Since the graph of $\sigma(\cdot)$ is invariant for (44), changing X into $\tilde{X} = X - \sigma(\mathbf{z})$ yields

$$\begin{aligned}\dot{\mathbf{z}} &= \mathbf{f}_0(\mathbf{z}),\\ \dot{\tilde{X}} &= F\tilde{X}.\end{aligned}$$

We see from this that the solution $X(t)$ of (44) passing through X_0 at time $t=0$ can be expressed as

$$X(t) = e^{Ft}[X_0 - \sigma(\mathbf{z}_0)] + \sigma(\mathbf{z}(t,\mathbf{z}_0)). \tag{49}$$

Using this result, it is easy to show that $\omega(\mathbf{B})$ assumes a very simple structure. As a matter of fact, the following result holds.

LEMMA 6. *Suppose that, in addition to Assumptions* (i), (ii), (iv), *also Assumption* (v) *holds. Then the values of* χ *and* $\tilde{\theta}$ *on any point of* $\omega(\mathbf{B})$ *are necessarily zero.*

Proof. By contradiction, suppose a point $\mathbf{p} = (\mathbf{z}, \chi_0, \tilde{\theta}_0, X)$ with either $\chi_0 \neq 0$ or $\tilde{\theta}_0 \neq 0$ is in $\omega(\mathbf{B})$. Since $\omega(\mathbf{B})$ is compact and invariant, in particular in backward time, the backward trajectory of (43) starting at this point is bounded. Along this trajectory, the function

$$V(t) := V(\chi_1(t), \zeta(t), \tilde{\theta}(t))$$

in (38) satisfies $V(t) \leq C$ for all $t \leq 0$ for some $C > 0$. Moreover, since $e(\mathbf{z}) = 0$ on $\omega(\mathbf{B})$, the same computations indicated in the proof of Lemma 4 show that

$$\dot{V}(t) \leq -\left(2\lambda - L_1 - \frac{1}{2}L_2^2\right)|\chi_1(t)|^2 - \frac{1}{2}|\zeta(t)|^2 - 2\tilde{\theta}(t)\mathrm{dzv}_\ell(\tilde{\theta}(t) + \theta(\varrho)),$$

in which L_1 and L_2 are the same constants introduced in the proof of Lemma 4. From this, using property (41), it turns out that if $\lambda \geq \lambda^\star$ (where $\lambda^\star$ is the same as in Lemma 4) then $V(t)$ is nonincreasing along trajectories. As a consequence, since $V(t)$ is bounded, there must exist a finite number V_α such that

$$\lim_{t \to -\infty} V(t) = V_\alpha \,.$$

The trajectory in question is attracted, in backward time, by its own α-limit set $\alpha(\mathbf{p})$, which, as it is well known, is nonempty, compact, and invariant. Moreover, by definition, the function $V(\chi_1, \zeta, \tilde{\theta})$ has the same value V_α at any point of $\alpha(\mathbf{p})$.

Now, as in the classical proof of LaSalle's invariance principle, pick an initial condition $\hat{\mathbf{p}}$ in the set $\alpha(\mathbf{p})$ and consider the corresponding trajectory of (43), which remains in $\alpha(\mathbf{p})$ for all times. Along such a trajectory, $V(t)$ is constantly equal to V_α, and hence

$$\chi_1(t) = 0, \qquad \zeta(t) = 0, \qquad \mathrm{dzv}_\ell(\tilde{\theta}(t) + \theta(\varrho)) = 0 \qquad \text{for all } t \in \mathbb{R}.$$

Entering these constraints in (43), and observing that the vector b is nonzero, it is seen that necessarily

$$\begin{aligned} &\tilde{\theta}^{\mathrm{T}}\beta(X, \tau_1(\mathbf{z})) = 0, \\ &\dot{\tilde{\theta}} = 0, \\ &\dot{X} = FX + G\Omega(\tau_1(\mathbf{z})). \end{aligned}$$

The second condition shows that $\tilde{\theta}(t)$ is a constant, say $\tilde{\theta}^*$, along such a trajectory. The third condition says that $X(t)$ is a solution of

$$\dot{X} = FX + G\Omega(\tau_1(\mathbf{z})).$$

Now, since F is Hurwitz, it is seen from (49) that $X(t)$ is bounded for $t \leq 0$ only if $X(0) = \sigma(\mathbf{z}(0))$, where $\sigma(\cdot)$ is the map introduced in Lemma 5, in which case $X(t) = \sigma(\mathbf{z}(t))$. Since $X(t)$ has to be bounded because $\alpha(\mathbf{p})$ is compact, it follows

that $X(t)$ is necessarily equal to $\sigma(\mathbf{z}(t))$. This being the case, bearing in mind the expression of β and the definition of the map $\varphi(\cdot)$, the first condition shows that necessarily

$$(\tilde{\theta}^*)^{\mathrm{T}}\beta(\sigma(\mathbf{z}(t,\mathbf{z}_0)),\tau_1(\mathbf{z}(t,\mathbf{z}_0))) = 0 \qquad \text{for all } t \in \mathbb{R}.$$

From this, in view of Assumption (v) and the result in Lemma 5, it follows that $\tilde{\theta}^* = 0$. It is seen in this way that $(\chi_1, \zeta, \tilde{\theta}) = (0,0,0)$ at any point of $\alpha(\mathbf{p})$, and this proves that $V_\alpha = 0$. But this is a contradiction because $V(t)$ is nonincreasing along trajectories and $V(0)$ is strictly positive if either $\chi_0 \neq 0$ or $\tilde{\theta}_0 \neq 0$. □

To complete the analysis, it remains to determine the values of X on points of $\omega(\mathbf{B})$. Knowing that $\chi_1 = 0$ on any such points, it follows from the previous analysis, and in particular from Lemma 5, that $X = \sigma(\mathbf{z})$. Altogether, bearing in mind how systems (30) and (43) are related, the following conclusion holds.

PROPOSITION 1. *Under Assumptions* (i), (ii), (iv), *and* (v) *the set* $\omega(\mathbf{B})$ *is the graph of a continuous map defined on* $\mathcal{A}_0$. *Any point of* $\omega(\mathbf{B})$ *is a point* $(\mathbf{z}, \eta, \tilde{\theta}, X)$ *in which* $\mathbf{z} \in \mathcal{A}_0$ *and*

$$\eta = \tau(\mathbf{z}), \qquad \tilde{\theta} = 0, \qquad X = \sigma(\mathbf{z})\,.$$

4.4. Exponential attractivity of the limit set of (30). Finally, we prove that the set $\omega(\mathbf{B})$ is locally exponentially attractive for the trajectories of the zero dynamics (30) of system (29) if the set $\mathcal{A}_0$ is locally exponentially attractive for the trajectories of (8). This fact is formalized in the next proposition.

PROPOSITION 2. *Suppose that, in addition to Assumptions* (i), (ii), (iv), *and* (v), *also Assumption* (iii) *holds. Then* $\omega(\mathbf{B})$ *is locally exponentially attractive for* (30).

Proof. Consider again the equivalent system (31); let the compact notation $\dot{\mathbf{z}} = \mathbf{f}_0(\mathbf{z})$ be used for the first three equations and let the variables (χ_1, ζ), introduced in the proof of Lemma 4, replace χ. Let $\bar{\sigma} : \mathcal{Z} \to \mathbb{R}^{(d-1)\times q}$ be any continuously differentiable map which on $\mathcal{A}_0$ agrees with the map σ introduced in Lemma 5, and change X into $\tilde{X} = X - \sigma(\mathbf{z})$. In this way, the last equation of (31) is transformed into an equation of the form

$$\dot{\tilde{X}} = F\tilde{X} + R(\chi_1, \mathbf{z}) + Q(\mathbf{z}) \tag{50}$$

in which

$$Q(\mathbf{z}) = F\bar{\sigma}(\mathbf{z}) + G\Omega(\tau_1(\mathbf{z})) - \frac{\partial \bar{\sigma}}{\partial \mathbf{z}}\mathbf{f}_0(\mathbf{z})$$

is a (matrix-valued) function vanishing on $\mathcal{A}_0$, while

$$R(\chi_1, \mathbf{z}) = G[\Omega(\chi_1 + \tau_1(\mathbf{z})) - \Omega(\tau_1(\mathbf{z}))]$$

is vanishing for $\chi_1 = 0$ for all $\mathbf{z} \in \mathcal{Z}$. Setting $\mathbf{x} = \mathrm{col}(\chi_1, \zeta, \tilde{\theta})$ and denoting by $\tilde{X}_1$ the first row of $\tilde{X}$, system (31) can be conveniently rewritten as the interconnection of system (50) with the system

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{f}_0(\mathbf{z}), \\ \dot{\mathbf{x}} &= \mathbf{A}(\mathbf{z})\mathbf{x} + \mathbf{h}(\mathbf{z}, \mathbf{x}) + G(\mathbf{z}, \mathbf{x}, \tilde{X})\nu(\mathbf{z}) + h(\mathbf{x}, \tilde{X}), \end{aligned} \tag{51}$$

in which

$$\mathbf{A}(\mathbf{z}) = \begin{pmatrix} -\lambda + r_1(\mathbf{z}) & \hat{c} & B^{\mathrm{T}}(\mathbf{z},0,0) \\ r_2(\mathbf{z}) & F & 0 \\ -B(\mathbf{z},0,0) & 0 & 0 \end{pmatrix},$$

$$G(\mathbf{z},\mathbf{x},\tilde{X}) = \begin{pmatrix} C & 0 & 0 \\ 0 & (\,\hat{b}\; I\,) & 0 \\ 0 & 0 & B(\mathbf{z},\mathbf{x},\tilde{X}) \end{pmatrix},$$

$$\nu(\mathbf{z}) = \begin{pmatrix} \varsigma_1(\mathbf{z}) \\ \varsigma_1(\mathbf{z}) \\ \varsigma_2(\mathbf{z}) \end{pmatrix}, \qquad h(\mathbf{x},\tilde{X}) = \begin{pmatrix} \tilde{X}_1\tilde{\theta} \\ 0 \\ -\tilde{X}_1^{\mathrm{T}}\chi_1 \end{pmatrix},$$

with

$$\begin{aligned} B(\mathbf{z},\mathbf{x},\tilde{X}) &= \beta(\tilde{X} + \sigma(\mathbf{z}), \chi_1 + \tau_1(\mathbf{z})), \\ r_1(\mathbf{z}) &= C\,[\partial\Delta/\partial\chi_1]_{\chi_1=0}\,, \\ r_2(\mathbf{z}) &= \begin{pmatrix} \hat{b} & I \end{pmatrix}[\partial\Delta/\partial\chi_1]_{\chi_1=0}\,, \end{aligned}$$

and where the vector $\mathbf{h}(\mathbf{z},\mathbf{x})$ is such that

$$\lim_{|\mathbf{x}|\to 0} \frac{|\mathbf{h}(\mathbf{z},\mathbf{x})|}{|\mathbf{x}|} = 0$$

uniformly in $\mathbf{z}$ (as the latter ranges over a compact set). Note that $\nu(\mathbf{z})$ is a vector of functions vanishing for $\mathbf{z} \in \mathcal{A}_0$ and that $h(\mathbf{x},\tilde{X})$ is a vector collecting higher order terms.

With this in mind, consider now the auxiliary system

$$\begin{aligned} \dot{\mathbf{z}} &= \mathbf{f}_0(\mathbf{z}), \\ \dot{\mathbf{y}} &= \mathbf{A}(\mathbf{z})\mathbf{y} \end{aligned} \tag{52}$$

with initial conditions $(\mathbf{z}(0),\mathbf{y}(0))$ in the compact set $\mathbf{Z}\times\mathbf{Y}$, where $\mathbf{Y} = \{\mathbf{y} : |\mathbf{y}| \le c\}$, with $c > 1$. Arguments identical to those used in the proof of Lemmas 4 and 6 make it possible to claim the existence of a $\lambda^\star > 0$ such that for all $\lambda > \lambda^\star$ the trajectories $(\mathbf{z}(t),\mathbf{y}(t))$ are bounded in positive time and that

$$\omega(\mathbf{Z}\times\mathbf{Y}) = \mathcal{A}_0 \times \{0\}\,.$$

We show now that $\mathcal{A}_0 \times \{0\}$ is locally exponentially attractive for (52). To this end, let $\mathbf{z}(t,\mathbf{z}_0)$ and $\mathbf{y}(t,\mathbf{y}_0)$ denote the solution pair of (52) passing through $\mathbf{z}_0$ and, respectively, $\mathbf{y}_0$ at time $t = 0$. Recall (see section 2) that $\mathcal{A}_0 \times \{0\}$ attracts the set $\mathbf{Z}\times\mathbf{Y}$ *uniformly*. Therefore, since $|\mathbf{y}| \le \mathrm{dist}((\mathbf{z},\mathbf{y}), \mathcal{A}_0 \times \{0\})$, for any $\varepsilon > 0$ there exists $T_\varepsilon > 0$ such that

$$|\mathbf{y}(t,\mathbf{y}_0)| \le \varepsilon \qquad \text{for all } \; t \ge T_\varepsilon \quad \text{and all} \quad (\mathbf{z}_0,\mathbf{y}_0) \in \mathbf{Z}\times\mathbf{Y}. \tag{53}$$

With this in mind, let δ be such that $\mathrm{dist}(\mathbf{z}_0,\mathcal{A}_0) \le \delta$ implies $\mathbf{z}(t,\mathbf{z}_0) \in \mathbf{Z}$ for all $t \ge 0$, which is always possible, since $\mathcal{A}_0$ is stable in the sense of Lyapunov for the upper equation of (52). Pick any $\mathbf{z}_0$ within a δ-distance from $\mathcal{A}_0$ and regard the bottom equation of (52) as a *time-varying linear* system

$$\dot{\mathbf{y}} = \mathbf{A}(\mathbf{z}(t,\mathbf{z}_0))\mathbf{y}\,. \tag{54}$$

Pick a pair $t \geq t_0 \geq 0$ and let $\Phi(t, t_0, \mathbf{z}_0)$ denote the associated state transition matrix (which, of course, depends on the choice of $\mathbf{z}_0$). By construction, the ith column $\phi_i(t, t_0, \mathbf{z}_0)$ of $\Phi(t, t_0, \mathbf{z}_0)$ is the solution of (54) which satisfies $\phi_i(t_0, t_0, \mathbf{z}_0) = \mathbf{v}_i$, where $\mathbf{v}_i$ is a vector in which all entries are zero, except the ith one, which is equal to 1. Consider now again (52) with initial conditions $\mathbf{z}(0) = \mathbf{z}(t_0, \mathbf{z}_0)$ and $\mathbf{y}(0) = \mathbf{v}_i$ (note that $(\mathbf{z}(t_0, \mathbf{z}_0), \mathbf{v}_i) \in \mathbf{Z} \times \mathbf{Y}$). Since (52) is time invariant, we observe that $\mathbf{y}(t, \mathbf{v}_i) = \phi_i(t + t_0, t_0, \mathbf{z}_0)$ for all $t \geq 0$. Thus, by appealing to (53), it is deduced that, for any ε, there exists T_ε such that

$$|\phi_i(t + t_0, t_0, \mathbf{z}_0)| \leq \varepsilon \quad \text{for all} \quad t \geq T_\varepsilon$$

and all $\mathbf{z}_0$, so long as $\operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0) \leq \delta$. This, in turn, by standard results (see, e.g., [23, page 92]) implies the existence of positive numbers M and a (independent of $\mathbf{z}_0$) such that

$$|\Phi(t, t_0, \mathbf{z}_0)| \leq M e^{-a(t-t_0)} \qquad \text{for all} \quad t \geq t_0 \geq 0 \tag{55}$$

and all $\mathbf{z}_0$, so long as $\operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0) \leq \delta$.

By a classical converse Lyapunov theorem (see Theorem 3.12 in [19]), we deduce from (55) the existence of a continuously differentiable and symmetric function $\bar{P}(t)$, a continuous and symmetric function $Q(t)$, and constants c_1, c_2, and c_3 such that

$$\frac{d\bar{P}(t)}{dt} + \bar{P}(t)\mathbf{A}(\mathbf{z}(t, \mathbf{z}_0)) + \mathbf{A}^{\mathrm{T}}(\mathbf{z}(t, \mathbf{z}_0))\bar{P}(t) = -Q(t) \tag{56}$$

with

$$0 < c_1 I \leq \bar{P}(t) \leq c_2 I \qquad \text{and} \qquad Q(t) \geq c_3 I > 0 \tag{57}$$

for all $t \geq 0$.

Bearing in mind this result, we return now to the overall system (50), (51) which can be more conveniently seen as a *time-varying nonlinear* system

$$\begin{aligned} \dot{\tilde{X}} &= F\tilde{X} + R(\mathbf{x}, \mathbf{z}(t, \mathbf{z}_0)) + Q(\mathbf{z}(t, \mathbf{z}_0)), \\ \dot{\mathbf{x}} &= \mathbf{A}(\mathbf{z}(t, \mathbf{z}_0))\mathbf{x} + \mathbf{h}(\mathbf{z}(t, \mathbf{z}_0), \mathbf{x}) + G(\mathbf{z}(t, \mathbf{z}_0), \mathbf{x}, \tilde{X})\nu(\mathbf{z}(t, \mathbf{z}_0)) + h(\mathbf{x}, \tilde{X}) \end{aligned} \tag{58}$$

in which, with a mild abuse of notation, we denote $R(\mathbf{x}, \mathbf{z}(t, \mathbf{z}_0)) = R(\chi_1, \mathbf{z}(t, \mathbf{z}_0))$.

In particular note that, as far as the higher order terms $\mathbf{h}(\mathbf{z}(t, \mathbf{z}_0), \mathbf{x})$ and $h(\mathbf{x}, \tilde{X})$ are concerned, for any $\epsilon > 0$ there is $\delta_\epsilon > 0$ such that

$$|(\mathbf{x}, \tilde{X})| \leq \delta_\epsilon \Rightarrow |\mathbf{h}(\mathbf{z}(t, \mathbf{z}_0), \mathbf{x})| \leq \epsilon|\mathbf{x}|\,, \quad |h(\mathbf{x}, \tilde{X})| \leq \epsilon|(\mathbf{x}, \tilde{X})|$$

for all $t \geq 0$ and all $\mathbf{z}_0 \in \mathbf{Z}$. Moreover, note that by Assumption (iii) and from the definitions of $\nu(\cdot)$ and $Q(\cdot)$ (and in particular from the fact that these functions are differentiable and vanishing on $\mathcal{A}_0$) it follows that there are positive constants M_z, a_z, and d such that

$$|\nu(\mathbf{z}(t, \mathbf{z}_0))| \leq M_z e^{-a_z t} \operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0)\,, \quad |Q(\mathbf{z}(t, \mathbf{z}_0))| \leq M_z e^{-a_z t} \operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0)$$

for all $t \geq 0$ and all $\mathbf{z}_0$ satisfying $\operatorname{dist}(\mathbf{z}_0, \mathcal{A}_0) \leq d$. Observe also that, since it is known that trajectories of (31) are bounded, there is a number $\bar{G}$ such that $|\mathbf{G}(\mathbf{z}(t, \mathbf{z}_0), \mathbf{x}(t), \tilde{X}(t))| \leq \bar{G}$ for all $t \geq 0$.

Using these facts, simple Lyapunov arguments can be adopted to show that system (58) is locally exponentially stable if it is the *cascade* system

$$\begin{aligned}\dot{\tilde{X}} &= F\tilde{X} + R(\mathbf{x}, \mathbf{z}(t, \mathbf{z}_0)),\\ \dot{\mathbf{x}} &= \mathbf{A}(\mathbf{z}(t, \mathbf{z}_0))\mathbf{x}\,.\end{aligned}$$

But local exponential stability of the previous system immediately follows by standard cascade arguments using the fact that F is Hurwitz, by the existence of $\bar{P}$ satisfying (56)–(57), and by definition of $R(\mathbf{x}, \mathbf{z})$. This concludes the proof of the proposition. □

5. Adaptive output regulation. We return now to the closed-loop system obtained from the interconnection of (20), (4), and (21). As mentioned before, this system, viewed as a system with input v and output $e = e_1$, has relative degree 1. To put it in "normal form" we use, instead of (23), the change of variables

$$\begin{aligned}\tilde{\theta} &= \hat{\theta} - \theta(\rho) - \int_0^e \beta(X, \xi_1 - CKe + CKs)ds,\\ \eta &= \xi - M(X)[\hat{\theta} - \theta(\rho)] - Ke,\end{aligned} \tag{59}$$

which is meant to cancel the residual input v from the η and $\tilde{\theta}$ dynamics (see [13, Proposition 9.1.1]). This, after some simple algebra and some obvious rearrangement of terms, yields a system of the form

$$\begin{aligned}
&\dot{\varrho} = 0,\\
&\dot{w} = s(\varrho, w),\\
&\dot{z} = f_0(\varrho, w, z) + f_1(\varrho, w, z, e)e,\\
&\dot{\eta} = A\eta + b\beta^{\mathrm{T}}\tilde{\theta} - K[q(\varrho, w, z, 0) + \eta_1] + \phi(\eta_1) + \Omega(\eta_1)\theta + \delta_1(\varrho, w, z, e, X, \eta_1)e,\\
&\dot{\tilde{\theta}} = -\beta[q(\varrho, w, z, 0) + \eta_1] + \delta_2(\varrho, w, z, e, X, \eta_1)\, e - \mathrm{dzv}_\ell(\tilde{\theta} + \theta(\varrho)),\\
&\dot{X} = FX + G\Omega(\eta_1) + \delta_3(\eta_1, e)e,\\
&\dot{e} = [q(\varrho, w, z, 0) + \eta_1] + \vartheta(\varrho, w, z, e)e + v,
\end{aligned} \tag{60}$$

in which $\delta_1(\cdot)$, $\delta_2(\cdot)$, $\delta_3(\cdot)$, and $\vartheta(\cdot)$ are continuously differentiable functions of their arguments.

A more succinct form can be obtained by setting $\mathbf{w}$ as in section 3 and setting

$$\mathbf{x} = \mathrm{col}(\eta, \tilde{\theta}, X_1, \ldots X_q),$$

(where X_i denotes the ith column of X) in which case the system in question can be rewritten in the form[3]

$$\begin{aligned}\dot{\mathbf{w}} &= \mathbf{s}(\mathbf{w}),\\ \dot{z} &= f_0(\mathbf{w}, z) + f_1(\mathbf{w}, z, e)e,\\ \dot{\mathbf{x}} &= \mathbf{g}_0(\mathbf{w}, z, \mathbf{x}) + \mathbf{g}_1(\mathbf{w}, z, \mathbf{x}, e)e,\\ \dot{e} &= \mathbf{h}(\mathbf{w}, z, \mathbf{x}) + \mathbf{k}(\mathbf{w}, z, \mathbf{x}, e)e + v.\end{aligned} \tag{61}$$

[3]By a minor abuse of notation we have replaced $f_0(\varrho, w, z)$ and $f_1(\varrho, w, z, e)$ with $f_0(\mathbf{w}, z)$ and, respectively, $f_1(\mathbf{w}, z, e)$.

In this notation, the set of equations

$$\begin{aligned} \dot{\mathbf{w}} &= \mathbf{s}(\mathbf{w}), \\ \dot{z} &= f_0(\mathbf{w}, z), \\ \dot{\mathbf{x}} &= \mathbf{g}_0(\mathbf{w}, z, \mathbf{x}) \end{aligned} \tag{62}$$

is a succinct version for the set of equations (30), whose asymptotic properties have been analyzed in the previous section. More precisely, under the hypotheses introduced earlier, the positive orbit of $\mathbf{Z} \times \Xi \times \Theta \times \mathbf{X}$ under the flow (62) is bounded and all trajectories uniformly converge to the compact invariant set $\omega(\mathbf{B})$ described in Proposition 1. Moreover, the function $\mathbf{h}(\mathbf{w}, z, \mathbf{x})$, which is a succinct version of the quantity $-[q(\varrho, w, z, 0) + \eta_1]$ in (60), vanishes on the set $\omega(\mathbf{B})$. With this in mind we are now in position to formulate our final result, which states that the controller (21) completed with

$$v = -ke \tag{63}$$

solves the problem of output regulation if k is chosen sufficiently large.

PROPOSITION 3. *Consider system* (20) *with exosystem* (4). *Let* $\mathbf{W}, Z, E$ *be fixed compact sets of initial conditions, for which Assumptions* (i)–(iv) *indicated in section* 3 *are supposed to hold. Suppose, in addition, that Assumption* (v) *introduced in section* 4.3 *holds. Consider the controller* (21) *completed with* (63) *and initial conditions in a fixed compact set* $\mathbf{K}$. *Then, there exists a number* $k^* > 0$ *such that if* $k \geq k^*$, *the positive orbit of* $\mathbf{W} \times Z \times E \times \mathbf{K}$ *in the closed loop system is bounded and* $e(t) \to 0$ *as* $t \to \infty$.

Proof. The result directly follows from Proposition 4 of the appendix. In particular it is easy to check that system (61)–(63) can be viewed as a system of the form (64), with the role of x in (64) being played here by the one-dimensional variable e. The properties established for (62) and the fact that $\mathbf{h}(\mathbf{w},z,\mathbf{x})$ vanishes on $\omega(\mathbf{B})$ show that all the assumptions of Proposition 4 are satisfied. Thus, the desired result follows by taking a large value of k. □

Remark. The previous proposition indicates that the proposed controller (21) completed with (63) solves the problem of output regulation for the relative degree 1 system (20). Bearing in mind the discussion at the end of section 3, though, it follows that a controller of the form (21), completed with

$$v = -k(e_r + g^{r-1}a_0e_1 + g^{r-2}a_1e_2 + \cdots + ga_{r-2}e_{r-1}),$$

is able to solve the problem of output regulation for the original plant (2) if g is large enough. In this respect, it is worth stressing that the assumptions under which the proposed controller solves the problem need only be checked on the original system (2) and not necessarily on the transformed, relative degree 1, system (12).

Appendix. A small-gain property. Consider a system of the form

$$\begin{aligned} \dot{\varrho} &= 0, \\ \dot{w} &= s(\varrho, w), \\ \dot{z} &= f_0(\varrho, w, z) + \ell(\varrho, w, z, x), \\ \dot{x} &= q_0(\varrho, w, z) + r(\varrho, w, z, x) + gAx, \end{aligned} \tag{64}$$

in which $(\varrho, w, z, x) \in \mathbb{R}^p \times \mathbb{R}^s \times \mathbb{R}^n \times \mathbb{R}^m$. Let the functions $f_0(\cdot), q_0(\cdot), g(\cdot), \ell(\cdot), s(\cdot)$ be continuously differentiable and, moreover, let $\ell(\varrho, w, z, 0) = 0$ and $r(\varrho, w, z, 0) = 0$

for all $(\varrho, w, z) \in \mathbb{R}^p \times \mathbb{R}^s \times \mathbb{R}^n$. A is a given Hurwitz matrix and g is a positive number. As in section 3, let $\mathcal{P} \subset \mathbb{R}^p$, $W \subset \mathbb{R}^s$, $Z \subset \mathbb{R}^n$ denote compact sets of initial conditions for ϱ, w, z, set $\mathbf{z} = \mathrm{col}(\varrho, w, z)$, and $\mathbf{Z} = \mathcal{P} \times W \times Z$. Suppose that the autonomous system

$$\begin{aligned} \dot{\varrho} &= 0, \\ \dot{w} &= s(\varrho, w), \\ \dot{z} &= f_0(\varrho, w, z), \end{aligned} \tag{65}$$

with initial conditions in the compact set $\mathbf{Z}$, satisfies Assumptions (i), (ii), (iii) of section 3 and, consistent with the abbreviated notation used throughout the paper, set $\mathcal{A}_0 = \omega(\mathbf{Z})$. The following lemma describes some relevant properties of (64), proven in [4] (see also [26, 27, 28]), which are instrumental in proving the desired results.

LEMMA 7. *Consider system* (64) *under the assumptions specified above, with initial conditions in* $\mathbf{Z} \times X$, *with* $X \subset \mathbb{R}^m$ *a compact set. Then the following hold:*

(a) *There exist positive numbers* d, M, a, *and* γ *such that if*

$$\mathrm{dist}(\mathbf{z}(0), \mathcal{A}_0) \le d \qquad \textit{and} \qquad |x(t)| \le d \quad \textit{for all } t \ge 0,$$

then

$$\mathrm{dist}(\mathbf{z}(t), \mathcal{A}_0) \le M e^{-at} \mathrm{dist}(\mathbf{z}(0), \mathcal{A}_0) + \gamma \max_{\tau \in [0,t]} |x(\tau)| \qquad \textit{for all } t \ge 0.$$

(b) *For all* $\epsilon > 0$ *there exist* $g^\star > 0$ *and* $T > 0$ *such that for all* $g \ge g^\star$ *the positive orbit of* $\mathbf{Z} \times X$ *under the flow of* (64) *is bounded and*

$$\mathrm{dist}(\mathbf{z}(t), \mathcal{A}_0) \le \epsilon, \qquad |x(t)| \le \epsilon \qquad \textit{for all } t \ge T.$$

The previous lemma provides the tools needed to study the asymptotic behavior of system (64) under the additional hypothesis that the function $q_0(\varrho, w, z)$ vanishes on $\mathcal{A}_0$ (or, equivalently, that $\mathcal{A}_0 \times \{0\}$ is invariant for (64)). This is specified in the next proposition.

PROPOSITION 4. *Consider system* (64) *under the assumptions specified above and assume, in addition, that* $q_0(\varrho, w, z) = 0$ *for all* $(\varrho, w, z) \in \mathcal{A}_0$. *Then for any compact set* X *there exists* $g_1^\star > 0$ *such that, for all* $g \ge g_1^\star$, *the positive orbit of* $\mathbf{Z} \times X$ *under the flow of* (64) *is bounded and* $\lim_{t \to \infty} x(t) = 0$.

Proof. The proof is an easy consequence of the results of Lemma 7 and of the small-gain theorem. As a matter of fact, we can pick $\epsilon \le d$ and set

$$\mathcal{S}_\epsilon = \{(\mathbf{z}, x) \in \mathbb{R}^{p+s+n} \times \mathbb{R}^m \ : \ \mathrm{dist}(\mathbf{z}, \mathcal{A}_0) \le \epsilon, \quad |x| \le \epsilon\}.$$

From property (b), it is seen that if $g \ge g^\star$, any initial condition in $\mathbf{Z} \times X$ produces a trajectory of (64) which is bounded in forward time and satisfies $(\mathbf{z}(t), x(t)) \in \mathcal{S}_\epsilon$ for all $t \ge T$. From property (a), it is seen that

$$\mathrm{dist}(\mathbf{z}(t-T), \mathcal{A}_0) \le M e^{-a(t-T)} \mathrm{dist}(\mathbf{z}(T), \mathcal{A}_0) + \gamma \max_{\tau \in [T, t-T]} |x(\tau)|$$

for all $t \ge T$. Note that the differentiable function $r(\varrho, w, z, x)$, which vanishes for $x = 0$, can be estimated as

$$|r(\varrho, w, z, x)| \le \alpha |x|$$

for all $(\mathbf{z}, x) \in \mathcal{S}_\varepsilon$, while the differentiable function $q_0(\varrho, w, z)$, which vanishes on $\mathcal{A}_0$, can be estimated as

$$|q_0(\varrho, w, z)| \leq \beta \operatorname{dist}(\mathbf{z}, \mathcal{A}_0)$$

for all $\mathbf{z} \in \mathcal{S}_\varepsilon$. Now, let $P > 0$ denote the solution of $PA + A^{\mathrm{T}}P = -I$, and let $\underline{\lambda}$ and $\bar{\lambda}$, respectively, denote the smallest and largest eigenvalue of P. Standard arguments can be used to show that, for all $t \geq T$,

$$|x(t)| \leq \sqrt{\frac{\bar{\lambda}}{\underline{\lambda}}}\, e^{-\frac{\lambda_g}{4\bar{\lambda}}(t-T)}|x(T)| + \frac{4\beta}{\lambda_g}|P| \max_{\tau \in [T, t-T]} \operatorname{dist}(\mathbf{z}, \mathcal{A}_0),$$

where $\lambda_g = g - 2|P|\alpha$. Hence the result follows by classical small-gain arguments if $g_1^\star \geq g^\star$ is picked so that the small-gain condition

$$\lambda_{g_1^\star} > 4\beta|P|\gamma$$

is fulfilled. This completes the proof of Proposition 4. □

Acknowledgments. The authors wish to thank Christopher I. Byrnes and Laurent Praly for fruitful discussion and helpful suggestions during the preparation of the paper.

REFERENCES

[1] G. Bastin and M. R. Gevers, *Stable adaptive observers for non-linear time-varying systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 650–658.
[2] I. Byrnes and A. Isidori, *Limit sets, zero dynamics, and internal models in the problem of nonlinear output regulation*, IEEE Trans. Automat. Control, 48 (2003), pp. 1712–1723.
[3] C. I. Byrnes and A. Isidori, *Nonlinear Internal Models for Output Regulation*, preprint math.OC/0311223, http://arxiv.org/abs/math.oc/0311223.
[4] C. I. Byrnes, A. Isidori, and L. Praly, *On the Asymptotic Properties of a System Arising in Non-Equilibrium Theory of Output Regulation*, preprint 18, Mittag-Leffler Institute, Stockholm, 2002–2003. Available online at http://www.mittag-leffler.se/preprints/0203.
[5] C. I. Byrnes, F. Delli Priscoli, A. Isidori, and W. Kang, *Structurally stable output regulation of nonlinear systems*, Automatica, 33 (1997), pp. 369–385.
[6] J. Huang and Z. Chen, *A general framework for tackling the output regulation problem*, IEEE Trans. Automat. Control, 49 (2004), pp. 2203–2218.
[7] E. J. Davison, *The robust control of a servomechanism problem for linear time-invariant multivariable systems*, IEEE Trans. Automatic Control, 21 (1976), pp. 25–34.
[8] B. A. Francis, *The linear multivariable regulator problem*, SIAM J. Control Optim., 15 (1977), pp. 486–505.
[9] B. A. Francis and W. M. Wonham, *The internal model principle of control theory*, Automatica, 12 (1976), pp. 457–465.
[10] J. K. Hale, L. T. Magalhães, and W. M. Oliva, *Dynamics in Infinite Dimensions*, Springer-Verlag, New York, 2002.
[11] J. Huang and C. F. Lin, *On a robust nonlinear servomechanism problem*, IEEE Trans. Automat. Control, 39 (1994), pp. 1510–1513.
[12] J. Huang and W. J. Rugh, *On a nonlinear multivariable servomechanism problem*, Automatica, 26 (1990), pp. 963–972.
[13] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, New York, 1995.
[14] A. Isidori, *Nonlinear Control Systems* II, Springer-Verlag, New York, 1999.
[15] A. Isidori, *A remark on the problem of semiglobal nonlinear output regulation*, IEEE Trans. Automat. Control, 42 (1997), pp. 1734–1738.
[16] A. Isidori and C. I. Byrnes, *Output regulation of nonlinear systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 131–140.
[17] P. Ioannou and J. Sun, *Robust Adaptive Control*, Prentice–Hall, Upper Saddle River, NJ, 1996.

[18] H. Khalil, *Robust servomechanism output feedback controllers for feedback linearizable systems*, Automatica, 30 (1994), pp. 1587–1599.
[19] H. K. Khalil, *Nonlinear Systems*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 1996.
[20] R. Marino and P. Tomei, *Nonlinear Control Design: Geometric, Adaptive, and Robust*, Prentice–Hall, London, 1995.
[21] R. Marino and P. Tomei, *Global adaptive observers for nonlinear systems via filtered transformations*, IEEE Trans. Automat. Control, 37 (1992), pp. 1239–1245.
[22] K. Narendra and A. Annaswamy, *Stable Adaptive Systems*, Inform. Systems Sci., Prentice–Hall, Englewood Cliffs, NJ, 1989.
[23] W. J. Rugh, *Linear System Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1996.
[24] A. Serrani, A. Isidori, and L. Marconi, *Semiglobal output regulation for minimum-phase systems*, Int. J. Robust Nonlinear Control, 10 (2000), pp. 379–396.
[25] A. Serrani, A. Isidori, and L. Marconi, *Semiglobal nonlinear output regulation with adaptive internal model*, IEEE Trans. Automat. Control, 46 (2001), pp. 1178–1194.
[26] J. Szarski, *Differential Inequalities*, Polska Akademia Nauk, Warszawa, 1967.
[27] A. Teel and L. Praly, *Tools for semiglobal stabilization by partial state and output feedback*, SIAM J. Control Optim., 33 (1995), pp. 1443–1488.
[28] F. W. Wilson, *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1969), pp. 413–428.
[29] X. D. Ye and J. Huang, *Decentralized adaptive output regulation for a class of large-scale nonlinear systems*, IEEE Trans. Automat. Control, 48 (2001), pp. 276–281.
[30] T. Yoshizawa, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, Springer-Verlag, New York, 1975.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 856–884

INTERCONNECTED SYSTEMS WITH UNCERTAIN COUPLINGS: EXPLICIT FORMULAE FOR μ-VALUES, SPECTRAL VALUE SETS, AND STABILITY RADII*

MICHAEL KAROW†, DIEDERICH HINRICHSEN‡, AND ANTHONY J. PRITCHARD§

Abstract. In this paper we study the variation of the spectrum of block-diagonal systems under perturbations of compatible block structure with fixed zero blocks at arbitrarily prescribed locations ("Gershgorin-type perturbations"). We derive explicit and computable formulae for the associated μ-values. The results are then applied to characterize spectral value sets and stability radii for such perturbed systems. By specializing our results to the scalar diagonal case, the classical eigenvalue inclusion theorems of Gershgorin, Brauer, and Brualdi are obtained as corollaries. Moreover it follows that the inclusion regions of Brauer and Brualdi are optimal for the corresponding perturbation structures.

Key words. linear systems, perturbations, spectral value sets, stability radii, μ-values

AMS subject classifications. 15A18, 93C05, 93C73, 93D09

DOI. 10.1137/050626508

1. Introduction. More than 20 years ago, various researchers recognized the importance of block-diagonal perturbations for describing structured uncertainties of interconnected systems where the overall model uncertainty is a consequence of those in its components; see [7] and [23]. Structured singular values (μ-values) were introduced in [7] as a means of analyzing the effect of block-diagonal perturbations. In recent years this concept has proved to be an effective tool in the robustness analysis of systems with structured uncertainties and in the synthesis of robust control systems; see, e.g., [1], [8], [19], [24].

Generalizing the definition in [7], the μ-value of a matrix $M \in \mathbb{C}^{q\times l}$ with respect to a given perturbation set $\mathbf{\Delta} \subset \mathbb{C}^{l\times q}$ and a given norm $\|\cdot\|$ on $\mathbb{C}^{l\times q}$ is the inverse of the smallest $\|\Delta\|$, $\Delta \in \mathbf{\Delta}$, such that 1 is an element of the spectrum of the matrix product ΔM; see [15]. The μ-value is denoted by $\mu_{\mathbf{\Delta}}(M)$. Explicit characterizations of $\mu_{\mathbf{\Delta}}(M)$, $M \in \mathbb{C}^{q\times l}$, have been obtained in the full block case where $\mathbf{\Delta} = \mathbb{C}^{l\times q}$ or $\mathbf{\Delta} = \mathbb{R}^{l\times q}$. For most other perturbation structures, e.g., block-diagonal, computable formulae are not available and so robust analysis/synthesis is usually based on upper bounds for the μ-value; see [18], [19]. In this paper we study the converse of the usual case in that we consider μ-problems where the *nominal matrix* M is block-diagonal and the *perturbations* $\Delta \in \mathbf{\Delta}$ are constrained only by the condition that they have zero blocks at certain fixed locations, e.g., on the diagonal ("Gershgorin type perturbations"). In contrast to the usual case we will be able to derive a number of computable exact formulae for the corresponding μ-values.

These formulae will then be applied to obtain computable characterizations of

*Received by the editors March 9, 2005; accepted for publication (in revised form) December 20, 2005; published electronically June 30, 2006.
http://www.siam.org/journals/sicon/45-3/62650.html
†Mathematics Institute, Berlin University of Technology, D-10623 Berlin, Germany (karow@math.TU-Berlin.de).
‡Mathematics Institute, University of Bremen, D-28334 Bremen, Germany (dh@math.uni-bremen.de).
§Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (ajp@maths.warwick.ac.uk).

spectral value sets and *stability radii* of block-diagonal systems under Gershgorin-type perturbations. Our objective is not only to prove new results but also to illustrate, on the methodological side, that the techniques of μ-*analysis* in combination with the concepts of *spectral values sets* and *stability radii* provide powerful tools for the spectral analysis of interconnected systems with uncertain couplings.

Pseudospectra (spectral value sets for unstructured complex perturbations) have been applied in various areas of the mathematical sciences, for instance, in numerical analysis [27] and the stability analysis of fluid flows [22], [27]. However, they have not found many applications in systems and control theory. For some papers in this field, see [12], [14], [15]. The *spectral value set* of a matrix A under perturbations $A \rightsquigarrow A_\Delta$, $\Delta \in \mathbf{\Delta}$, consists of all eigenvalues of the perturbed matrices A_Δ with $\Delta \in \mathbf{\Delta}$ constrained by $\|\Delta\| < \delta$. Here δ reflects the level of uncertainty of the nominal matrix measured in terms of some norm $\|\cdot\|$. By visualizing spectral value sets as the perturbation level changes, one obtains insight into the mobility of the eigenvalues under the perturbations in question. This is particularly useful for the stability analysis of uncertain linear systems.

A linear system is said to be stable with respect to a given stability region $\mathbb{C}_g$ in the complex plane if all the eigenvalues of the system matrix lie in $\mathbb{C}_g$. The nominal matrix A is regarded as an approximation to a system matrix whose exact value is unknown. If $\sigma(A) \subset \mathbb{C}_g$ and a bound for the level of uncertainty is known, then the exact system matrix will also be stable provided that the associated spectral value set is contained in $\mathbb{C}_g$.

An alternative but related approach is through the concept of a *stability radius* [13], [15]. This is defined to be the smallest perturbation level for which at least one of the perturbed matrices A_Δ with $\Delta \in \mathbf{\Delta}$, $\|\Delta\| \leq \delta$ becomes unstable. It is therefore a robustness measure of the $\mathbb{C}_g$-stability of the nominal matrix A. We will see that spectral value sets and stability radii can be expressed in terms of μ-values (section 2).

In this paper we consider perturbations of the form $A \rightsquigarrow A_\Delta = A + B\Delta C$, where A, B, C are given block-diagonal matrices and $\Delta \in \mathbf{\Delta}$. The perturbed matrices A_Δ can be viewed as the system matrices of composite systems obtained by the interconnection of subsystems via couplings determined by the Δ's; see section 3. The overall transfer matrix of the block-diagonal system is the direct sum of the transfer matrices of its subsystems and thus the formulae we obtain for μ-values of block-diagonal matrices can be applied to this transfer matrix to yield computable formulae for the corresponding spectral value sets and stability radii.

In the decentralized control of large scale systems it is common to adopt a *decomposition principle* where the overall system is regarded as the interconnection of decoupled subsystems. For such systems a notion of *connective stability* has been introduced where the decoupled subsystems are assumed to be stable and the system is said to be connectively stable if the overall system is stable for all interconnections in a set E which reflects the size and structure of the interconnections; see [25]. We will see that the results we develop for the stability radii of systems of the form A_Δ can be used to obtain precise statements for the connective stability of large scale systems.

The organization of the paper is as follows. In section 2 we give definitions of spectral value sets and stability radii and establish their connection to μ-values. In section 3 we introduce the perturbation structures to be considered and interpret them in the context of interconnected systems. Sections 4 and 5 contain the main results of this paper. Here we provide formulae for the computation of μ-values

with respect to Gershgorin-type perturbations and apply them to obtain computable characterizations of spectral value sets and stability radii. Two different types of norms will be considered on the perturbation spaces. In section 6 we specialize our results to the full class of all off-diagonal perturbations. Finally in section 7 we relate our results to the classical eigenvalue inclusion theorems of Gershgorin, Brauer and Brualdi; see [16].

2. The framework. In this section we introduce some basic concepts and fix the notation. The symbols $\mathbb{N}, \mathbb{R}, \mathbb{R}_+, \mathbb{C}$ denote the sets of positive integers, real numbers, nonnegative real numbers, and complex numbers, respectively. For $a \in \mathbb{C}$ the closed disk of radius $r > 0$ in $\mathbb{C}$ is $\mathcal{D}(a,r) = \{s \in \mathbb{C}; |s-a| \le r\}$. By $\mathbb{K}^{n\times m}$ we denote the set of $n \times m$ matrices with entries in $\mathbb{K}$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Furthermore, $\mathbb{K}^n = \mathbb{K}^{n\times 1}$ is the set of column vectors of length n. The transpose of $A \in \mathbb{K}^{n\times m}$ is denoted by $A^\top$. If A is square, then $\sigma(A)$, $\rho(A) = \mathbb{C} \setminus \sigma(A)$, and $\varrho(A)$ denote its spectrum, its resolvent set, and its spectral radius, respectively, $\varrho(A) = \max\{|s|;\, s \in \sigma(A)\}$. We let $L_{n,l,q}$ be the set of triples of matrices (A,B,C) with $A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{n\times l}, C \in \mathbb{C}^{q\times n},\ n,l,q \in \mathbb{N}$.

By $\partial\mathcal{S}$ we denote the boundary of the set $\mathcal{S} \subseteq \mathbb{C}$. We use the conventions

$$0^{-1} = \infty, \quad \infty^{-1} = 0, \quad \inf \emptyset = \infty, \tag{1}$$

where $\emptyset$ stands for the empty set. Throughout the paper we will consider the following perturbation structures; see [15].

DEFINITION 2.1. *Let $l,q \in \mathbb{N}$. By $\mathcal{P}_{l,q}$ we denote the set of pairs $(\mathbf{\Delta}, \|\cdot\|)$, where*

- *$\mathbf{\Delta} \neq \{0\}$ is a nonempty closed subset of $\mathbb{C}^{l\times q}$ which is star-shaped with respect to 0; i.e., $\Delta \in \mathbf{\Delta}$ implies $t\Delta \in \mathbf{\Delta}$ for every $t \in [0,1]$;*
- *$\|\cdot\|$ is a norm on the real vector space $\operatorname{span}_{\mathbb{R}} \mathbf{\Delta} \subseteq \mathbb{C}^{l\times q}$.*

By $\mathcal{P}^{\mathbb{C}}_{l,q}$ we denote the set of pairs $(\mathbf{\Delta}, \|\cdot\|)$, where

- *$\mathbf{\Delta} \neq \{0\}$ is a nonempty closed subset of $\mathbb{C}^{l\times q}$ which satisfies $\mathbb{C}\mathbf{\Delta} = \mathbf{\Delta}$; i.e., $\Delta \in \mathbf{\Delta}$ implies that $s\Delta \in \mathbf{\Delta}$ for every $s \in \mathbb{C}$;*
- *$\|\cdot\|$ is a norm on the complex vector space $\operatorname{span}_{\mathbb{C}} \mathbf{\Delta} \subseteq \mathbb{C}^{l\times q}$.*

The pairs $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ are called perturbation structures*, and the pairs $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}^{\mathbb{C}}_{l,q}$ are called* complex perturbation structures*.*

By definition we have $\mathcal{P}^{\mathbb{C}}_{l,q} \subset \mathcal{P}_{l,q}$. Given any triple $(A,B,C) \in L_{n,l,q}$ and a perturbation structure $(\mathbf{\Delta}, \|\cdot\|)$, we consider perturbations of A of the form

$$A \rightsquigarrow A_\Delta = A + B\Delta C, \quad \Delta \in \mathbf{\Delta}. \tag{2}$$

DEFINITION 2.2. *Let $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ be a perturbation structure. The spectral value set of the triple $(A,B,C) \in L_{n,l,q}$ with respect to $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ and perturbation level $\delta > 0$ is the following subset of the complex plane:*

$$\begin{aligned} \sigma_{\mathbf{\Delta}}(A,B,C;\delta) &:= \bigcup_{\Delta\in\mathbf{\Delta},\ \|\Delta\|<\delta} \sigma(A + B\Delta C) \\ &= \{s \in \mathbb{C};\ \exists\, \Delta \in \mathbf{\Delta} : \|\Delta\| < \delta,\ \textit{and } \det(sI_n - (A + B\Delta C)) = 0\}. \end{aligned} \tag{3}$$

Thus the spectral value set $\sigma_{\mathbf{\Delta}}(A,B,C;\delta)$ is the union of all the spectra of the perturbed matrices A_Δ where $\Delta \in \mathbf{\Delta}$, $\|\Delta\| < \delta$. The assumption that the perturbation class $\mathbf{\Delta}$ is star-shaped with respect to 0 guarantees that each connected component of $\sigma_{\mathbf{\Delta}}(A,B,C;\delta)$ contains an eigenvalue of A.

A concept closely related to the notion of *spectral value set* is that of *stability radius*. It presupposes that a stability region $\mathbb{C}_g \subset \mathbb{C}$ is given and measures the robustness of $\mathbb{C}_g$-stability of a matrix A with respect to perturbations of the form (2).

DEFINITION 2.3. *Let $\mathbb{C}_g$ be a nonempty open subset of $\mathbb{C}$. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be $\mathbb{C}_g$-stable if $\sigma(A) \subset \mathbb{C}_g$. The $\mathbb{C}_g$-stability radius of $(A, B, C) \in L_{n,l,q}$ with respect to $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ is defined as follows:*

$$r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g) := \inf\{\|\Delta\| \,;\, \Delta \in \mathbf{\Delta}, A + B\Delta C \text{ is not } \mathbb{C}_g\text{-stable}\}$$

$$= \inf\{\|\Delta\| \,;\, \Delta \in \mathbf{\Delta},\ \sigma(A + B\Delta C) \not\subset \mathbb{C}_g\}. \tag{4}$$

If A is not $\mathbb{C}_g$-stable, then $r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g) = 0$. It is easily seen that a minimum in (4) always exists if $r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g)$ is finite. Obviously,

$$r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g) = \inf\{\delta > 0 \,;\, \sigma_{\mathbf{\Delta}}(A, B, C; \delta) \not\subseteq \mathbb{C}_g\}.$$

Next, we give the definition of μ-values.

DEFINITION 2.4. *For $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$ the corresponding μ-value of $M \in \mathbb{C}^{q\times l}$ is given by*

$$\mu_{\mathbf{\Delta}}(M) := [\,\inf\{\|\Delta\| \,;\, \Delta \in \mathbf{\Delta},\ 1 \in \sigma(\Delta M)\,\}\,]^{-1}. \tag{5}$$

Note that the set $\mathbf{\Delta}_M = \{\Delta \in \mathbf{\Delta};\ 1 \in \sigma(\Delta M)\}$ is closed and does not contain the zero matrix. Thus a minimum in (5) is attained and is nonzero unless $\mathbf{\Delta}_M = \emptyset$. Hence, with the conventions (1), $\mu_{\mathbf{\Delta}}(M)$ is always well defined and $\mu_{\mathbf{\Delta}}(M) = 0$ if and only if $\mathbf{\Delta}_M = \emptyset$.

The following theorem specifies the relationship between spectral value sets, stability radii, and μ-values.

THEOREM 2.5. *Let $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$, $(A, B, C) \in L_{n,l,q}$, and $G(s) = C(sI_n - A)^{-1}B$. Then*

$$\mu_{\mathbf{\Delta}}(G(s)) = [\inf\{\|\Delta\| \mid \Delta \in \mathbf{\Delta},\ s \in \sigma(A + B\Delta C)\,\}]^{-1}, \quad s \in \rho(A); \tag{6}$$

$$\sigma_{\mathbf{\Delta}}(A, B, C; \delta) = \sigma(A) \cup \{s \in \rho(A);\ \mu_{\mathbf{\Delta}}(G(s)) > \delta^{-1}\}, \quad \delta > 0; \tag{7}$$

$$r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g) = \left(\sup_{s\in\partial\mathbb{C}_g} \mu_{\mathbf{\Delta}}(G(s))\right)^{-1} \quad \text{if } A \text{ is } \mathbb{C}_g\text{-stable.} \tag{8}$$

Proof. (6) follows from the definition of $\mu_{\mathbf{\Delta}}(\cdot)$ and the equivalence

$$s \in \sigma(A + B\Delta C) \ \Leftrightarrow\ 1 \in \sigma(\Delta G(s)), \tag{9}$$

which holds for all $s \in \rho(A)$ and all $\Delta \in \mathbb{C}^{l\times q}$; see [13, Proposition 2.3]. Then the characterizations (7), (8) are immediate consequences of (6). ∎

Theorem 2.5 is the basis for our further development. It shows that spectral value sets and stability radii can be calculated by evaluating the function $s \mapsto \mu_{\mathbf{\Delta}}(G(s))$. For completeness we mention some facts related to the characterization (7). The proofs can be found in [15], [17].

Remark 2.6. Let $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}^{\mathbb{C}}$. Then, for any $\delta > 0$,

(i) the sets $\sigma_{\mathbf{\Delta}}(A,B,C;\delta) \setminus \sigma(A) = \{\, s \in \rho(A) \,;\, \mu_{\mathbf{\Delta}}(G(s)) > \delta^{-1} \,\}$ are open;

(ii) the closure of $\sigma_{\mathbf{\Delta}}(A,B,C;\delta)$ is given by

$$\operatorname{cl}\left(\sigma_{\mathbf{\Delta}}(A,B,C;\delta)\right) = \bigcup_{\substack{\Delta \in \mathbf{\Delta} \\ \|\Delta\| \le \delta}} \sigma(A + B\Delta C) = \sigma(A) \cup \{s \in \rho(A) \,;\, \mu_{\mathbf{\Delta}}(G(s)) \ge \delta^{-1}\}; \tag{10}$$

(iii) the boundary of $\sigma_{\mathbf{\Delta}}(A,B,C;\delta)$ satisfies

$$\partial\sigma_{\mathbf{\Delta}}(A,B,C;\delta) \setminus \sigma(A) = \{\, s \in \rho(A) \,;\, \mu_{\mathbf{\Delta}}(G(s)) = \delta^{-1} \}.$$

Note that these statements do not hold for all perturbation structures $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}$.

Next, we give a useful characterization of $\mu_{\mathbf{\Delta}}(\cdot)$ via the spectral radius. It generalizes a result of [19].

LEMMA 2.7. *Let $M \in \mathbb{C}^{q\times l}$ and $(\mathbf{\Delta}, \|\cdot\|) \in \mathcal{P}_{l,q}^{\mathbb{C}}$. Then*

$$\mu_{\mathbf{\Delta}}(M) = \max\{\, \varrho(\Delta M) \,;\, \Delta \in \mathbf{\Delta}, \|\Delta\| = 1 \,\}. \tag{11}$$

Suppose that the maximum in (11) *is nonzero and is attained at $\Delta \in \mathbf{\Delta}$, $\|\Delta\| = 1$. Let $\Delta_1 = s^{-1}\Delta$, where $s \in \sigma(\Delta M)$ and $|s| = \varrho(\Delta M) \neq 0$. Then $\Delta_1 \in \mathbf{\Delta}$, $1 \in \sigma(\Delta_1 M)$, and $\|\Delta_1\| = \mu_{\mathbf{\Delta}}(M)^{-1}$.*

Proof. Let ϱ_0 denote the maximum on the right-hand side of (11). For any nonzero $\Delta \in \mathbf{\Delta}$ we have $\varrho(\Delta M) = \|\Delta\|\, \varrho(\frac{\Delta}{\|\Delta\|} M) \le \|\Delta\|\, \varrho_0$. Hence, the condition $1 \in \sigma(\Delta M)$ implies that $1 \le \|\Delta\|\, \varrho_0$. This yields $\mu_{\mathbf{\Delta}}(M) \le \varrho_0$. Equality holds if $\varrho_0 = 0$. Suppose $\varrho_0 \neq 0$. Then the matrix Δ_1 satisfies $\|\Delta_1\| = \varrho_0^{-1}$ and $\|\Delta_1\| \ge \mu_{\mathbf{\Delta}}(M)^{-1}$. Thus $\varrho_0 = \mu_{\mathbf{\Delta}}(M)$. ∎

We now determine $\mu_{\mathbf{\Delta}}(M)$ for the case that $\mathbf{\Delta} = \mathbb{C}^{l\times q}$ and the underlying norm is an operator norm. Let $\|\cdot\|_\alpha, \|\cdot\|_\beta$ be norms on $\mathbb{C}^q$ and $\mathbb{C}^l$, respectively. Then the induced operator norms on $\mathbb{C}^{l\times q}$ (resp., $\mathbb{C}^{q\times l}$) are defined by

$$\|\Delta\|_{\alpha,\beta} = \max_{y \in \mathbb{C}^q \setminus \{0\}} \frac{\|\Delta y\|_\beta}{\|y\|_\alpha},\ \Delta \in \mathbb{C}^{l\times q}, \quad \text{and} \quad \|M\|_{\beta,\alpha} = \max_{u \in \mathbb{C}^l \setminus \{0\}} \frac{\|Mu\|_\alpha}{\|u\|_\beta},\ M \in \mathbb{C}^{q\times l}.$$

Recall that, for every $\Delta \in \mathbb{C}^{l\times q}$, there exist $y \in \mathbb{C}^q$, $u \in \mathbb{C}^l$, with $\|y\|_\alpha = \|u\|_\beta^D = 1$ and

$$\|\Delta\|_{\alpha,\beta} = u^\top \Delta y.$$

Here $\|\cdot\|_\beta^D$ denotes the dual of the norm $\|\cdot\|_\beta$,

$$\|u\|_\beta^D = \max_{z \in \mathbb{C}^l \setminus \{0\}} \frac{|u^\top z|}{\|z\|_\beta}, \quad u \in \mathbb{C}^l.$$

PROPOSITION 2.8. *Let $\|\cdot\|_\alpha, \|\cdot\|_\beta$ be norms on $\mathbb{C}^q$ and $\mathbb{C}^l$, respectively. Let $\|\cdot\| = \|\cdot\|_{\alpha,\beta}$ be the induced operator norm and $\mathbf{\Delta} := \mathbb{C}^{l\times q}$. Then the following hold:*

(a) *For any $M \in \mathbb{C}^{q\times l}$, $\mu_{\mathbf{\Delta}}(M) = \|M\|_{\beta,\alpha}$.*

(b) *Suppose $M \neq 0$. Let $u \in \mathbb{C}^l$, $y \in \mathbb{C}^q$ be such that $\|u\|_\beta = \|y\|_\alpha^D = 1$ and $y^\top M u = \|M\|_{\beta,\alpha}$. Then the matrix $\Delta_0 := \|M\|_{\beta,\alpha}^{-1} u y^\top$ satisfies $1 \in \sigma(\Delta_0 M)$ and $\|\Delta_0\| = \mu_{\mathbf{\Delta}}(M)^{-1}$.*

Proof. If $1 \in \sigma(\Delta M)$, then there is $u \neq 0$ with $u = \Delta M u$. Hence,

$$0 \neq \|u\|_\beta = \|\Delta M u\|_\beta \leq \|\Delta\|_{\alpha,\beta}\, \|M\|_{\beta,\alpha}\, \|u\|_\beta.$$

Thus $1 \leq \|\Delta\|_{\alpha,\beta}\, \|M\|_{\beta,\alpha}$. This implies $\mu_{\mathbf{\Delta}}(M) \leq \|M\|_{\beta,\alpha}$. Equality holds if $M = 0$. Let $M \neq 0$. Then the matrix Δ_0 satisfies $\|\Delta_0\|_{\alpha,\beta} = \|M\|^{-1}_{\beta,\alpha}$. Furthermore, $\Delta_0 M u = u$. Thus $1 \in \sigma(\Delta_0 M)$. It follows that $\mu_{\mathbf{\Delta}}(M) \geq \|\Delta_0\|^{-1}_{\alpha,\beta} = \|M\|_{\beta,\alpha}$. So $\mu_{\mathbf{\Delta}}(M) = \|M\|_{\beta,\alpha} = \|\Delta_0\|^{-1}_{\alpha,\beta}$. ∎

Remark 2.9. Throughout the rest of this paper we consider only complex perturbation structures. There are some results available for real perturbation structures. For example, if $M \in \mathbb{C}^{q\times l}$ and $\mathbf{\Delta} = \mathbb{R}^{l\times q}$, there are formulae for $\mu_{\mathbf{\Delta}}(M)$ (and hence for spectral value sets and stability radii) if $\mathbb{R}^l$ and $\mathbb{R}^q$ are normed with Euclidean norms; see [15], [17], and [21]. In [15] formulae are proved for stability radii of a real diagonal matrix with respect to real off-diagonal perturbations; see Corollary 6.6 for the complex case. Also in [20] a formula is given for the stability radius of real symmetric systems with respect to real symmetric (or diagonal) dynamic perturbations.

Remark 2.10. $\sigma_{\mathbf{\Delta}}(A, B, C; \delta)$, $r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g)$, and $\mu_{\mathbf{\Delta}}(G(s))$ depend strongly upon the perturbation norm $\|\cdot\|$ on $\mathbf{\Delta}$. Also the problem of evaluating numerically the formulae in Theorem 2.5 depends strongly on this norm. In a given application one should therefore carefully choose the norm on the perturbation space in such a way that it reflects the parametric uncertainty of the application and is also suitable from a computational viewpoint. In order to provide greater flexibility we have stated the results in the following sections for different classes of norms from which one can choose the most appropriate norm in a given case. In general, the (approximate) computation of $\sigma_{\mathbf{\Delta}}(A, B, C; \delta)$, $r_{\mathbf{\Delta}}(A, B, C; \mathbb{C}_g)$, $\mu_{\mathbf{\Delta}}(G(s))$ is a difficult problem, but under specific conditions efficient algorithms and estimation procedures are available; see, e.g., [5], [1], [18], [9], [26], [11].

3. Composite systems. Let us introduce some additional notation. In the following, q, l are finite sequences $q = (q_1, \ldots, q_m)$, $l = (l_1, \ldots, l_m)$. We write $\underline{m} := \{1, 2, \ldots, m\}$ and denote by $\mathbb{C}^{q\times l} := \{[M_{jk}];\, M_{jk} \in \mathbb{C}^{q_j \times l_k} \text{ for } (j,k) \in \underline{m} \times \underline{m}\}$ the set of $m \times m$ block matrices

$$[M_{jk}] = [M_{jk}]_{j\in\underline{m},k\in\underline{m}} = \begin{bmatrix} M_{11} & \cdots & M_{1m} \\ \vdots & & \vdots \\ M_{m1} & \cdots & M_{mm} \end{bmatrix}. \tag{12}$$

The block-diagonal matrix with blocks $M_j \in \mathbb{C}^{q_j\times l_j}$, $j \in \underline{m}$ is denoted by

$$M = \oplus_{j=1}^m M_j := \operatorname{diag}(M_1, \ldots, M_m) = \begin{bmatrix} M_1 & & & 0 \\ & M_2 & & \\ & & \ddots & \\ 0 & & & M_m \end{bmatrix} \in \mathbb{C}^{q\times l}.$$

For any index set $\mathcal{I} \subseteq \underline{m} \times \underline{m}$ we denote by $\mathbf{\Delta}_{\mathcal{I},q,l}$ the set of block matrices Δ of the form

$$\Delta = [\Delta_{jk}] := \begin{bmatrix} \Delta_{11} & \cdots & \Delta_{1m} \\ \vdots & & \vdots \\ \Delta_{m1} & \cdots & \Delta_{mm} \end{bmatrix}, \quad \Delta_{jk} \in \mathbb{C}^{l_j\times q_k} \text{ and } \Delta_{jk} = 0 \text{ if } (j,k) \notin \mathcal{I}. \tag{13}$$

Given $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$, $j \in \underline{m}$, the object of this paper is to study the variation of the spectrum of the block-diagonal matrix $A = \oplus_{j=1}^m A_j$ under perturbations of the form

$$A \rightsquigarrow A_\Delta := A + B\Delta C, \quad \Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}, \tag{14}$$

where B, C are the block-diagonal matrices $B = \oplus_{j=1}^m B_j$, $C = \oplus_{j=1}^m C_j$.

The matrices A_Δ have the following system theoretic interpretation. Consider the system

$$\Sigma : \qquad \dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) \tag{15}$$

which is the direct sum of the m subsystems

$$\Sigma_j : \qquad \dot{x}_j(t) = A_j x_j(t) + B_j u_j(t), \qquad y_j(t) = C_j x_j(t), \qquad j \in \underline{m}. \tag{16}$$

The transfer matrix of Σ is the direct sum of the transfer matrices of these subsystems:

$$G(s) = C(sI - A)^{-1}B = \oplus_{j=1}^m G_j(s), \qquad G_j(s) := C_j(sI_{n_j} - A_j)^{-1}B_j, \ \ j \in \underline{m}. \tag{17}$$

Introducing the couplings

$$u_j(t) = \sum_{k \in \underline{m},\, (j,k) \in \mathcal{I}} \Delta_{jk}\, y_k(t), \quad j \in \underline{m}, \tag{18}$$

one obtains the composite system

$$\Sigma_\Delta : \qquad \begin{bmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_m \end{bmatrix} = (A + B\Delta C) \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = A_\Delta \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}. \tag{19}$$

Thus the perturbed system Σ_Δ with system matrix A_Δ can be viewed as the composite system obtained by interconnecting the subsystems Σ_j via the couplings (18) defined by the perturbation blocks Δ_{jk}. The unperturbed ("nominal") system $\Sigma_0 : \ \dot{x} = Ax$ obtained by setting $\Delta = 0$ is simply the direct sum of the subsystems $\dot{x}_j = A_j x_j$.

The pairs $(j,k) \in \mathcal{I}$ can be regarded as the oriented edges of a directed graph $\Gamma(\underline{m}, \mathcal{I})$ whose vertices are the numbers $1, \ldots, m$. This is illustrated in Example 3.1 for the case where $m = 3$. Observe that in the directed graph the endpoint of the edge (j,k) is the first component, j.[1] This orientation reflects the interconnection structure (18).

Example 3.1. Consider the index set $\mathcal{I} = \{(1,2), (1,3), (2,1), (3,2), (3,3)\}$. Then the matrices $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$ take the form

$$\Delta = \begin{bmatrix} 0 & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & 0 & 0 \\ 0 & \Delta_{32} & \Delta_{33} \end{bmatrix}.$$

The directed graph $\Gamma(\underline{3}, \mathcal{I})$ and the block diagram of the closed loop system (19) are shown in Figure 1.

[1] Note that this is the reverse of standard notation in graph theory. We have used our notation to be in harmony with the system theoretic interpretation.

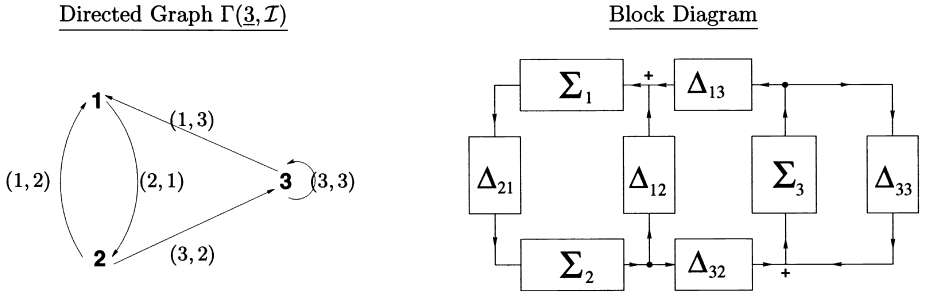


FIG. 1. *Composite system.*

Applying Theorem 2.5, the spectral value sets and stability radii of the system (A, B, C) under perturbations of the form (14) are given by

$$\sigma_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C;\delta) = \sigma(A) \cup \{s \in \rho(A)\,;\, \mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(G(s)) > \delta^{-1}\} \tag{20}$$

and

$$r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C;\mathbb{C}_g) = \left(\sup_{s\in\partial\mathbb{C}_g} \mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(G(s))\right)^{-1}. \tag{21}$$

In order to determine the spectral value sets and stability radii via (20), (21) we need to study the μ-values of block-diagonal matrices $M = \oplus_{j=1}^{m} M_j$ with respect to perturbations $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$. Note that this is just the inverse situation of traditional μ-analysis where block-diagonal perturbations of arbitrary matrices are considered; see [7]. Applying Proposition 2.7, we obtain

$$\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) = \max_{\substack{\Delta\in\mathbf{\Delta}_{\mathcal{I},q,l}\\ \|\Delta\|=1}} \varrho(\Delta M), \quad M \in \mathbb{C}^{q\times l}. \tag{22}$$

The size of the perturbations $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$ will be measured by two types of norms: a weighted maximum of the nonzero block norms $\|\Delta_{jk}\|$, $(j,k) \in \mathcal{I}$, and mixed operator norms of the overall matrix Δ. In the next two sections we derive formulae for $\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M)$ with respect to these types of norms.

Remark 3.2. A composite system Σ of the form (15) which is the direct sum of subsystems Σ_i of the form (16) is said to be *connectively stable* with respect to a given set of interconnections E (possibly time-varying and/or nonlinear) if $\sigma(A_j) \subset \mathbb{C}_-$, $j \in \underline{m}$, and the origin of the interconnected system obtained from the block-diagonal system Σ by the feedback $u(t) = e(t, y(t))$ is globally asymptotically stable for all $e \in E$; see [25]. In the literature many different methods have been put forward for obtaining sufficient criteria of connective stability based on knowledge of the subsystems Σ_i and their interconnection structure. Input-output and passivity methods have been used, but the most popular seem to be Liapunov methods; see [25]. The advantage of these methods is that time-varying and nonlinear interconnections can be considered. However, the robustness results obtained in this way are in general quite conservative. This is in contrast with the full block case (where $\mathbf{\Delta}_{\mathcal{I},q,l} = \mathbb{C}^{l\times q}$; i.e., $\mathcal{I} = \underline{m} \times \underline{m}$). In this case a quadratic Liapunov function of optimal robustness can be constructed which secures asymptotic stability for all time-varying nonlinearities with gain strictly smaller than $r_{\mathbb{C}^{l\times q}}(A, B, C; \mathbb{C}_-)$; see [13] and [15, section 5.6].

It remains an open problem to determine those perturbation structures $\mathbf{\Delta}_{\mathcal{I},q,l}$ for which it is possible to construct a joint quadratic Liapunov function for all perturbed systems Σ_Δ, $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$, $\|\Delta\| < r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C;\mathbb{C}_-)$. Such a Liapunov function would secure the connective stability for all time-varying nonlinearities with gain strictly smaller than $r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C;\mathbb{C}_-)$.

4. Weighted maximum norms. We consider the same basic framework as that in section 3. Let $\|\cdot\|_{\alpha_j}$ be a norm on $\mathbb{C}^{q_j}$ and $\|\cdot\|_{\beta_k}$ be a norm on $\mathbb{C}^{l_k}$. We assume that we are given a nonnegative weight matrix $R = [r_{jk}] \in \mathbb{R}_+^{m\times m}$ and introduce the index set

$$\mathcal{I} = \mathcal{I}_R := \{(j,k) \in \underline{m} \times \underline{m}\,;\, r_{jk} > 0\}. \tag{23}$$

With these data we associate a normed perturbation space $(\mathbf{\Delta}_{\mathcal{I},q,l}, \|\cdot\|)$, where (see (13))

$$\mathbf{\Delta}_{\mathcal{I},q,l} = \{[\Delta_{jk}]\,;\, \Delta_{jk} \in \mathbb{C}^{l_j\times q_k} \text{ for } j,k \in \underline{m} \text{ and } \Delta_{jk} = 0 \text{ if } (j,k) \notin \mathcal{I}\} \tag{24}$$

and $\|\cdot\|$ is the weighted maximum norm

$$\|\Delta\| := \max_{(j,k)\in\mathcal{I}} r_{jk}^{-1} \|\Delta_{jk}\|_{\alpha_k,\beta_j}, \quad \Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}. \tag{25}$$

Note that the following equivalence holds for $\Delta \in \mathbb{C}^{l\times q}$:

$$(\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l} \text{ and } \|\Delta\| \le 1) \quad \Leftrightarrow \quad \|\Delta_{jk}\|_{\alpha_k,\beta_j} \le r_{jk} \text{ for all } (j,k) \in \underline{m} \times \underline{m}. \tag{26}$$

In this section we determine the μ-value of block-diagonal matrices with respect to the perturbation structure $(\mathbf{\Delta}_{\mathcal{I},q,l}, \|\cdot\|)$ and apply it to obtain formulae for spectral value sets and stability radii. We will make use of the following well-known results from the theory of nonnegative matrices; see [2], [10], [16].

- (ϱ1) If $A \in \mathbb{R}^{n\times n}$ is nonnegative, the spectral radius $\varrho(A)$ is an eigenvalue of A and there exists a nonnegative eigenvector corresponding to $\varrho(A)$.
- (ϱ2) Let $A_1, A_2 \in \mathbb{R}^{n\times n}$. If $0 \le A_1 \le A_2$, then $\varrho(A_1) \le \varrho(A_2)$.
- (ϱ3) If $\alpha v \le Av$ and $v \ge 0$, $v \ne 0$, then $\alpha \le \varrho(A)$. If $Av \le \beta v$ and $v_i > 0$ for $i \in \underline{n}$, then $\varrho(A) \le \beta$.

The next lemma is a consequence of (ϱ3).

LEMMA 4.1. *Let $Y_{jk} \in \mathbb{C}^{l_j\times l_k}$, $j,k \in \underline{m}$, and $\|\cdot\|_{\beta_j}$ be a norm on $\mathbb{C}^{l_j}$. Then*

$$\varrho\left(\begin{bmatrix} Y_{11} & \dots & Y_{1m} \\ \vdots & & \vdots \\ Y_{m1} & \dots & Y_{mm} \end{bmatrix}\right) \le \varrho\left(\begin{bmatrix} \|Y_{11}\|_{\beta_1,\beta_1} & \dots & \|Y_{1m}\|_{\beta_m,\beta_1} \\ \vdots & & \vdots \\ \|Y_{m1}\|_{\beta_1,\beta_m} & \dots & \|Y_{mm}\|_{\beta_m,\beta_m} \end{bmatrix}\right). \tag{27}$$

Proof. Let $\lambda \in \mathbb{C}$ be an eigenvalue of the block matrix $[Y_{jk}]$, i.e.,

$$\begin{bmatrix} Y_{11} & \dots & Y_{1m} \\ \vdots & & \vdots \\ Y_{m1} & \dots & Y_{mm} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}, \qquad x_j \in \mathbb{C}^{l_j}, \quad j \in \underline{m},$$

and $x_j \ne 0$ for at least one j. Then $\lambda x_j = \sum_{k=1}^m Y_{jk}x_k$ for all $j \in \underline{m}$. This implies

$$|\lambda|\,\|x_j\|_{\beta_j} \le \sum_{k=1}^m \|Y_{jk}\|_{\beta_k,\beta_j}\,\|x_k\|_{\beta_k}, \qquad j \in \underline{m}.$$

Setting $A = (\|Y_{jk}\|_{\beta_k,\beta_j})_{j,k\in\underline{m}}$ and $v = (v_j)_{j\in\underline{m}}$, where $v_j = \|x_j\|_{\beta_j}$, we have

$$|\lambda|\, v_j \le (Av)_j, \qquad j \in \underline{m}.$$

Hence $|\lambda| \le \varrho(A)$ by (ϱ3), and this implies (27). □

We associate with any given block matrix $M = [M_{jk}] \in \mathbb{C}^{q\times l}$ of the form (12) the following nonnegative $m \times m$ matrix of block norms:

$$\tilde{M} = \begin{bmatrix} \|M_{11}\|_{\beta_1,\alpha_1} & \dots & \|M_{1m}\|_{\beta_m,\alpha_1} \\ \vdots & & \vdots \\ \|M_{m1}\|_{\beta_1,\alpha_m} & \dots & \|M_{mm}\|_{\beta_m,\alpha_m} \end{bmatrix}. \tag{28}$$

We now prove the main result of this section.

THEOREM 4.2. *Suppose $R = [r_{jk}] \in \mathbb{R}^{m\times m}$ is a nonnegative matrix and $\mathcal{I} = \mathcal{I}_R$ is given by* (23). *If $M = [M_{jk}] \in \mathbb{C}^{q\times l}$ and $\tilde{M} = (\|M_{jk}\|_{\beta_k,\alpha_j})_{j,k\in\underline{m}}$ is the associated matrix of block norms, then, with respect to the norm* (25),

$$\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) \le \varrho(R\,\tilde{M}). \tag{29}$$

Equality holds in (29) *if $M = \oplus_{j=1}^m M_j$, $M_j \in \mathbb{C}^{q_j\times l_j}$, $j \in \underline{m}$, is block-diagonal, viz.*

$$\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) = \varrho(R\,\mathrm{diag}(\|M_1\|_{\beta_1,\alpha_1},\dots,\|M_m\|_{\beta_m,\alpha_m})). \tag{30}$$

Proof. We have already seen that by Lemma 2.7,

$$\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) = \max_{\substack{\Delta\in\mathbf{\Delta}_{\mathcal{I},q,l}\\ \|\Delta\|=1}} \varrho(\Delta\, M)\,. \tag{31}$$

Moreover, if $(\Delta M)_{jk}$ is the (j,k)-entry of the $m \times m$ block matrix $\Delta M \in \mathbb{C}^{l\times l}$, then

$$\|(\Delta M)_{jk}\|_{\beta_k,\beta_j} = \left\|\sum_{i=1}^m \Delta_{ji}M_{ik}\right\|_{\beta_k,\beta_j} \le \sum_{i=1}^m \|\Delta_{ji}\|_{\alpha_i,\beta_j}\,\|M_{ik}\|_{\beta_k,\alpha_i} = (\tilde{\Delta}\tilde{M})_{jk},\ j,k\in\underline{m}, \tag{32}$$

where $\tilde{\Delta} := [\|\Delta_{jk}\|_{\alpha_k,\beta_j}]_{j,k\in\underline{m}}$. Now let $\Delta = [\Delta_{jk}] \in \mathbf{\Delta}_{\mathcal{I},q,l}$ with $\|\Delta\| \le 1$. Then $\tilde{\Delta} \le R$ by (26) and

$$\begin{aligned} \varrho(\Delta M) &\le \varrho([\,\|\,(\Delta M)_{jk}\,\|_{\beta_k,\beta_j}\,]) && \text{(by Lemma 4.1)}\\ &\le \varrho(\tilde{\Delta}\tilde{M}) && \text{(by (32) and property (ϱ2))}\\ &\le \varrho(R\tilde{M}) && \text{(by (26) and property (ϱ2)).} \end{aligned}$$

Thus,

$$\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) = \max_{\substack{\Delta\in\mathbf{\Delta}_{\mathcal{I},q,l}\\ \|\Delta\|=1}} \varrho(\Delta\, M) \le \varrho(R\tilde{M}).$$

It remains to show that the latter inequality is actually an equality if $M = \oplus_{j=1}^m M_j$ is block-diagonal. For $k\in\underline{m}$ let $y_k \in \mathbb{C}^{q_k}$ and $u_k \in \mathbb{C}^{l_k}$ be such that $\|u_k\|_{\beta_k} = \|y_k\|_{\alpha_k}^D = 1$ and $y_k^\top M_k u_k = \|M_k\|_{\beta_k,\alpha_k}$. Let $\Delta^0 = [\Delta^0_{jk}]$, where $\Delta^0_{jk} = r_{jk}u_j y_k^\top$. Then $\Delta^0 \in \mathbf{\Delta}_{\mathcal{I},q,l}$ and $\|\Delta^0\| = 1$. Since $R\tilde{M} = R\,\mathrm{diag}(\|M_1\|_{\beta_1,\alpha_1},\dots,\|M_m\|_{\beta_m,\alpha_m}) \in$

$\mathbb{R}^{m\times m}$ is nonnegative, there is a nonnegative vector $\xi = [\xi_1, \dots, \xi_m]^\top \in \mathbb{R}^m$ such that $R\tilde{M}\xi = \varrho(R\tilde{M})\xi$. Define $w = [\xi_1 u_1^\top, \dots, \xi_m u_m^\top]^\top$. Then a straightforward computation yields $\Delta^0(\oplus_{k=1}^m M_k)\, w = \varrho(R\tilde{M})w$. Thus $\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) \geq \varrho(\Delta^0(\oplus_{k=1}^m M_k)) \geq \varrho(R\tilde{M})$, and the proof is complete. ∎

For the case where all the α_j and β_k are 2-norms, a characterization of $\mu_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M)$ is given in Part I of [6] which shows that the μ-value can be obtained as the solution of a smooth constrained optimization problem. Associated computational aspects are discussed in Part II of [6].

We will now apply the above theorem to determine the spectral value sets and stability radii of block-diagonal matrices A with respect to perturbations of the form (14). Let $\mathcal{B}_R(\delta)$ denote the open ball with radius $\delta > 0$ about the origin in the perturbation space $\mathbf{\Delta}_{\mathcal{I},q,l}$ provided with the norm (25),

$$\mathcal{B}_R(\delta) = \{\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}; \|\Delta\| < \delta\}. \tag{33}$$

It follows from (26) that $\mathcal{B}_R(\delta)$ is the set of block matrices $\Delta = [\Delta_{jk}]$ satisfying

$$\|\Delta_{jk}\|_{\alpha_k,\beta_j} < \delta\, r_{jk}, \quad (j,k) \in \mathcal{I} = \mathcal{I}_R, \quad \|\Delta_{jk}\|_{\alpha_k,\beta_j} = 0 \quad \text{otherwise.}$$

COROLLARY 4.3. *Suppose $R = [r_{jk}] \in \mathbb{R}^{m\times m}$ is a nonnegative matrix and $\mathcal{I} = \mathcal{I}_R$ is given by* (23). *Let $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$, $j \in \underline{m}$, and consider perturbations* (14) *of the block-diagonal matrix A. If $\mathbf{\Delta}_{\mathcal{I},q,l}$ is provided with the norm* (25) *and $G_j(s)$ is defined by* (17), *then the following hold:*

(a) *The spectral value set $\sigma_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A, B, C; \delta)$ is given by*

$$\bigcup_{\Delta\in\mathcal{B}_R(\delta)} \sigma(A_\Delta) = \sigma(A) \cup \left\{ s \in \rho(A)\,;\, \varrho(R\,\mathrm{diag}(\|G_1(s)\|_{\beta_1,\alpha_1}, \dots, \|G_m(s)\|_{\beta_m,\alpha_m})) > \delta^{-1} \right\}.$$

(b) *Let $\mathbb{C}_g$ be an open subset of $\mathbb{C}$ and suppose $A_1, \dots, A_m$ are $\mathbb{C}_g$-stable (i.e., $\sigma(A) \subset \mathbb{C}_g$). Then the stability radius is given by*

$$r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A, B, C; \mathbb{C}_g) = \left(\sup_{s\in\partial\mathbb{C}_g} \varrho(R\,\mathrm{diag}(\|G_1(s)\|_{\beta_1,\alpha_1}, \dots, \|G_m(s)\|_{\beta_m,\alpha_m})) \right)^{-1}. \tag{34}$$

Proof. Applying Theorem 4.2, (a) follows directly from (20) and (b) follows from (21). ∎

We conclude this section by specializing the previous results to the scalar *diagonal* case where $A = \mathrm{diag}(a_1, \dots, a_n)$ is perturbed to $A_\Delta = A + \Delta$ with

$$\Delta \in \mathbf{\Delta}_{\mathcal{I}} := \{\Delta \in \mathbb{C}^{n\times n}\,;\, \Delta_{jk} = 0 \text{ if } r_{jk} = 0\}. \tag{35}$$

Here $R = (r_{jk})_{j,k\in\underline{n}}$ is a given nonnegative $n \times n$ matrix, and the perturbation space $\mathbf{\Delta}_{\mathcal{I}}$ is provided with the norm

$$\|\Delta\| = \max_{(j,k)\in\mathcal{I}} r_{jk}^{-1}|\Delta_{jk}|, \quad \Delta \in \mathbf{\Delta}_{\mathcal{I}}, \quad \text{where } \mathcal{I} := \mathcal{I}_R = \{(j,k) \in \underline{n} \times \underline{n}\,;\, r_{jk} > 0\}. \tag{36}$$

This can be subsumed into the above framework by setting $m = n$, $l_j = q_j = 1$ for $j \in \underline{m}$; $(A_j, B_j, C_j) = (a_j, 1, 1)$, $j \in \underline{m}$; and $\|\Delta_{jk}\|_{\alpha_k,\beta_j} = |\Delta_{jk}|$, $j, k \in \underline{m}$. Note that for this special case $G_j(s) = (s - a_j)^{-1}$, $j \in \underline{m}$.

COROLLARY 4.4. *Suppose $R = [r_{jk}] \in \mathbb{R}^{n\times n}$ is a nonnegative matrix with associated index set $\mathcal{I} = \mathcal{I}_R$ defined by* (36) *and normed perturbation space $(\mathbf{\Delta}_{\mathcal{I}}, \|\cdot\|)$ defined by* (35) *and* (36). *Let $a_1, \dots, a_n \in \mathbb{C}$, $\sigma_0 = \{a_1, \dots, a_n\}$, and set $A_\Delta = \mathrm{diag}(a_1, \dots, a_n) + \Delta$ for arbitrary $\Delta \in \mathbb{C}^{n\times n}$. Then the following hold:*

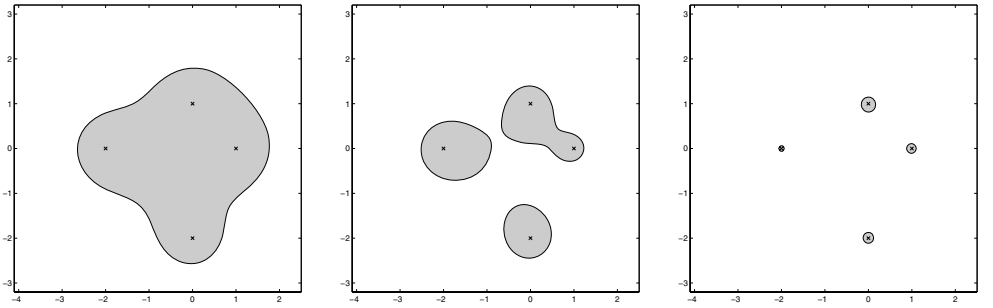


FIG. 2. *The sets S_j defined in* (40).

(a) $$\bigcup_{\Delta\in\mathbb{C}^{n\times n},\,|\Delta|\le R} \sigma(A_\Delta) = \sigma_0 \cup \{s\in\mathbb{C}\setminus\sigma_0\,;\ \varrho(R\,\mathrm{diag}(|s-a_1|^{-1},\dots,|s-a_n|^{-1}))\ge 1\ \}.$$

(b) *If $\mathbb{C}_g$ is an open subset of $\mathbb{C}$, $\sigma_0\subset\mathbb{C}_g$, then*

(37)
$$r_{\mathbf{\Delta}_{\mathcal{I}_R}}(\mathrm{diag}(a_1,\dots,a_n),I_n,I_n;\mathbb{C}_g) = \left(\sup_{s\in\partial\mathbb{C}_g}\varrho\left(R\,\mathrm{diag}(|s-a_1|^{-1},\dots,|s-a_n|^{-1})\right)\right)^{-1}.$$

(c) *In particular, if $\mathbb{C}_g=\mathbb{C}_-:=\{s\in\mathbb{C}\,;\,\Re s<0\}$ and $a_1,\dots,a_n<0$, then*

(38)
$$r_{\mathbf{\Delta}_{\mathcal{I}_R}}(\mathrm{diag}(a_1,\dots,a_n),I_n,I_n;\mathbb{C}_-) = \left(\varrho(R\,\mathrm{diag}(|a_1|^{-1},\dots,|a_n|^{-1}))\right)^{-1}.$$

Proof. (a) follows directly from Corollary 4.3(a) since $\|\Delta\|\le 1$ if and only if $\Delta\in\mathcal{B}_R(\delta)$ for all $\delta>1$. Equation (37) is a special case of (34) since $G_j(s)=(s-a_j)^{-1}$. To verify (38) note that by assumption, $a_1,\dots,a_n\in\mathbb{R}$ and so the functions $\omega\mapsto|i\,\omega-a_k|^{-1}$ attain their maxima on $\mathbb{R}$ at $\omega=0$. Hence, the monotonicity property ($\varrho 2$) of the spectral radius yields

$$\sup_{s\in i\mathbb{R}}\varrho\left(R\,\mathrm{diag}(|s-a_1|^{-1},\dots,|s-a_n|^{-1})\right) = \varrho\left(R\,\mathrm{diag}(|a_1|^{-1},\dots,|a_n|^{-1})\right).$$

Thus, (38) is a consequence of (37). ∎

Example 4.5. Suppose $A=\mathrm{diag}(1,i,-2,-2i)$ and

$$R_1=\frac{1}{4}\begin{bmatrix}0&6&2&2\\2&0&2&2\\2&2&0&8\\1&1&2&0\end{bmatrix},\quad R_2=\frac{1}{4}\begin{bmatrix}0&6&0&0\\0&0&2&2\\2&2&0&8\\1&1&2&0\end{bmatrix},\quad R_3=\frac{1}{4}\begin{bmatrix}0&6&0&0\\0&0&2&2\\0&0&0&8\\1&1&0&0\end{bmatrix}. \tag{39}$$

Figure 2 shows the sets

$$S_j := \bigcup_{\Delta\in\mathbb{C}^{4\times4},\,|\Delta|\le R_j}\sigma(A+\Delta),\qquad j=1,2,3. \tag{40}$$

Note that R_1 (resp., R_2) is obtained from R_3 by replacing all (resp., some) off-diagonal zeros of R_3 with $1/2$. Since $R_1\ge R_2\ge R_3$, the sets S_j decrease as j varies from 1 to 3. The pictures have been obtained via Corollary 4.4(a).

5. Mixed operator norms. We consider the same basic framework as that in the previous two sections. Let $\|\cdot\|_{\mathbb{C}^m}$ be an absolute norm on $\mathbb{C}^m$ which is invariant with respect to a permutation of the coordinates (for instance, a p-norm, $1 \le p \le \infty$), and let $\mathcal{N}(\cdot)$ be the induced operator norm on $\mathbb{C}^{m\times m}$. For $j,k \in \underline{m}$ let $\|\cdot\|_{\alpha_j}$ be a norm on $\mathbb{C}^{q_j}$ and let $\|\cdot\|_{\beta_k}$ be a norm on $\mathbb{C}^{l_k}$. Given any index set $\mathcal{I} \subseteq \underline{m}\times\underline{m}$, we define a norm on the perturbation space $\mathbf{\Delta}_{\mathcal{I},q,l}$ (24) by the formula

$$\|\Delta\| := \mathcal{N}\left(\begin{bmatrix} \|\Delta_{11}\|_{\alpha_1,\beta_1} & \cdots & \|\Delta_{1m}\|_{\alpha_m,\beta_1} \\ \vdots & & \vdots \\ \|\Delta_{m1}\|_{\alpha_1,\beta_m} & \cdots & \|\Delta_{mm}\|_{\alpha_m,\beta_m} \end{bmatrix}\right), \quad \Delta = [\Delta_{jk}] \in \mathbf{\Delta}_{\mathcal{I},q,l}. \tag{41}$$

In this section we derive a formula for the μ-value of *block-diagonal* matrices M with respect to the perturbation space $\mathbf{\Delta}_{\mathcal{I},q,l}$ provided with the norm (41). As a preparation we consider the general case of an arbitrary block matrix $M \in \mathbb{C}^{q\times l}$ of the form (12) and determine an upper bound for $\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M)$ [2] in terms of the associated nonnegative $m \times m$ matrix $\tilde{M}$; see (28).

PROPOSITION 5.1. *Let $M_{jk} \in \mathbb{C}^{q_j\times l_k}$, $j,k \in \underline{m}$, $M = [M_{jk}]$, and $\mathcal{I} \subseteq \underline{m}\times\underline{m}$. Then with respect to the norm* (41),

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) \le \mu^{\mathcal{N}}_{\mathbf{\Delta}_{\mathcal{I}}}(\tilde{M}), \tag{42}$$

where $\tilde{M}$ is defined by (28) *and*

$$\mathbf{\Delta}_{\mathcal{I}} := \{[\delta_{ij}] \in \mathbb{C}^{m\times m}; \delta_{ij} = 0 \text{ for } (i,j) \notin \mathcal{I}\}. \tag{43}$$

Proof. To prove (42) it suffices by Lemma 2.7 to show that for each $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$ with $\|\Delta\| = 1$ there exists $\tilde{\Delta} \in \mathbf{\Delta}_{\mathcal{I}}$ such that $\mathcal{N}(\tilde{\Delta}) = 1$ and $\varrho(\Delta M) \le \varrho(\tilde{\Delta}\tilde{M})$. Given any $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$ with $\|\Delta\| = 1$, let $\tilde{\Delta} \in \mathbb{R}^{m\times m}$ be the matrix in the parentheses on the right-hand side of (41). Then $\tilde{\Delta} \in \mathbf{\Delta}_{\mathcal{I}}$ and $\mathcal{N}(\tilde{\Delta}) = \|\Delta\| = 1$ by (41). Let $u = (u^j)_{j\in\underline{m}} \in \oplus_{j=1}^m \mathbb{C}^{l_j}$, $u \ne 0$, and $\lambda \in \mathbb{C}$ be such that $\Delta M u = \lambda u$ and $|\lambda| = \varrho(\Delta M)$. We set $\tilde{u} = (\|u^j\|_{\beta_j})_{j\in\underline{m}}$. If $\Delta M u = ((\Delta M u)^j)_{j\in\underline{m}} \in \oplus_{j=1}^m \mathbb{C}^{l_j}$ is partitioned as u, then, for every $j \in \underline{m}$,

$$\begin{aligned} \|(\Delta M u)^j\|_{\beta_j} &= \left\| \sum_{k=1}^m \sum_{i=1}^m \Delta_{ji} M_{ik} u^k \right\|_{\beta_j} \le \sum_{k=1}^m \sum_{i=1}^m \|\Delta_{ji}\|_{\alpha_i,\beta_j} \|M_{ik}\|_{\beta_k,\alpha_i} \|u^k\|_{\beta_k} \\ &= (\tilde{\Delta}\tilde{M}\tilde{u})^j. \end{aligned}$$

It follows that we have the componentwise inequality

$$\tilde{\Delta}\tilde{M}\tilde{u} \ge \left(\|(\Delta M u)^j\|_{\beta_j}\right)_{j\in\underline{m}} = |\lambda|\tilde{u}$$

and so $\varrho(\tilde{\Delta}\tilde{M}) \ge |\lambda| = \varrho(\Delta M)$ by (ϱ3). This concludes the proof. ∎

We will now prove that equality holds in (42) if $M = [M_{jk}]$ is block-diagonal, i.e., $M_{jk} = 0$ for $j,k \in \underline{m}$, $j \ne k$. In the proof we will make use of some elementary notions from graph theory [16], [2] which are summarized in the following remark.

[2] Since in this section we will consider μ-values with respect to more than one norm, we use the notation $\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}$ where there may be a risk of confusion.

Remark 5.2. A finite sequence $\gamma = (j_1, \dots, j_\ell)$ of integers is said to be a path from j_1 to j_ℓ in the directed graph $\Gamma(\underline{m}, \mathcal{I})$ if $(j_i, j_{i+1}) \in \mathcal{I}$ for all $i \in \underline{\ell - 1}$. Two nodes j, k of $\Gamma(\underline{m}, \mathcal{I})$ are said to be strongly connected if there exists a path from j to k and a path from k to j in $\Gamma(\underline{m}, \mathcal{I})$. A subset $J \subset \underline{m}$ is said to be strongly connected if any two distinct nodes in J are strongly connected in $\Gamma(\underline{m}, \mathcal{I})$. The maximal strongly connected subsets of $\underline{m}$ are called the strongly connected components of the directed graph $\Gamma(\underline{m}, \mathcal{I})$. They form a partition of $\underline{m}$. A finite sequence $\gamma = (j_1, \dots, j_\ell)$ of mutually distinct integers is said to be a cycle of length $|\gamma| := \ell \geq 1$ of the directed graph $\Gamma(\underline{m}, \mathcal{I})$ if $(j_i, j_{i+1}) \in \mathcal{I}$ for all $i \in \underline{\ell - 1}$ and $(j_\ell, j_1) \in \mathcal{I}$. We will write $j \in \gamma$ if $j = j_i$ for some $i \in \underline{\ell}$. By $\mathcal{Z}(\mathcal{I})$ we denote the set of all cycles in $\Gamma(\underline{m}, \mathcal{I})$. A cycle $\gamma \in \mathcal{Z}(\mathcal{I})$ is said to be nontrivial if $|\gamma| \geq 2$. If for a given $j_0 \in \underline{m}$ there does not exist a nontrivial cycle $\gamma \in \mathcal{Z}(\mathcal{I})$ such that $j_0 \in \gamma$, then $\{j_0\}$ is a strongly connected component of $\Gamma(\underline{m}, \mathcal{I})$.

For any $A = [a_{jk}] \in \mathbb{C}^{m \times m}$ we set $\mathcal{I}_A := \{(j,k) \in \underline{m} \times \underline{m}\,;\, a_{jk} \neq 0\}$. Let $\gamma = (j_1, \dots, j_\ell) \in \mathcal{Z}(\underline{m} \times \underline{m})$. Then the cycle product of A over γ is defined as

$$\textstyle\prod_\gamma A := \prod_{i=1}^{\ell} a_{j_i j_{i+1}}, \qquad \text{where } j_{\ell+1} := j_1.$$

Note that if $\gamma = (j)$ is a cycle of length 1, then $\prod_\gamma A = a_{jj}$.

If $A = [a_{jk}], B = [b_{jk}] \in \mathbb{C}^{m \times m}$ we denote by $A \circ B$ the Hadamard product of A and B, $A \circ B = [a_{jk} b_{jk}] \in \mathbb{C}^{m \times m}$. For nonnegative matrices the Hadamard product satisfies the following inequality which is a corollary of Theorem 5.7.21 in [16].

LEMMA 5.3. *Let $A, B \in \mathbb{R}_+^{m \times m}$. If $\mathcal{Z}(\mathcal{I}_A) = \emptyset$, then $\varrho(B \circ A) = 0$. Otherwise we have*

$$\varrho(B \circ A) \leq \varrho(B) \max_{\gamma \in \mathcal{Z}(\mathcal{I}_A)} \left(\textstyle\prod_\gamma A \right)^{\frac{1}{|\gamma|}}.$$

Given $(m_1, \dots, m_\ell) \in \mathbb{N}^\ell$, $\ell \geq 1$, and matrices $C_j \in \mathbb{C}^{m_j \times m_{j+1}}$, $j \in \underline{\ell}$, where $m_{\ell+1} := m_1$, the associated block cyclic matrix is defined by

$$Z(C_1, \dots, C_\ell) := \begin{bmatrix} 0 & C_1 & 0 & \cdots & 0 \\ 0 & \ddots & C_2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & C_{\ell-1} \\ C_\ell & 0 & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{(\sum_{j=1}^{\ell} m_j) \times (\sum_{j=1}^{\ell} m_j)}.$$

The next result which follows from the Frobenius theorem (see [10, Chapter XIII, section 2]) determines the spectral radius of nonnegative cyclic matrices with scalar blocks.

LEMMA 5.4. *Let $c_1, \dots, c_\ell \geq 0$, $\ell \in \mathbb{N}$. Then the spectrum of the cyclic matrix $Z(c_1, \dots, c_\ell)$ is given by*

$$\sigma\left(Z(c_1, \dots, c_\ell)\right) = \{ e^{2\pi i \frac{k-1}{\ell}} \varrho \,;\, k \in \underline{\ell} \}, \tag{44}$$

where $\varrho = \varrho(Z(c_1, \dots, c_\ell)) = (c_1 c_2 \cdots c_\ell)^{1/\ell}$.

The following theorem is the main result of this section.

THEOREM 5.5. *Suppose $M_j \in \mathbb{C}^{q_j \times l_j}$ for $j \in \underline{m}$, $M = \oplus_{j=1}^m M_j$, and $\mathcal{I} \subseteq \underline{m} \times \underline{m}$ are given and let $\mathcal{I}_0 := \{(j,k) \in \mathcal{I}\,;\, M_j \neq 0 \text{ and } M_k \neq 0\}$. If $(\mathbf{\Delta}_{\mathcal{I},q,l}, \|\cdot\|)$ is the perturbation structure defined by (24), (41) and $\mathbf{\Delta}_{\mathcal{I}}$ defined by (43) is provided with the norm $\mathcal{N}$, then*

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) = \mu^{\mathcal{N}}_{\mathbf{\Delta}_{\mathcal{I}}}(\tilde{M}) = \begin{cases} \max\limits_{\gamma \in \mathcal{Z}(\mathcal{I}_0)} \big(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j} \big)^{\frac{1}{|\gamma|}} & \text{if } \mathcal{Z}(\mathcal{I}_0) \neq \emptyset, \\ 0 & \text{if } \mathcal{Z}(\mathcal{I}_0) = \emptyset, \end{cases} \tag{45}$$

where $\tilde{M} = \operatorname{diag}(\|M_1\|_{\beta_1,\alpha_1}, \ldots, \|M_m\|_{\beta_m,\alpha_m})$.

Proof. Let c denote the right-hand-side of (45). We first show that

$$\mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I},q,l}}(M) \leq \mu^{\mathcal{N}}_{\mathbf{\Delta}_{\mathcal{I}}}(\tilde{M}) = \max_{\substack{\tilde{\Delta} \in \mathbf{\Delta}_{\mathcal{I}} \\ \mathcal{N}(\tilde{\Delta})=1}} \varrho\left(\tilde{\Delta}\,\tilde{M}\right) \leq c. \tag{46}$$

The first inequality in (46) follows directly from Proposition 5.1. To prove the second inequality in (46) let $E = [e_{jk}] \in \mathbb{R}^{m\times m}$, where $e_{jk} = 1$ if $(j,k) \in \mathcal{I}$ and $e_{jk} = 0$ otherwise. Set

$$A := E\tilde{M} = E \operatorname{diag}(\|M_1\|_{\beta_1,\alpha_1} \ldots \|M_m\|_{\beta_m,\alpha_m}) = [\, e_{jk}\|M_k\|_{\beta_k,\alpha_k}\,] \in \mathbb{R}^{m\times m}_+.$$

Then $\mathcal{I}_A \subseteq \mathcal{I}$, $\mathcal{Z}(\mathcal{I}_A) = \mathcal{Z}(\mathcal{I}_0)$, and we have $\prod_\gamma A = \prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}$ for every cycle $\gamma \in \mathcal{Z}(\mathcal{I})$. Thus

$$c = \begin{cases} \max\limits_{\gamma \in \mathcal{Z}(\mathcal{I}_A)} \big(\prod_\gamma A\big)^{\frac{1}{|\gamma|}} & \text{if } \mathcal{Z}(\mathcal{I}_A) \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

Let $\tilde{\Delta} = [\tilde{\Delta}_{jk}] \in \mathbf{\Delta}_{\mathcal{I}}$ with $\mathcal{N}(\tilde{\Delta}) = 1$. Then $\tilde{\Delta}\tilde{M} = [\tilde{\Delta}_{jk}\, e_{jk}\, \|M_k\|_{\beta_k,\alpha_k}] = \tilde{\Delta} \circ (E\tilde{M}) = \tilde{\Delta} \circ A$, and so by Lemma 5.3 and (47)

$$\varrho(\tilde{\Delta}\tilde{M}) = \varrho(\tilde{\Delta} \circ A) \leq \varrho(\tilde{\Delta})\, c \leq \mathcal{N}(\tilde{\Delta}) c = c.$$

This proves (46). If $c = 0$, then equality holds in (46) and hence (45). By Lemma 2.7 it remains to construct, for each cycle $\gamma \in \mathcal{Z}(\mathcal{I}_0)$, a matrix $\Delta^\gamma \in \mathbf{\Delta}_{\mathcal{I},q,l}$ such that $\|\Delta^\gamma\| = 1$ and

$$\varrho\left(\Delta^\gamma \left(\oplus_{j=1}^m M_j\right)\right) \geq \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}. \tag{48}$$

The construction of Δ^γ is as follows. Suppose that $\gamma = (j_1, \ldots, j_\ell)$. For $j \in \underline{m}$ let $u_j \in \mathbb{C}^{\ell_j}$ and $y_j \in \mathbb{C}^{q_j}$ be such that $\|u_j\|_{\beta_j} = \|y_j\|^D_{\alpha_j} = 1$ and $y_j^\top M_j u_j = \|M_j\|_{\beta_j,\alpha_j}$. Let $\Delta^\gamma := [\Delta^\gamma_{jk}]$, where

$$\Delta^\gamma_{jk} := \begin{cases} u_{j_i} y_{j_{i+1}}^\top & \text{if } (j,k) = (j_i, j_{i+1}),\ i \in \underline{\ell-1}, \\ u_{j_\ell} y_{j_1}^\top & \text{if } (j,k) = (j_\ell, j_1), \\ 0 \in \mathbb{C}^{l_j \times q_k} & \text{otherwise.} \end{cases} \qquad j,k \in \underline{m}.$$

For instance, if m=4, then

$$\Delta^{(1,3,4)} = \begin{bmatrix} 0 & 0 & u_1 y_3^\top & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_3 y_4^\top \\ u_4 y_1^\top & 0 & 0 & 0 \end{bmatrix}, \qquad \Delta^{(3,4,2,1)} = \begin{bmatrix} 0 & 0 & u_1 y_3^\top & 0 \\ u_2 y_1^\top & 0 & 0 & 0 \\ 0 & 0 & 0 & u_3 y_4^\top \\ 0 & u_4 y_2^\top & 0 & 0 \end{bmatrix}.$$

We claim that Δ^γ has the required properties. Obviously, $\Delta^\gamma \in \mathbf{\Delta}_{\mathcal{I},q,l}$. To see that $\|\Delta^\gamma\| = 1$, note that Δ^γ contains $\ell \geq 1$ nonzero blocks and in each block row and each block column of Δ^γ there is at most one nonzero block. All nonzero blocks have norm 1. Since the norm $\|\cdot\|_{\mathbb{C}^m}$ is absolute and invariant with respect to a permutation of the coordinates, it follows that

$$\|\Delta^\gamma\| = \mathcal{N}(\,[\,\|\Delta^\gamma_{jk}\|_{\alpha_j,\beta_k}\,]\,) = 1.$$

Let us show (48). Observe that the principal block submatrix of $\Delta^\gamma M$ corresponding to the block rows and columns with numbers $j_1, \dots, j_\ell$ is permutation similar to the block cyclic matrix

$$Z(u_{j_1} y_{j_2}^\top M_{j_2}, u_{j_2} y_{j_3}^\top M_{j_3}, \dots, u_{j_\ell} y_{j_1}^\top M_{j_1}) = Z(u_{j_1} y_{j_2}^\top, u_{j_2} y_{j_3}^\top, \dots, u_{j_\ell} y_{j_1}^\top)\left(\oplus_{i=1}^\ell M_{j_i}\right).$$

By Lemma 5.4 the product $\varrho := \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}$ is the spectral radius of the matrix

$$Z = Z(\|M_{j_1}\|_{\beta_{j_1},\alpha_{j_1}}, \dots, \|M_{j_\ell}\|_{\beta_{j_\ell},\alpha_{j_\ell}}).$$

Let $\xi = [\xi_{j_1}, \dots, \xi_{j_\ell}]^\top \in \mathbb{R}^\ell_+$ be an eigenvector of Z corresponding to ϱ; see (ϱ1). Then the following relations hold:

$$\|M_{j_i}\|_{\beta_{j_i},\alpha_{j_i}} \xi_{j_{i+1}} = \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}} \xi_{j_i}, \qquad i \in \underline{\ell},\ j_{\ell+1} := j_1. \tag{49}$$

Now set $w := [w_1^\top, \dots, w_m^\top]^\top$, where

$$w_j = \begin{cases} \xi_{j_{i+1}} u_{j_i} & \text{if } j = j_i,\ i \in \underline{\ell},\ j_{\ell+1} := j_1, \\ 0 \in \mathbb{C}^{\ell_j} & \text{otherwise,} \end{cases} \qquad j \in \underline{m},$$

and let $\widehat{w} := [w_{j_1}^\top, w_{j_2}^\top, \dots, w_{j_\ell}^\top]^\top$. Then using (49), we have

$$\begin{aligned} &Z(u_{j_1} y_{j_2}^\top, u_{j_2} y_{j_3}^\top, \dots, u_{j_\ell} y_{j_1}^\top)\left(\oplus_{i=1}^\ell M_{j_i}\right)\widehat{w} \\ &\quad = \begin{bmatrix} 0 & u_{j_1} y_{j_2}^\top M_{j_2} & 0 & \dots & 0 \\ 0 & 0 & u_{j_2} y_{j_3}^\top M_{j_3} & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & & \ddots & u_{j_{\ell-1}} y_{j_\ell}^\top M_{j_\ell} \\ u_{j_\ell} y_{j_1}^\top M_{j_1} & 0 & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} \xi_{j_2} u_{j_1} \\ \xi_{j_3} u_{j_2} \\ \vdots \\ \xi_{j_\ell} u_{j_{\ell-1}} \\ \xi_{j_1} u_{j_\ell} \end{bmatrix} \\ &\quad = \begin{bmatrix} \|M_{j_2}\|_{\beta_{j_2},\alpha_{j_2}} \xi_{j_3} u_{j_1} \\ \|M_{j_3}\|_{\beta_{j_3},\alpha_{j_3}} \xi_{j_4} u_{j_2} \\ \vdots \\ \vdots \\ \|M_{j_\ell}\|_{\beta_{j_\ell},\alpha_{j_\ell}} \xi_{j_1} u_{j_{\ell-1}} \\ \|M_{j_1}\|_{\beta_{j_1},\alpha_{j_1}} \xi_{j_2} u_{j_\ell} \end{bmatrix} = \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}} \begin{bmatrix} \xi_{j_2} u_{j_1} \\ \xi_{j_3} u_{j_2} \\ \vdots \\ \vdots \\ \xi_{j_\ell} u_{j_{\ell-1}} \\ \xi_{j_1} u_{j_\ell} \end{bmatrix} \\ &\quad = \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}} \widehat{w}. \end{aligned}$$

This implies that $\Delta^\gamma \left(\oplus_{j=1}^m M_j\right) w = \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}} w$. Thus (48) holds, and the proof is complete. ∎

As a corollary we obtain the following characterization of spectral value sets and stability radii for the perturbation space $(\mathbf{\Delta}_{\mathcal{I},q,l}, \|\cdot\|)$.

COROLLARY 5.6. *Suppose* $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$ *for* $j \in \underline{m}$, $\delta > 0$, $\mathcal{I} \subset \underline{m} \times \underline{m}$ *and* $\mathbf{\Delta} = \mathbf{\Delta}_{\mathcal{I},q,l}$ *is provided with the norm* (41). *Let* $A = \oplus_{j=1}^m A_j$, $B = \oplus_{j=1}^m B_j$, $C = \oplus_{j=1}^m C_j$, *and*

$$G_j(s) := C_j(sI_{n_j} - A_j)^{-1}B_j,\ j \in \underline{m}, \quad \mathcal{I}_0 := \{(j,k) \in \mathcal{I};\ B_j \neq 0, C_k \neq 0\}.$$

Then the following hold:

(a) *If* $\mathcal{Z}(\mathcal{I}_0) \neq \emptyset$, *the spectral value set of* A *with respect to perturbations of the form* (14) *is given by*

(50)
$$\bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I},q,l},\,\|\Delta\|<\delta} \sigma(A_\Delta) = \sigma(A) \cup \left\{ s \in \rho(A);\ \max_{\gamma\in\mathcal{Z}(\mathcal{I}_0)} \prod_{j\in\gamma} \|G_j(s)\|_{\beta_j,\alpha_j} > \delta^{-|\gamma|} \right\}.$$

If $\mathcal{Z}(\mathcal{I}_0) = \emptyset$, *then all the eigenvalues of* A *are fixed under perturbations of the form* (14); *i.e.,*

$$\bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I},q,l}} \sigma(A_\Delta) = \sigma(A).$$

(b) *If* $j_0 \in \underline{m}$, *and there does not exist any cycle* $\gamma \in \mathcal{Z}(\mathcal{I}_0)$ *such that* $j_0 \in \gamma$, *then*

$$\sigma(A_{j_0}) \subseteq \sigma(A_\Delta), \quad \Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}.$$

(c) *Let* $\mathbb{C}_g$ *be an open subset of* $\mathbb{C}$ *and suppose* $A_1, \ldots, A_m$ *are* $\mathbb{C}_g$*-stable. Then the stability radius* $r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A, B, C; \mathbb{C}_g)$ *is given by*

$$r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A, B, C; \mathbb{C}_g)$$
(51)
$$= \begin{cases} \left[\sup_{s\in\partial\mathbb{C}_g} \max_{\gamma\in\mathcal{Z}(\mathcal{I}_0)} \left(\prod_{j\in\gamma} \|G_j(s)\|_{\beta_j,\alpha_j} \right)^{\frac{1}{|\gamma|}} \right]^{-1} & \text{if } \mathcal{Z}(\mathcal{I}_0) \neq \emptyset, \\ \infty & \text{if } \mathcal{Z}(\mathcal{I}_0) = \emptyset. \end{cases}$$

Proof. (a) Since $G(s) = \oplus_{j=1}^m G_j(s)$ is the transfer function of (A, B, C), (a) is a direct consequence of Theorems 2.5 and 5.5.

(b) Suppose $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$. Since $A_\Delta = [A_{jk} + B_j\Delta_{jk}C_k]$ and $B_j\Delta_{jk}C_k = 0$ if $(j,k) \notin \mathcal{I}$, $B_j = 0$, or $C_k = 0$, we have $\mathcal{I}_{B\Delta C} \subseteq \mathcal{I}_0$ and $\mathcal{I}_{A_\Delta} \subseteq \mathcal{I}_A \cup \mathcal{I}_0 \subseteq \{(k,k);\ k \in \underline{m}\} \cup \mathcal{I}_0$. Now assume that there does not exist any cycle $\gamma \in \mathcal{Z}(\mathcal{I}_0)$ such that $j_0 \in \gamma$. Then $(j_0, j_0) \notin \mathcal{I}_0$; hence $B_{j_0}\Delta_{j_0j_0}C_{j_0} = 0$, and $\{j_0\}$ is a strongly connected component of the directed graph $\Gamma(\underline{m}, \mathcal{I}_{A_\Delta})$. This implies that A_Δ is permutation similar to a matrix $\tilde{A}$ of block upper triangular form:

$$A_\Delta \sim \tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \cdot & \cdots & \tilde{A}_{1r} \\ & \tilde{A}_{22} & \cdots & \tilde{A}_{2r} \\ & & \ddots & \vdots \\ & & & \tilde{A}_{rr} \end{bmatrix},$$

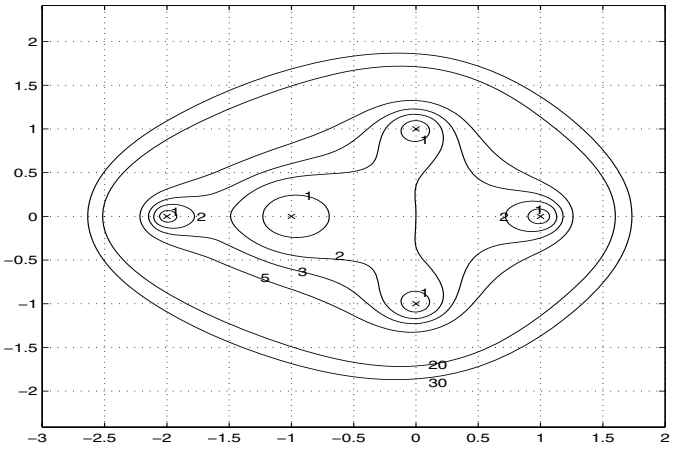


FIG. 3. *The boundaries of the Brualdi sets* $\mathcal{B}(-2,-1,1,-i,i;\delta)$, $\delta = 1,2,3,5,20,30$.

where the diagonal blocks correspond to the connected components of the graph $\Gamma(\underline{m},\mathcal{I}_{A_\Delta})$; see [2, section 2.3]. Hence $A_{j_0} = A_{j_0} + B_{j_0}\Delta_{j_0 j_0}C_{j_0} = \tilde{A}_{kk}$ for some $k \in \underline{r}$ and this shows that $\sigma(A_{j_0}) \subset \sigma(A_\Delta)$.

(c) If $\mathcal{Z}(\mathcal{I}_0) = \emptyset$, then the last statement of (a) implies $r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C;\mathbb{C}_g) = \infty$. Now suppose $\mathcal{Z}(\mathcal{I}_0) \neq \emptyset$. By the continuity of the spectrum, $r_{\mathbf{\Delta}_{\mathcal{I},q,l}}(A,B,C;\mathbb{C}_g)$ is the largest value of δ such that $\sigma(A_\Delta) \cap \partial\mathbb{C}_g = \emptyset$ for all $\Delta \in \mathbf{\Delta}_{\mathcal{I},q,l}$ of norm $\|\Delta\| < \delta$. By (a) this condition is equivalent to

$$\max_{\gamma\in\mathcal{Z}(\mathcal{I}_0)} \prod_{j\in\gamma} \|G_j(s)\|_{\beta_j,\alpha_j} \leq \delta^{-|\gamma|}, \quad s \in \partial\mathbb{C}_g,$$

or, equivalently,

$$\sup_{s\in\partial\mathbb{C}_g} \max_{\gamma\in\mathcal{Z}(\mathcal{I}_0)} \left(\prod_{j\in\gamma} \|G_j(s)\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}} \leq \delta^{-1}.$$

This concludes the proof of (51). ∎

We will now specialize the previous result to the case where the blocks are reduced to scalars, i.e., $l_i = q_i = 1$ for $i \in \underline{m}$. In this case a more concrete version of the formula (50) is obtained in which the spectral value sets are expressed as a finite union of sets of the following form:

(52)
$$\mathcal{B}(z_1,\ldots,z_\ell;\delta) := \left\{s\in\mathbb{C}\,;\, \prod_{j\in\underline{\ell}}|s-z_j| \leq \delta\right\}, \qquad \ell\in\mathbb{N},\ z_1,\ldots,z_\ell\in\mathbb{C}, \quad \delta\geq 0.$$

These sets are called *Brualdi sets* in honor of Brualdi who introduced them in [4]. They will be further discussed in what follows. Note that $\mathcal{B}(z_1,\ldots,z_\ell;0) = \{z_1,\ldots,z_\ell\}$ and $\mathcal{B}(z;\delta) = \mathcal{D}(z;\delta)$ is the closed disk of radius δ about z. For an illustration, see Figure 3.

COROLLARY 5.7. *Suppose* $a_j, b_j, c_j \in \mathbb{C}$, $j \in \underline{n}$, $A = \operatorname{diag}(a_1,\ldots,a_n)$, $B = \operatorname{diag}(b_1,\ldots,b_n)$, $C = \operatorname{diag}(c_1,\ldots,c_n)$, $\mathcal{I} \subseteq \underline{n}\times\underline{n}$, *and* $\|\cdot\|_{\mathbb{C}^n}$ *is an arbitrary norm on*

$\mathbb{C}^n$ *with induced operator norm* $\mathcal{N}(\cdot)$ *on* $\mathbb{C}^{n\times n}$. *Let*

$$\begin{aligned}
&\mathbf{\Delta}_{\mathcal{I}} := \{\,[\Delta_{jk}]\in\mathbb{C}^{n\times n}\,;\,\Delta_{jk}\in\mathbb{C} \text{ and } \Delta_{jk}=0 \text{ if } (j,k)\notin\mathcal{I}\},\\
&\|\Delta\| := \mathcal{N}(|\Delta|) = \mathcal{N}\left(\begin{bmatrix} |\Delta_{11}| & \dots & |\Delta_{1n}| \\ \vdots & & \vdots \\ |\Delta_{n1}| & \dots & |\Delta_{nn}| \end{bmatrix}\right), \quad \Delta\in\mathbf{\Delta}_{\mathcal{I}}\,, \qquad (53)\\
&A_\Delta := A + B\Delta C,\\
&\mathcal{I}_0 := \{(j,k)\in\mathcal{I}\,;\,b_jc_k\neq 0\}.
\end{aligned}$$

Then the following statements hold.

(a) *For all* $\delta>0$,

$$\bigcup_{\substack{\Delta\in\mathbf{\Delta}_{\mathcal{I}}\\ \|\Delta\|\le\delta}} \sigma(A_\Delta) = \{a_1,\dots,a_n\} \cup \bigcup_{(j_1,\dots,j_\ell)\in\mathcal{Z}(\mathcal{I}_0)} \mathcal{B}\left(a_{j_1},\dots,a_{j_\ell}\,;\,\delta^\ell\textstyle\prod_{i=1}^{\ell}|b_{j_i}c_{j_i}|\right). \tag{54}$$

(b) *Let* $j_0\in\underline{n}$ *and suppose there does not exist any cycle* $\gamma\in\mathcal{Z}(\mathcal{I}_0)$ *such that* $j_0\in\gamma$. *Then* $a_{j_0}\in\sigma(A_\Delta)$ *for all* $\Delta\in\mathbf{\Delta}_{\mathcal{I}}$.

(c) *If* $\mathbb{C}_g$ *is an open subset of* $\mathbb{C}$, $a_1,\dots,a_n\in\mathbb{C}_g$, *then*

$$\begin{aligned}
&r_{\mathbf{\Delta}_{\mathcal{I}}}(A,B,C;\mathbb{C}_g)\\
&\quad= \begin{cases} \inf\limits_{s\in\partial\mathbb{C}_g}\min\limits_{\gamma\in\mathcal{Z}(\mathcal{I}_0)} \left(\prod_{j\in\gamma}|s-a_j|/|b_jc_j|\right)^{\frac{1}{|\gamma|}} & \text{if } \mathcal{Z}(\mathcal{I}_0)\neq\emptyset,\\ \infty & \text{if } \mathcal{Z}(\mathcal{I}_0)=\emptyset. \end{cases}
\end{aligned} \tag{55}$$

Proof. (a) If $\mathcal{Z}(\mathcal{I}_0)=\emptyset$, then Corollary 5.6 (a) implies

$$\bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I}},\,\|\Delta\|\le\delta}\sigma(A_\Delta) = \{a_1,\dots,a_n\}\,.$$

Now suppose that $\mathcal{Z}(\mathcal{I}_0)\neq\emptyset$ and let $\delta>0$. Since

$$G(s) = \operatorname{diag}(g_1(s),\dots,g_n(s)), \quad g_j(s) = c_j(s-a_j)^{-1}b_j,\ j\in\underline{n},$$

is the transfer function of the system (A,B,C), we have by (10) and Theorem 5.5 the following equivalences for $s\in\mathbb{C}\setminus\{a_1,\dots,a_n\}$:

$$\begin{aligned}
s\in&\bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I}},\,\|\Delta\|\le\delta}\sigma(A_\Delta)\\
&\Leftrightarrow \mu^{\|\cdot\|}_{\mathbf{\Delta}_{\mathcal{I}}}(\operatorname{diag}(c_1(s-a_1)^{-1}b_1,\dots,c_n(s-a_n)^{-1}b_n))\ge\delta^{-1}\\
&\Leftrightarrow \max_{\gamma\in\mathcal{Z}(\mathcal{I}_0)}\left(\textstyle\prod_{j\in\gamma}|c_j(s-a_j)^{-1}b_j|\right)^{\frac{1}{|\gamma|}}\ge\delta^{-1}\\
&\Leftrightarrow \exists\,\gamma\in\mathcal{Z}(\mathcal{I}_0):\ \left(\textstyle\prod_{j\in\gamma}|c_j(s-a_j)^{-1}b_j|\right)^{\frac{1}{|\gamma|}}\ge\delta^{-1}\\
&\Leftrightarrow \exists\,\gamma\in\mathcal{Z}(\mathcal{I}_0):\ \textstyle\prod_{j\in\gamma}|s-a_j|\le\delta^{|\gamma|}\prod_{j\in\gamma}|b_jc_j|.
\end{aligned}$$

Hence (54) holds.

(b) is a special case of Corollary 5.6(b).

(c) Since $|g_j(s)|^{-1}=|s-a_j|/|b_jc_j|$ if $b_jc_j\neq 0$, formula (55) is a special case of (51). ∎

Remark 5.8. (i) Corollaries 5.6 and 5.7 show that the spectral value sets and stability radii of block-diagonal and diagonal matrices with respect to the normed perturbation structure $(\mathbf{\Delta}_{\mathcal{I},q,l}, \|\cdot\|)$ (see (41)) are independent of the norm $\mathcal{N}$.

(ii) For the special case that $\mathbb{C}_g = \mathbb{C}_-$ and $a_1, \dots, a_n < 0$ it follows from (55) that

$$\begin{aligned} r_{\mathbf{\Delta}_{\mathcal{I}}}(A, B, C; \mathbb{C}_-) &= \min_{\gamma \in \mathcal{Z}(\mathcal{I}_0)} \left(\prod_{j \in \gamma} |a_j| / |b_j c_j| \right)^{\frac{1}{|\gamma|}} \\ &= \min_{\gamma \in \mathcal{Z}(\mathcal{I}_0)} \left(\prod_{j \in \gamma} r_{\mathbb{C}}(a_j, b_j, c_j; \mathbb{C}_-) \right)^{\frac{1}{|\gamma|}}. \end{aligned}$$

By Corollary 5.7 Brualdi sets play a fundamental role in determining the spectral value sets of diagonal matrices with respect to perturbations $\Delta \in \mathbf{\Delta}_{\mathcal{I}}$. We conclude this section with some remarks concerning these sets. Each Brualdi set (52) can be represented as the intersection of a family of sets which are unions of ℓ closed disks of centers z_i, $i \in \underline{\ell}$. More precisely, we have the following proposition.

PROPOSITION 5.9. *Let $z_1, \dots, z_\ell \in \mathbb{C}$ and $\delta > 0$. Then*

$$\mathcal{B}(z_1, \dots, z_\ell; \delta) = \bigcap_{\substack{r_1, \dots, r_\ell > 0, \\ \prod_{j \in \underline{\ell}} r_j = 1}} \left(\bigcup_{j \in \underline{\ell}} \mathcal{D}(z_j; \delta^{\frac{1}{\ell}} r_j) \right). \tag{56}$$

Proof. Let D denote the set on the right-hand side of (56). Suppose that $s \notin D$. Then there are $r_1, \dots, r_\ell > 0$ such that $\prod_{j \in \underline{\ell}} r_j = 1$ and $|s - z_j| > \delta^{\frac{1}{\ell}} r_j$ for all $j \in \underline{\ell}$. Multiplying the latter inequalities we obtain that $\prod_{j \in \underline{\ell}} |s - z_j| > \delta$. Thus $s \notin \mathcal{B}(z_1, \dots, z_\ell; \delta)$. Hence $\mathcal{B}(z_1, \dots, z_\ell; \delta) \subseteq D$. Now suppose that $s \notin \mathcal{B}(z_1, \dots, z_\ell; \delta)$; then $\delta_1 := \prod_{j \in \underline{\ell}} |s - z_j| > \delta$. If we define $r_j > 0$ by $|s - z_j| = \delta_1^{\frac{1}{\ell}} r_j$, $j \in \underline{\ell}$, then $s \notin \bigcup_{j \in \underline{\ell}} \mathcal{D}(z_j; \delta^{\frac{1}{\ell}} r_j)$ and $\prod_{j \in \underline{\ell}} r_j = 1$. So $s \notin D$. ∎

From the relation (56) one can derive an upper bound for the connected components of Brualdi sets.

PROPOSITION 5.10. *Let $z_1, \dots, z_\ell \in \mathbb{C}$ and $\delta > 0$. Then the following hold:*

(a) *Each connected component of $\mathcal{B}(z_1, \dots, z_\ell; \delta)$ contains at least one of the points z_j, $j \in \underline{\ell}$.*

(b) *Let $\epsilon > 0$ and suppose that for a given $j \in \underline{\ell}$*

$$\min_{k \in \underline{\ell}, k \neq j} |z_j - z_k| > \delta^{\frac{1}{\ell}} \left(\epsilon + \epsilon^{-\frac{1}{\ell - 1}} \right). \tag{57}$$

Then the connected component K_j of $\mathcal{B}(z_1, \dots, z_\ell; \delta)$ with $z_j \in K_j$ is contained in $\mathcal{D}(z_j; \epsilon \delta^{\frac{1}{\ell}})$.

Proof. (a) Set $f(s) := \prod_{j \in \underline{\ell}} (s - z_j)$. Let K be a connected component of $\mathcal{B}(z_1, \dots, z_\ell; \delta)$. Then K is compact. Hence there exists $s_0 \in K$ such that $|f(s_0)| = \min_{s \in K} |f(s)|$. Let U be an open neighborhood of s_0 such that $U \cap \mathcal{B}(z_1, \dots, z_\ell; \delta) = U \cap K$. By the definition of $\mathcal{B}(z_1, \dots, z_\ell; \delta)$ we have $|f(s)| > \delta$ for all $s \in U \setminus K$. Thus $|f(s_0)| = \min_{s \in U} |f(s)|$ and this implies $f(s_0) = 0$ since f is holomorphic and nonconstant. Thus $s_0 = z_j$ for some $j \in \underline{\ell}$.

(b) For $i \in \underline{\ell}$ set

$$r_i = \begin{cases} \epsilon & \text{if } i = j, \\ \epsilon^{-\frac{1}{\ell - 1}} & \text{otherwise.} \end{cases}$$

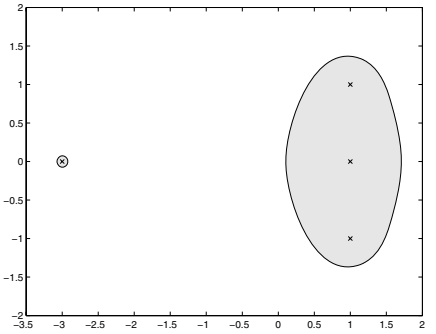


FIG. 4. *The Brualdi set* $\mathcal{B}(-3, 1, 1+\imath, 1-\imath; \delta)$, $\delta = 5$.

Then $\prod_{i\in\underline{\ell}} r_i = 1$ and Proposition 5.9 yields that $\mathcal{B}(z_1, \dots, z_\ell; \delta) \subseteq \bigcup_{i\in\underline{\ell}} \mathcal{D}(z_i; \delta^{\frac{1}{\ell}} r_i)$. The condition (57) implies that $\mathcal{D}(z_j; \delta^{\frac{1}{\ell}} r_j) \cap \mathcal{D}(z_k; \delta^{\frac{1}{\ell}} r_k) = \emptyset$ for all $k \neq j$. Thus $K_j \subseteq \mathcal{D}(z_j; \delta^{\frac{1}{\ell}} r_j)$. ∎

Roughly speaking, the above proposition states that if the distance of $z_j \in \mathbb{C}$ from the numbers $z_k \in \mathbb{C}$, $k \neq j$, is large, then the connected component K_j of $\mathcal{B}(z_1, \dots, z_\ell; \delta)$ is a small set. This is illustrated in Figure 4.

6. Off-diagonal perturbation structures. We consider the same basic framework as that in section 5 but now the index set is off-diagonal:

$$\mathcal{I}_{\text{off}} := \{(j,k) \in \underline{m} \times \underline{m};\ j \neq k\}. \tag{58}$$

The corresponding perturbation class $\mathbf{\Delta}_{\mathcal{I}_{\text{off}},q,l}$ is the set of all $m \times m$ block matrices $\Delta = [\Delta_{jk}]$ such that $\Delta_{jk} \in \mathbb{C}^{l_j \times q_k}$ and $\Delta_{jj} = 0$ for all $j, k \in \underline{m}$. In this section we derive formulae for the corresponding μ-function, spectral value sets, and stability radii.

Recall the following inequality for the geometric mean.

LEMMA 6.1. *Let* $c_1 \geq c_2 \geq \cdots \geq c_\ell \geq 0$. *Then, for all* $k \in \underline{\ell}$, $(\prod_{j=1}^{\ell} c_j)^{\frac{1}{\ell}} \leq (\prod_{j=1}^{k} c_j)^{\frac{1}{k}}$.

For $\mathcal{I} = \mathcal{I}_{\text{off}}$ the following proposition is a special case of Theorem 5.5.

PROPOSITION 6.2. *Let* $M_j \in \mathbb{C}^{q_j \times l_j}$, $j \in \underline{m}$, $M = \bigoplus_{j=1}^{m} M_j$. *Then with respect to the norm* (41),

$$\mu_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}},q,\ell}}(M) = \max_{1\leq j<k\leq m} \sqrt{\|M_j\|_{\beta_j,\alpha_j}\ \|M_k\|_{\beta_k,\alpha_k}}.$$

Proof. Each pair $(j,k) \in \underline{m} \times \underline{m}$, $j \neq k$, is a cycle in the graph associated with $\mathcal{I}_{\text{off}}$. Thus

$$\max_{1\leq j<k\leq m} \sqrt{\|M_j\|_{\beta_j,\alpha_j}\ \|M_k\|_{\beta_k,\alpha_k}} \leq \max_{\gamma\in\mathcal{Z}(\mathcal{I}_{\text{off}})} \left(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j}\right)^{\frac{1}{|\gamma|}}. \tag{59}$$

Each cycle $\gamma \in \mathcal{Z}(\mathcal{I}_{\text{off}})$ has length $|\gamma| \geq 2$. Let $\gamma = (j_1, \dots, j_\ell)$, and let $j_i, j_r \in \gamma$ be such that $i \neq r$ and $\|M_{j_i}\|_{\beta_{j_i},\alpha_{j_i}} \geq \|M_{j_r}\|_{\beta_{j_r},\alpha_{j_r}} \geq \|M_{j_\nu}\|_{\beta_{j_\nu},\alpha_{j_\nu}}$ for all $\nu \in \underline{\ell} \setminus \{i, r\}$. By Lemma 6.1 we have that $(\prod_{j\in\gamma} \|M_j\|_{\beta_j,\alpha_j})^{\frac{1}{|\gamma|}} \leq \sqrt{\|M_{j_i}\|_{\beta_{j_i},\alpha_{j_i}}\ \|M_{j_r}\|_{\beta_{j_r},\alpha_{j_r}}}$. Thus, equality holds in (59). Now, the proposition follows from Theorem 5.5. ∎

An analogous result holds if the underlying norm on $\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}$ is the operator norm induced by p-norms on $\mathbb{C}^{l_1+\cdots+l_m}$ and $\mathbb{C}^{q_1+\cdots+q_m}$. The corresponding μ-function will be denoted by $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}}(\cdot)$. To prove the result we need the following lemma.

LEMMA 6.3. *Let $m_j, n_j \in \mathbb{N}$, $j = 1, 2$, and suppose $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are absolute norms on $\mathbb{C}^{m_1+m_2}$ and $\mathbb{C}^{n_1+n_2}$, respectively. If $X \in \mathbb{C}^{m_1\times n_2}$, $Y \in \mathbb{C}^{m_2\times n_1}$, and $Z \in \mathbb{C}^{m_2\times n_2}$, then*

$$\left\| \begin{bmatrix} 0 & tX \\ tY & Z \end{bmatrix} \right\|_{\beta,\alpha} \le t \left\| \begin{bmatrix} 0 & X \\ Y & Z \end{bmatrix} \right\|_{\beta,\alpha}, \quad t \ge 1.$$

Proof. The function

$$\zeta \mapsto f(\zeta) := \left\| \begin{bmatrix} 0 & X \\ Y & \zeta Z \end{bmatrix} \right\|_{\beta,\alpha}$$

is convex on $\mathbb{R}$, and since $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are absolute, we have that

$$f(-\zeta) = \left\| \begin{bmatrix} I_{m_1} & 0 \\ 0 & -I_{m_2} \end{bmatrix} \begin{bmatrix} 0 & X \\ Y & \zeta Z \end{bmatrix} \begin{bmatrix} -I_{n_1} & 0 \\ 0 & I_{n_2} \end{bmatrix} \right\|_{\beta,\alpha} = f(\zeta)$$

for all $\zeta \in \mathbb{R}$. From this it follows that f is a nondecreasing function on $[0, \infty)$. Thus for all $t \ge 1$,

$$\left\| \begin{bmatrix} 0 & tX \\ tY & Z \end{bmatrix} \right\|_{\beta,\alpha} = tf\left(\frac{1}{t}\right) \le tf(1) = t \left\| \begin{bmatrix} 0 & X \\ Y & Z \end{bmatrix} \right\|_{\beta,\alpha}. \qquad \square$$

In the following theorem $\|\cdot\|_p$ denotes the operator norm induced by p-norms on the corresponding vector spaces.

THEOREM 6.4. *Suppose $p \in [1, \infty]$, $M_j \in \mathbb{C}^{q_j\times l_j}$ for $j \in \underline{m}$ and $M = \bigoplus_{j=1}^m M_j$. Then*

$$\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = \max_{1\le j<k\le m} \sqrt{\|M_j\|_p \, \|M_k\|_p}\,.$$

Proof. Without loss of generality we may assume that $\|M_1\|_p \ge \|M_2\|_p \ge \|M_j\|_p$ for $j \ge 3$. Let $\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}$. Then $\Delta = \begin{bmatrix} 0 & X \\ Y & Z \end{bmatrix}$ for some $X \in \mathbb{C}^{l_1\times Q'}, Y \in \mathbb{C}^{L'\times q_1}$, $Z \in \mathbb{C}^{L'\times Q'}$, where $Q' = \sum_{j=2}^m q_j$, $L' = \sum_{j=2}^m l_j$. Suppose first that $M_2 = 0$. Then all eigenvalues of $\Delta M = \begin{bmatrix} 0 & 0 \\ YM_1 & 0 \end{bmatrix}$ are zero for all $\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}$. Consequently, $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}}(M) = 0 = \sqrt{\|M_1\|_p \|M_2\|_p}$. Suppose now that $M_2 \ne 0$ and let

$$F_t := \mathrm{diag}(tI_{q_1}, I_{Q'}), \qquad G_t := \mathrm{diag}(tI_{l_1}, I_{L'}), \qquad t := \sqrt{\frac{\|M_1\|_p}{\|M_2\|_p}} \ge 1.$$

Then $\|t^{-2}M_1\|_p = \|M_2\|_p$ and therefore

$$\|F_t^{-1} M G_t^{-1}\|_p = \|\operatorname{diag}\left(t^{-2}M_1, M_2, M_3, \ldots, M_m\right)\|_p = \|M_2\|_p.$$

By Lemma 6.3 $\|G_t \Delta F_t\|_p \le t\,\|\Delta\|_p$ for all $\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}$. Suppose that $\|\Delta\|_p = 1$. Then we have

$$\begin{aligned} \varrho(\Delta M) &\le \|G_t \Delta M G_t^{-1}\|_p = \|(G_t \Delta F_t)(F_t^{-1} M G_t^{-1})\|_p \\ &\le \|G_t \Delta F_t\|_p \, \|F_t^{-1} M G_t^{-1}\|_p = \|G_t \Delta F_t\|_p \, \|M_2\|_p \\ &\le \|\Delta\|_p \, t \, \|M_2\|_p = \sqrt{\|M_1\|_p \|M_2\|_p}\,. \end{aligned}$$

Therefore $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}}(M) \le \sqrt{\|M_1\|_p \|M_2\|_p}$. To see that equality holds, let $u_j \in \mathbb{C}^{l_j}$, $y_j \in \mathbb{C}^{q_j}$, $j = 1, 2$, be such that $\|u_j\|_p = \|y_j\|_p^D = 1$ and $y_j^\top M_j u_j = \|M_j\|_p$, $j = 1, 2$. Define

$$\Delta_0 := \operatorname{diag}\left(\begin{bmatrix} 0 & u_1 y_2^\top \\ u_2 y_1^\top & 0 \end{bmatrix}, 0\right) \in \mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}\,, \qquad u := \begin{bmatrix} \sqrt{\|M_2\|_p}\, u_1 \\ \sqrt{\|M_1\|_p}\, u_2 \\ 0 \end{bmatrix} \in \mathbb{C}^{l_1+\cdots+l_m}.$$

Then $\|\Delta_0\|_p = 1$, and an easy calculation yields $\Delta_0 M u = \sqrt{\|M_1\|_p \, \|M_2\|_p}\, u$. Thus $\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}}(M) \ge \varrho(\Delta_0 M) \ge \sqrt{\|M_1\|_p \, \|M_2\|_p}$, and this concludes the proof. ∎

We have not been able to extend Theorem 6.4 to arbitrary index sets $\mathcal{I} \subseteq \underline{m} \times \underline{m}$ and so we formulate a conjecture in this respect as an open question.

Open question: Does the identity

$$\mu^{(p)}_{\mathbf{\Delta}_{\mathcal{I},q,l}}\left(\bigoplus_{j=1}^m M_j\right) = \max_{\gamma \in \mathcal{Z}(\mathcal{I})} \left(\prod_{j\in\gamma} \|M_j\|_p\right)^{\frac{1}{|\gamma|}}$$

hold for arbitrary index sets $\mathcal{I} \subseteq \underline{m} \times \underline{m}$?

As corollaries of Proposition 6.2 and Theorem 6.4 we obtain the following formulae for spectral value sets and stability radii. The first corollary deals with the general block-diagonal case, and the second deals with the (scalar) diagonal case.

COROLLARY 6.5. *Suppose* $(A_j, B_j, C_j) \in L_{n_j,l_j,q_j}$, $j \in \underline{m}$, $A = \oplus_{j=1}^m A_j$, $B = \oplus_{j=1}^m B_j$, $C = \oplus_{j=1}^m C_j$, $\delta > 0$, *and* $\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}$ *is provided with the norm* (41) *or with the operator norm induced by some p-norm,* $1 \le p \le \infty$.

Let $G_j(s) = C_j(sI_{n_j} - A_j)^{-1}B_j$, $j \in \underline{m}$, *and define* $\|G_i(s)\| = \|G_i(s)\|_{\beta_i,\alpha_i}$ *or* $\|G_i(s)\| = \|G_i(s)\|_p$, $i \in \underline{m}$, *respectively. Then the following hold:*

(a) *The spectral value set of A with respect to perturbations of the form*

$$A \rightsquigarrow A_\Delta = A + B\Delta C, \quad \Delta \in \mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l},\ \|\Delta\| < \delta, \tag{60}$$

is given by

$$\begin{aligned} &\bigcup_{\substack{\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l} \\ \|\Delta\| < \delta}} \sigma(A_\Delta) \\ &\quad = \sigma(A) \cup \left\{ s \in \rho(A);\ \max_{1\le j<k\le m} \sqrt{\|G_j(s)\|\|G_k(s)\|} > \delta^{-1} \right\}. \end{aligned} \tag{61}$$

(b) *Let* $\mathbb{C}_g$ *be an open subset of* $\mathbb{C}$ *and suppose* $A_1, \ldots, A_m$ *are* $\mathbb{C}_g$*-stable. Then the stability radius is given by*

$$\begin{aligned} r_{\mathbf{\Delta}_{\mathcal{I}_{\mathrm{off}},q,l}}(A, B, C; \mathbb{C}_g) &= \left[\sup_{s\in\partial\mathbb{C}_g} \max_{1\le j<k\le m} \sqrt{\|G_j(s)\|\|G_k(s)\|} \right]^{-1} \\ &\ge \min_{1\le j<k\le m} \sqrt{r_{\mathbb{C}}(A_j, B_j, C_j; \mathbb{C}_g)\, r_{\mathbb{C}}(A_k, B_k, C_k; \mathbb{C}_g)}. \end{aligned} \tag{62}$$

Proof. Making use of Theorem 2.5 and Proposition 6.2 (resp., Theorem 6.4), the corollary can be proved in a way similar to that of Corollary 5.6(a),(c). ∎

COROLLARY 6.6. *Suppose* $a_j, b_j, c_j \in \mathbb{C}$, $j \in \underline{n}$, $A = \operatorname{diag}(a_1, \dots, a_n)$, $B = \operatorname{diag}(b_1, \dots, b_n)$, $C = \operatorname{diag}(c_1, \dots, c_n)$, *and* $\mathcal{N}(\cdot)$ *is an arbitrary operator norm on* $\mathbb{C}^{n\times n}$. *Let*

$$\mathbf{\Delta}_{\mathcal{I}_{\text{off}}} := \left\{[\Delta_{jk}] \in \mathbb{C}^{n\times n}\,;\, \Delta_{11} = \cdots = \Delta_{nn} = 0\right\}$$

be provided with the norm $\|\cdot\| = \|\cdot\|_p$, $1 \le p \le \infty$, *or*

$$\|\Delta\| := \mathcal{N}(|\Delta|) = \mathcal{N}\left(\begin{bmatrix} |\Delta_{11}| & \dots & |\Delta_{1n}| \\ \vdots & & \vdots \\ |\Delta_{n1}| & \dots & |\Delta_{nn}| \end{bmatrix}\right), \quad \Delta \in \mathbf{\Delta}_{\mathcal{I}_{\text{off}}}, \tag{63}$$

$$A_\Delta := A + B\Delta C.$$

If $A = \operatorname{diag}(a_1, \dots, a_n)$, $B = \operatorname{diag}(b_1, \dots, b_n)$, $C = \operatorname{diag}(c_1, \dots, c_n)$, *then the following statements hold:*

(a) *For all* $\delta > 0$,

$$\bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I}_{\text{off}}},\, \|\Delta\|\le\delta} \sigma(A_\Delta) = \Big\{s \in \mathbb{C}; \min_{1\le j<k\le n} |s - a_j||s - a_k| \le \delta^2 |b_j c_j b_k c_k|\Big\}. \tag{64}$$

(b) *If* $\mathbb{C}_g$ *is an open subset of* $\mathbb{C}$ *and* $a_1, \dots, a_n \in \mathbb{C}_g$, *then*

$$r_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}}}}(A, B, C; \mathbb{C}_g) = \inf_{s\in\partial\mathbb{C}_g} \min_{1\le j<k\le n} \left(\frac{|s - a_j|}{|b_j c_j|}\frac{|s - a_k|}{|b_k c_k|}\right)^{1/2}. \tag{65}$$

Proof. Making use of Theorem 2.5 and Proposition 6.2 (resp., Theorem 6.4) the corollary can be proved in a way similar to that of Corollary 5.7. □

If $\mathbb{C}_g = \mathbb{C}_-$ and $a_1, \dots, a_n < 0$, then $|\imath\omega - a_j| \ge |a_j|$ for all $\omega \in \mathbb{R}$, $j \in \underline{n}$, so that (65) implies

$$r_{\mathbf{\Delta}_{\mathcal{I}_{\text{off}}}}(A, B, C; \mathbb{C}_-) = \min_{1\le j<k\le n} \left(\frac{|a_j a_k|}{|b_j c_j b_k c_k|}\right)^{1/2}.$$

7. Application: Inclusion theorems. An arbitrary matrix $A = [a_{jk}] \in \mathbb{C}^{n\times n}$ can be represented as a perturbation of the diagonal matrix $D_A = \operatorname{diag}(a_{11}, \dots, a_{nn})$ by an off-diagonal perturbation matrix Δ_A:

$$A = D_A + \Delta_A, \quad \text{where } \Delta_A = A - D_A \in \mathbf{\Delta}_{\mathcal{I}_{\text{off}}} := \left\{[\Delta_{jk}] \in \mathbb{C}^{n\times n}\,;\, \Delta_{11} = \cdots = \Delta_{nn} = 0\right\}.$$

Hence, setting

$$\mathcal{I} := \mathcal{I}_A \cap \mathcal{I}_{\text{off}} = \{(j, k) \in \underline{n} \times \underline{n}\,;\, j \ne k \text{ and } a_{jk} \ne 0\}, \tag{66}$$

we have by Remark 2.6(ii) that

$$\sigma(A) \subseteq \bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I}},\, \|\Delta\|\le\|\Delta_A\|} \sigma(D_A + \Delta) = \overline{\sigma_{\mathbf{\Delta}_{\mathcal{I}}}(D_A, I_n, I_n; \|\Delta_A\|)}. \tag{67}$$

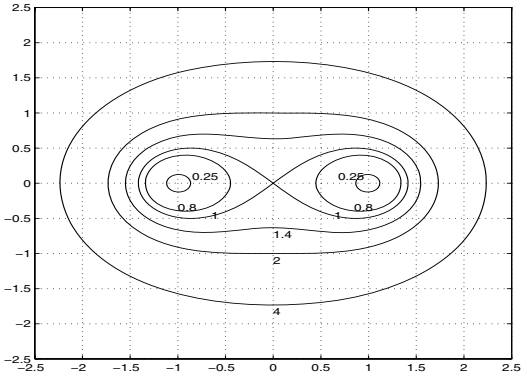


FIG. 5. *The ovals of Cassini* $\mathcal{C}(1,-1;\rho)$, $\rho = 0.25, 0.8, 1, 1.4, 2, 4$.

Applying the previous results about spectral value sets of diagonal matrices, one obtains different estimates for the location of the spectrum of A (depending on whether one chooses the perturbation norm to be (36) or (63)). In this section we recall the classical eigenvalue inclusion theorems of Gershgorin, Brauer, and Brualdi and show how they can be obtained as corollaries of the results in the previous sections.

Gershgorin's theorem states that for all $A = [a_{jk}] \in \mathbb{C}^{n\times n}$

$$\sigma(A) \subset \mathcal{G}_A := \bigcup_{j\in\underline{n}} \mathcal{D}(a_{jj}; R_j(A)), \quad \text{where } R_j(A) := \sum_{k\in\underline{n},k\neq j} |a_{jk}|. \tag{68}$$

Gershgorin's theorem was improved by Brauer [3]. He used inclusion regions for the eigenvalues of the following type:

$$\mathcal{C}(z_1, z_2; \rho) := \{ s\in\mathbb{C} \,;\, |s - z_1|\,|s - z_2| \leq \rho \}, \qquad z_1, z_2 \in \mathbb{C},\ \rho \geq 0.$$

The sets $\mathcal{C}(z_1, z_2; \rho)$ and their boundaries are called the *ovals of Cassini*. For an illustration, see Figure 5. Brauer's theorem states that

$$\sigma(A) \subseteq \mathcal{C}_A := \bigcup_{1\leq j<k\leq n} \mathcal{C}(a_{jj}, a_{kk}; R_j(A)R_k(A)), \quad A = [a_{jk}] \in \mathbb{C}^{n\times n}. \tag{69}$$

A further refinement has been obtained by Brualdi [4] who gave more precise information about the location of the eigenvalues by taking into account the zero structure of A. For this he introduced sets of the form (52) which now carry his name. With every matrix $A = [a_{jk}] \in \mathbb{C}^{n\times n}$ we associate the following union of Brualdi sets:

$$\mathcal{B}_A := \bigcup_{(j_1,\ldots,j_\ell)\in\mathcal{Z}(\mathcal{I})} \mathcal{B}\left(a_{j_1j_1}, \ldots, a_{j_\ell j_\ell}; \textstyle\prod_{i=1}^{\ell} R_{j_i}(A)\right), \tag{70}$$

where R_j are as in (68) and $\mathcal{I} := \mathcal{I}_A \cap \mathcal{I}_{\text{off}}$; see (66). Brualdi's theorem states that

$$\sigma(A) \subseteq \mathcal{B}_A \tag{71}$$

provided that each index $j \in \underline{n}$ is contained in some cycle $\gamma \in \mathcal{Z}(\mathcal{I})$. From Corollary 5.7 we obtain the following slight extension of this result.

COROLLARY 7.1. *Let* $A \in \mathbb{C}^{n\times n}$ *and set*

$$\begin{aligned}\sigma_0(A) &:= \{a_{jj}\,;\, j\in\underline{n} \text{ and } \forall \gamma\in\mathcal{Z}(\mathcal{I}) : j\notin\gamma\}\\ &= \{a_{jj}\,;\, j\in\underline{n} \text{ and } (j\in\gamma\in\mathcal{Z}(\mathcal{I}_A)\Rightarrow \gamma=(j))\}.\end{aligned}$$

Then $\sigma_0(A)\subseteq\sigma(A)$ *and* $\sigma(A)\setminus\sigma_0(A)\subseteq\mathcal{B}_A$. *In particular, if* $\sigma_0(A)=\emptyset$, *then* $\sigma(A)\subseteq\mathcal{B}_A$.

Proof. If $A\in\mathbb{C}^{n\times n}$ is a diagonal matrix, then there is nothing to prove. Assume that A is nondiagonal and has off-diagonal row sums $R_j(A)$, $j\in\underline{n}$. Set $\tilde{\Delta}=[\tilde{\Delta}_{jk}]\in\mathbb{C}^{n\times n}$, where

$$\tilde{\Delta}_{jk} := \begin{cases} R_j(A)^{-1}a_{jk} & \text{if } j\neq k \text{ and } R_j(A)\neq 0,\\ 0 & \text{otherwise.}\end{cases}$$

Then $A = A_{\tilde{\Delta}} := \operatorname{diag}(a_{11},\dots,a_{nn}) + \operatorname{diag}(R_1(A),\dots,R_n(A))\,\tilde{\Delta}$, and $\|\tilde{\Delta}\|_1 = 1$. Note that $\|\Delta\|_1 = \|\,|\Delta|\,\|_1$ is a norm of the form (63). Furthermore, we have that $\tilde{\Delta}\in\mathbf{\Delta}_{\mathcal{I}}$, where $\mathcal{I}=\mathcal{I}_A\cap\mathcal{I}_{\text{off}}$. Thus

$$\begin{aligned}\sigma(A)\subseteq &\bigcup_{\Delta\in\mathbf{\Delta}_{\mathcal{I}},\,\|\Delta\|_1\le 1}\sigma(A_\Delta),\\ &\text{where } A_\Delta := \operatorname{diag}(a_{11},\dots,a_{nn}) + \operatorname{diag}(R_1(A),\dots,R_n(A))\,\Delta.\end{aligned}$$

Hence, applying Corollary 5.7 with the norm $\mathcal{N}(\cdot)=\|\cdot\|_1$, $\delta=1$, $b_i=R_i(A)$, and $c_i=1$, $i\in\underline{n}$, we obtain the result: $\sigma_0(A)\subseteq\sigma(A)$ follows from (b) and $\sigma(A)\backslash\sigma_0(A)\subseteq\mathcal{B}_A$ follows from (a). ∎

An equivalent extension of Brualdi's result can be found in [29, Theorem 2.5] (note that the definition of $\mathcal{B}_A$ in [29] is different from ours).

We conclude this paper with a brief discussion of the relationship between the above results of Gershgorin, Brauer, and Brualdi. First note that $\mathcal{B}(z_1;\rho)=\mathcal{D}(z_1;\rho)$ and $\mathcal{B}(z_1,z_2;\rho)=\mathcal{C}(z_1,z_2;\rho)$. The following proposition yields a useful tool for establishing inclusion relations between these sets.

PROPOSITION 7.2. *Let* $z_1,\dots,z_\ell\in\mathbb{C}$ *and* $\rho_1,\dots,\rho_\ell\ge 0$. *Then*

$$\mathcal{B}\left(z_1,\dots,z_\ell;\textstyle\prod_{j\in\underline{\ell}}\rho_j\right)\subseteq\bigcup_{k\in\underline{\ell}}\mathcal{B}\left(z_1,\dots,\widehat{z}_k,\dots,z_\ell;\textstyle\prod_{j\in\underline{\ell},\,j\neq k}\rho_j\right),$$

where $\mathcal{B}\left(z_1,\dots,\widehat{z}_k,\dots,z_\ell;\prod_{j\in\underline{\ell},\,j\neq k}\rho_j\right)=\left\{\, s\in\mathbb{C}\,;\ \prod_{j\in\underline{\ell},\,j\neq k}|s-z_j|\le\prod_{j\in\underline{\ell},j\neq k}\rho_j\,\right\}$.

Proof. Suppose that $s\notin\bigcup_{k\in\underline{\ell}}\mathcal{B}\,(z_1,\dots,\widehat{z}_k,\dots,z_\ell;\prod_{j\in\underline{\ell},\,j\neq k}\,\rho_j)$. Then we have, for all $k\in\underline{\ell}$, $\prod_{j\in\underline{\ell},\,j\neq k}|s-z_j|>\prod_{j\in\underline{\ell},\,j\neq k}\rho_j$. By multiplying these ℓ inequalities we obtain $(\prod_{j\in\underline{\ell}}|s-z_j|\,)^{\ell-1}>(\prod_{j\in\underline{\ell}}\rho_j\,)^{\ell-1}$. Thus $s\notin\mathcal{B}(z_1,\dots,z_\ell;\prod_{j\in\underline{\ell}}\rho_j)$. ∎

By induction we obtain from Proposition 7.2 the following corollary.

COROLLARY 7.3. *Let* $z_1,\dots,z_\ell\in\mathbb{C}$ *and* $\rho_1,\dots,\rho_\ell\ge 0$. *Then*

$$\mathcal{B}\left(z_1,\dots,z_\ell;\textstyle\prod_{j\in\underline{\ell}}\rho_j\right)\subseteq\bigcup_{1\le j<k\le\ell}\mathcal{C}(z_j,z_k;\rho_j\rho_k)\subseteq\bigcup_{j\in\underline{\ell}}\mathcal{D}(z_j;\rho_j).$$

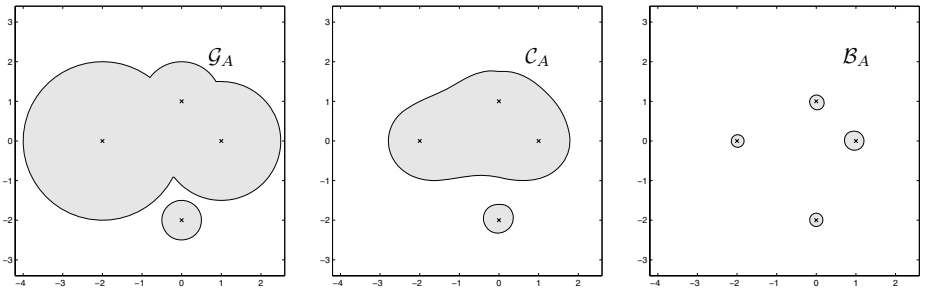


FIG. 6. *Comparison of the regions* $\mathcal{G}_A$, $\mathcal{C}_A$, *and* $\mathcal{B}_A$.

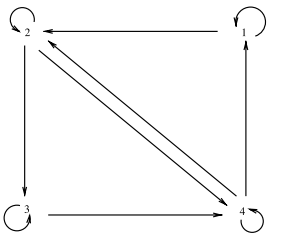
Corollary 7.3 implies that for all $A \in \mathbb{C}^{n\times n}$, $n \geq 2$,

$$\mathcal{B}_A \subseteq \mathcal{C}_A \subseteq \mathcal{G}_A. \tag{72}$$

Thus the theorems of Brauer and Gershgorin are consequences of Corollary 7.1. The first inclusion in (72) has been shown by Varga [28], the second by Brauer [3]. Note that each of the three sets, $\mathcal{B}_A, \mathcal{C}_A, \mathcal{G}_A$, is closed (as a finite union of closed sets).

Example 7.4. Consider the following matrix and the corresponding incidence graph:

$$A = \begin{bmatrix} 1 & 1.5 & 0 & 0 \\ 0 & \imath & 0.5 & -0.5\imath \\ 0 & 0 & -2 & 2 \\ 0.25 & 0.25\imath & 0 & -2\imath \end{bmatrix}$$



The matrix A can be represented as a sum of the diagonal matrix $A_0 = \mathrm{diag}(1, \imath, -2, -2\imath)$ and the off-diagonal perturbation $R_3 \in \mathcal{I}_{\mathrm{off}}$ defined in Example 4.5. Figure 6 illustrates the eigenvalue inclusion regions for A due to Gershgorin, Brauer, and Brualdi.

The crosses mark the diagonal elements of A. Comparing the right-hand figures in Figures 2 and 6, we see that the inclusion provided by Corollary 4.4(a) is somewhat tighter than the estimate provided by Brualdi's theorem; see (71).

The above example should not convey the impression that Brualdi sets are always considerably smaller then the corresponding Brauer sets. In fact, one can easily see that they are equal if all off-diagonal entries of A are nonzero, i.e., $\mathcal{I}_A \supset \mathcal{I}_{\mathrm{off}}$. For more details on this, see [29, section 2.3].

Although the theorems of Gershgorin, Brauer, and Brualdi follow directly from our main results, we emphasize that the problems underlying the inclusion theorems and those underlying our results are quite different. The inclusion theorems consider the matrix $A = D_A + \Delta_A$ as *given* and establish *upper bounds* for $\sigma(A)$ viewing A as the result of a (known) off-diagonal perturbation of D_A. On the contrary, Corollaries 5.7 and 4.4 provide precise formulae for the union of the spectra of *all* the matrices $A_\Delta = D_A + \Delta$ where Δ is an arbitrary complex matrix of norm $\leq \delta$ with the zero

structure determined by $\mathcal{I}$ (resp., an arbitrary complex matrix satisfying $|\Delta| \le R$). In these corollaries the diagonal matrix D_A, the index set $\mathcal{I}$, and the uncertainty level $\delta > 0$ (resp., the diagonal matrix D_A and the nonnegative matrix R) are the only data. It follows from Corollary 6.6 that

(73)
$$\sigma(A) \subseteq \bigcup_{\substack{\Delta \in \mathbf{\Delta}_{\mathcal{I}_{\text{off}}} \\ \|\Delta\|_1 \le 1}} \sigma(D_A + B\Delta) = \left\{ s \in \mathbb{C}; \min_{1 \le j < k \le m} |s - a_{jj}||s - a_{kk}| \le R_j(A) R_k(A) \right\} = \mathcal{C}_A,$$

where $B = \operatorname{diag}(R_1(A), \dots, R_n(A))$; see (69). Under the assumptions of Brualdi's theorem we have

(74)
$$\sigma(A) \subseteq \bigcup_{\substack{\Delta \in \mathbf{\Delta}_{\mathcal{I}} \\ \|\Delta\|_1 \le 1}} \sigma(D_A + B\Delta) = \bigcup_{(j_1, \dots, j_\ell) \in \mathcal{Z}(\mathcal{I})} \mathcal{B}\left(a_{j_1}, \dots, a_{j_\ell}; \textstyle\prod_{i=1}^{\ell} R_i(A) \right) = \mathcal{B}_A,$$

where $\mathcal{I} := \mathcal{I}_A \cap \mathcal{I}_{\text{off}}$; see the proof of Corollary 7.1, (54), and (70). Hence the upper bounds in the inclusion theorems of Brauer and Brualdi, respectively, are *tight* estimates which cannot be improved if we presuppose as the only a priori knowledge the diagonal of A and the off-diagonal row sums $R_j(A)$ (resp., the diagonal of A, the zero pattern of A, and the off-diagonal row sums $R_j(A)$). To make this more precise we note that

$$\{D_A + B\Delta;\ \Delta \in \mathbf{\Delta}_{\mathcal{I}_{\text{off}}} \text{ and } \|\Delta\|_1 \le 1\}$$

is the set of all matrices $\tilde{A} \in \mathbb{C}^{n \times n}$ with the same diagonal as A and with off-diagonal row sums $R_j(\tilde{A}) \le R_j(A)$. By (73) the Brauer set $\mathcal{C}_A$ is exactly the union of the spectra of all these matrices. Similarly,

$$\{D_A + B\Delta;\ \Delta \in \mathbf{\Delta}_{\mathcal{I}} \text{ and } \|\Delta\|_1 \le 1\}$$

is the set of all matrices $\tilde{A} \in \mathbb{C}^{n \times n}$ with the same diagonal as A, with row sums $R_j(\tilde{A}) \le R_j(A)$, and with $\mathcal{I}_{\tilde{A}} \subset \mathcal{I}_A$. Under the assumptions of Brualdi's theorem it follows from (74) that the Brualdi set $\mathcal{B}_A$ is exactly the union of the spectra of all these matrices. A more detailed discussion on the sharpness of the Brualdi inclusion theorem can be found in [29, section 2.4].

Remark 7.5. In the same way as in the proof of Corollary 5.7 one could derive from Corollaries 5.6 and 6.5 inclusion theorems for the eigenvalues of a block matrix. Such results are obtained in [29] by a different approach from ours.

REFERENCES

[1] G. J. Balas, J. C. Doyle, K. Glover, A. Packard, and R. Smith, *μ-Analysis and Synthesis Toolbox. User's Guide*, The MathWorks, South Natick, MA, 1991.

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.

[3] A. Brauer, *Limits for the characteristic roots of a matrix*. II, Duke Math. J., 14 (1947), pp. 21–26.

[4] R. A. Brualdi, *Matrices, eigenvalues, and directed graphs*, Linear and Multilinear Algebra, 11 (1982), pp. 143–165.

[5] N. A. Bruinsma and M. Steinbuch, *A fast algorithm to compute the H^∞-norm of a transfer function matrix*, Systems Control Lett., 14 (1990), pp. 287–293.

[6] M. J. Chen, K. N. Fan, and C. N. Nett, *Structured singular values with nondiagonal structures. I. Characterizations; Part II. Computation*, IEEE Trans. Automat. Control, 41 (1996), pp. 1507–1516.
[7] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, Proc. IEE-D, 129 (1982), pp. 242–250.
[8] G. E. Dullerud and F. Paganini, *A Course in Robust Control Theory*, Texts Appl. Math. 36, Springer-Verlag, New York, 2000.
[9] E. Gallestey, *Computing spectral value sets using the subharmonicity of the norm of rational matrices*, BIT, 38 (1998), pp. 22–33.
[10] F. R. Gantmacher, *The Theory of Matrices*, Vol. 2, Chelsea, New York, 1959.
[11] N. J. Higham and F. Tisseur, *A block algorithm for matrix* 1*-norm estimation, with an application to* 1*-norm pseudospectra*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1185–1201.
[12] D. Hinrichsen and B. Kelb, *Spectral value sets: A graphical tool for robustness analysis*, Systems Control Lett., 21 (1993), pp. 127–136.
[13] D. Hinrichsen and A. J. Pritchard, *Real and complex stability radii: A survey*, in Control of Uncertain Systems, D. Hinrichsen and B. Mårtensson, eds., Birkhäuser-Verlag, Basel, Switzerland, 1990, pp. 119–162.
[14] D. Hinrichsen and A. J. Pritchard, *On the robustness of stable discrete-time linear systems*, in New Trends in Systems Theory (Genoa, 1990), Birkhäuser-Verlag, Basel, Switzerland, 1991, pp. 393–400.
[15] D. Hinrichsen and A. J. Pritchard, *Mathematical Systems Theory. I. Modelling, State Space Analysis, Stability and Robustness*, Springer-Verlag, Berlin, 2005.
[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
[17] M. Karow, *Geometry of Spectral Value Sets*, Ph.D. thesis, University of Bremen, Bremen, Germany, 2003.
[18] T. C. Lawrence, A. L. Tits, and P. Van Dooren, *A fast algorithm for the computation of an upper bound on the* μ*-norm*, Automatica, 36 (2000), pp. 449–456.
[19] A. Packard and J. C. Doyle, *The complex structured singular value*, Automatica J. IFAC, 29 (1993), pp. 71–109.
[20] L. Qiu, *On the robustness of symmetric systems*, Systems Control Lett., 27 (1996), pp. 187–190.
[21] L. Qiu, B. Bernhardsson, A. Rantzer, E. J. Davison, P. M. Young, and J. C. Doyle, *A formula for computation of the real stability radius*, Automatica J. IFAC, 31 (1995), pp. 879–890.
[22] S. C. Reddy, P. J. Schmid, and D. S. Henningson, *Pseudospectra of the Orr–Sommerfeld operator*, SIAM J. Appl. Math., 53 (1993), pp. 15–47.
[23] M. G. Safonov. *Tight bounds on the response of multivariable systems with component uncertainty*, in 16th Annual Allerton Conference on Communication, Control, and Computing, University of Illinois, Urbana, IL, 1978, pp. 451–460.
[24] M. G. Safonov and K.-H. M. Fan, *Special issue: Multivariable stability margin*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 97–226.
[25] D. D. Šiljak, *Large Scale Dynamic Systems. Stability and Structure*, North–Holland Ser. System Sci. Engrg., North–Holland, New York, 1978.
[26] L. N. Trefethen, *Computation of pseudospectra*, Acta Numer., 8 (1999), pp. 247–295.
[27] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.
[28] R. S. Varga, *Geršgorin-type eigenvalue inclusion theorems and their sharpness*, Electron. Trans. Numer. Anal., 12 (2001), pp. 113–133.
[29] R. S. Varga, *Geršgorin and His Circles*, Springer Ser. Comput. Math. 36, Springer-Verlag, Berlin, 2004.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 885–926

PRACTICAL OUTPUT-FEEDBACK RISK-SENSITIVE CONTROL FOR STOCHASTIC NONLINEAR SYSTEMS WITH STABLE ZERO-DYNAMICS*

YUN-GANG LIU† AND JI-FENG ZHANG‡

Abstract. This paper addresses the design problem of practical (or satisfaction) output-feedback controls for stochastic strict-feedback nonlinear systems in observer canonical form with stable zero-dynamics under long-term average tracking risk-sensitive cost criteria. The cost function adopted here is of the quadratic-integral type usually encountered in practice, rather than the quartic-integral one used to avoid difficulty in control design and performance analysis of the closed-loop system. A sequence of coordinate diffeomorphisms is introduced to separate the zero-dynamics from the entire system, so that the transformed system has an appropriate form suitable for integrator backstepping design. For any given risk-sensitivity parameter and desired cost value, by using the integrator backstepping methodology, an output-feedback control is constructively designed such that (a) the closed-loop system is bounded in probability and (b) the long-term average risk-sensitive cost is upper bounded by the desired value. In addition, this paper does not require the uniform boundedness of the gain functions of the system noise. Furthermore, an example is given to show the effectiveness of the theory.

Key words. nonlinear system, stochastic system, integrator backstepping methodology, risk-sensitive control, output-feedback control, zero dynamics

AMS subject classifications. 93B52, 93C10, 93E03, 93E15, 93E20

DOI. 10.1137/S0363012903439185

1. Introduction.

Research on global stabilization control design for nonlinear systems has been accelerated over the last two decades. After the celebrated characterization of the feedback linearizable systems (see [13]), a breakthrough was achieved with the introduction of the integrator backstepping design methodology (see [20]), which provides a general constructive tool for designing global stabilization controls for nonlinear systems in or feedback equivalent to strict-feedback form. Since the early 1990s, a series of research results on strict-feedback systems have been obtained by using this methodology together with other design tools, such as nonlinear damping, tuning functions, and MT filters (see, e.g., [8], [15], [18], [19], [22], [23], [32], [34], and [38]).

The research on risk-sensitive control can be traced back to the early 1970s, when Jacobson introduced the linear exponential quadratic Gaussian (LEQG) problem (see [14]). Then, Whittle put a risk-sensitivity parameter into the cost, and solved the linear discrete-time problem (see [39]). Bensoussan and van Schuppen considered the continuous-time case in their paper [4]. But the significance of the risk-sensitive control was not fully realized until the 1990s. It has been known that risk-sensitive control is more general than H_∞ control and H_2 control, and closely related to differential game problems (see, e.g., [9], [10], [17], [31], [37], and [40]). For example, when

*Received by the editors December 30, 2003; accepted for publication (in revised form) February 18, 2006; published electronically July 17, 2006. This work was supported by the National Natural Science Foundation of China under grants 60274021, 60221301, 60334040, 60304002, and the Science and Technical Development Plan of Shandong Province under grant 2004GG4204014.

http://www.siam.org/journals/sicon/45-3/43918.html

†School of Control Science and Engineering, Shandong University, Jinan 250061, China (lygfr@sdu.edu.cn).

‡Corresponding author. Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China (jif@iss.ac.cn).

the noise vanishes, the large deviation limit of the risk-sensitive control is nothing but a deterministic differential game problem. These connections have initialized and accelerated the research on stochastic risk-sensitive controls over the last decade.

The design of controls for strict-feedback stochastic nonlinear systems has received intense investigation recently (see, e.g., [1], [5], [6], [7], [11], [26], [27], [28], [29], [33], and [35]), where [7], [11], and [33] considered full state-feedback control design, and [1], [5], [6], [26], [27], [28], [29], and [35] considered output-feedback control design. Under the assumption (A), "the disturbance vector field vanishes at the origin," [5], [7], and [11] studied the problem of designing a control to asymptotically stabilize the closed-loop systems in the large. Meanwhile, [1], [6], [26], [27], [28], [29], [33], and [35] considered the control design to achieve the boundedness in probability of the closed-loop system without using assumption (A). Specifically, [7] considered the disturbance attenuation problem; [35] considered the stabilization problem of systems with stable zero-dynamics; [33], [26], [1], and [29] considered the design of satisfaction control under a quadratic, a quartic regulation, and a quadratic tracking risk-sensitive cost criterion, respectively. [1] used the assumption (B), "the gain functions of stochastic noise are uniformly bounded," while [26], [29], and [33] did not; [27] and [28] considered the reduced-order observer-based stabilization control design of the single-input multioutput stochastic nonlinear systems.

This paper studies the problem of output-feedback control design for a class of stochastic nonlinear systems in observer canonical form with stable zero-dynamics under a quadratic tracking risk-sensitive cost criterion. In general, the design of output-feedback control is more difficult and challenging than that of full state-feedback control. Since the early 1990s, a general framework for studying output-feedback control problems has been developed. The key thought is to first introduce the so-called *information state*, which is a generalization of observer or filter, and then, by a measure transformation, to change the output-feedback control design problem into a full state-feedback problem of an augmented system (see, e.g., [2], [3], [12], [16], and [17]). However, generally speaking, the equality (or inequality) of the information state satisfied is infinite-dimensional, to which an explicit finite-dimensional solution exists only for linear or special nonlinear systems (see [2]). The method of this paper is different from the information state one and can be used to deal with more general inherently nonlinear systems. The objective of this paper is very practical: to search for a *satisfaction* control rather than an optimal one. This makes it possible to avoid the measure transformation. In order to get the explicit formula of the control, strict-feedback nonlinear systems are considered. The main results of this paper indicate that for any given risk-sensitivity parameter and desired tracking risk-sensitive cost value, a dynamic output-feedback control can always be constructively designed so that the closed-loop system is bounded in probability and the long-time average risk-sensitive cost is upper bounded by the desired value. While [1] considered assumption (B) to be essential, the current paper does not use this assumption. In addition, the value range of the characteristic parameter of the value function used for backstepping design is enlarged from $\frac{2}{3}$ (see [26]) to set $(\frac{1}{2}, 1)$. This provides control designers with a freedom in choosing the value function.

The remainder of the paper is organized as follows. Section 2 provides some notation. Section 3 describes the system model and formulates the control objective to be studied. Section 4 describes the constructive design procedure of the control by employing an integrator backstepping approach, and presents several important lemmas for the closed-loop performance analysis. Section 5 addresses the main results

of this paper. Section 6 gives a simulation example to illustrate our theoretical findings. Section 7 gives some concluding remarks. The paper ends with two appendices. Appendix A introduces the definitions of stability notions *asymptotically stable in the large* and *bounded in probability*, and gives a key theorem of sufficient conditions for the solvability of the control problem. Appendix B provides some technical lemmas that play an important role in the control design and performance analysis.

2. Notation. Throughout this paper, $\mathbb{N}$ denotes the set of all natural numbers; $\mathbb{R}$ denotes the set of all real numbers, and $\mathbb{R}^n$ denotes the real n-dimensional space, $n \in \mathbb{N}$; $\mathcal{C}^i$ denotes the set of all functions with continuous partial derivative up to ith order, $i \in \mathbb{N}$, and $\mathcal{C}^\infty$ denotes the set of all smooth functions; for a given vector or matrix W, we use $W^\top$ to denote its transpose; $\mathrm{Tr}(W)$ denotes its trace when W is square, i.e., the sum of all elements on the main diagonal line; we use $|W|$ to denote the absolute value for scalar numbers, and $\|W\|$ to denote the Euclidean norm for vectors and the corresponding induced norm for matrices; we also introduce the Frobenius norm of W defined by $\|W\|_F = \sqrt{\mathrm{Tr}(W^\top W)}$ with properties: $\|W\| \leq \|W\|_F$ and $\|WV\|_F \leq \|W\|\|V\|_F$ for any matrix V with appropriate dimension; for any $x \in \mathbb{R}^n$, x_i denotes its ith element, $x_{[i]}$ denotes the column vector consisting of the first i elements of x in the original order, i.e., $x_{[i]} = [x_1, \ldots, x_i]^\top$; for any given ith continuously differentiable function $y_d(t)$, $y_d^{(i)}(t)$ denotes the ith derivative with respect to the time variable t, the first and second derivatives are denoted by $\dot{y}_d$ and $\ddot{y}_d$, respectively, and $y_d^{[i]}$ denotes the $(i+1)$-dimensional column vector consisting of y_d, $\dot{y}_d, \ldots, y_d^{(i)}$, i.e., $y_d^{[i]} = [y_d, \dot{y}_d, \ddot{y}_d, \ldots, y_d^{(i)}]^\top$. Obviously, $x_{[1]} = x_1$, $x_{[n]} = x$, $y_d^{[0]} = y_d$. $0_{i\times j}$ denotes the $(i \times j)$-dimensional matrix with all zero elements and will be written as 0 for brevity when there is no confusion caused. We use I_i to denote the $i \times i$ identity matrix. For a set A, I_A denotes the characteristic function of the set. For any given symmetric matrix P, $\lambda_{\max}(P)$ and $\lambda_{\min}(P)$ denote its maximum and minimum eigenvalue, respectively.

In addition, when a function shows up for the first time, we will clearly write out its arguments, and then, for simplicity of expression in later use, we sometimes drop the arguments when no confusion is caused.

For a given stochastic system

$$\mathrm{d}x = f(t,x)\,\mathrm{d}t + h(t,x)\,\mathrm{d}w, \quad x(t_0) = x_0,$$

define a differential operator $\mathcal{L}$:

$$\mathcal{L}V(t,\,x) = \frac{\partial V(t,\,x)}{\partial t} + \frac{\partial V(t,\,x)}{\partial x} f(t,\,x) + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V(t,\,x)}{\partial x^2} h(t,\,x)\,(h(t,\,x))^\top\right),$$

where x is an n-dimensional state vector, $n \in \mathbb{N}$; $f : [0,\ \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ and $h : [0,\ \infty) \times \mathbb{R}^n \to \mathbb{R}^{n\times s}$, $s \in \mathbb{N}$, are assumed to be continuous in t and locally Lipschitz in x; w is an s-dimensional vector-valued Brownian motion defined on a probability space $(\Omega,\ \mathcal{F},\ \mathcal{P})$; and $V : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^1$ in t and $\mathcal{C}^2$ in x.

3. Problem formulation.

3.1. System model. We consider the stochastic nonlinear systems in observer canonical form with zero-dynamics of the form (see [35]):

$$
\begin{aligned}
\mathrm{d}x_1 &= x_2\,\mathrm{d}t + f_1(y)\,\mathrm{d}t + h_1(y)\,\mathrm{d}w,\\
&\vdots\\
\mathrm{d}x_{\rho-1} &= x_\rho\,\mathrm{d}t + f_{\rho-1}(y)\,\mathrm{d}t + h_{\rho-1}(y)\,\mathrm{d}w,\\
\mathrm{d}x_\rho &= x_{\rho+1}\,\mathrm{d}t + f_\rho(y)\,\mathrm{d}t + b_m g(y)u\,\mathrm{d}t + h_\rho(y)\,\mathrm{d}w,\\
&\vdots\\
\mathrm{d}x_{n-1} &= x_n\,\mathrm{d}t + f_{n-1}(y)\,\mathrm{d}t + b_1 g(y)u\,\mathrm{d}t + h_{n-1}(y)\,\mathrm{d}w,\\
\mathrm{d}x_n &= f_n(y)\,\mathrm{d}t + b_0 g(y)u\,\mathrm{d}t + h_n(y)\,\mathrm{d}w,\\
y &= x_1,
\end{aligned}
\tag{3.1}
$$

where $x = [x_1, \ldots, x_n]^\top$ is the n-dimensional state vector, $n \in \mathbb{N}$, and its initial value $x(t_0) = x_0$ is fixed but unknown; u is the scalar control input; y is the scalar measurable output; $f_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$, are the system nonlinearities depending only on output y; $h_i : \mathbb{R} \to \mathbb{R}^{1\times s}$, $i = 1, \ldots, n$, are the gain functions of the system noise depending only on y, $s \in \mathbb{N}$; $g : \mathbb{R} \to \mathbb{R}$ is the nonlinear gain function of the control input u depending only on y; $w \in \mathbb{R}^s$ is a vector-valued standard Brownian motion defined on probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with Ω being a sample space, $\mathcal{F}$ being a filtration, and $\mathcal{P}$ being the probability measure, $s \in \mathbb{N}$; $m \in \mathbb{N}$ satisfies $0 \le m < n$; and $\rho = n - m \in \mathbb{N}$ is the relative degree of the system.

The main results of this paper are based on the following assumptions:

A1. The nonlinear functions f_i and h_i $(i = 1, \ldots, n)$ are smooth. That is, $f_i \in \mathcal{C}^\infty$ and $h_i \in \mathcal{C}^\infty$; the nonlinear function g is continuous; and, for any $y \in \mathbb{R}$, $g(y) \neq 0$.

A2. All the roots of the polynomial $b_m s^m + \cdots + b_1 s + b_0$, $b_m \neq 0$, have negative real parts.

A3. Desired system output y_d is deterministic, and it and its derivatives $\dot{y}_d, \ldots, y_d^{(\rho)}$ are known and bounded; i.e., there exist known positive constants $C_{y_d^{(i)}}$, $i = 0, \ldots, \rho$, that bound the reference trajectory y_d and its derivatives.

Assumption A1 is standard for this class of control problems, to ensure that f_i and h_i $(i = 1, \ldots, n)$ are local Lipschitz functions and, together with assumption A3, to ensure the global boundedness of $h_i(y_d)$ $(i = 1, \ldots, n)$. Assumption A2 ensures that the zero-dynamics are stable.

Unlike the problem of feedback stabilization, there is no need to require that the origin $x = \mathbf{0}_{n\times 1}$ be the equilibrium point of the open-loop system. This is because the purpose of the tracking control is to make the system output conform to the time-varying desired system output $y_d(t)$, rather than to steer the system state to the origin $x = \mathbf{0}_{n\times 1}$.

3.2. Control objective. The goal of control design is to make the solution process of the system (3.1) be bounded in probability and the following quadratic tracking risk-sensitive cost criterion achieve a predefined long-term cost value:

$$
J_\theta(y) = \limsup_{T\to\infty} \frac{1}{T}\frac{2}{\theta}\ln\left(E\left(\exp\left(\frac{\theta}{2}\int_0^T (y - y_d)^2\,\mathrm{d}t\right)\right)\right). \tag{3.2}
$$

That is, for any given positive cost value R_l (arbitrarily close to zero), the risk-sensitive cost $J_\theta(y)$ is not greater than R_l, where θ is called the risk-sensitivity parameter and $y - y_d$ is called the output tracking error. When $\theta > 0$, the cost function weights

heavily the large deviation of $y - y_d$ through the exponential operator, which leads to a risk-averse control design problem. The greater the value of θ, the more conservative is the controller. Actually, by the value of θ, the risk-sensitive problem can be classified (see [10] and [31]) as follows: (i) when $\theta > 0$, it is a risk-averse problem; (ii) when $\theta < 0$, it is a risk-seeking problem; (iii) when $\theta \to 0$, the cost function converges to a standard integral cost, and so it is known as a risk-neutral problem.

In this paper, we will study only the case where θ is positive.

For convenience of expression, we give the following definition.

DEFINITION 3.1. *For a given positive risk-sensitivity parameter θ, a controller u is said to achieve a guaranteed risk-sensitive cost R_l ($R_l > 0$) if the following inequality holds for the output of the closed-loop system:*

$$J_\theta(y) \le R_l.$$

In addition to the purposes of cost upper bound, we are also interested in achieving *boundedness in probability* for the closed-loop system. This notion, together with the *asymptotical stability in the large*, was introduced in the classical book [21] and has now been widely used. For the sake of the self-containedness of this paper, we will restate these two notions in Appendix A.

The system (3.1) can be rewritten into the following compact form:

$$\mathrm{d}x = f(x)\,\mathrm{d}t + Bg(x)u\,\mathrm{d}t + h(x)\,\mathrm{d}w, \tag{3.3}$$

where

$$f(x) = \begin{bmatrix} x_2 + f_1(x_1) \\ \vdots \\ x_n + f_{n-1}(x_1) \\ f_n(x_1) \end{bmatrix}, \qquad g(x) = g(x_1),$$

$$B = \begin{bmatrix} \mathbf{0}_{(n-m-1)\times 1} \\ b_m \\ \vdots \\ b_0 \end{bmatrix}, \qquad h(x) = \begin{bmatrix} h_1(x_1) \\ h_2(x_1) \\ \vdots \\ h_n(x_1) \end{bmatrix}.$$

If $\rho = 1$, then $m = n - \rho = n - 1$. For this special case, vector B defined above is simply $[b_{n-1}, \ldots, b_1, b_0]^\top$.

For tracking purposes, the controller to be designed is time-varying in general, and so is the resulting closed-loop system, even though the original system is not. Thus, as in [26] and [33], with the long-term average risk-sensitive cost criterion $J_\theta(y)$, for a given desired cost value $R_l > 0$, a practical risk-sensitive output-feedback tracking control is designed as

$$\begin{cases} \dot{\xi} = \alpha\,(t,\, \xi,\, y)\,, & \alpha \in \mathcal{C}^1, \\ u = \mu\,(t,\, \xi,\, y)\,, & \mu \in \mathcal{C}^1, \end{cases} \tag{3.4}$$

so that there exists a nonnegative value function $V(t, x, \xi)$, which is $\mathcal{C}^1$ in t and $\mathcal{C}^2$ in (x, ξ) and radially unbounded with respect to x and ξ, satisfying the following Hamilton–Jacobi–Bellman (HJB) inequality:

$$\begin{aligned} \frac{\partial V}{\partial t} + \begin{bmatrix} \dfrac{\partial V}{\partial x} & \dfrac{\partial V}{\partial \xi} \end{bmatrix} \begin{bmatrix} f + Bg\mu \\ \alpha \end{bmatrix} + \frac{\theta}{4} \frac{\partial V}{\partial x} hh^\top \left(\frac{\partial V}{\partial x} \right)^\top \\ + \frac{1}{2} \mathrm{Tr} \left(\frac{\partial^2 V}{\partial x^2} hh^\top \right) + (y - y_d)^2 \;\le\; R_l. \end{aligned} \tag{3.5}$$

From (3.5), it is easy to see that the essential difference between the stochastic HJB and deterministic HJB equations is that the former has the Itô term $\frac{1}{2}\mathrm{Tr}(\frac{\partial^2 V}{\partial x^2}hh^\top)$. How to deal with this term is the key to the control design and performance analysis.

4. Output-feedback risk-sensitive control design. We shall design the output-feedback tracking controller in three steps. First, we introduce an observer to rebuild the system states. With the observer dynamics in the loop, we introduce a sequence of coordinate diffeomorphisms transforming the system into a lower triangular structure which is amenable to the application of integrator backstepping methodology. Then, we describe the control design procedure and present several lemmas, which will be used for the performance analysis of the closed-loop systems in the next section.

4.1. Observer design. Since the states of (3.1), except for the state x_1 which can be obtained directly since $y = x_1$, are unknown and need an observer to rebuild,

$$\begin{aligned}
\dot{\widehat{x}}_1 &= \widehat{x}_2 + k_1(y - \widehat{x}_1) + f_1(y),\\
&\vdots\\
\dot{\widehat{x}}_{\rho-1} &= \widehat{x}_\rho + k_{\rho-1}(y - \widehat{x}_1) + f_{\rho-1}(y),\\
\dot{\widehat{x}}_\rho &= \widehat{x}_{\rho+1} + k_\rho(y - \widehat{x}_1) + f_\rho(y) + b_m g(y)u,\\
&\vdots\\
\dot{\widehat{x}}_{n-1} &= \widehat{x}_n + k_{n-1}(y - \widehat{x}_1) + f_{n-1}(y) + b_1 g(y)u,\\
\dot{\widehat{x}}_n &= k_n(y - \widehat{x}_1) + f_n(y) + b_0 g(y)u,
\end{aligned} \tag{4.1}$$

where $k_1, k_2, \ldots, k_n$ are design constants such that all the roots of polynomial $s^n + k_1 s^{n-1} + \cdots + k_n$ have negative real parts. The initial condition for observer (4.1) is set by certain value $\widehat{x}(t_0) = \widehat{x}_0$.

Let $\widehat{x} = [\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n]^\top$. Both system output y and observer state vector $\widehat{x}$ are available for control design. Denote the state estimation error as $\widetilde{x} = x - \widehat{x}$. Then we have

$$\mathrm{d}\widetilde{x} = \begin{bmatrix} -k_1 & & & \\ \vdots & & I_{n-1} & \\ -k_n & 0 & \cdots & 0 \end{bmatrix} \widetilde{x}\,\mathrm{d}t + h(y)\,\mathrm{d}w \triangleq A\widetilde{x}\,\mathrm{d}t + h(y)\,\mathrm{d}w. \tag{4.2}$$

Thus, with observer dynamics (4.1) in the loop, we have the following entire system:

$$\begin{aligned}
\mathrm{d}\widetilde{x} &= A\widetilde{x}\,\mathrm{d}t + h(y)\,\mathrm{d}w,\\
\mathrm{d}y &= (\widehat{x}_2 + \widetilde{x}_2)\,\mathrm{d}t + f_1(y)\,\mathrm{d}t + h_1(y)\,\mathrm{d}w,\\
\dot{\widehat{x}}_2 &= \widehat{x}_3 + k_2(y - \widehat{x}_1) + f_2(y),\\
&\vdots\\
\dot{\widehat{x}}_{\rho-1} &= \widehat{x}_\rho + k_{\rho-1}(y - \widehat{x}_1) + f_{\rho-1}(y),\\
\dot{\widehat{x}}_\rho &= \widehat{x}_{\rho+1} + k_\rho(y - \widehat{x}_1) + f_\rho(y) + b_m g(y)u,\\
&\vdots\\
\dot{\widehat{x}}_{n-1} &= \widehat{x}_n + k_{n-1}(y - \widehat{x}_1) + f_{n-1}(y) + b_1 g(y)u,\\
\dot{\widehat{x}}_n &= k_n(y - \widehat{x}_1) + f_n(y) + b_0 g(y)u.
\end{aligned} \tag{4.3}$$

System (4.3) has three parts, which are associated with the estimation error $\widetilde{x}$, system output y, and observer states $\widehat{x}_2, \ldots, \widehat{x}_\rho$, respectively. In particular, when $\rho = 1$, then $m = n - \rho = n - 1$. For this case, the second subequation of (4.3) shall be replaced by the following equation:

$$\mathrm{d}y = (\widehat{x}_2 + \widetilde{x}_2)\,\mathrm{d}t + f_1(y)\,\mathrm{d}t + b_{n-1}g(y)u\mathrm{d}t + h_1(y)\,\mathrm{d}w.$$

In the next subsections, with this entire system as starting point, we shall search for the desired controller.

4.2. Coordinate diffeomorphisms. To prepare for the backstepping design in the next subsection, we introduce a series of ρ coordinate diffeomorphisms (see [36]) so as to convert the entire system (4.3) into zero-dynamics canonical form, which is amenable to the application of integrator backstepping methodology.

The idea of such coordinate diffeomorphisms was first introduced in [30] and significantly modified in Chapter 8 of [24]. Our presentation, including the two cases of $\rho = 1$ and $\rho > 1$, is much more direct and easier to implement.

4.2.1. Case of $\boldsymbol{\rho = 1}$. When $\rho = 1$, $m = n-\rho = n-1$. This means that control input appears in every subequation of (3.1) and (4.1). In this case, one coordinate transformation is sufficient to obtain the desired structure.

Let $\varsigma_0 = [y, \widehat{x}_2, \ldots, \widehat{x}_n]^\top$. Then, by (4.3), we have the following dynamics for ς_0:

$$\mathrm{d}\varsigma_0 = D_0\varsigma_0\,\mathrm{d}t + G_0(y, \widetilde{x}_1)\,\mathrm{d}t + [1,\, \mathbf{0}_{1\times(n-1)}]^\top \widetilde{x}_2\,\mathrm{d}t + g(y)B_0u\,\mathrm{d}t + H_0(y)\,\mathrm{d}w, \tag{4.4}$$

where

$$\begin{aligned}
D_0 &= \begin{bmatrix} 0 & & \\ \vdots & & I_{n-1} \\ 0 & 0 & \cdots\ 0 \end{bmatrix},\\
G_0 &= [f_1(y),\, f_2(y) + k_2\widetilde{x}_1, \ldots, f_n(y) + k_n\widetilde{x}_1]^\top,\\
B_0 &= [b_{n-1}, \ldots, b_1, b_0]^\top,\\
H_0 &= [(h_1(y))^\top,\, \mathbf{0}_{s\times(n-1)}]^\top.
\end{aligned}$$

By coordinate transformation we would like to transform the vector B_0 into one with all elements being zero except the first element, b_{n-1}. Let $\varsigma_1 = T_1\varsigma_0$, where T_1 is the same as I_n, except with the first column replaced by $[1, -b_{n-2}/b_{n-1}, \ldots, -b_0/b_{n-1}]^\top$. Then, T_1^{-1} is also the same as I_n, except with the first column replaced by $\left[1, \frac{b_{n-2}}{b_{n-1}}, \ldots, \frac{b_0}{b_{n-1}}\right]^\top$.

Then we have

$$\mathrm{d}\varsigma_1 = D_1\varsigma_1\,\mathrm{d}t + G_1(y, \widetilde{x}_1)\,\mathrm{d}t + [1, L_2^\top]^\top \widetilde{x}_2\,\mathrm{d}t + g(y)B_1u\,\mathrm{d}t + H_1(y)\,\mathrm{d}w,$$

where $D_1 = T_1D_0T_1^{-1}$ is the same as D_0, except with the first and second columns replaced by $[d_{11}, \ldots, d_{1,n}]^\top$ and $[1, -b_{n-2}/b_{n-1}, \ldots, -b_0/b_{n-1}]^\top$, respectively:

$$\begin{aligned}
G_1 &= T_1G_0(y, \widetilde{x}_1) \triangleq \left[\overline{g}_{11}(y) + \overline{d}_{11}\widetilde{x}_1, \ldots, \overline{g}_{1n}(y) + \overline{d}_{1n}\widetilde{x}_1\right]^\top,\\
L_2 &= [-b_{n-2}/b_{n-1}, \ldots, -b_1/b_{n-1},\, -b_0/b_{n-1}]^\top,\\
B_1 &= T_1B_0 = [b_{n-1},\, \mathbf{0}_{1\times m}]^\top,\\
H_1 &= T_1H_0(y) \triangleq \left[(\widehat{h}_1(y))^\top, \ldots, (\widehat{h}_n(y))^\top\right]^\top.
\end{aligned}$$

Denote $\eta = \varsigma_1 = [\eta_1, \dots, \eta_n]^\top$ and $\zeta = [\eta_2, \eta_3, \dots, \eta_n]^\top$. Then, the dynamics of $\widetilde{x}$, ζ, and η_1 can be expressed as follows:

$$
\begin{aligned}
\mathrm{d}\widetilde{x} &= A\widetilde{x}\,\mathrm{d}t + h(y)\,\mathrm{d}w,\\
\mathrm{d}\zeta &= \begin{bmatrix} -\frac{b_{n-2}}{b_{n-1}} & & \\ \vdots & & I_{n-2} \\ -\frac{b_1}{b_{n-1}} & & \\ -\frac{b_0}{b_{n-1}} & 0 \quad \cdots & 0 \end{bmatrix} \zeta\,\mathrm{d}t + \begin{bmatrix} d_{12} \\ d_{13} \\ \vdots \\ d_{\rho n} \end{bmatrix} y\,\mathrm{d}t \\
&\quad + \begin{bmatrix} \overline{g}_{12}(y) + \overline{d}_{12}\widetilde{x}_1 \\ \overline{g}_{13}(y) + \overline{d}_{13}\widetilde{x}_1 \\ \vdots \\ \overline{g}_{1n}(y) + \overline{d}_{1n}\widetilde{x}_1 \end{bmatrix} \mathrm{d}t - \frac{1}{b_{n-1}} \begin{bmatrix} b_{n-2} \\ \vdots \\ b_1 \\ b_0 \end{bmatrix} \widetilde{x}_2 \mathrm{d}t + \begin{bmatrix} \widehat{h}_2(y) \\ \widehat{h}_3(y) \\ \vdots \\ \widehat{h}_n(y) \end{bmatrix} \mathrm{d}w \qquad (4.5)\\
&\triangleq E\zeta\,\mathrm{d}t + L_1\widetilde{x}_1\,\mathrm{d}t + L_2\widetilde{x}_2\,\mathrm{d}t + G(y)\,\mathrm{d}t + \Psi(y)\,\mathrm{d}w,\\
\mathrm{d}\eta_1 &= \left[1,\, \mathbf{0}_{1\times(n-2)}\right]\zeta\,\mathrm{d}t + (\overline{g}_{11}(y) + d_{11}y)\,\mathrm{d}t\\
&\quad + \overline{d}_{11}\widetilde{x}_1\,\mathrm{d}t + \widetilde{x}_2\,\mathrm{d}t + b_{n-1}g(y)u\,\mathrm{d}t + \widehat{h}_1(y)\,\mathrm{d}w,
\end{aligned}
$$

where

$$
\begin{aligned}
y &= \eta_1,\\
L_1 &= [\overline{d}_{12},\, \overline{d}_{13}, \dots, \overline{d}_{1n}]^\top,\\
G &= [\overline{g}_{12}(y) + d_{12}y,\, \overline{g}_{13}(y) + d_{13}y, \dots, \overline{g}_{1n}(y) + d_{1n}y]^\top.
\end{aligned}
$$

This system is equivalent to the entire system (4.3) under the transformation $[\eta_1,\, \zeta^\top]^\top = T_1[y, \widehat{x}_2, \dots, \widehat{x}_n]^\top$. The structure of (4.5) makes the design of an output feedback controller much easier (see the latter design procedure for details).

4.2.2. Case of $\rho > 1$. Let us now give the coordinate transformations for the case of $\rho > 1$. From the ρ transformations below one can see that there exist some essential differences between this case and the case of $\rho = 1$.

Let $\overline{g}_{01}(y) = f_1(y)$, $\overline{d}_{0i} = k_i$, and $\overline{g}_{0i}(y) = f_i(y)$ $(i = 2, \dots, n)$. Then we have the following dynamics for $\varsigma_0 = [y, \widehat{x}_2, \dots, \widehat{x}_n]^\top$:

$$
\mathrm{d}\varsigma_0 = D_0\varsigma_0\,\mathrm{d}t + G_0(y, \widetilde{x}_1)\,\mathrm{d}t + [1,\, \mathbf{0}_{1\times(n-1)}]^\top\widetilde{x}_2\,\mathrm{d}t + g(y)B_0u\,\mathrm{d}t + H_0(y)\,\mathrm{d}w,
$$

where matrix D_0 and function H_0 are the same as those of (4.4), and

$$
\begin{aligned}
G_0 &= [f_1(y),\, f_2(y) + k_2\widetilde{x}_1, \dots, f_n(y) + k_n\widetilde{x}_1]^\top\\
&\triangleq [\overline{g}_{01}(y) + \overline{d}_{01}\widetilde{x}_1,\, \overline{g}_{02}(y) + \overline{d}_{02}\widetilde{x}_1, \dots, \overline{g}_{0n}(y) + \overline{d}_{0n}\widetilde{x}_1]^\top\\
&\triangleq [g_{01}(y, \widetilde{x}_1),\, g_{02}(y, \widetilde{x}_1), \dots, g_{0n}(y, \widetilde{x}_1)]^\top,\\
B_0 &= [\mathbf{0}_{1\times(\rho-1)}, b_m, b_{m-1}, \dots, b_0]^\top.
\end{aligned}
$$

By the first transformation, we would like to transform the matrix B_0 into one with all elements being zero except the ρth element, b_m. Let $\varsigma_1 = T_1\varsigma_0$, where T_1 is the same as I_n, except with the ρth column replaced by $\left[\mathbf{0}_{1\times(\rho-1)}, 1, -\frac{b_{m-1}}{b_m}, \dots, -\frac{b_0}{b_m}\right]^\top$.

Then T_1^{-1} is also the same as I_n, except with the ρth column replaced by $\big[\mathbf{0}_{1\times(\rho-1)}, 1, -\frac{b_{m-1}}{b_m}, \ldots, -\frac{b_0}{b_m}\big]^\top$.

Then we have

$$\mathrm{d}\varsigma_1 = D_1\varsigma_1\,\mathrm{d}t + G_1(y,\widetilde{x}_1)\,\mathrm{d}t + \big[1, \mathbf{0}_{1\times(n-1)}\big]^\top \widetilde{x}_2\,\mathrm{d}t + g(y)B_1u\,\mathrm{d}t + H_1(y)\,\mathrm{d}w,$$

where $D_1 = T_1D_0T_1^{-1}$ is the same as D_0 except with the ρth and $(\rho+1)$st columns replaced by $[\mathbf{0}_{1\times(\rho-2)}, 1, d_{11}, \ldots, d_{1,m+1}]^\top$ and $[\mathbf{0}_{1\times(\rho-1)}, 1, -b_{m-1}/b_m, \ldots, -b_0/b_m]^\top$, respectively,

$$\begin{aligned}
G_1 &= T_1G_0(y,\widetilde{x}_1)\\
&\triangleq [\overline{g}_{11}(y) + \overline{d}_{11}\widetilde{x}_1, \ldots, \overline{g}_{1n}(y) + \overline{d}_{1n}\widetilde{x}_1]^\top\\
&\triangleq [g_{11}(y,\widetilde{x}_1), \ldots, g_{1n}(y,\widetilde{x}_1)]^\top,\\
B_1 &= T_1B_0 = [\mathbf{0}_{1\times(\rho-1)},\, b_m,\, \mathbf{0}_{1\times m}]^\top,\\
H_1 &= T_1H_0(y) = H_0(y).
\end{aligned}$$

By the ith $(i = 2, \ldots, \rho-1)$ transformation, we would like to transform the $(\rho-i+2)$nd column of the matrix D_{i-1} into the $(\rho-i+1)$st unit vector. Let $\varsigma_i = T_i\varsigma_{i-1}$, where T_i is the same as I_n except with the $(\rho-i+1)$st column replaced by $[\mathbf{0}_{1\times(\rho-i)}, 1, -d_{i-1,1}, \ldots, -d_{i-1,m+i-1}]^\top$. Then, T_i^{-1} is also the same as I_n except with the $(\rho-i+1)$st column replaced by $[\mathbf{0}_{1\times(\rho-i)}, 1, d_{i-1,1}, \ldots, d_{i-1,m+i-1}]^\top$.

This leads to

$$\mathrm{d}\varsigma_i = D_i\varsigma_i\,\mathrm{d}t + G_i(y,\widetilde{x}_1)\,\mathrm{d}t + [1, \mathbf{0}_{1\times(n-1)}]^\top\widetilde{x}_2\,\mathrm{d}t + g(y)B_iu\,\mathrm{d}t + H_i(y)\,\mathrm{d}w,$$

where $D_i = T_iD_{i-1}T_i^{-1}$ is the same as D_0 except with the $(\rho-i+1)$st and $(\rho+1)$st columns replaced by $[\mathbf{0}_{1\times(\rho-i-1)}, 1, d_{i1}, \ldots, d_{i,i+m}]^\top$ and $[\mathbf{0}_{1\times(\rho-1)}, 1, -b_{m-1}/b_m, \ldots, -b_0/b_m]^\top$, respectively,

$$\begin{aligned}
G_i &= T_iG_{i-1}(y,\widetilde{x}_1)\\
&\triangleq [\overline{g}_{i1}(y) + \overline{d}_{i1}\widetilde{x}_1, \ldots, \overline{g}_{in}(y) + \overline{d}_{in}\widetilde{x}_1]^\top\\
&\triangleq [g_{i1}(y,\widetilde{x}_1), \ldots, g_{in}(y,\widetilde{x}_1)]^\top,\\
B_i &= T_iB_{i-1} = B_1 = [\mathbf{0}_{1\times(\rho-1)},\, b_m,\, \mathbf{0}_{1\times m}]^\top,\\
H_i &= T_iH_{i-1}(y) = H_0(y).
\end{aligned}$$

Finally, by the last transformation, we would like to transform the second column of the matrix $D_{\rho-1}$ into the first unit vector. Let $\varsigma_\rho = T_\rho\varsigma_{\rho-1}$, where T_ρ is the same as I_n except with the first column replaced by $[1, -d_{\rho-1,1}, \ldots, -d_{\rho-1,n-1}]^\top$. Then, T_ρ^{-1} is also the same as I_n except with the first column replaced by $[1, d_{\rho-1,1}, \ldots, d_{\rho-1,n-1}]^\top$.

This leads to

$$\begin{aligned}
\mathrm{d}\varsigma_\rho = {} & D_\rho\varsigma_\rho\,\mathrm{d}t + G_\rho(y,\widetilde{x}_1)\,\mathrm{d}t + [1, -d_{\rho-1,1}, \ldots, -d_{\rho-1,n-1}]^\top\widetilde{x}_2\,\mathrm{d}t\\
& + g(y)B_\rho u\,\mathrm{d}t + H_\rho(y)\,\mathrm{d}w,
\end{aligned}$$

where $D_\rho = T_\rho D_{\rho-1} T_\rho^{-1}$ is the same as D_0 except with the first and $(\rho+1)$st column replaced by $[d_{\rho 1}, \dots, d_{\rho n}]^\top$ and $[\mathbf{0}_{1\times(\rho-1)}, 1, -b_{m-1}/b_m, \dots, -b_0/b_m]^\top$, respectively,

$$\begin{aligned}
G_\rho &= T_\rho G_{\rho-1}(y, \widetilde{x}_1) \\
&\triangleq \left[\overline{g}_{\rho 1}(y) + \overline{d}_{\rho 1}\widetilde{x}_1, \dots, \overline{g}_{\rho n}(y) + \overline{d}_{\rho n}\widetilde{x}_1\right]^\top \\
&\triangleq \left[g_{\rho 1}(y, \widetilde{x}_1), \dots, g_{\rho n}(y, \widetilde{x}_1)\right]^\top, \\
B_\rho &= T_\rho B_{\rho-1} = B_1, \\
H_\rho &= T_\rho H_{\rho-1} = T_\rho H_0 \triangleq \left[(\widehat{h}_1(y))^\top, \dots, (\widehat{h}_n(y))^\top\right]^\top.
\end{aligned}$$

Denote $\eta = \varsigma_\rho = [\eta_1, \dots, \eta_n]^\top$ and $\zeta = [\eta_{\rho+1}, \dots, \eta_n]^\top$. Then $\eta_1 = y$, and the dynamics of estimation error, the zero-dynamics of ζ, and the lower triangular form for the dynamics of $\eta_1, \dots, \eta_\rho$ can be expressed as follows:

$$\begin{aligned}
\mathrm{d}\widetilde{x} &= A\widetilde{x}\,\mathrm{d}t + h(y)\,\mathrm{d}w, \\
\mathrm{d}\zeta &= \begin{bmatrix} -\frac{b_{m-1}}{b_m} & & \\ \vdots & & I_{m-1} \\ -\frac{b_1}{b_m} & & \\ -\frac{b_0}{b_m} & 0 \quad \cdots \quad 0 & \end{bmatrix} \zeta\,\mathrm{d}t + \begin{bmatrix} d_{\rho,\rho+1} \\ \vdots \\ d_{\rho n} \end{bmatrix} y\,\mathrm{d}t \\
&\quad + \begin{bmatrix} -d_{\rho-1,\rho} \\ \vdots \\ -d_{\rho-1,n-1} \end{bmatrix} \widetilde{x}_2\,\mathrm{d}t + \begin{bmatrix} g_{\rho,\rho+1}(y, \widetilde{x}_1) \\ \vdots \\ g_{\rho n}(y, \widetilde{x}_1) \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \widehat{h}_{\rho+1}(y) \\ \vdots \\ \widehat{h}_n(y) \end{bmatrix} \mathrm{d}w \qquad (4.6) \\
&\triangleq E\zeta\,\mathrm{d}t + L_1\widetilde{x}_1\,\mathrm{d}t + L_2\widetilde{x}_2\,\mathrm{d}t + G(y)\,\mathrm{d}t + \Psi(y)\,\mathrm{d}w, \\
\mathrm{d}\eta_1 &= d_{\rho 1} y\,\mathrm{d}t + (\eta_2 + \widetilde{x}_2)\,\mathrm{d}t + g_{\rho 1}(y, \widetilde{x}_1)\,\mathrm{d}t + \widehat{h}_1(y)\,\mathrm{d}w, \\
\mathrm{d}\eta_2 &= [d_{\rho 2} y + \eta_3 + g_{\rho 2}(y, \widetilde{x}_1) - d_{\rho-1,1}\widetilde{x}_2]\,\mathrm{d}t + \widehat{h}_2(y)\,\mathrm{d}w, \\
&\vdots \\
\mathrm{d}\eta_{\rho-1} &= [d_{\rho,\rho-1} y + \eta_\rho + g_{\rho,\rho-1}(y, \widetilde{x}_1) - d_{\rho-1,\rho-2}\widetilde{x}_2]\,\mathrm{d}t + \widehat{h}_{\rho-1}(y)\,\mathrm{d}w, \\
\mathrm{d}\eta_\rho &= \left[1, \mathbf{0}_{1\times(m-1)}\right]\zeta\,\mathrm{d}t + d_{\rho\rho} y\,\mathrm{d}t + g_{\rho\rho}(y, \widetilde{x}_1)\,\mathrm{d}t \\
&\quad + b_m g(y) u\,\mathrm{d}t - d_{\rho-1,\rho-1}\widetilde{x}_2\,\mathrm{d}t + \widehat{h}_\rho(y)\,\mathrm{d}w,
\end{aligned}$$

where

$$\begin{aligned}
y &= \eta_1, \\
L_1 &= [\overline{d}_{\rho,\rho+1}, \dots, \overline{d}_{\rho n}]^\top, \\
G &= [\overline{g}_{\rho,\rho+1}(y) + d_{\rho,\rho+1} y, \dots, \overline{g}_{\rho,n}(y) + d_{\rho n} y]^\top.
\end{aligned}$$

This system is equivalent to the entire system (4.3) under the transformation $[\eta_1, \dots, \eta_\rho, \zeta^\top]^\top = T_\rho \cdots T_1 [y, \widehat{x}_2, \dots, \widehat{x}_n]^\top$. The structure of (4.6) allows the design of an output feedback controller by using integrator backstepping methodology.

4.3. Control design procedure. We now start to design the desired controller with the estimation error $\widetilde{x}$ and the zero-dynamics ζ (given by (4.5) for the case of

$\rho = 1$ and (4.6) for the case of $\rho > 1$). To do so, let $\chi = [\zeta^\top,\ \widetilde{x}^\top]^\top \in \mathbb{R}^{n+m}$. Then for both $\rho = 1$ and $\rho > 1$ we have

$$\begin{aligned} \mathrm{d}\chi &= \begin{bmatrix} E & L \\ \mathbf{0}_{n\times m} & A \end{bmatrix} \chi\,\mathrm{d}t + \begin{bmatrix} G(y) \\ \mathbf{0}_{n\times 1} \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \Psi(y) \\ h(y) \end{bmatrix} \mathrm{d}w \\ &\triangleq W\chi\,\mathrm{d}t + F(y)\,\mathrm{d}t + \Phi(y)\,\mathrm{d}w, \end{aligned} \tag{4.7}$$

where $L = [L_1,\ L_2,\ \mathbf{0}_{m\times(n-2)}]$, and F and Φ are $\mathcal{C}^\infty$.

For the objective of a tight controller, the dynamics χ would be partitioned as $\chi = [\chi_a^\top,\ \chi_b^\top]^\top$, where $\chi_a = [\zeta^\top,\ \widetilde{x}_1]^\top \in \mathbb{R}^{m+1}$ is available for feedback design, while $\chi_b = [\widetilde{x}_2, \ldots, \widetilde{x}_n]^\top \in \mathbb{R}^{n-1}$ is not. Furthermore, χ_a and χ_b satisfy the following stochastic differential equations, respectively:

$$\begin{aligned} \mathrm{d}\chi_a &= \begin{bmatrix} E & L_1 \\ \mathbf{0}_{1\times m} & -k_1 \end{bmatrix} \chi_a\,\mathrm{d}t + \begin{bmatrix} L_2 & \mathbf{0}_{m\times(n-2)} \\ 1 & \mathbf{0}_{1\times(n-2)} \end{bmatrix} \chi_b\,\mathrm{d}t \\ &\quad + \begin{bmatrix} G(y) \\ 0 \end{bmatrix} \mathrm{d}t + \begin{bmatrix} \Psi(y) \\ h_1(y) \end{bmatrix} \mathrm{d}w \\ &\triangleq W_a\chi_a\,\mathrm{d}t + L_a\chi_b\,\mathrm{d}t + F_a(y)\,\mathrm{d}t + \Phi_a(y)\,\mathrm{d}w, \end{aligned} \tag{4.8a}$$

$$\begin{aligned} \mathrm{d}\chi_b &= \begin{bmatrix} 0 & & \\ \vdots & I_{n-2} & \\ 0 & 0 \cdots & 0 \end{bmatrix} \chi_b\,\mathrm{d}t + \begin{bmatrix} 0 & \cdots & 0 & -k_2 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & -k_n \end{bmatrix} \chi_a\,\mathrm{d}t \\ &\quad + \begin{bmatrix} h_2(y) \\ \vdots \\ h_n(y) \end{bmatrix} \mathrm{d}w \\ &\triangleq W_b\chi_b\,\mathrm{d}t + L_b\chi_a\,\mathrm{d}t + \Phi_b(y)\,\mathrm{d}w, \end{aligned} \tag{4.8b}$$

where F_a, Φ_a, and Φ_b are $\mathcal{C}^\infty$.

Remark 4.1. From subsection 4.2 we know that E, G, L, and Ψ in (4.7)–(4.8) are differently defined with respect to $\rho = 1$ and $\rho > 1$, respectively, and so are W, W_a, W_b, F, F_a, F_b, L_a, L_b, and Φ_a. Thus, for the sake of the unambiguousness, these two cases will be separately handled below.

We are now in a position to develop a recursive construction procedure for the desired risk-sensitive controller.

4.3.1. Initial assignment. First, we present the initial assignment for the entire design procedure.

By assumption A2, we know that matrix E is Hurwitz. This, together with the Hurwitz property of matrix A, implies that W is also Hurwitz. Therefore, there exists a symmetric and positive definite matrix P such that

$$W^\top P + PW = -I_{n+m}. \tag{4.9}$$

We introduce a value function (or *Lyapunov* function) for the χ system:

$$V_0(\chi) = \phi(\xi(\chi)) = \delta(c + \xi(\chi))^\gamma - \delta c^\gamma, \quad \xi = \chi^\top P\chi, \tag{4.10}$$

where $0 < \delta \le 1$, $c > 0$, $\frac{1}{2} < \gamma < 1$. Design constants c and δ will be specified later. Constant γ is pregiven and called the *characteristic parameter* of value function V_0. Clearly, $V_0(\chi)$ is positive definite and radially unbounded, and it vanishes at the origin $\chi = \mathbf{0}_{(n+m)\times 1}$.

Remark 4.2. Risk-sensitive control is much different from stochastic stabilization, and thus the methods developed by [5], [6], and [35] are not suitable for our control objective. Therefore, here we introduce a subquadratic function V_0 characterized by γ (see (4.10)), by which a method suitable for output-feedback risk-sensitive control design is developed.

Let $z_1 = y - y_d$ be the tracking error. Then, by assumptions A1 and A3, there exist a vector-valued smooth function $\overline{F}(y_d, z_1)$ and a matrix-valued smooth function $\overline{\Phi}(y_d, z_1)$ such that

$$F(y) = F(z_1 + y_d) = F(y_d) + z_1\overline{F}(y_d, z_1), \tag{4.11a}$$

$$\Phi(y) = \Phi(z_1 + y_d) = \Phi(y_d) + z_1\overline{\Phi}(y_d, z_1). \tag{4.11b}$$

LEMMA 4.1. *There exist a smooth vector-valued function $\sigma_0(\chi, y)$, a smooth function $N_0(y_d, \chi_a, z_1)$, and smooth $r_0(y_d)$, $C_0(y_d)$ such that*

$$\mathrm{d}V_0 \le \sigma_0\,\mathrm{d}w - \frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t - \frac{r_0\,\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t + N_0 z_1\,\mathrm{d}t + C_0\,\mathrm{d}t. \tag{4.12}$$

Proof. By (4.7) and the Itô formula we have

$$\begin{aligned}\mathrm{d}V_0 = {}& -\frac{\partial\phi}{\partial\xi}\|\chi\|^2\,\mathrm{d}t + \sigma_0(y,\chi)\,\mathrm{d}w - \frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t + \frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t \\ & + 2\frac{\partial\phi}{\partial\xi}\chi^\top PF(y)\,\mathrm{d}t + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2 V_0}{\partial\chi^2}\Phi(y)(\Phi(y))^\top\right)\mathrm{d}t,\end{aligned} \tag{4.13}$$

where $\sigma_0 = \frac{\partial V_0}{\partial\chi}\Phi$ is a row vector-valued function. In the above equality, we have used the technique of subtracting from and adding term $\frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t$ to its right-hand side. Notice that $\frac{\partial V_0}{\partial\chi} = \frac{\partial\phi}{\partial\xi}\cdot\frac{\partial\xi}{\partial\chi}$ and $\frac{\partial\phi}{\partial\xi} = \frac{\delta\gamma}{(c+\xi)^{1-\gamma}}$. Then we have

$$\sigma_0 = \frac{\partial\phi}{\partial\xi}\cdot\frac{\partial\xi}{\partial\chi}\Phi = \frac{2\delta\gamma}{(c+\xi)^{1-\gamma}}\chi^\top P\Phi(y). \tag{4.14}$$

Let $\xi_a = \chi_a^\top(P_1 - P_2P_3^{-1}P_2^\top)\chi_a$, where $P = \begin{bmatrix} P_1 & P_2 \\ P_2^\top & P_3\end{bmatrix}$, $P_1 \in \mathbb{R}^{(m+1)\times(m+1)}$, $P_2 \in \mathbb{R}^{(m+1)\times(n-1)}$, $P_3 \in \mathbb{R}^{(n-1)\times(n-1)}$. Clearly, since P is positive definite, so is $P_1 - P_2P_3^{-1}P_2^\top$. Then ξ_a is available for feedback design and satisfies $0 \le \xi_a \le \xi$. For the first term of the second line on the right-hand side of (4.13), by using (4.11a), we have

$$\begin{aligned}2\frac{\partial\phi}{\partial\xi}\chi^\top PF(y) &= 2\frac{\partial\phi}{\partial\xi}\chi^\top P(F(y_d) + \overline{F}(y_d, z_1)z_1) \\ &= \frac{2\delta\gamma\chi^\top PF(y_d)}{(c+\xi)^{1-\gamma}} + \frac{2\delta\gamma\chi^\top P\overline{F}(y_d, z_1)}{(c+\xi)^{1-\gamma}}z_1 \\ &\le \frac{\delta\gamma\varepsilon_{01}^2\|\chi\|^2}{(c+\xi)^{1-\gamma}} + \frac{\delta\gamma\|PF(y_d)\|^2}{\varepsilon_{01}^2(c+\xi)^{1-\gamma}} + \frac{\delta\gamma\|P\overline{F}(y_d, z_1)\|^2}{\varepsilon_{02}^2(c+\xi)^{1-\gamma}}z_1^2 + \frac{\delta\gamma\varepsilon_{02}^2\|\chi\|^2}{(c+\xi)^{1-\gamma}} \\ &\le \frac{\delta\gamma(\varepsilon_{01}^2+\varepsilon_{02}^2)\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}} + \frac{\delta\gamma\|PF(y_d)\|^2}{\varepsilon_{01}^2 c^{1-\gamma}} + \frac{\delta\gamma\|P\overline{F}(y_d, z_1)\|^2}{\varepsilon_{02}^2(c+\xi_a)^{1-\gamma}}z_1^2.\end{aligned} \tag{4.15}$$

Here and hereafter, ε_{01}, ε_{02}, ε_{03}, and ε_{04} are positive design constants to be determined later.

For the term $\frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t$ on the right-hand side of (4.13), by (4.11b) we have

$$
\begin{aligned}
\frac{\theta}{4}\sigma_0\sigma_0^\top &= \frac{\theta\delta^2\gamma^2\chi^\top P\Phi(y)(\Phi(y))^\top P\chi}{(c+\chi^\top P\chi)^{2-2\gamma}}\\
&= \frac{\theta\delta^2\gamma^2\chi^\top P\left(\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)\cdot\left(\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)^\top P\chi}{(c+\chi^\top P\chi)^{2-2\gamma}}\\
&= \frac{\theta\delta^2\gamma^2\chi^\top P\Phi(y_d)(\Phi(y_d))^\top P\chi}{(c+\chi^\top P\chi)^{2-2\gamma}}\\
&\quad+\frac{\theta\delta^2\gamma^2\left[\chi_a^\top,\,\chi_b^\top\right]P\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)(\overline{\Phi}(y_d,\,z_1))^\top P[\chi_a^\top,\,\chi_b^\top]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1 \qquad (4.16)\\
&= \frac{\theta\delta^2\gamma^2\chi^\top P\Phi(y_d)(\Phi(y_d))^\top P\chi}{(c+\chi^\top P\chi)^{2-2\gamma}}\\
&\quad+\frac{\theta\delta^2\gamma^2\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]P\overline{\Phi}(y_d,\,z_1)(\overline{\Phi}(y_d,\,z_1))^\top P\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1^2\\
&\quad+\frac{\theta\delta^2\gamma^2\chi^\top P\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)(\overline{\Phi}(y_d,\,z_1))^\top P\left[\mathbf{0}_{1\times(m+1)},\,\chi_b^\top\right]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1\\
&\quad+\frac{2\theta\delta^2\gamma^2\chi^\top P\Phi(y_d)(\overline{\Phi}(y_d,\,z_1))^\top P\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1\\
&\quad+\frac{\theta\delta^2\gamma^2\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]P\overline{\Phi}(y_d,\,z_1)(\overline{\Phi}(y_d,\,z_1))^\top P[\mathbf{0}_{1\times(m+1)},\,\chi_b^\top]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1^2\\
&\le \frac{\theta\delta^2\gamma^2\|P\Phi(y_d)\|_F^2}{c^{1-\gamma}}\frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\\
&\quad+\frac{\theta\delta^2\gamma^2\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]P\overline{\Phi}(y_d,\,z_1)(\overline{\Phi}(y_d,\,z_1))^\top P\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]^\top}{(c+\xi_a)^{2-2\gamma}}z_1^2\\
&\quad+\frac{\theta\delta^2\gamma^2\|P\|^2\left\|\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)(\overline{\Phi}(y_d,\,z_1))^\top\right\|_F^p}{p\varepsilon_{03}^p(c+\xi_a)^{p-\gamma(p+1)}}z_1^p\\
&\quad+\frac{(p-1)\theta\delta^2\gamma^2\|P\|^2\varepsilon_{03}^{\frac{p}{p-1}}}{p\lambda_{\min}^{\frac{1}{p-1}}(P)}\frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\\
&\quad+\frac{\theta\delta^2\gamma^2\|P\|^4}{\varepsilon_{04}^2(c+\xi_a)^{3-3\gamma}}\|\chi_a\|^2\left(\|\Phi(y_d)\|_F^2+\frac{z_1^2}{2}\|\overline{\Phi}(y_d,\,z_1)\|_F^2\right)\|\overline{\Phi}(y_d,\,z_1)\|_F^2z_1^2\\
&\quad+\frac{3\theta\delta^2\gamma^2\varepsilon_{04}^2\|\chi\|^2}{2(c+\chi^\top P\chi)^{1-\gamma}},
\end{aligned}
$$

where p is a positive even integer (that is, it takes values in set $\{2,4,6,8,\ldots\}$) and satisfies the inequality $p\ge\frac{\gamma}{1-\gamma}$ (or $\gamma\le\frac{p}{p+1}$). Let $q=\frac{p}{p-1}$. Then, p and q satisfy $\frac{1}{p}+\frac{1}{q}=1$. In the inequality (4.16), we have used the Young's inequality

$$
x^\top y\le\frac{\varepsilon^pW^p(x,\,y)}{p}\|x\|^p+\frac{1}{q\varepsilon^qW^q(x,\,y)}\|y\|^q\qquad\forall\,x,y\in\mathbb{R}^n,\,\varepsilon>0,\,W(x,\,y)>0,
$$

to get

$$\begin{aligned}
&\frac{\theta\delta^2\gamma^2\chi^\top P\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)(\overline{\Phi}(y_d,\,z_1))^\top P[\mathbf{0}_{1\times(m+1)},\,\chi_b^\top]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1\\
&\le\frac{\theta\delta^2\gamma^2\|P\|^2W^p(\chi)\left\|\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)(\overline{\Phi}(y_d,\,z_1))^\top\right\|^p}{p\varepsilon_{03}^p(c+\xi)^{2p(1-\gamma)}}z_1^p\\
&\quad+\frac{\theta\delta^2\gamma^2\|P\|^2\varepsilon_{03}^q\|\chi\|^{2q}}{qW^q(\chi)}\\
&\le\frac{\theta\delta^2\gamma^2\|P\|^2\left\|\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\,z_1)\right)(\overline{\Phi}(y_d,\,z_1))^\top\right\|_F^p}{p\varepsilon_{03}^p(c+\xi_a)^{p-\gamma(p+1)}}z_1^p\\
&\quad+\frac{(p-1)\theta\delta^2\gamma^2\|P\|^2\varepsilon_{03}^{\frac{p}{p-1}}}{p\lambda_{\min}^{\frac{1}{p-1}}(P)}\frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}},
\end{aligned}$$

with $W(\chi)=(c+\xi)^{1+\frac{\gamma}{p}-\gamma}$, and

$$\begin{aligned}
&\frac{2\theta\delta^2\gamma^2\chi^\top P\Phi(y_d)(\overline{\Phi}(y_d,\,z_1))^\top P\left[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}\right]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1\\
&\le\frac{\theta\delta^2\gamma^2\left(\frac{1}{\varepsilon_{04}^2(W(\chi))^2}\left\|P\Phi(y_d)(\overline{\Phi}(y_d,z_1))^\top P[\chi_a^\top,\mathbf{0}_{1\times(n-1)}]^\top\right\|^2z_1^2+\varepsilon_{04}^2(W(\chi))^2\|\chi\|^2\right)}{(c+\chi^\top P\chi)^{2-2\gamma}}\\
&\le\frac{\theta\delta^2\gamma^2\|P\|^4}{\varepsilon_{04}^2(c+\xi_a)^{3-3\gamma}}\|\chi_a\|^2\left\|\Phi(y_d)\right\|_F^2\left\|\overline{\Phi}(y_d,\,z_1)\right\|_F^2z_1^2+\frac{\theta\delta^2\gamma^2\varepsilon_{04}^2\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}},
\end{aligned}$$

$$\begin{aligned}
&\frac{\theta\delta^2\gamma^2[\chi_a^\top,\,\mathbf{0}_{1\times(n-1)}]P\overline{\Phi}(y_d,\,z_1)(\overline{\Phi}(y_d,\,z_1))^\top P[\mathbf{0}_{1\times(m+1)},\,\chi_b^\top]^\top}{(c+\chi^\top P\chi)^{2-2\gamma}}z_1^2\\
&\le\frac{\theta\delta^2\gamma^2\left(\frac{1}{\varepsilon_{04}^2W^2(\chi)}\left\|[\chi_a^\top,\mathbf{0}_{1\times(n-1)}]P\overline{\Phi}(y_d,z_1)(\overline{\Phi}(y_d,z_1))^\top P\right\|^2z_1^4+\varepsilon_{04}^2W^2(\chi)\|\chi_b\|^2\right)}{2(c+\chi^\top P\chi)^{2-2\gamma}}\\
&\le\frac{\theta\delta^2\gamma^2\|P\|^4}{2\varepsilon_{04}^2(c+\xi_a)^{3-3\gamma}}\|\chi_a\|^2\left\|\overline{\Phi}(y_d,\,z_1)\right\|_F^4z_1^4+\frac{\theta\delta^2\gamma^2\varepsilon_{04}^2\|\chi\|^2}{2(c+\chi^\top P\chi)^{1-\gamma}},
\end{aligned}$$

with $W(\chi)=(c+\xi)^{\frac{1}{2}-\frac{1}{2}\gamma}$.

For the last term on the right-hand side of (4.13), by (4.11b), we have

$$\begin{aligned}
&\frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2V_0}{\partial\chi^2}\Phi\Phi^\top\right)\\
&=\mathrm{Tr}\left(\left(\frac{\delta\gamma P}{(c+\chi^\top P\chi)^{1-\gamma}}-\frac{2\delta\gamma(1-\gamma)P\chi\chi^\top P}{(c+\chi^\top P\chi)^{2-\gamma}}\right)\Phi\Phi^\top\right)\\
&\le\frac{2\delta\gamma}{c^{1-\gamma}}\mathrm{Tr}\left((\Phi(y_d))^\top P\Phi(y_d)\right)+\frac{2\delta\gamma\mathrm{Tr}\left((\overline{\Phi}(y_d,\,z_1))^\top P\overline{\Phi}(y_d,\,z_1)\right)}{(c+\xi_a)^{1-\gamma}}z_1^2.
\end{aligned}\tag{4.17}$$

Substituting (4.15), (4.16), and (4.17) into (4.13), we get (4.12) with

$$r_0 = \delta\gamma(1-\varepsilon_{01}^2-\varepsilon_{02}^2) - \theta\delta^2\gamma^2\left(\frac{3}{2}\varepsilon_{04}^2 + \frac{\|P\Phi(y_d)\|_F^2}{c^{1-\gamma}} + \frac{(p-1)\varepsilon_{03}^{\frac{p}{p-1}}\|P\|^2}{p\lambda_{\min}^{\frac{1}{p-1}}(P)}\right), \tag{4.18}$$

$$\begin{aligned}
N_0 = {} & \frac{\delta\gamma\left\|P\overline{F}(y_d,\, z_1)\right\|^2}{\varepsilon_{02}^2(c+\xi_a)^{1-\gamma}}z_1 \\
& + \frac{\theta\delta^2\gamma^2\left[\chi_a^\top,\, \mathbf{0}_{1\times(n-1)}\right]P\overline{\Phi}(y_d,\, z_1)(\overline{\Phi}(y_d,\, z_1))^\top P\left[\chi_a^\top,\, \mathbf{0}_{1\times(n-1)}\right]^\top}{(c+\xi_a)^{2-2\gamma}}z_1 \\
& + \frac{\theta\delta^2\gamma^2\|P\|^2\left\|\left(2\Phi(y_d)+z_1\overline{\Phi}(y_d,\, z_1)\right)(\overline{\Phi}(y_d,\, z_1))^\top\right\|_F^p}{p\varepsilon_{03}^p(c+\xi_a)^{p-\gamma(p+1)}}z_1^{p-1} \\
& + \frac{\theta\delta^2\gamma^2\|P\|^4}{\varepsilon_{04}^2(c+\xi_a)^{3-3\gamma}}\|\chi_a\|^2\left(\|\Phi(y_d)\|_F^2 + \frac{z_1^2}{2}\|\overline{\Phi}(y_d,\, z_1)\|_F^2\right)\|\overline{\Phi}(y_d,\, z_1)\|_F^2 z_1 \\
& + \frac{2\delta\gamma\mathrm{Tr}\left((\overline{\Phi}(y_d,\, z_1))^\top P\overline{\Phi}(y_d,\, z_1)\right)}{(c+\xi_a)^{1-\gamma}}z_1,
\end{aligned}$$

$$C_0 = \frac{\delta\gamma}{c^{1-\gamma}}\left(\frac{\|PF(y_d)\|^2}{\varepsilon_{01}^2} + 2\mathrm{Tr}\left(P\Phi(y_d)(\Phi(y_d))^\top\right)\right). \tag{4.19}$$

□

The control design procedure will be presented for the two cases of $\rho = 1$ and $\rho > 1$ separately in subsections 4.3.2 and 4.3.3 below.

4.3.2. Control design for the case of $\rho = 1$. Let us now present the control design for the system (3.1) with $\rho = 1$. From (4.5) and (4.7) we obtain the following overall systems:

$$\mathrm{d}\chi = W\chi\,\mathrm{d}t + F(y)\,\mathrm{d}t + \Phi(y)\,\mathrm{d}w, \tag{4.20a}$$

$$\mathrm{d}\eta_1 = b_{n-1}g(y)u\,\mathrm{d}t + d_1\chi_b\,\mathrm{d}t + g_1(y,\, \chi_a)\,\mathrm{d}t + \widehat{h}_1(y)\,\mathrm{d}w, \tag{4.20b}$$

where

$$\begin{aligned}
& y = \eta_1, \\
& d_1 = [1,\ \mathbf{0}_{1\times(n-2)}], \\
& g_1 = \overline{g}_{11}(y) + \overline{d}_{11}[\mathbf{0}_{1\times(n-1)},\ 1]\chi_a + d_{11}y + \left[1,\ \mathbf{0}_{1\times(n-1)}\right]\chi_a.
\end{aligned}$$

It is easy to check that g_1 and $\widehat{h}_1$ are $\mathcal{C}^\infty$.

Let $\alpha_1 = b_{n-1}g(y)u$, $S_1 = d_1$, $F(y_d^{[1]},\, \chi_1,\, \eta_1) = g_1(y,\, \chi_a) - \dot{y}_d$, and $\Psi_1(y) = \widehat{h}_1(y)$. Then, by (4.20), we have the dynamics of tracking error $z_1 = \eta_1 - y_d$:

$$\mathrm{d}z_1 = (\alpha_1 + S_1\chi_b + F_1)\,\mathrm{d}t + \Psi_1\,\mathrm{d}w. \tag{4.21}$$

Let $V_1 = V_0 + \Xi_1(y_d)z_1^2$, where V_0 is defined by (4.10) and Ξ_1 is to be specified later.

Then, by (4.12) and (4.21), we have

(4.22)

$$\begin{aligned}
\mathrm{d}V_1 &= \mathrm{d}V_0 + 2z_1\Xi_1(\alpha_1 + F_1 + S_1\chi_b)\,\mathrm{d}t + \frac{\partial \Xi_1}{\partial t}z_1^2\,\mathrm{d}t + \Xi_1\Psi_1\Psi_1^\top\,\mathrm{d}t + 2\Xi_1\Psi_1 z_1\,\mathrm{d}w \\
&\le \sigma_1(y_d,\,\chi,\,z_1)\,\mathrm{d}w - \frac{\theta}{4}\sigma_1\sigma_1^\top\,\mathrm{d}t - \beta_1\Xi_1 z_1^2\,\mathrm{d}t + \beta_1\Xi_1 z_1^2\,\mathrm{d}t - r_0(y_d)\frac{\|\chi\|^2}{(c+\xi)^{1-\gamma}}\,\mathrm{d}t \\
&\quad + 2z_1\Xi_1(y_d)(\alpha_1 + F_1)\,\mathrm{d}t + \frac{\partial \Xi_1}{\partial t}z_1^2\,\mathrm{d}t + N_0(y_d,\,\chi_a,\,z_1)z_1\,\mathrm{d}t \\
&\quad + \frac{\theta}{4}\sigma_1\sigma_1^\top\,\mathrm{d}t - \frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t + M_1\chi_b z_1\,\mathrm{d}t + \Xi_1\Psi_1\Psi_1^\top\,\mathrm{d}t + C_0(y_d)\,\mathrm{d}t,
\end{aligned}$$

where $M_1 = 2\Xi_1 S_1$, $\sigma_1 = \sigma_0 + 2\Xi_1\Psi_1 z_1$. In the above inequality, we have used the technique of subtracting from and adding to its right-hand side the terms $\frac{\theta}{4}\sigma_1\sigma_1^\top\,\mathrm{d}t$ and $\beta_1\Xi_1 z_1^2\,\mathrm{d}t$. Here and hereafter, $\beta_1,\,\beta_1,\ldots,\beta_\rho$ are positive design constants to be determined.

Since $\sigma_0 = \frac{\partial V_0}{\partial \chi}\Phi(y)$ is unavailable for feedback design, so is the term $\frac{\theta}{4}\sigma_1\sigma_1^\top\,\mathrm{d}t - \frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t$ on the right-hand side of (4.22). Therefore, we give the following estimate:

(4.23)

$$\begin{aligned}
\frac{\theta}{4}(\sigma_1\sigma_1^\top - \sigma_0\sigma_0^\top) &= \theta\Xi_1\sigma_0\Psi_1^\top z_1 + \theta\Xi_1^2\Psi_1\Psi_1^\top z_1^2 \\
&= \theta\Xi_1^2\Psi_1\Psi_1^\top z_1^2 + \theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial \chi}\right)^\top\right\|^2 \\
&\quad + \frac{\theta}{4\varepsilon_{11}^2}\Xi_1^2\Psi_1\Phi^\top\Phi\Psi_1^\top z_1^2 - \theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial \chi}\right)^\top - \frac{\Xi_1}{2\varepsilon_{11}^2}\Phi\Psi_1^\top z_1\right\|^2 \\
&\le -\theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial \chi}\right)^\top - \frac{\Xi_1}{2\varepsilon_{11}^2}\Phi\Psi_1^\top z_1\right\|^2 + \frac{4\theta\delta^2\gamma^2\varepsilon_{11}^2\|P\|^2}{c^{1-\gamma}} \\
&\quad \cdot \frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}} + \theta\Xi_1^2\left(\Psi_1\Psi_1^\top + \frac{1}{4\varepsilon_{11}^2}\Psi_1\Phi^\top\Phi\Psi_1^\top\right)z_1^2,
\end{aligned}$$

where (and whereafter) ε_{11} and ε_1 are positive design constants to be specified.

Define

$$\Delta_{11}(y_d,\chi,z_1) = \theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial \chi}\right)^\top - \frac{\Xi_1}{2\varepsilon_{11}^2}\Phi\Psi_1^\top z_1\right\|^2, \tag{4.24a}$$

$$\Delta_{12}(y_d,\chi,z_1) = \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\|\chi_b\|^{\frac{p_1}{p_1-1}} + \frac{\|M_1^\top\|^{p_1}}{p_1\varepsilon_1^{p_1}}z_1^{p_1} - M_1\chi_b z_1. \tag{4.24b}$$

Clearly, $\Delta_{11} \ge 0$. Thus also, by Young's inequality, it is easy to see that $\Delta_{12} \ge 0$.

If p_1 takes values in set $\{4,\,6,\,8,\,10,\ldots\}$ and satisfies $p_1 \ge \frac{2\gamma}{2\gamma-1}$, then we can give an upper bound for "$M_1\chi_b z_1$":

$$M_1\chi_b z_1 = -\Delta_{12} + \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\|\chi_b\|^{\frac{p_1}{p_1-1}} + \frac{\|M_1^\top\|^{p_1}}{p_1\varepsilon_1^{p_1}}z_1^{p_1} \tag{4.25}$$
$$\le -\Delta_{12} + \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\lambda_{\max}^{1-\gamma}(P)\left(\mathcal{M}_\gamma(c) + \frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\right)$$
$$+ \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\mathcal{K}\left(\frac{p_1}{p_1-1}, 2\gamma\right) + \frac{\|M_1^\top\|^{p_1}}{p_1\varepsilon_1^{p_1}}z_1^{p_1},$$

where $\mathcal{K}(a_1, a_2)$ is defined in Lemma B.3.

Since there exist smooth functions $\overline{\Psi}_1(y_d, z_1)$ and $\overline{\Psi}_{11}(y_d, z_1)$ satisfying

$$\Psi_1 = \overline{\Psi}_1(y_d, z_1) = \overline{\Psi}_1(y_d) + z_1\overline{\Psi}_{11}(y_d, z_1),$$

for the fourth term of the last line on the right-hand side of (4.22) we have

$$\Xi_1\Psi_1\Psi_1^\top = \Xi_1\|(\overline{\Psi}_1(y_d) + z_1\overline{\Psi}_{11}(y_d, z_1))^\top\|^2 \tag{4.26}$$
$$\le 2\Xi_1\|(\overline{\Psi}_1(y_d))^\top\|^2 + 2\Xi_1\|(\overline{\Psi}_{11}(y_d, z_1))^\top\|^2 z_1^2.$$

Choose

$$\Xi_1 = \frac{\kappa_1}{1 + \|(\overline{\Psi}_1(y_d))^\top\|^2}, \tag{4.27}$$

where (and whereafter) $\kappa_1, \kappa_2, \ldots, \kappa_\rho$ are positive design constants to be determined.

Thus, by substituting (4.23), (4.25), (4.26), and (4.27) into (4.22), and via some straightforward calculations, we get

$$\mathrm{d}V_1 \le -z_1^2\,\mathrm{d}t + \sigma_1\,\mathrm{d}w - \frac{\theta}{4}\sigma_1\sigma_1^\top\,\mathrm{d}t - \frac{r_1(y_d)\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t \tag{4.28}$$
$$- \Xi_1\beta_1 z_1^2\,\mathrm{d}t + 2z_1\Xi_1\left(\alpha_1 - \overline{\alpha}_1(y_d^{[1]}, \chi_a, \eta_1)\right)\mathrm{d}t$$
$$- \Delta_1(y_d, \chi, z_1)\,\mathrm{d}t + 2\Xi_1 z_1 z_2\,\mathrm{d}t + C_1(y_d^{[1]})\,\mathrm{d}t,$$

where

$$r_1 = r_0(y_d) - \frac{4\theta\delta^2\gamma^2\varepsilon_{11}^2\|P\|^2}{c^{1-\gamma}} - \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\lambda_{\max}^{1-\gamma}(P), \tag{4.29}$$

$$N_1 = F_1 + \frac{\beta_1 z_1}{2} + \frac{z_1}{2\Xi_1}\frac{\partial\Xi_1}{\partial y_d}\dot{y}_d + \frac{\|M_1^\top\|^{p_1}}{2p_1\Xi_1\varepsilon_1^{p_1}}z_1^{p_1-1} \tag{4.30}$$
$$+ \theta\Xi_1\left(\frac{\Psi_1\Psi_1^\top}{2} + \frac{\Psi_1\Phi^\top\Phi\Psi_1^\top}{8\varepsilon_{11}^2}\right)z_1 + \|(\overline{\Psi}_{11}(y_d, z_1))^\top\|^2 z_1,$$

$$\overline{\alpha}_1 = \left\{-N_1 - \frac{z_1}{2\Xi_1} - \frac{N_0}{2\Xi_1}\right\}\Big|_{z_1=\eta_1-y_d}, \tag{4.31}$$

$$\Delta_1 = \Delta_{11} + \Delta_{12}, \quad \text{with } \Delta_{11} \text{ and } \Delta_{12} \text{ being defined by (4.24)},$$

$$C_1 = C_0(y_d) + \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\lambda_{\max}^{1-\gamma}(P)\mathcal{M}_\gamma(c) + 2\Xi_1\|(\overline{\Psi}_1(y_d))^\top\|^2 \tag{4.32}$$
$$+ \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\mathcal{K}\left(\frac{p_1}{p_1-1}, 2\gamma\right).$$

It is easy to check that C_1, r_1, and $\overline{\alpha}_1$ are $\mathcal{C}^\infty$.

Thus, we can choose the function $\alpha_1(y_d^{[1]}, \chi_a, \eta_1)$ in the following form:

$$\alpha_1 = \overline{\alpha}_1(y_d^{[1]}, \chi_a, \eta_1). \tag{4.33}$$

From this and the definition of α_1, i.e., $\alpha_1 = b_{n-1}g(y)u$, we immediately obtain the following risk-sensitive controller:

$$u = \frac{\alpha_1}{b_{n-1}g(y)} = \frac{1}{b_{n-1}g(y)}\,\overline{\alpha}_1(y_d^{[1]}, \chi_a, \eta_1). \tag{4.34}$$

Then, by (4.28) and (4.33), we have

$$\begin{aligned} \mathrm{d}V_1 \le {} & -z_1^2\,\mathrm{d}t + \sigma_1\,\mathrm{d}w - \frac{\theta}{4}\sigma_1\sigma_1^\top\,\mathrm{d}t - \frac{r_1(y_d)\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t \\ & - \Xi_1\beta_1 z_1^2\,\mathrm{d}t - \Delta_1(y_d, \chi, z_1)\,\mathrm{d}t + C_1(y_d^{[1]})\,\mathrm{d}t. \end{aligned} \tag{4.35}$$

4.3.3. Control design for the case of $\rho > 1$. This subsection investigates the control design for the system (3.1) with $\rho > 1$. From the procedure addressed below, we know that the control design for this case is more complicated than that for the case of $\rho = 1$ given in subsection 4.3.2.

First, from (4.6) and (4.7) we obtain the following overall systems amenable for integrator backstepping design:

$$\begin{aligned} \mathrm{d}\chi &= W\chi\,\mathrm{d}t + F(y)\,\mathrm{d}t + \Phi(y)\,\mathrm{d}w, \\ \mathrm{d}\eta_1 &= \eta_2\,\mathrm{d}t + d_1\chi_b\,\mathrm{d}t + g_1(y, \chi_a)\,\mathrm{d}t + \widehat{h}_1(y)\,\mathrm{d}w, \\ \mathrm{d}\eta_2 &= \eta_3\,\mathrm{d}t + d_2\chi_b\,\mathrm{d}t + g_2(y, \chi_a)\,\mathrm{d}t + \widehat{h}_2(y)\,\mathrm{d}w, \\ &\vdots \\ \mathrm{d}\eta_{\rho-1} &= \eta_\rho\,\mathrm{d}t + d_{\rho-1}\chi_b\,\mathrm{d}t + g_{\rho-1}(y, \chi_a)\,\mathrm{d}t + \widehat{h}_{\rho-1}(y)\,\mathrm{d}w, \\ \mathrm{d}\eta_\rho &= b_m g(y)u\,\mathrm{d}t + d_\rho\chi_b\,\mathrm{d}t + g_\rho(y, \chi_a)\,\mathrm{d}t + \widehat{h}_\rho(y)\,\mathrm{d}w, \end{aligned} \tag{4.36}$$

where

$$\begin{aligned} & y = \eta_1, \\ & d_1 = [1, \mathbf{0}_{1\times(n-2)}], \quad d_i = [-d_{\rho-1,i-1}, \mathbf{0}_{1\times(n-2)}],\ i = 2, \dots, \rho, \\ & g_i = g_{\rho i}\,(y, [\mathbf{0}_{1\times m}, 1]\chi_a) + d_{\rho i}y, \quad i = 1, \dots, \rho-1, \\ & g_\rho = g_{\rho\rho}\,(y, [\mathbf{0}_{1\times m}, 1]\chi_a) + d_{\rho\rho}y + [1, \mathbf{0}_{1\times m}]\,\chi_a. \end{aligned}$$

It is easy to check that g_i, $\widehat{h}_i$, $i = 1, \dots, \rho$, are $\mathcal{C}^\infty$.

Below is the backstepping design procedure, which involves ρ steps in all.

Step 1. Define variable $z_2 = \eta_2 - \alpha_1(y_d^{[1]}, \chi_a, \eta_1)$ and value function $V_1 = V_0 + \Xi_1(y_d)z_1^2$ for this step, where α_1 is a smooth function known as a virtual control law and Ξ_1 is a positive and smooth function. Both α_1 and Ξ_1 will be specified in this step.

From (4.36) it follows that

$$\mathrm{d}z_1 = (z_2 + \alpha_1 + F_1(y_d^{[1]}, \chi_a, \eta_1))\,\mathrm{d}t + S_1\chi_b\,\mathrm{d}t + \Psi_1(\eta_1)\,\mathrm{d}w, \tag{4.37}$$

where $F_1 = g_1(y, \chi_a) - \dot{y}_d$, $S_1 = d_1$, $\Psi_1 = \widehat{h}_1(y)$.

Clearly, (4.37) has the same structure as that of (4.21). Then, as in the case of $\rho = 1$, the virtual controller α_1 can be given by (4.33), which is such that

$$\begin{aligned}(4.38)\qquad \mathrm{d}V_1 \le{}& -z_1^2\,\mathrm{d}t + \sigma_1(y_d,\,\chi,\,z_1)\,\mathrm{d}w - \frac{\theta}{4}\sigma_1\sigma_1^{\top}\,\mathrm{d}t - \frac{r_1(y_d)\|\chi\|^2}{(c+\chi^{\top}P\chi)^{1-\gamma}}\,\mathrm{d}t\\ &- \Xi_1\beta_1 z_1^2\,\mathrm{d}t - \Delta_1(y_d,\,\chi,\,z_1)\,\mathrm{d}t + 2\Xi_1 z_1 z_2\,\mathrm{d}t + C_1(y_d^{[1]})\,\mathrm{d}t,\end{aligned}$$

where σ_1, r_1, Ξ_1, Δ_1, C_1 are defined as in the case of $\rho = 1$.

This completes *Step* 1.

Step i ($i = 2, \ldots, \rho-1$). Suppose that from step 1 through to step $i-1$ we have obtained $z_j = \eta_j - \alpha_{j-1}(y_d^{[j-1]}, \chi_a, \eta_{[j-1]})$, $j = 1, \ldots, i$, and value function

$$V_{i-1} = V_0 + \sum_{j=1}^{i-1}\Xi_j(y_d^{[j-1]},\,\chi_a,\,z_{[j-1]})z_j^2$$

satisfying

$$\begin{aligned}(4.39)\qquad \mathrm{d}V_{i-1} \le{}& -z_1^2\,\mathrm{d}t + \sigma_{i-1}(y_d^{[i-2]},\,\chi,\,z_{[i-1]})\,\mathrm{d}w - \frac{\theta}{4}\sigma_{i-1}\sigma_{i-1}^{\top}\,\mathrm{d}t\\ &- r_{i-1}(y_d)\frac{\|\chi\|^2}{(c+\chi^{\top}P\chi)^{1-\gamma}}\,\mathrm{d}t - \Delta_{i-1}(y_d^{[i-2]},\,\chi,\,z_{[i-1]})\,\mathrm{d}t\\ &- \sum_{j=1}^{i-1}\Xi_j\left(\beta_j - 2\sum_{m=j+1}^{i-1}(m-1)\,\kappa_m\right)z_j^2\,\mathrm{d}t\\ &+ 2\Xi_{i-1}z_{i-1}z_i\,\mathrm{d}t + C_{i-1}(y_d^{[i-1]},\,\chi_a,\,z_{[i-1]})\,\mathrm{d}t,\end{aligned}$$

where

$$\begin{aligned}\mathrm{d}z_1 &= (z_2 + \alpha_1 + F_1(y_d^{[1]},\,\chi_a,\,\eta_1))\mathrm{d}t + S_1\chi_b\mathrm{d}t + \Psi_1(\eta_1)\mathrm{d}w,\\ \mathrm{d}z_j &= (z_{j+1} + \alpha_j + F_j(y_d^{[j]},\,\chi_a,\,\eta_{[j]}))\mathrm{d}t + S_j(y_d^{[j-1]},\,\chi_a,\,\eta_{[j-1]})\chi_b\mathrm{d}t\\ &\quad + \Psi_j(y_d^{[j-1]},\,\chi_a,\,\eta_{[j-1]})\mathrm{d}w,\quad j = 2,\,3,\ldots,i-1,\\ \sigma_{i-1} &= \sigma_0 + \sum_{j=1}^{i-1}\left(2\Xi_j z_j\Psi_j + z_j^2\frac{\partial\Xi_j}{\partial\chi_a}\Phi_a + z_j^2\sum_{k=1}^{j-1}\frac{\partial\Xi_j}{\partial z_k}\Psi_k\right),\end{aligned}$$

$$(4.40)\qquad r_{i-1} = r_1(y_d),$$

$$(4.41)\qquad C_{i-1} = C_1(y_d^{[1]}) + \sum_{j=2}^{i-1}2\Xi_j\|(\overline{\Psi}_j(y_d^{[j-1]},\,\chi_a,\,\mathbf{0}_{(j-1)\times 1}))^{\top}\|^2,$$

and $\Delta_{i-1} = \Delta_{i-1,1}(y_d^{[i-2]},\,\chi,\,z_{[i-1]}) + \Delta_{i-1,2}(y_d^{[i-2]},\,\chi_b,\,z_{[i-1]})$. Here $\Delta_{i-1,1}$ and $\Delta_{i-1,2}$ are given as follows:

$$\begin{aligned}\Gamma_j(y_d^{[j-1]},\,\chi_a,\,z_{[j]}) &= \Psi_j + \frac{z_j}{2\Xi_j}\frac{\partial\Xi_j}{\partial\chi_a}\Phi_a + \frac{z_j}{2\Xi_j}\sum_{k=1}^{j-1}\frac{\partial\Xi_j}{\partial z_k}\Psi_k,\quad j = 1,\ldots,i-1,\\ \Delta_{i-1,1} &= \theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial\chi}\right)^{\top} - \frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Phi\Gamma_j^{\top}z_j\right\|^2,\\ M_k(y_d^{[k-1]},\,\chi_a,\,z_{[k]}) &= 2\Xi_k S_k + z_k\sum_{j=1}^{k-1}\frac{\partial\Xi_k}{\partial z_j}S_j + z_k\frac{\partial\Xi_k}{\partial\chi_a}L_a,\quad 1\le k\le i-1,\end{aligned}$$

$$(4.42)\qquad \Delta_{i-1,2} = \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\|\chi_b\|^{\frac{p_1}{p_1-1}} + \frac{\left\|\sum_{k=1}^{i-1}M_k^{\top}z_k\right\|^{p_1}}{p_1\varepsilon_1^{p_1}} - \left(\sum_{k=1}^{i-1}M_k z_k\right)\chi_b.$$

It should be noted that Ξ_i, F_j, S_j, Γ_j, Ψ_j, M_j, $j = 1, \ldots, i-1$, α_j, $\overline{\alpha}_j$, $i = 0, 1, \ldots, i-1$, σ_{i-1}, r_{i-1}, C_{i-1} are $\mathcal{C}^\infty$. By Young's inequality, it is easy to see that $\Delta_{i-1,\,2} \geq 0$.

Let $\varrho_{i-1}(y_d^{[i-2]}, \chi_a, z_{[i-2]},\, z_{i-1}) = \frac{1}{p_1 \varepsilon_1^{p_1}} \| \sum_{k=1}^{i-2} M_k^\top z_k + M_{i-1}^\top z_{i-1} \|^{p_1}$. Then

$$\begin{aligned}\Upsilon_{i-1}(y_d^{[i-2]}, \chi_a, z_{[i-1]}) &\triangleq \frac{1}{p_1 \varepsilon_1^{p_1}} \left(\left\| \sum_{k=1}^{i-2} M_k^\top z_k + M_{i-1}^\top z_{i-1} \right\|^{p_1} - \left\| \sum_{k=1}^{i-2} M_k^\top z_k \right\|^{p_1} \right) \\ &= \varrho_{i-1}(y_d^{[i-2]}, \chi_a, z_{[i-2]}, z_{i-1}) - \varrho_{i-1}(y_d^{[i-2]}, \chi_a, z_{[i-2]}, 0).\end{aligned}$$

Thus, by using the identity (see [25])

$$f(X) - f(0) = \left(\int_0^1 \left. \frac{\partial f(s)}{\partial s} \right|_{s=\beta X} d\beta \right) X,$$

we have

$$\Upsilon_{i-1} = z_{i-1} \overline{\Upsilon}_{i-1}(y_d^{[i-2]}, \chi_a, z_{[i-1]}), \tag{4.43}$$

where

$$\overline{\Upsilon}_{i-1} = \int_0^1 \left. \frac{\partial \varrho_{i-1}(\cdot,\, s)}{\partial s} \right|_{s=\alpha z_{i-1}} d\alpha.$$

Let $z_{i+1} = \eta_{i+1} - \alpha_i(y_d^{[i]}, \chi_a, \eta_{[i]})$, where α_i is a $\mathcal{C}^\infty$ function to be defined later. Then we have

$$\begin{aligned}\mathrm{d}z_i = {} & (z_{i+1} + \alpha_i + F_i(y_d^{[i]},\, \chi_a,\, \eta_{[i]}))\, \mathrm{d}t \\ & + S_i(y_d^{[i-1]},\, \chi_a,\, \eta_{[i-1]})\chi_b\, \mathrm{d}t + \Psi_i(y_d^{[i-1]},\, \chi_a,\, \eta_{[i-1]})\, \mathrm{d}w,\end{aligned} \tag{4.44}$$

where

$$\begin{aligned}F_i = {} & g_i(y,\, \chi_a) - \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial \eta_j} (\eta_{j+1} + g_j(y,\, \chi_a)) - \frac{\partial \alpha_{i-1}}{\partial \chi_a} (W_a \chi_a + F_a(y)) \\ & - \sum_{j=0}^{i-1} \frac{\partial \alpha_{i-1}}{\partial y_d^{(j)}} y_d^{(j+1)} - \frac{1}{2} \mathrm{Tr} \left(\frac{\partial^2 \alpha_{i-1}}{\partial ([\chi_a^\top,\, \eta_{[i-1]}^\top]^\top)^2} \begin{bmatrix} \Phi_a \\ \widehat{h}_1 \\ \vdots \\ \widehat{h}_{i-1} \end{bmatrix} \begin{bmatrix} \Phi_a \\ \widehat{h}_1 \\ \vdots \\ \widehat{h}_{i-1} \end{bmatrix}^\top \right), \\ S_i = {} & d_i - \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial \eta_j} d_j - \frac{\partial \alpha_{i-1}}{\partial \chi_a} L_a, \\ \Psi_i = {} & \widehat{h}_i(y) - \sum_{j=1}^{i-1} \frac{\partial \alpha_{i-1}}{\partial \eta_j} \widehat{h}_j(y) - \frac{\partial \alpha_{i-1}}{\partial \chi_a} \Phi_a(y)\end{aligned}$$

all are smooth functions.

Now we introduce the value function for this step as follows:

$$V_i = V_{i-1} + \Xi_i(y_d^{[i-1]},\, \chi_a,\, z_{[i-1]})z_i^2, \tag{4.45}$$

where Ξ_i is a positive smooth weighting function to be determined below in this step. By (4.39) and (4.44), we have

$$\begin{aligned}
\mathrm{d}V_i &= \mathrm{d}V_{i-1} + 2z_i\Xi_i(z_{i+1} + \alpha_i + F_i + S_i\chi_b)\,\mathrm{d}t + z_i^2\sum_{j=0}^{i-1}\frac{\partial\Xi_i}{\partial y_d^{(j)}}y_d^{(j+1)}\,\mathrm{d}t \\
&\quad + z_i^2\left(\frac{\partial\Xi_i}{\partial\chi_a}(W_a\chi_a + F_a + L_a\chi_b) + \sum_{j=1}^{i-1}\frac{\partial\Xi_i}{\partial z_j}(z_{j+1} + \alpha_j + F_j + S_j\chi_b)\right)\mathrm{d}t \\
&\quad + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2(\Xi_i z_i^2)}{\partial([\chi_a^\top,\, z_{[i]}^\top]^\top)^2}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}^\top\right)\mathrm{d}t \\
&\quad + 2\Xi_i z_i\Psi_i\,\mathrm{d}w + z_i^2\left(\frac{\partial\Xi_i}{\partial\chi_a}\Phi_a + \sum_{j=1}^{i-1}\frac{\partial\Xi_i}{\partial z_j}\Psi_j\right)\mathrm{d}w \\
&\le -z_1^2\,\mathrm{d}t + \sigma_i(y_d^{[i-1]}, \chi, z_{[i]})\,\mathrm{d}w - \frac{\theta}{4}\sigma_i\sigma_i^\top\,\mathrm{d}t + \frac{\theta}{4}(\sigma_i\sigma_i^\top - \sigma_{i-1}\sigma_{i-1}^\top)\,\mathrm{d}t \\
&\quad - \beta_i\Xi_i z_i^2\,\mathrm{d}t + \beta_i\Xi_i z_i^2\,\mathrm{d}t - \frac{r_{i-1}(y_d)\,\|\chi\|^2}{(c + \chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t \\
&\quad - \sum_{j=1}^{i-1}\Xi_j\left(\beta_j - 2\sum_{m=j+1}^{i-1}(m-1)\,\kappa_m\right)z_j^2\,\mathrm{d}t - \Delta_{i-1}\,\mathrm{d}t \\
&\quad + 2z_i\Xi_i\big(z_{i+1} + \alpha_i + \overline{F}_i(y_d^{[i]}, \chi_a, z_{[i]})\big)\,\mathrm{d}t + z_i M_i(y_d^{[i-1]}, \chi_a, z_{[i]})\chi_b\,\mathrm{d}t \\
&\quad + C_{i-1}\,\mathrm{d}t + \frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2(\Xi_i z_i^2)}{\partial([\chi_a^\top,\, z_{[i]}^\top]^\top)^2}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}^\top\right)\mathrm{d}t,
\end{aligned} \tag{4.46}$$

where

$$\begin{aligned}
\sigma_i &= \sigma_{i-1} + 2\Xi_i z_i\Psi_i + z_i^2\frac{\partial\Xi_i}{\partial\chi_a}\Phi_a + z_i^2\sum_{j=1}^{i-1}\frac{\partial\Xi_i}{\partial z_j}\Psi_j, \\
\overline{F}_i &= F_i + \frac{\Xi_{i-1}}{\Xi_i}z_{i-1} + \frac{z_i}{2\Xi_i}\left(\sum_{j=0}^{i-1}\frac{\partial\Xi_i}{\partial y_d^{(j)}}y_d^{(j+1)}\right. \\
&\qquad\qquad \left. + \sum_{j=1}^{i-1}\frac{\partial\Xi_i}{\partial z_j}(z_{j+1} + \alpha_j + F_j) + \frac{\partial\Xi_i}{\partial\chi_a}(W_a\chi_a + F_a)\right), \\
M_i &= 2\Xi_i S_i + z_i\sum_{j=1}^{i-1}\frac{\partial\Xi_i}{\partial z_j}S_j + z_i\frac{\partial\Xi_i}{\partial\chi_a}L_a
\end{aligned}$$

all are smooth functions. Here, we have used the technique of subtracting and adding terms $\frac{\theta}{4}\sigma_i\sigma_i^\top\,\mathrm{d}t$ and $\beta_i\Xi_i z_i^2\,\mathrm{d}t$ to the right-hand side of inequality (4.46).

Let $\overline{\sigma}_{i-1}(y_d^{[i-2]},\chi,z_{[i-1]})=\sigma_{i-1}-\frac{\partial V_0}{\partial\chi}\Phi=\sum_{j=1}^{i-1}(2\Xi_j z_j\Psi_j+z_j^2\frac{\partial\Xi_j}{\partial\chi_a}\Phi_a+z_j^2\sum_{k=1}^{j-1}\frac{\partial\Xi_j}{\partial z_k}\Psi_k)$. Then by noticing that $\overline{\sigma}_{i-1}$ is independent of χ_b, we have

(4.47)

$$
\begin{aligned}
&-\Delta_{i-1,1}+\frac{\theta}{4}(\sigma_i\sigma_i^\top-\sigma_{i-1}\sigma_{i-1}^\top)\\
&\quad=-\theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial\chi}\right)^\top-\frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Phi\Gamma_j^\top z_j\right\|^2+\theta\Xi_i\sigma_{i-1}\Gamma_i^\top z_i+\theta\Xi_i^2\Gamma_i\Gamma_i^\top z_i^2\\
&\quad=-\theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial\chi}\right)^\top-\frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Phi\Gamma_j^\top z_j\right\|^2\\
&\qquad+\theta\Xi_i\frac{\partial V_0}{\partial\chi}\Phi\Gamma_i^\top z_i+\theta\Xi_i(\overline{\sigma}_{i-1}+\Xi_i\Gamma_i z_i)\Gamma_i^\top z_i\\
&\quad=-\theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial\chi}\right)^\top-\frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Phi\Gamma_j^\top z_j\right\|^2\\
&\qquad+\left(\theta\left(\frac{\partial V_0}{\partial\chi}-\frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Gamma_j\Phi^\top z_j\right)\Xi_i\Phi\Gamma_i^\top z_i-\frac{\theta}{4\varepsilon_{11}^2}\Xi_i^2\Gamma_i\Phi^\top\Phi\Gamma_i^\top z_i^2\right)\\
&\qquad+\frac{\theta}{4\varepsilon_{11}^2}\Xi_i^2\Gamma_i\Phi^\top\Phi\Gamma_i^\top z_i^2+\frac{\theta}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Gamma_j\Phi^\top z_j\Xi_i\Phi\Gamma_i^\top z_i\\
&\qquad+\theta\Xi_i(\overline{\sigma}_{i-1}+\Xi_i\Gamma_i z_i)\Gamma_i^\top z_i\\
&\quad=-\Delta_{i1}+\theta\Xi_i\Bigg(\frac{1}{4\varepsilon_{11}^2}\Xi_i\Gamma_i\Phi^\top\Phi z_i\\
&\qquad\qquad+\frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Gamma_j\Phi^\top\Phi z_j+\overline{\sigma}_{i-1}+\Xi_i\Gamma_i z_i\Bigg)\Gamma_i^\top z_i,
\end{aligned}
$$

where

$$
\Gamma_i(y_d^{[i-1]},\,\chi_a,\,z_{[i]})=\Psi_i+\frac{z_i}{2\Xi_i}\frac{\partial\Xi_i}{\partial\chi_a}\Phi_a+\frac{z_i}{2\Xi_i}\sum_{j=1}^{i-1}\frac{\partial\Xi_i}{\partial z_j}\Psi_j,
$$

$$
\Delta_{i1}(y_d^{[i-1]},\,\chi,\,z_{[i]})=\theta\varepsilon_{11}^2\left\|\left(\frac{\partial V_0}{\partial\chi}\right)^\top-\frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i}\Xi_j\Phi\Gamma_j^\top z_j\right\|^2. \tag{4.48}
$$

Similar to (4.25), by using (4.42) we have

$$
\begin{aligned}
(4.49)\qquad & M_i\chi_b z_i - \Delta_{i-1,2} \\
&= \left(\sum_{k=1}^{i} M_k z_k\right)\chi_b - \frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\|\chi_b\|^{\frac{p_1}{p_1-1}} - \frac{\left\|\sum_{j=1}^{i-1} M_j^\top z_j\right\|^{p_1}}{p_1\varepsilon_1^{p_1}} z_1^{p_1} \\
&= -\left(\frac{\varepsilon_1^{\frac{p_1}{p_1-1}}(p_1-1)}{p_1}\|\chi_b\|^{\frac{p_1}{p_1-1}} + \frac{\left\|\sum_{j=1}^{i} M_j^\top z_j\right\|^{p_1}}{p_1\varepsilon_1^{p_1}} - \sum_{j=1}^{i} M_j z_j \chi_b\right) \\
&\quad + \frac{1}{p_1\varepsilon_1^{p_1}}\left(\left\|\sum_{j=1}^{i-1} M_j^\top z_j + M_i^\top z_i\right\|^{p_1} - \left\|\sum_{j=1}^{i-1} M_j^\top z_j\right\|^{p_1}\right) \\
&= -\Delta_{i2} + \Upsilon_i,
\end{aligned}
$$

where

$$
\begin{aligned}
(4.50)\qquad \Delta_{i2}(y_d^{[i-1]}, \chi_b, z_{[i]}) &= \frac{(p_1-1)\varepsilon_1^{\frac{p_1}{p_1-1}}}{p_1}\|\chi_b\|^{\frac{p_1}{p_1-1}} + \frac{1}{p_1\varepsilon_1^{p_1}}\left\|\sum_{j=1}^{i} M_j^\top z_j\right\|^{p_1} \\
&\quad - \sum_{j=1}^{i} M_j z_j \chi_b, \\
\Upsilon_i(y_d^{[i-1]}, \chi_a, z_{[i]}) &= \frac{1}{p_1\varepsilon_1^{p_1}}\left(\left\|\sum_{j=1}^{i-1} M_j^\top z_j + M_i^\top z_i\right\|^{p_1} - \left\|\sum_{j=1}^{i-1} M_j^\top z_j\right\|^{p_1}\right).
\end{aligned}
$$

By Young's inequality, it is easy to see that $\Delta_{i2} \geq 0$. And similar to (4.43), there exists a smooth function $\overline{\Upsilon}_i(y_d^{[i-1]}, \chi_a, z_{[i]})$ such that

$$
\Upsilon_i = z_i\overline{\Upsilon}_i(y_d^{[i-1]}, \chi_a, z_{[i]}).
$$

By assumptions A1 and A3, we know that there exist vector-valued smooth functions $\overline{\Psi}_i(y_d^{[i-1]}, \chi_a, \eta_{[i-1]})$ and $\overline{\Psi}_{ij}(y_d^{[i-1]}, \chi_a, z_{[j]})$, $j = 1, \ldots, i-1$, such that

$$
\begin{aligned}
\Psi_i(y_d^{[i-1]}, \chi_a, \eta_{[i-1]}) &= \overline{\Psi}_i(y_d^{[i-1]}, \chi_a, z_{[i-1]}) \\
&= \overline{\Psi}_i(y_d^{[i-1]}, \chi_a, \mathbf{0}_{(i-1)\times 1}) + \sum_{j=1}^{i-1}\overline{\Psi}_{ij}(y_d^{[i-1]}, \chi_a, z_{[j]})z_j.
\end{aligned}
$$

Then, for the last term on the right-hand side of (4.46), we have

(4.51)

$$
\begin{aligned}
&\frac{1}{2}\mathrm{Tr}\left(\frac{\partial^2(\Xi_i z_i^2)}{\partial([\chi_a^\top, z_{[i]}^\top]^\top)^2}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}^\top\right) \\
&= \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} \dfrac{\partial^2\Xi_i}{\partial([\chi_a^\top, z_{[i-1]}^\top]^\top)^2}z_i^2 & 2\left(\dfrac{\partial\Xi_i}{\partial[\chi_a^\top, z_{[i-1]}^\top]^\top}\right)^\top z_i \\ 2\dfrac{\partial\Xi_i}{\partial[\chi_a^\top, z_{[i-1]}^\top]^\top}z_i & 2\Xi_i \end{bmatrix}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}\begin{bmatrix}\Phi_a\\ \Psi_1\\ \vdots\\ \Psi_i\end{bmatrix}^\top\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} \dfrac{\partial^2 \Xi_i}{\partial([\chi_a^\top,\, z_{[i-1]}^\top]^\top)^2} z_i & 2\left(\dfrac{\partial \Xi_i}{\partial[\chi_a^\top,\, z_{[i-1]}^\top]^\top}\right)^\top \\ 2\dfrac{\partial \Xi_i}{\partial[\chi_a^\top,\, z_{[i-1]}^\top]^\top} & 0 \end{bmatrix} \begin{bmatrix} \Phi_a \\ \Psi_1 \\ \vdots \\ \Psi_i \end{bmatrix} \begin{bmatrix} \Phi_a \\ \Psi_1 \\ \vdots \\ \Psi_i \end{bmatrix}^\top\right) z_i \\
&\quad + \Xi_i \left(\overline{\Psi}_i(y_d^{[i-1]},\, \chi_a,\, \mathbf{0}_{(i-1)\times 1}) + \sum_{j=1}^{i-1} \overline{\Psi}_{ij}(y_d^{[i-1]},\, \chi_a,\, z_{[j]}) z_j\right) \\
&\qquad \times \left(\overline{\Psi}_i(y_d^{[i-1]},\, \chi_a,\, \mathbf{0}_{(i-1)\times 1}) + \sum_{j=1}^{i-1} \overline{\Psi}_{ij}(y_d^{[i-1]},\, \chi_a,\, z_{[j]}) z_j\right)^\top \\
&\le \frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix} \dfrac{\partial^2 \Xi_i}{\partial([\chi_a^\top,\, z_{[i-1]}^\top]^\top)^2} z_i & 2\left(\dfrac{\partial \Xi_i}{\partial[\chi_a^\top,\, z_{[i-1]}^\top]^\top}\right)^\top \\ 2\dfrac{\partial \Xi_i}{\partial[\chi_a^\top,\, z_{[i-1]}^\top]^\top} & 0 \end{bmatrix} \begin{bmatrix} \Phi_a \\ \Psi_1 \\ \vdots \\ \Psi_i \end{bmatrix} \begin{bmatrix} \Phi_a \\ \Psi_1 \\ \vdots \\ \Psi_i \end{bmatrix}^\top\right) z_i \\
&\quad + 2\Xi_i \left\| (\overline{\Psi}_i(y_d^{[i-1]},\, \chi_a,\, \mathbf{0}_{(i-1)\times 1}))^\top \right\|^2 \\
&\quad + 2(i-1)\Xi_i \sum_{j=1}^{i-1} \Xi_j^{-1} \left\| (\overline{\Psi}_{ij}(y_d^{[i-1]},\, \chi_a,\, z_{[j]}))^\top \right\|^2 \Xi_j z_j^2.
\end{aligned}
$$

Choose

(4.52)
$$
\Xi_i = \frac{\kappa_i}{1 + \left\| (\overline{\Psi}_i(y_d^{[i-1]},\, \chi_a,\, \mathbf{0}_{(i-1)\times 1}))^\top \right\|^2 + \sum_{j=1}^{i-1} \Xi_j^{-1} \left\| (\overline{\Psi}_{ij}(y_d^{[i-1]},\, \chi_a,\, z_{[j]}))^\top \right\|^2}.
$$

Then we have

$$
\begin{aligned}
&2\Xi_i \left\| (\overline{\Psi}_i(y_d^{[i-1]},\, \chi_a,\, \mathbf{0}_{(i-1)\times 1}))^\top \right\|^2 \le 2\kappa_i, \\
&2(i-1)\Xi_i \sum_{j=1}^{i-1} \Xi_j^{-1} \left\| (\overline{\Psi}_{ij}(y_d^{[i-1]},\, \chi_a,\, z_{[j]}))^\top \right\|^2 \Xi_j z_j^2 \le 2(i-1)\kappa_i \sum_{j=1}^{i-1} \Xi_j z_j^2.
\end{aligned}
$$

By substituting (4.47)–(4.52) into (4.46), we get

$$
\begin{aligned}
\mathrm{d}V_i \le\ & -z_1^2\,\mathrm{d}t + \sigma_i\,\mathrm{d}w - \frac{\theta}{4}\sigma_i\sigma_i^\top\,\mathrm{d}t - r_i(y_d)\frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t \\
& - \sum_{j=1}^{i} \Xi_j \left(\beta_j - 2\sum_{m=j+1}^{i} (m-1)\,\kappa_m\right) z_j^2\,\mathrm{d}t - \Delta_i(y_d^{[i-1]}, \chi, z_{[i]})\,\mathrm{d}t \\
& + 2\Xi_i z_i z_{i+1}\,\mathrm{d}t + 2z_i\Xi_i(\alpha_i - \overline{\alpha}_i(y_d^{[i]},\, \chi_a,\, \eta_{[i]}))\,\mathrm{d}t \\
& + C_i(y_d^{[i]},\, \chi_a,\, z_{[i-1]})\,\mathrm{d}t,
\end{aligned} \tag{4.53}
$$

where

(4.54)

$r_i = r_{i-1}(y_d)$, with r_{i-1} being defined by (4.40),

$\Delta_i = \Delta_{i1} + \Delta_{i2}$, with Δ_{i1} and Δ_{i2} being defined by (4.48) and (4.50), respectively,

$$N_i(y_d^{[i]}, \chi_a, z_{[i]}) = \overline{F}_i + \frac{\overline{\Upsilon}_i}{2\Xi_i} + \frac{\beta_i z_i}{2}$$
$$+ \frac{\theta}{2}\left(\frac{1}{4\varepsilon_{11}^2}\Xi_i\Gamma_i\Phi^\top\Phi z_i + \frac{1}{2\varepsilon_{11}^2}\sum_{j=1}^{i-1}\Xi_j\Gamma_j\Phi^\top\Phi z_j + \overline{\sigma}_{i-1} + \Xi_i\Gamma_i z_i\right)\Gamma_i^\top$$
$$+ \frac{1}{4\Xi_i}\mathrm{Tr}\left(\begin{bmatrix} \dfrac{\partial^2\Xi_i}{\partial([\chi_a^\top,\, z_{[i-1]}^\top]^\top)^2} z_i & 2\left(\dfrac{\partial\Xi_i}{\partial[\chi_a^\top,\, z_{[i-1]}^\top]^\top}\right)^\top \\ 2\dfrac{\partial\Xi_i}{\partial[\chi_a^\top,\, z_{[i-1]}^\top]^\top} & 0 \end{bmatrix} \begin{bmatrix} \Phi_a \\ \Psi_1 \\ \vdots \\ \Psi_i \end{bmatrix} \begin{bmatrix} \Phi_a \\ \Psi_1 \\ \vdots \\ \Psi_i \end{bmatrix}^\top\right),$$

$$\overline{\alpha}_i = \left\{ - N_i(y_d^{[i]}, \chi_a, z_{[i]})\right\}\Big|_{z_j = \eta_j - \alpha_{j-1},\, j=1,\ldots,i},$$

$$C_i = C_{i-1} + 2\Xi_i\left\|(\overline{\Psi}_i(y_d^{[i-1]}, \chi_a, \mathbf{0}_{(i-1)\times 1}))^\top\right\|^2, \quad C_{i-1} \text{ is given by (4.41).} \tag{4.55}$$

Now, we choose virtual controller $\alpha_i(y_d^{[i]}, \chi_a, \eta_{[i]})$ as follows:

$$\alpha_i = \overline{\alpha}_i(y_d^{[i]}, \chi_a, \eta_{[i]}). \tag{4.56}$$

Substituting (4.56) into (4.53), we have

$$\begin{aligned} \mathrm{d}V_i \le {} & -z_1^2\,\mathrm{d}t + \sigma_i\,\mathrm{d}w - \frac{\theta}{4}\sigma_i\sigma_i^\top\,\mathrm{d}t - r_i(y_d)\frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t - \Delta_i\,\mathrm{d}t \\ & - \sum_{j=1}^{i}\Xi_j\left(\beta_j - 2\sum_{m=j+1}^{i}(m-1)\,\kappa_m\right)z_j^2\,\mathrm{d}t \\ & + 2\Xi_i z_i z_{i+1}\,\mathrm{d}t + C_i(y_d^{[i]}, \chi_a, z_{[i]})\,\mathrm{d}t. \end{aligned} \tag{4.57}$$

This completes *Step i*.

Step ρ. It is easy to see that the results of Step i hold also for $i = \rho$, where $\eta_{\rho+1} = b_m g(y)u$. Define the value function V_ρ as in (4.45) with $i = \rho$ for this step. Then, V_ρ satisfies (4.57) with $i = \rho$. Set $z_{\rho+1} = 0$. Then, we arrive at the controller

$$u(y_d^{[\rho]}, \chi_a, \eta_{[\rho]}) = \frac{1}{b_m g(y)}\alpha_\rho(y_d^{[\rho]}, \chi_a, \eta_{[\rho]}), \tag{4.58}$$

where α_ρ is defined by letting $i = \rho$ in (4.56). Let $V_\rho = V_{\rho-1} + \Xi(y_d^{[\rho-1]}, \chi_a, z_{[\rho-1]})z_\rho^2$. Then we have

$$\begin{aligned} \mathrm{d}V_\rho \le {} & -z_1^2\,\mathrm{d}t + \sigma_\rho(y_d^{[\rho-1]}, \chi_a, z)\,\mathrm{d}w - \frac{\theta}{4}\sigma_\rho\sigma_\rho^\top\,\mathrm{d}t \\ & - r_\rho(y_d)\frac{\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t - \Delta_\rho(y_d^{[\rho-1]}, \chi, z)\,\mathrm{d}t \\ & - \sum_{j=1}^{\rho}\Xi_j\left(\beta_j - 2\sum_{m=j+1}^{\rho}(m-1)\,\kappa_m\right)z_j^2\,\mathrm{d}t + C_\rho(y_d^{[\rho]}, \chi_a, z_{[\rho]})\,\mathrm{d}t, \end{aligned} \tag{4.59}$$

where Ξ_ρ, σ_ρ, r_ρ, Δ_ρ, and C_ρ are defined in the same way as in Step i ($i = 2, \ldots, \rho-1$), with i being replaced by ρ.

So far, we have completed the entire backstepping design.

4.4. Properties of the design procedure. In this subsection, we give several properties of the design procedure above. To avoid duplication of the expression, here only the case of $\rho > 1$ is considered, since the case of $\rho = 1$ has the same properties.

By Lemmas B.1 and B.2, we have

$$\begin{aligned} r_\rho(y_d)\frac{\|\chi\|^2}{(c+\xi)^{1-\gamma}} &\geq \frac{r_\rho(y_d)}{(\lambda_{\max}(P))^{1-\gamma}}\|\chi\|^{2\gamma} - r_\rho(y_d)\mathcal{M}_\gamma(c) \\ &\geq \frac{r_\rho(y_d)}{\lambda_{\max}(P)}\left((c+\xi)^\gamma - c^\gamma\right) - r_\rho(y_d)\mathcal{M}_\gamma(c) \\ &= \frac{r_\rho(y_d)}{\delta\lambda_{\max}(P)}\phi(\xi) - r_\rho(y_d)\mathcal{M}_\gamma(c), \end{aligned} \tag{4.60}$$

where $r_\rho(y_d) = r_1(y_d)$ is defined by (4.54) and ξ and $\phi(\xi)$ are defined in (4.10).

Define

$$\overline{r}(y_d) = \frac{r_\rho(y_d)}{\delta\lambda_{\max}(P)}, \tag{4.61}$$

$$\overline{\beta}_i = \beta_i - 2\sum_{j=i+1}^{\rho}(j-1)\,\kappa_j, \quad i = 1, \ldots, \rho, \tag{4.62}$$

$$C(y_d^{[\rho]}, \chi_a, z_{[\rho]}) = C_\rho + r_\rho(y_d)\mathcal{M}_\gamma(c). \tag{4.63}$$

Then, by (4.59) (or (4.35) for the case of $\rho = 1$) and (4.60)–(4.63), we have

$$\mathrm{d}V_\rho \leq \sigma_\rho\,\mathrm{d}w - \frac{\theta}{4}\sigma_\rho\sigma_\rho^\top\,\mathrm{d}t - z_1^2\,\mathrm{d}t - \overline{r}\,\phi(\xi)\,\mathrm{d}t - \sum_{i=1}^{\rho}\Xi_i\overline{\beta}_i z_i^2\,\mathrm{d}t - \Delta_\rho\,\mathrm{d}t + C\,\mathrm{d}t. \tag{4.64}$$

The following lemma presents the method specifying the design constants.

LEMMA 4.2. *For any given cost value $R_l > 0$, risk-sensitivity parameter $\theta > 0$, and characteristic parameter $\gamma \in (\frac{1}{2}, 1)$, there always exist positive design constants $\delta, c, \varepsilon_{01}, \varepsilon_{02}, \varepsilon_{03}, \varepsilon_{04}, \varepsilon_1, \varepsilon_{11}, \beta_1, \ldots, \beta_\rho, \kappa_1, \ldots, \kappa_\rho$, such that the following inequalities hold:*

$$r_\rho \geq r > 0, \quad \overline{\beta}_1 > 0, \ldots, \overline{\beta}_\rho > 0, \quad \textit{and} \quad C_\rho \leq R_l, \tag{4.65}$$

where r is constant.

Proof. The proof can be accomplished by properly selecting a set of design constants.

Design constants δ, ε_{01}, and ε_{02} are chosen such that

$$0 < \delta \leq 1, \tag{4.66}$$

$$0 < \varepsilon_{01} < \frac{\sqrt{2}}{4}, \tag{4.67}$$

$$0 < \varepsilon_{02} < \frac{\sqrt{2}}{4}. \tag{4.68}$$

For given $\gamma \in (\frac{1}{2}, 1)$, even numbers $p \in \{2, 4, 6, \ldots\}$ and $p_1 \in \{4, 6, 8, \ldots\}$ are chosen such that

$$\frac{p_1}{2(p_1-1)} \le \gamma \le \frac{p}{p+1}.$$

For example, when $\gamma = \frac{2}{3}$, even numbers $p \ge 2$ and $p_1 \ge 4$ are proper; when $\gamma = \frac{4}{5}$, then even numbers $p \ge 4$ and $p_1 \ge 4$ are proper; when $\gamma = \frac{3}{5}$, then even numbers $p \ge 2$ and $p_1 \ge 6$ are proper.

Then, for given $\gamma \in (\frac{1}{2}, 1)$, risk-sensitivity parameter θ, desired positive risk-sensitive cost value R_l, given output y_d, selected δ, ε_{01}, p, and p_1, and design constants ε_{03}, ε_{04}, c, and ε_1 are chosen such that

$$0 < \varepsilon_{03} < \left(\frac{p \lambda_{\min}^{\frac{1}{p-1}}(P)}{12\theta\delta\gamma(p-1)\|P\|^2} \right)^{\frac{p-1}{p}}, \tag{4.69}$$

$$0 < \varepsilon_{04} < \sqrt{\frac{1}{18\theta\delta\gamma}}, \tag{4.70}$$

(4.71)

$$c \ge \max\left\{1, \left(\max\left\{ \frac{5\delta\gamma \max_{|y_d|\le C_{y_d}} \|PF(y_d)\|^2}{R_l \varepsilon_{01}^2}, 12\theta\delta\gamma \max_{|y_d|\le C_{y_d}} \left(\|P\Phi(y_d)\|_F^2\right), \right.\right.\right.$$
$$\left.\left.\left. \frac{10\delta\gamma \max_{|y_d|\le C_{y_d}} \mathrm{Tr}\left(P\Phi(y_d)\Phi^\top(y_d)\right)}{R_l} \right\}\right)^{\frac{1}{1-\gamma}}\right\},$$

$$0 < \varepsilon_1 < \delta \left(\min\left\{ \frac{p_1 R_l}{5(p_1-1)(1+\mathcal{K}(\frac{p_1}{p_1-1}, 2\gamma))}, \quad \frac{p_1 R_l}{5(p_1-1)\lambda_{\max}^{1-\gamma}(P)\mathcal{M}_\gamma(c)}, \right.\right.$$
$$\left.\left. \frac{\gamma p_1}{8(p_1-1)\lambda_{\max}^{1-\gamma}(P)} \right\}\right)^{\frac{p_1-1}{p_1}}. \tag{4.72}$$

For given γ and selected constants c and δ, constant ε_{11} is chosen such that

$$0 < \varepsilon_{11} < \sqrt{\frac{c^{1-\gamma}}{16\theta\delta\gamma\|P\|^2}}. \tag{4.73}$$

Design constants $\kappa_1, \ldots, \kappa_\rho$ are chosen such that

$$0 < \kappa_i = \min\left\{1, \quad \frac{R_l}{10\rho}\right\}, \quad i = 1, \ldots, \rho. \tag{4.74}$$

For given κ_i's, design constants $\beta_1, \ldots, \beta_\rho$ are chosen such that

$$\beta_i > 2 \sum_{m=i+1}^{\rho} (m-1)\,\kappa_m, \quad i = 1, \ldots, \rho. \tag{4.75}$$

Thus, by (4.54) with $i = \rho$, (4.29), (4.18), (4.67)–(4.73), we have $r_\rho(y_d) > \frac{1}{8}\delta\gamma > 0$; by (4.62) and (4.75) we have $\overline{\beta}_1 > 0, \ldots, \overline{\beta}_\rho > 0$; by (4.59) (or (4.35) for the case of $\rho = 1$), (4.32), (4.19), (4.67), (4.71), (4.72), and (4.75), we have $C_\rho \le R_l$. □

For the value function V_ρ, we have the following lemma.

LEMMA 4.3. *There are positive definite, continuous, and radially unbounded functions* $W_1(\chi, z)$ *and* $W_2(\chi, z)$ *such that*

$$W_1(\chi, z) \le V_\rho(y_d^{[\rho-1]}, \chi, z) \le W_2(\chi, z). \tag{4.76}$$

Proof. Define

$$W_1(\chi, z) = V_0(\chi) + \sum_{i=1}^{\rho} \min_{|y_d|\le C_{y_d},\ldots,|y_d^{(i-1)}|\le C_{y_d^{(i-1)}}} \Xi_i(y_d^{[i-1]}, \chi_a, z_{[i-1]}) z_i^2,$$

$$W_2(\chi, z) = V_0(\chi) + \sum_{i=1}^{\rho} \max_{|y_d|\le C_{y_d},\ldots,|y_d^{(i-1)}|\le C_{y_d^{(i-1)}}} \Xi_i(y_d^{[i-1]}, \chi_a, z_{[i-1]}) z_i^2.$$

Then inequality (4.76) holds.

We now show that $W_1(\chi, z)$ and $W_2(\chi, z)$ are positive definite, continuous, and radially unbounded. Clearly, W_1 and W_2 are continuous. In fact, based on $W_1(\chi, z) \le W_2(\chi, z)$ and $W_2(\mathbf{0}_{(n+m)\times 1}, \mathbf{0}_{\rho\times 1}) = 0$, it suffices to show that so is $W_1(\chi, z)$.

Let us next prove the positive definition and radial unboundedness of W_1 by induction. It is clear that $V_0(\chi)$ is positive definite and radially unbounded by the definition (4.10).

By the definition (4.27) of Ξ_1, assumption A3, and the smoothness of $\|\overline{\Psi}_1\|^2$, we see that $\min_{|y_d|\le C_{y_d}} \Xi_1(y_d)$ is existent and positive. Thus, $V_0(\chi)+\min_{|y_d|\le C_{y_d}} \Xi_1(y_d) z_1^2$ is positive definite and radially unbounded.

Clearly, $\min_{|y_d|\le C_{y_d}, |y_d^{(1)}|\le C_{y_d^{(1)}}} \Xi_2(y_d^{[1]}, \chi_a, z_1)$ is positive and continuous with respect to (χ_a, z_1). Thus, by Lemma B.4 in Appendix B we can get the positive definiteness and radial unboundedness of $V_0(\chi)+\min_{|y_d|\le C_{y_d}} \Xi_1(y_d) z_1^2 + \min_{|y_d|\le C_{y_d}, |y_d^{(1)}|\le C_{y_d^{(1)}}} \Xi_2(y_d^{[1]}, \chi_a, z_1) z_2^2$.

Suppose that $V_0(\chi) + \sum_{j=1}^{i-1} \min_{|y_d|\le C_{y_d},\ldots,|y_d^{(j-1)}|\le C_{y_d^{(j-1)}}} \Xi_j(y_d^{[j-1]}, \chi_a, z_{[j-1]}) z_j^2$ $(i = 3, 4, \ldots, \rho)$ is positive definite and radially unbounded. Then, from the positiveness and continuity of $\min_{|y_d|\le C_{y_d},\ldots,|y_d^{(i-1)}|\le C_{y_d^{(i-1)}}} \Xi_i(y_d^{[i-1]}, \chi_a, z_{[i-1]})$ and Lemma B.4 in Appendix B, we obtain the positive definiteness and radial unboundedness of

$$V_0(\chi) + \sum_{j=1}^{i} \min_{|y_d|\le C_{y_d},\ldots,|y_d^{(j-1)}|\le C_{y_d^{(j-1)}}} \Xi_j(y_d^{[j-1]}, \chi_a, z_{[j-1]}) z_i^2.$$

Thus, by induction, $W_1(\chi, z)$ is positive definite and radially unbounded. ∎

The following two properties are largely straightforward.

PROPERTY 4.1. *If the design constants are chosen such that inequalities* (4.65) *hold, then we have*

$$\begin{aligned} \mathrm{d}V_\rho \le & -(y - y_d)^2\,\mathrm{d}t + \sigma_\rho\,\mathrm{d}w - \frac{\theta}{4}\sigma_\rho\sigma_\rho^\top\,\mathrm{d}t \\ & - l\big(y_d^{[\rho-1]}, \chi, z_{[\rho]}\big)\,\mathrm{d}t + R_l\,\mathrm{d}t + r(y_d)\mathcal{M}_\gamma(c)\mathrm{d}t, \end{aligned} \tag{4.77}$$

where $l = \overline{r}(y_d)\phi(\xi) + \sum_{i=1}^{\rho} \Xi_i \overline{\beta}_i z_i^2 + \Delta_\rho$ *is nonnegative.*

Proof. By (4.59) (or (4.35) for the case of $\rho = 1$) and (4.65) one can easily get (4.77). □

PROPERTY 4.2. *If design constants are chosen such that inequalities* (4.65) *hold, then*

$$\mathcal{L}V_\rho \le -z_1^2 + R_l, \tag{4.78}$$

$$\mathcal{L}V_\rho \le -c_1 V_\rho + c_2, \tag{4.79}$$

where constants c_1 *and* c_2 *satisfy*

$$\begin{cases} c_1 = \min\left\{ \frac{r}{\delta\lambda_{\max}(P)}, \overline{\beta}_1, \overline{\beta}_2, \dots, \overline{\beta}_\rho \right\} > 0, \\ c_2 = R_l + \max_{y_d} r_\rho(y_d)\mathcal{M}_\gamma(c). \end{cases} \tag{4.80}$$

Proof. When design constants are chosen such that inequalities (4.65) hold, then (4.78) comes from (4.77), and (4.79) comes from (4.61)–(4.64), $V_\rho = \phi(\xi) + \sum_{i=1}^{\rho} \Xi_i z_i^2$, and

$$\overline{r}(y_d) \ge c_1, \overline{\beta}_1 \ge c_1, \dots, \overline{\beta}_\rho \ge c_1, \quad \text{and} \quad C \le c_2. \qquad \square$$

5. Main results. In this section, we summarize the main results of this paper as a theorem.

THEOREM 5.1. *Consider the system* (3.1) *and the tracking risk-sensitive cost criterion* (3.2). *Suppose that assumptions* A1–A3 *hold. Then, for any given risk-sensitivity parameter* $\theta > 0$ *and desired cost value* $R_l > 0$, *there exists an output-feedback controller such that the closed-loop system*

1. *has a unique solution on* $[0, \infty)$ *almost surely,*
2. *admits a guaranteed cost value* R_l *for the risk-sensitive cost criterion* (3.2),
3. *is bounded in probability.*

Proof. We prove this theorem only for the case of $\rho > 1$ by construction. The proof for the case $\rho = 1$ is similar and straightforward, and so is omitted here.

For any given risk-sensitivity parameter $\theta > 0$ and desired cost value $R_l > 0$, section 4 provides a constructive design procedure of an output-feedback risk-sensitive controller. From Lemma 4.2, it is easily known that there are design constants such that inequalities (4.65) hold. Then, Lemma 4.3, (4.59), and the first two statements of Theorem A.1 imply statements 1 and 2.

Property 4.2, together with the third statement of Theorem A.1, leads directly to the boundedness in probability of $[\chi^\top, z_{[\rho]}^\top]^\top$. To show statement 3, let us first show the boundedness in probability of $[\eta_1, \dots, \eta_\rho]^\top$.

By $\eta_1 = y = z_1 + y_d$ and assumption A3, it is easy to see that η_1 is bounded in probability. Suppose that $[\eta_1, \dots, \eta_{k-1}]^\top$ is bounded in probability for k ($k = 2, \dots, \rho$). Then by $\eta_k = z_k + \alpha_{k-1}(y_d^{[k-1]}, \chi_a, \eta_{[k-1]})$, the smoothness of α_{k-1}, and assumption A3, we know that η_k, and hence $[\eta_1, \dots, \eta_k]^\top$, is bounded in probability. Therefore, by induction, $[\eta_1, \dots, \eta_\rho]^\top$ is bounded in probability.

Thus, by $[\eta_1, \dots, \eta_\rho, \zeta^\top]^\top = T_\rho, \dots, T_1[y, \widehat{x}_2, \dots, \widehat{x}_n]^\top$ and $\chi = [\zeta^\top, \widetilde{x}^\top]^\top$, it is easy to derive that $\widetilde{x}$ and $[y, \widehat{x}_2, \dots, \widehat{x}_n]^\top$ are bounded in probability. This, together with $y = x_1$, $\widehat{x}_1 = y - \widetilde{x}_1$, and $[x_2, \dots, x_n]^\top = [\widetilde{x}_2 + \widehat{x}_2, \dots, \widetilde{x}_n + \widehat{x}_n]$, leads to the boundedness in probability of $[x^\top, \widehat{x}^\top]^\top$. That is, statement 3 is true. □

Remark 5.1. As for the value range of characteristic parameter γ in value function V_ρ (or V_0 given by (4.10)), the following two points are considered. First, since χ_b is unknown, in order to guarantee stability of the closed-loop system, we use

$-\frac{r_0(y_d)\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t$ to dominate the term $M_1\chi_b z_1\,\mathrm{d}t$ on the right-hand side of (4.22). This requires that the power 2γ of χ_b (or χ) in the term $\frac{r_0(y_d)\|\chi\|^2}{(c+\chi^\top P\chi)^{1-\gamma}}\,\mathrm{d}t$ on the right-hand side of (4.22) be greater than 1, the power of χ_b of $M_1\chi_b z_1\,\mathrm{d}t$, that is, $\gamma > 1/2$. Second, to deal with the term $\sigma_0\,\mathrm{d}w$ on the right-hand side of (4.13), we use the technique of subtracting and adding $\frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t$ on the right-hand side of (4.13). The negative term $-\frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t$ is used to control the term $\sigma_0\,\mathrm{d}w$, while the positive term $\frac{\theta}{4}\sigma_0\sigma_0^\top\,\mathrm{d}t$ is dominated by term $-\frac{\partial\phi}{\partial\xi}\|\chi\|^2\,\mathrm{d}t$ and the system input. Thus, it is natural to require that the power 2γ of χ in $\frac{\partial\phi}{\partial\xi}\|\chi\|^2$ be greater than $4\gamma-2$, the power of χ in $\sigma_0\sigma_0^\top$; that is, $2\gamma > 4\gamma - 2$, or equivalently, $\gamma < 1$.

6. Example. Consider the second-order system

$$\begin{aligned}\mathrm{d}x_1 &= x_2\mathrm{d}t + u\mathrm{d}t + \frac{1}{2}y^2\mathrm{d}w,\\ \mathrm{d}x_2 &= u\,\mathrm{d}t,\\ y &= x_1.\end{aligned}$$

The purpose is to design u based on only y such that the output y of the closed-loop system tracks the sinusoidal signal:

$$y_d(t) = a\sin(\omega t), \quad a = 2,\ \omega = 2.$$

Clearly, in this case, we have $n = 2$, $m = 1$, $\rho = 1$, and $h(y) = [\frac{1}{2}y^2,\ 0]^\top$.

Design the following state observer:

$$\begin{aligned}\dot{\widehat{x}}_1 &= \widehat{x}_2 + k_1(y - \widehat{x}_1) + u, \quad k_1 = 1,\\ \dot{\widehat{x}}_2 &= k_2(y - \widehat{x}_1) + u, \qquad\quad k_2 = 1.\end{aligned}$$

Then, the estimation error $\widetilde{x} = [x_1 - \widehat{x}_1,\ x_2 - \widehat{x}_2]^\top$ satisfies the following equation:

$$\mathrm{d}\widetilde{x} = \begin{bmatrix} -1 & 1\\ -1 & 0\end{bmatrix}\widetilde{x}\mathrm{d}t + h(y)\mathrm{d}w \triangleq A\widetilde{x}\mathrm{d}t + h(y)\mathrm{d}w.$$

Set $\varsigma_0 = [y,\ \widehat{x}_2]^\top$. Then we have the following dynamical equation for ς_0:

$$\mathrm{d}\varsigma_0 = \begin{bmatrix} 0 & 1\\ 0 & 0\end{bmatrix}\varsigma_0\mathrm{d}t + \begin{bmatrix} 0\\ \widetilde{x}_1\end{bmatrix}\mathrm{d}t + \begin{bmatrix} 1\\ 0\end{bmatrix}\widetilde{x}_2\mathrm{d}t + \begin{bmatrix} 1\\ 1\end{bmatrix}u\,\mathrm{d}t + \begin{bmatrix} \frac{1}{2}y^2\\ 0\end{bmatrix}\mathrm{d}w.$$

Let $\varsigma_1 = T_1\varsigma_0$, $T_1 = \begin{bmatrix} 1 & 0\\ -1 & 1\end{bmatrix}$, $T_1^{-1} = \begin{bmatrix} 1 & 0\\ 1 & 1\end{bmatrix}$. Then we have

$$\mathrm{d}\varsigma_1 = \begin{bmatrix} 1 & 1\\ -1 & -1\end{bmatrix}\varsigma_1\mathrm{d}t + \begin{bmatrix} 0\\ \widetilde{x}_1\end{bmatrix}\mathrm{d}t + \begin{bmatrix} 1\\ -1\end{bmatrix}\widetilde{x}_2\mathrm{d}t + \begin{bmatrix} 1\\ 0\end{bmatrix}u\,\mathrm{d}t + \begin{bmatrix} \frac{1}{2}y^2\\ -\frac{1}{2}y^2\end{bmatrix}\mathrm{d}w.$$

Let $\eta = [\eta_1,\ \eta_2]^\top = \varsigma_1$ and $\zeta = \eta_2$. Then we have $\eta_1 = y$ and the following dynamics used to control design:

$$\begin{aligned}\mathrm{d}\widetilde{x} &= A\widetilde{x}\mathrm{d}t + h(y)\mathrm{d}w,\\ \mathrm{d}\zeta &= (-\zeta - y + \widetilde{x}_1 - \widetilde{x}_2)\mathrm{d}t - \frac{1}{2}y^2\,\mathrm{d}w,\\ \mathrm{d}y &= (\zeta + y + u + \widetilde{x}_2)\mathrm{d}t + \frac{1}{2}y^2\,\mathrm{d}w.\end{aligned}$$

In this case, we have

$$W = \begin{bmatrix} -1 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & -1 & 0 \end{bmatrix}.$$

By solving Lyapunov equation $W^\top P + PW = -I_3$, we get

$$P = \begin{bmatrix} 0.5000 & 0.3333 & -0.1667 \\ 0.3333 & 1.5000 & -0.6667 \\ -0.1667 & -0.6667 & 1.6667 \end{bmatrix}.$$

Clearly, P is symmetric and positive definite. The eigenvalues of P are 0.3978, 0.9476, 2.3213, and thus, $\lambda_{\min}(P) = 0.3978$, $\lambda_{\max}(P) = 2.3213$.

Let $\gamma = \frac{2}{3}$, $p = 2$, $p_1 = 4$, $z_1 = y - y_d$, $\chi_a = [\zeta, \widetilde{x}_1]^\top$, $\xi_a = \chi_a^\top (P_1 - P_2 P^{-1} P_2^\top)\chi_a = 0.4833\zeta^2 + 1.2333\widetilde{x}_1^2 + 0.5332\zeta\widetilde{x}_1$. Then we design the controller $u(y_d^{[1]}, \chi_a, \eta_1) = \alpha_1(y_d^{[1]}, \chi_a, \eta_1)$ as follows:

$$\alpha_1 = \left\{ -N_1 - \frac{z_1}{2\Xi_1} - \frac{N_0}{2\Xi_1} \right\}\Big|_{z_1 = \eta_1 - y_d}$$

with $\Xi_1 = \frac{4\kappa_1}{4 + y_d^4}$ and

$$\begin{aligned} N_0 = {} & \frac{0.2593\delta}{\varepsilon_{02}^2} \cdot \frac{1}{(c + \xi_a)^{\frac{1}{3}}} + \frac{4\theta\delta^2}{9} \cdot \frac{(y + y_d)^2 (0.5834\widetilde{x}_1 - 0.0833\zeta)^2 z_1}{(c + \xi_a)^{\frac{2}{3}}} \\ & + \frac{3.2261\theta\delta^2}{\varepsilon_{04}^2} \cdot \frac{(\zeta^2 + \widetilde{x}_1^2)\left(2y_d^4 + z_1^2 (y + y_d)^2\right)(y + y_d)^2 z_1}{c + \xi_a} \\ & + \frac{0.2993\theta\delta^2}{\varepsilon_{03}^2} \cdot (y + y_d)^2 (y^2 + y_d^2)^2 z_1 + \frac{\delta}{3} \cdot \frac{(y + y_d)^2 z_1}{(c + \xi_a)^{\frac{1}{3}}}, \\ N_1 = {} & \zeta - \dot{y}_d + \frac{\beta_1}{2} z_1 + \frac{z_1}{2\Xi_1} \frac{\partial \Xi_1}{\partial y_d} \dot{y}_d + \frac{2}{\varepsilon_1^4} \Xi_1^3 z_1^3 + \frac{\theta}{8} \Xi_1 \left(1 + \frac{y^4}{8\varepsilon_{11}^2}\right) y^4 z_1 \\ & + \frac{1}{4}(y + y_d)^2 z_1. \end{aligned}$$

Here, the desired cost value R_l is set to 0.5. Accordingly, the design constants in Lemma 4.2 are chosen as $\delta = 0.9$, $\theta = 0.2$, $\varepsilon_{01} = 0.3$, $\varepsilon_{02} = 0.3$, $\varepsilon_{03} = 0.32$, $\varepsilon_{04} = 0.68$, $c = 100$, $\varepsilon_1 = 0.1$, $\varepsilon_{11} = 0.29$, $\kappa_1 = 0.05$, $\beta_1 = 40$; the stochastic disturbance $\frac{\mathrm{d}w}{\mathrm{d}t}$ is chosen to be Gaussian white noise with power 1; and the initial conditions are simply set to $x_1(0) = 0.8$, $x_2(0) = 0$, $\widehat{x}_1(0) = 0$, $\widehat{x}_2(0) = 0$.

The simulation results are shown in Figures 1–4 given below. In particular, Figure 1 is about x_1 (solid line) and its estimation $\widehat{x}_1$ (dashdotted line); Figure 2 is about x_2 (solid line) and its estimation $\widehat{x}_2$ (dashdotted line); Figure 3 is about desired output y_d (solid line), system output y (dashdotted line), and tracking error $y - y_d$ (dashed line); Figure 4 is about control input u; Figure 5 gives a diagram of $\frac{1}{t}\int_0^t (y(s) - y_d(s))^2 \mathrm{d}s$, used to demonstrate the validity of the design. From Figure 3 and Figure 5 we can see that the system output tracks the desired output ideally.

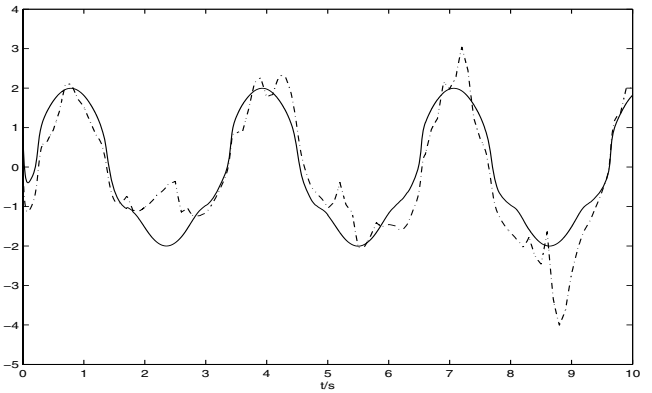


FIG. 1. *System state x_1 (solid) and observer state $\hat{x}_1$ (dashdotted).*

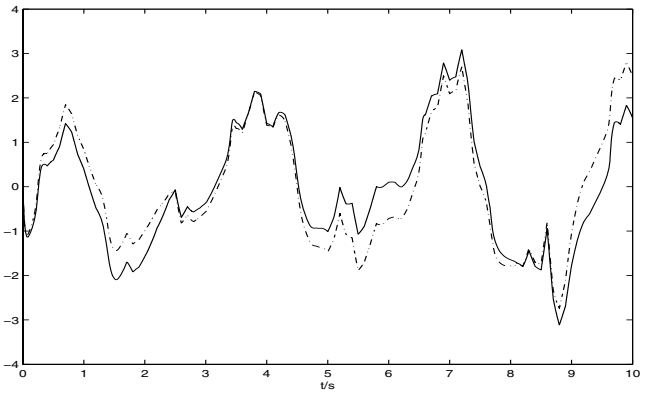


FIG. 2. *System state x_2 (solid) and observer state $\hat{x}_2$ (dashdotted).*

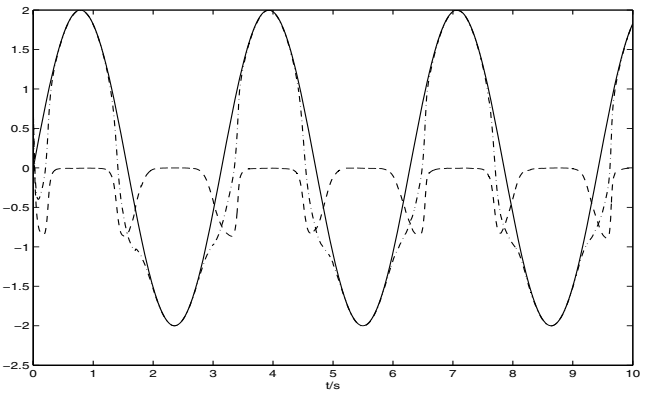


FIG. 3. *Desired output y_d (solid), system output y (dashdotted), and tracking error $y - y_d$ (dashed).*

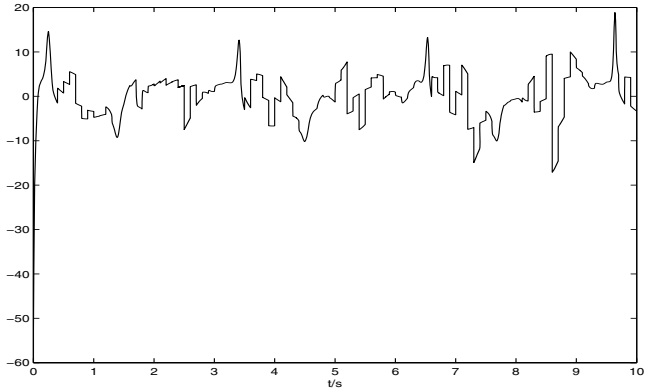


FIG. 4. *Control input* u.

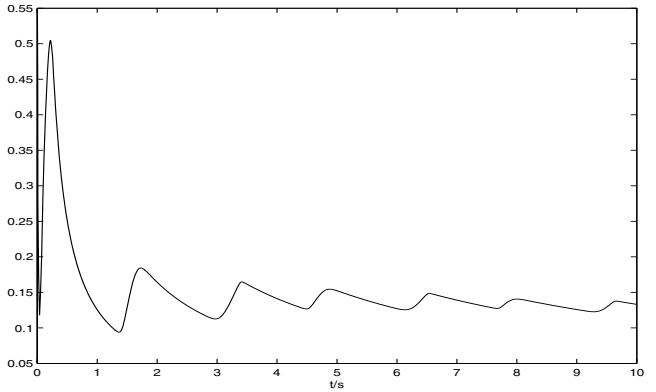


FIG. 5. $\frac{1}{t}\int_0^t (y(s) - y_d(s))^2 \mathrm{d}s$.

7. Concluding remarks. In this paper, the practical output-feedback control design problem of stochastic nonlinear strict-feedback systems in observer canonical form with stable zero-dynamics under a long-term tracking risk-sensitive cost criterion is investigated. A state observer is designed to guarantee an exponentially convergent state estimate when there is no disturbance. By introducing a state-transformation, we transform the system with the state observer in the loop into a lower triangular structure. And then, for any given risk-sensitivity parameter and desired cost value, by using an integrator backstepping method, we present constructively the output-feedback control design algorithm. The cost function adopted here is of quadratic form usually encountered in practice, rather than the quartic one used to avoid difficulty on controller design and performance analysis of the closed-loop systems. It is shown that under our control design (a) the closed-loop system is bounded in probability, and (b) the long-term average risk-sensitive cost of the closed-loop systems is upper bounded by the desired value. Besides, the value range of the characteristic parameters of the value function is investigated. As a special case when system vector nonlinearity and stochastic disturbance vector field vanish at the desired output y_d, it can be expected that there exists a control such that the closed-loop system is asymptotically stable in the large and admits a zero risk-sensitive cost. This question is now under study.

Appendix A. Preliminary results. In this appendix, we give the definitions of *bounded in probability* and *asymptotically stable in the large*, as well as a key theorem to present the sufficient conditions for these two stability notions.

For a general control-free stochastic nonlinear system,

$$\mathrm{d}x = f(t,x)\,\mathrm{d}t + h(t,x)\mathrm{d}w, \quad x(t_0) = x_0, \tag{A.1}$$

where x is an n-dimensional state vector, $n \in \mathbb{N}$; $f: [0,\ \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ and $h: [0,\ \infty) \times \mathbb{R}^n \to \mathbb{R}^{n\times s}$, $s \in \mathbb{N}$, are assumed to be continuous in t and locally Lipschitz in x; w is an s-dimensional vector-valued Brownian motion defined on a probability space $(\Omega,\ \mathcal{F},\ \mathcal{P})$; $t_0 \geq 0$ and $x_0 \in \mathbb{R}^n$. Denote the solution to (A.1) by $x_{t_0,x_0}(t)$.

Definition A.1. *The solution process* $\{x_{t_0,x_0}(t),\ t \geq t_0\}$ *is said to be* bounded in probability *if*

$$\lim_{\varepsilon\to\infty} \sup_{t\in[t_0,\,\infty)} \mathcal{P}\left\{\|x_{t_0,x_0}(t)\| > \varepsilon\right\} = 0.$$

Definition A.2. *Consider the system* (A.1)*, with* $f(t,\mathbf{0}_{n\times 1}) = \mathbf{0}_{n\times 1}$ *and* $h(t,\mathbf{0}_{n\times 1}) = \mathbf{0}_{n\times s}$ $\forall t \geq 0$*. The identically zero solution process is said to be* asymptotically stable in the large *if* $\forall \varepsilon > 0$*,* $t_0 \in [0,\ \infty)$*,*

$$\lim_{\|x_0\|\to 0^+} \mathcal{P}\left\{\sup_{t\geq 0} \|x_{t_0,x_0}(t)\| \geq \varepsilon\right\} = 0$$

and $\forall x_0 \in \mathbb{R}^n$*,* $\forall t_0 \in [0,\ \infty)$*,*

$$\mathcal{P}\left\{\lim_{t\to\infty} x_{t_0,x_0}(t) = \mathbf{0}_{n\times 1}\right\} = 1.$$

The following theorem gives the sufficient conditions for the above two stochastic stability concepts.

Theorem A.1. *Consider stochastic nonlinear system* (A.1) *and the following risk-sensitive cost criterion:*

$$J_\theta = \limsup_{T\to\infty} \frac{1}{T}\,\frac{2}{\theta}\ln\left(E\left(\exp\left(\frac{\theta}{2}\int_0^T q(t,x_{0,x_0}(t))\,\mathrm{d}t\right)\right)\right), \tag{A.2}$$

where $\theta > 0$ *is the risk-sensitive parameter and* $q: [0,\ \infty)\times\mathbb{R}^n \to \mathbb{R}$ *is a nonnegative continuous function. For any* $\theta > 0$ *and any desired cost value* $R_l > 0$*, if there exists a nonnegative value function* $V: [0,\ \infty)\times\mathbb{R}^n \to \mathbb{R}$*, which is* $\mathcal{C}^1$ *in the first argument and* $\mathcal{C}^2$ *in the second argument; a continuous function* $\sigma: [0,\ \infty)\times\mathbb{R}^n \to \mathbb{R}^{1\times s}$*; a nonnegative continuous function* $l: [0,\ \infty)\times\mathbb{R}^n \to \mathbb{R}$*; and a nonnegative, continuous, and radially unbounded function* $W_1: \mathbb{R}^n \to \mathbb{R}$ *such that*

$$W_1(x) \leq V(t,x) \ \ \forall (t,\,x) \in [0,\,\infty)\times\mathbb{R}^n, \tag{A.3}$$

$$\begin{aligned}\mathrm{d}V(t,\,x) = {}& \sigma(t,\,x)\,\mathrm{d}w - \frac{\theta}{4}\sigma(t,\,x)(\sigma(t,\,x))^\top\,\mathrm{d}t - l(t,\,x)\,\mathrm{d}t\\ & - q(t,\,x)\,\mathrm{d}t + R_l\,\mathrm{d}t \qquad \forall (t,\,x) \in [0,\,\infty)\times\mathbb{R}^n,\end{aligned} \tag{A.4}$$

then the following statements hold:

1. *The system* (A.1) *has a unique solution on* $[t_0,\,\infty)$ *almost surely* $\forall t_0 \geq 0$.

2. $J_\theta(x_0) \le R_l \ \forall x_0 \in \mathbb{R}^n$.
3. *If, in addition, there are constants* $c_{l1} \in (0, \infty)$ *and* $c_{l2} \in (0, \infty)$ *such that*

$$\mathcal{L}V(t,x) \le -c_{l1}V(t,\,x) + c_{l2} \quad \forall (t,\,x) \in [0,\,\infty) \times \mathbb{R}^n, \tag{A.5}$$

then the solution of system (A.1) *is bounded in probability.*

4. *If, in addition,* $f(t,\,\mathbf{0}_{n\times 1}) = \mathbf{0}_{n\times 1}$, $h(t,\,\mathbf{0}_{n\times 1}) = \mathbf{0}_{n\times s}\ \forall t \ge 0$, W_1 *is positive definite,* $R_l = 0$, *and there exist a continuous and positive definite function* $W_2 : \mathbb{R}^n \to \mathbb{R}$ *and a positive definite function* $W_3 : \mathbb{R}^n \to \mathbb{R}$ *such that*

$$V(t,\,x) \le W_2(x) \quad \forall (t,\,x) \in [0,\,\infty) \times \mathbb{R}^n, \tag{A.6}$$

$$\frac{\theta}{4}\sigma(t,\,x)(\sigma(t,\,x))^\top + l(t,\,x) + q(t,\,x) \ge W_3(x) \quad \forall (t,\,x) \in [0,\,\infty) \times \mathbb{R}^n, \tag{A.7}$$

then the zero solution of the system (A.1) *is asymptotically stable in the large.*

Proof. Define $\widehat{V} \triangleq V + R_l$. Clearly, $\widehat{V}$ is nonnegative and satisfies

$$\mathcal{L}\widehat{V} = -\frac{\theta}{4}\sigma(t,\,x)(\sigma(t,x))^\top - l(t,\,x) - q(t,\,x) + R_l \le \widehat{V}$$

and

$$\lim_{r\to\infty} \inf_{\|x\|>r} \widehat{V}(t,x) \ge \lim_{r\to\infty} \inf_{\|x\|>r} W_1(x) = \infty \quad \forall t \in [0,\,\infty).$$

Then, by Theorem 4.1 of Chapter III of [21], statement 1 follows.

For statement 2, fix $t_0 = 0$ and $x_0 \in \mathbb{R}^n$. By (A.4), we have

$$\begin{aligned} &V(T,\,x_{0,x_0}(T)) + \int_0^T (q(t,\,x_{0,x_0}(t)) + l(t,\,x_{0,x_0}(t)))\,\mathrm{d}t \\ &\le V(0,\,x_0) + \int_0^T \sigma(t,\,x_{0,x_0}(t))\,\mathrm{d}w \\ &\quad - \frac{\theta}{4}\int_0^T \sigma(t,\,x_{0,x_0}(t))(\sigma(t,x_{0,x_0}(t)))^\top\,\mathrm{d}t + R_l\,T \quad \forall T \ge 0. \end{aligned}$$

This implies that

$$\begin{aligned} &\frac{1}{T}\frac{2}{\theta}\ln\Bigg(E\Bigg(\exp\left(\frac{\theta}{2}\int_0^T q(t,\,x_{0,x_0}(t))\,\mathrm{d}t\right)\Bigg)\Bigg) \\ &\le \frac{1}{T}\frac{2}{\theta}\ln\Bigg(E\Bigg(\exp\left(\frac{\theta}{2}\left(V(T,\,x_{0,x_0}(T)) + \int_0^T \left(q(t,\,x_{0,x_0}(t)) + l(t,\,x_{0,x_0}(t))\right)\mathrm{d}t\right)\right)\Bigg)\Bigg) \\ &\le \frac{V(0,\,x_0)}{T} + \frac{1}{T}\frac{2}{\theta}\ln\Bigg(E\Bigg(\exp\left(\frac{\theta}{2}\left(\int_0^T \sigma(t,\,x_{0,x_0}(t))\,\mathrm{d}w\right.\right. \\ &\qquad\qquad \left.\left. - \frac{\theta}{4}\int_0^T \sigma(t,\,x_{0,x_0}(t))(\sigma(t,\,x_{0,x_0}(t)))^\top\,\mathrm{d}t\right)\right)\Bigg)\Bigg) + R_l. \end{aligned}$$

Let, $\forall T \ge 0$,

$$\zeta(T) \triangleq \exp\left(\int_0^T \frac{\theta}{2}\sigma(t,\,x_{0,x_0}(t))\,\mathrm{d}w - \frac{\theta^2}{8}\int_0^T \sigma(t,\,x_{0,x_0}(t))(\sigma(t,x_{0,x_0}(t)))^\top\,\mathrm{d}t\right).$$

Then $\zeta(T)$ is a supermartingale (see [33]), and $E(\zeta(T)) \le E(\zeta(0)) = 1\ \forall T \ge 0$. Thus, we have

$$J_\theta(x_0) \le \limsup_{T\to\infty}\left(\frac{V(0,\,x_0)}{T} + \frac{1}{T}\frac{2}{\theta}\ln\left(E\left(\zeta(T)\right)\right) + R_l\right) \le R_l.$$

This establishes statement 2.

For statement 3, fix $t_0 \in [0,\,\infty)$ and $x_0 \in \mathbb{R}^n$. Let

$$\alpha(c) = \inf_{\|x\|>c,\, x\in\mathbb{R}^n} W_1(x) \quad \forall c \in [0,\,\infty).$$

Then, by (A.3) and (A.5), we have, for sufficiently large $c \in [0,\,\infty)$ and any $t \ge t_0$,

$$\begin{aligned}
\mathcal{P}(\|x_{t_0,x_0}(t)\| > c) &= \int_\Omega I_{\{\|x_{t_0,x_0}(t)\|>c\}}(\omega)\mathcal{P}(d\omega)\\
&= \int_\Omega \frac{I_{\{\|x_{t_0,x_0}(t)\|>c\}}W_1(x_{t_0,x_0}(t))}{W_1(x_{t_0,x_0}(t))}\mathcal{P}(d\omega)\\
&\le \frac{E(W_1(x_{t_0,x_0}(t)))}{\alpha(c)} \le \frac{E(V(t,\,x_{t_0,x_0}(t)))}{\alpha(c)}\\
&\le \frac{V(t_0,\,x_0) + c_{l2}/c_{l1}}{\alpha(c)}.
\end{aligned}$$

Since the fact that W_1 is radially unbounded implies that $\alpha(c) \to \infty$ as $c \to \infty$, then statement 3 follows.

For statement 4, we note that V is clearly positive definite,

$$0 \le \lim_{x\to\mathbf{0}_{n\times 1}} \sup_{t\ge 0} V(t,\,x) \le \lim_{x\to\mathbf{0}_{n\times 1}} W_2(x) = 0,$$

which implies that V has infinitesimal upper limit, $\mathcal{L}V(t,\,x) = -\frac{\theta}{4}\sigma(t,\,x)(\sigma(t,\,x))^\top - l(t,\,x) - q(t,\,x) \le -W_3(x)\ \forall (t,\,x) \in [0,\,\infty)\times\mathbb{R}^n$, which is negative definite, and

$$\lim_{R\to\infty} \inf_{\|x\|>R,\, x\in\mathbb{R}^n} \inf_{t>0} V(t,\,x) \ge \lim_{R\to\infty} \inf_{\|x\|>R,\, x\in\mathbb{R}^n} W_1(x) = \infty.$$

By Theorem 4.4 in Chapter V of [21], the zero solution of system (A.1) is asymptotically stable in the large. ∎

Appendix B. Technical lemmas.

LEMMA B.1. *Let $n \in \mathbb{N}$, P be an $n \times n$-dimensional symmetric positive definite matrix, $\gamma \in (\frac{1}{2},\,1)$,*

$$\Pi_\gamma(x,c) = \|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c + x^\top Px)^{\gamma-1}\|x\|^2 \qquad \forall x \in \mathbb{R}^n,\ \forall c \in (0,\,\infty),$$

and

$$\mathcal{M}_\gamma(c) = \sup_{x\in\mathbb{R}^n} \Pi_\gamma(x,c) \quad \forall c \in (0,\,\infty).$$

Then $\Pi_\gamma(x,c) \ge 0\ \forall x \in \mathbb{R}^n,\ \forall c \in (0,\,\infty)$; and $\mathcal{M}_\gamma$ is strictly increasing on $(0,\,\infty)$; and

$$\lim_{c\to 0^+} \mathcal{M}_\gamma(c) = 0, \qquad \lim_{c\to+\infty} \mathcal{M}_\gamma(c) = +\infty.$$

Proof. Clearly, for any nonzero vector $x \in \mathbb{R}^n$, we have

$$\Pi(x,\, c) > \Pi(x,\, 0) = 0 \quad \forall c \in (0,\, \infty).$$

From this together with $\Pi(\mathbf{0}_{n\times 1},\, c) = 0\ \forall c \in (0,\, \infty)$, it follows that $\Pi_\gamma(x,c) \geq 0$ $\forall x \in \mathbb{R}^n$, $\forall c \in (0,\, \infty)$.

Let us next prove the properties of $\mathcal{M}_\gamma$. From the definitions of $\lambda_{\max}(P)$ and $\lambda_{\min}(P)$, it follows that for any nonzero vector $x \in \mathbb{R}^n$,

$$\lambda_{\min}(P) \,\leq\, \frac{x^\top P x}{\|x\|^2} \,\leq\, \lambda_{\max}(P).$$

Further, by $\gamma < 1$ we have for any nonzero vector $x \in \mathbb{R}^n$,

$$\begin{aligned}
\Pi_\gamma(x,c) \;&\leq\; \|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c+\lambda_{\max}(P)\|x\|^2)^{\gamma-1}\|x\|^2 \\
&=\; \lambda_{\max}^{\gamma-1}(P)\|x\|^{2\gamma}\left(1-\left(1+\frac{c}{\lambda_{\max}(P)\|x\|^2}\right)^{\gamma-1}\right) \\
&=\; \frac{\lambda_{\max}^{\gamma-1}(P)\|x\|^{2\gamma}}{\left(1+\frac{c}{\lambda_{\max}(P)\|x\|^2}\right)^{\gamma-1}}\left(\left(1+\frac{c}{\lambda_{\max}(P)\|x\|^2}\right)^{1-\gamma}-1\right) \\
&\leq\; \lambda_{\max}^{\gamma-1}(P)\|x\|^{2\gamma}\left(1+\frac{c(1-\gamma)}{\lambda_{\max}(P)\|x\|^2}-1\right) \\
&=\; c(1-\gamma)\lambda_{\max}^{\gamma-2}(P)\|x\|^{2(\gamma-1)} \\
&\longrightarrow 0 \qquad \text{as} \quad \|x\| \to \infty,
\end{aligned}$$

where we have used the following inequality: $(1+a)^r \leq 1+ar\ \forall a \in [0,\, \infty)$, $\forall r \in (0,\, 1)$.

Let

$$\mathcal{X}_0 = \{x \in \mathbb{R}^n : Px = \lambda_{\max}(P)x\}\,.$$

Then it can be shown that for any constant $c > 0$,

$$\mathcal{M}_\gamma(c) \,\geq \sup_{x\in\mathcal{X}_0,\|x\|=1} \Pi_\gamma(x,c) = \lambda_{\max}^{\gamma-1}(P) - (c+\lambda_{\max}(P))^{\gamma-1} \,>\, 0.$$

Therefore, there is a nonzero $x_1 \in \mathbb{R}^n$ at which $\Pi_\gamma(x,c)$ reaches its maximum. Furthermore, we can show that $x_1 \in \mathcal{X}_0$, since otherwise there would be $x_1^\top P x_1 < \lambda_{\max}(P)\|x_1\|^2$. Take $x_0 \in \mathcal{X}_0$ such that $\|x_0\| = \|x_1\|$. Then,

$$\begin{aligned}
\Pi_\gamma(x_0,c) &= \|x_0\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c+x_0^\top P x_0)^{\gamma-1}\|x_0\|^2 \\
&= \|x_0\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c+\lambda_{\max}(P)\|x_0\|^2)^{\gamma-1}\|x_0\|^2 \\
&= \|x_1\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c+\lambda_{\max}(P)\|x_1\|^2)^{\gamma-1}\|x_1\|^2 \\
&> \|x_1\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c+x_1^\top P x_1)^{\gamma-1}\|x_1\|^2.
\end{aligned}$$

This contradicts the fact that x_1 is the maximum point of $\Pi_\gamma(x,c)$.

Thus, there must be

$$\begin{aligned}
\mathcal{M}_\gamma(c) &= \sup_{x\in\mathbb{R}^n} \Pi_\gamma(x,c) = \sup_{x\in\mathcal{X}_0} \Pi_\gamma(x,c) \\
&= \sup_{x\in\mathcal{X}_0} \left[\|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c+\lambda_{\max}(P)\|x\|^2)^{\gamma-1}\|x\|^2\right].
\end{aligned}$$

In other words, the vector maximization problem has been transformed into a scalar one in α of the following two-variable function $f(\alpha, c)$:

$$f(\alpha, c) = \alpha^{\gamma}\lambda_{\max}^{\gamma-1}(P) - (c + \lambda_{\max}(P)\alpha)^{\gamma-1}\alpha.$$

That is, for any given $c \geq 0$, we have

$$\mathcal{M}_{\gamma}(c) = \sup_{\alpha \geq 0} f(\alpha, c).$$

Noticing that

$$\frac{\partial f(\alpha, c)}{\partial c} = (1-\gamma)(c + \lambda_{\max}(P)\alpha)^{\gamma-2}\alpha \ > \ 0 \quad \forall \alpha > 0, \ \forall c > 0,$$

we know that $f(\alpha, c)$ is a strictly increasing function of c for fixed $\alpha \in (0, \infty)$. It can also be shown that $\mathcal{M}_{\gamma}(c)$ is strictly increasing; i.e., for any $0 < c_1 < c_2$ there is always $0 < \mathcal{M}_{\gamma}(c_1) < \mathcal{M}_{\gamma}(c_2)$ due to the following argument. Let α_{c_1} maximize $f(\alpha, c_1)$, or $\mathcal{M}_{\gamma}(\alpha_{c_1}, c_1) = f(\alpha_{c_1}, c_1) = \sup_{\alpha \geq 0} f(\alpha, c_1)$. Then by the monotonicity of $f(\alpha, \cdot)$, we have

$$\mathcal{M}_{\gamma}(c_1) = f(\alpha_{c_1}, c_1) \ < \ f(\alpha_{c_1}, c_2) \ \leq \ \sup_{\alpha \geq 0} f(\alpha, c_2) = \mathcal{M}_{\gamma}(c_2).$$

Note that, $\forall c \in (0, \infty)$ and $\forall x \in \mathbb{R}^n$ with $x \neq \mathbf{0}_{n\times 1}$,

$$\begin{aligned}
\Pi_{\gamma}(x, c) &\leq \|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P) - (c + \lambda_{\max}(P)\|x\|^2)^{\gamma-1}\|x\|^2 \\
&= \frac{\|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P)}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}} \left((c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma} - \lambda_{\max}^{1-\gamma}(P)\|x\|^{2(1-\gamma)} \right) \\
&= \frac{\|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P)}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}} \left(\left(1 + \frac{c}{\lambda_{\max}(P)\|x\|^2} \right)^{1-\gamma} \lambda_{\max}^{1-\gamma}(P)\|x\|^{2(1-\gamma)} \right. \\
&\qquad \left. - \lambda_{\max}^{1-\gamma}(P)\|x\|^{2(1-\gamma)} \right) \\
&= \frac{\|x\|^2}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}} \left(\left(1 + \frac{c}{\lambda_{\max}(P)\|x\|^2} \right)^{1-\gamma} - 1 \right) \\
&< \frac{\|x\|^2}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}} \frac{c(1-\gamma)}{\lambda_{\max}(P)\|x\|^2} \\
&= \frac{c(1-\gamma)(\lambda_{\max}(P))^{-1}}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}}.
\end{aligned}$$

Clearly, we have $\Pi_{\gamma}(\mathbf{0}_{n\times 1}, c) = 0 < \frac{c(1-\gamma)(\lambda_{\max}(P))^{-1}}{(c)^{1-\gamma}}$ $\forall c \in (0, \infty)$. Then,

$$\Pi_{\gamma}(x, c) < \frac{c(1-\gamma)(\lambda_{\max}(P))^{-1}}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}} \quad \forall c \in (0, \infty), \ \forall x \in \mathbb{R}^n.$$

This implies that, $\forall c \in (0, \infty)$,

$$\mathcal{M}_{\gamma}(c) \ = \ \sup_{x \in \mathbb{R}^n} \Pi_g(x, c) \leq \sup_{x \in \mathbb{R}^n} \frac{c(1-\gamma)(\lambda_{\max}(P))^{-1}}{(c + \lambda_{\max}(P)\|x\|^2)^{1-\gamma}} \ = \ c^{\gamma}(1-\gamma)(\lambda_{\max}(P))^{-1}.$$

Clearly, $\mathcal{M}_{\gamma}(c) > 0$ $\forall c \in (0, \infty)$. Moreover, we have

$$0 \leq \lim_{c \to 0^+} \mathcal{M}_g(c) \leq \lim_{c \to 0^+} c^{\gamma}(1-\gamma)(\lambda_{\max}(P))^{-1} = 0.$$

We now show

$$\lim_{c\to+\infty} \mathcal{M}_\gamma(c) = +\infty, \tag{B.1}$$

since otherwise, by the fact that $\mathcal{M}_\gamma(c)$ is strictly increasing in $[0, \infty)$, $\lim_{c\to+\infty} \mathcal{M}_\gamma(c)$ would be existent and finite. Let $\Upsilon = \lim_{c\to+\infty} \mathcal{M}_\gamma(c)$. Then for $c = \left(1 - \frac{1}{2^{1-\gamma}}\right)^{-\frac{1}{\gamma}} \cdot (2\lambda_{\max}(P)\Upsilon)^{\frac{1}{\gamma}}$, we would have $\mathcal{M}_\gamma(c) \geq f(c\lambda_{\max}^{-1}(P), c) = 2\Upsilon > \Upsilon$. This contradicts $\Upsilon = \lim_{c\to+\infty} \mathcal{M}_\gamma(c)$ and the fact that $\mathcal{M}_\gamma(c)$ is strictly increasing in $[0, \infty)$. Thus, (B.1) is true. □

Remark B.1. Lemma B.1 means that the difference between term $\|x\|^{2\gamma}\lambda_{\max}^{\gamma-1}(P)$ and term $(c + x^\top Px)^{\gamma-1}\ \|x\|^2$ is less than or equal to $\mathcal{M}_\gamma(c)$, which is arbitrarily close to zero as constant c is.

LEMMA B.2. *Let $n \in \mathbb{N}$, $P \in \mathbb{R}^{n\times n}$ be symmetric and positive definite, $c \in [0, \infty)$, and $\gamma \in (\frac{1}{2}, 1)$. Define the function*

$$\Delta_\gamma(x, c) = (c + x^\top Px)^\gamma - c^\gamma - \lambda_{\max}^\gamma(P)\|x\|^{2\gamma}, \quad \forall c \geq 0,\ \forall x \in \mathbb{R}^n.$$

Then we have

$$\Delta_\gamma(x, c) \leq 0 \quad \forall x \in \mathbb{R}^n,\ \forall c \geq 0. \tag{B.2}$$

Proof. Let $z = \|x\|^2$ and $\overline{\Delta}_\gamma(z, c) = (c + \lambda_{\max}(P)z)^\gamma - c^\gamma - \lambda_{\max}^\gamma(P)z^\gamma$. Then, $\overline{\Delta}_\gamma(z, c) \geq \Delta_\gamma(x, c)\,\forall\, c \geq 0$, and for any $z > 0$ and $c \geq 0$,

$$\begin{aligned}\frac{\partial\overline{\Delta}_\gamma(z,\, c)}{\partial z} &= \frac{\gamma\lambda_{\max}(P)}{(c + \lambda_{\max}(P)z)^{1-\gamma}} - \frac{\gamma\lambda_{\max}^\gamma(P)}{z^{1-\gamma}} \\ &\leq \frac{\gamma\lambda_{\max}^\gamma(P)}{(c\lambda_{\max}^{-1}(P) + z)^{1-\gamma}} - \frac{\gamma\lambda_{\max}^\gamma(P)}{z^{1-\gamma}} \leq 0.\end{aligned}$$

This together with $\overline{\Delta}_\gamma(0, c) = \Delta_\gamma(0, c) = 0$ gives (B.2). □

Remark B.2. From Lemmas B.1 and B.2, we know that for any $\chi \in \mathbb{R}^{n+m}$ there exist the following inequalities:

$$\begin{aligned}\|\chi\|^{2\gamma} &\leq \lambda_{\max}^{1-\gamma}(P)\left(\mathcal{M}_\gamma(c) + \frac{\|\chi\|^2}{(c + \chi^\top P\chi)^{1-\gamma}}\right), \\ (c + \chi^\top P\chi)^\gamma - c^\gamma &\leq \lambda_{\max}(P)\left(\mathcal{M}_\gamma(c) + \frac{\|\chi\|^2}{(c + \chi^\top P\chi)^{1-\gamma}}\right).\end{aligned} \tag{B.3}$$

LEMMA B.3. *For any given constants a_1 and a_2 satisfying $1 \leq a_1 \leq a_2$, set $f_{a_1,\, a_2}(x) = x^{a_1} - x^{\alpha_2}\ \forall x \geq 0$, and*

$$\begin{cases}\mathcal{K}(a_1, a_2) = \frac{a_2 - a_1}{a_2}\left(\frac{a_1}{a_2}\right)^{\frac{a_1}{a_2 - a_1}} & \text{if } a_1 < a_2, \\ \mathcal{K}(a_1, a_2) = 0 & \text{if } a_1 = a_2.\end{cases}$$

Then

$$\sup_{x\geq 0} f_{a_1,\, a_2}(x) = \mathcal{K}(a_1, a_2). \tag{B.4}$$

Proof. When $1 \leq a_1 < a_2$, by $\frac{\mathrm{d}}{\mathrm{d}x} f_{a_1,\, a_2}(x) = a_1x^{a_1-1} - a_2x^{\alpha_2-1} = 0$, we see that $f_{a_1,\, a_2}(x)$ achieves its maximum at $x_m = (\frac{a_1}{a_2})^{\frac{1}{a_2-a_1}}$. Substituting x_m into $f_{a_1,\, a_2}(x)$ leads to (B.4).

When $a_2 = a_1 \geq 1$, (B.4) is obvious since, in this case, $f_{a_1,a_2}(x) \equiv 0$ and $\mathcal{K}(a_1, a_2) = 0$. □

LEMMA B.4. *For positive definite and radially unbounded functions $V_1 : \mathbb{R}^m \to \mathbb{R}$ and $V_2 : \mathbb{R} \to \mathbb{R}$, and a positive continuous function $\Xi : \mathbb{R}^m \to \mathbb{R}^+$, $m \in \mathbb{N}$, define $V(X, x) = V_1(X) + \Xi(X)V_2(x)$. Then, for any $[X^\top, x]^\top \neq \mathbf{0}_{(m+1)\times 1}$, $V(X, x) > 0$, and in addition, if $\sqrt{\|X\|^2 + x^2} \to \infty$, $V(X, x) \to \infty$. That is, $V(X, x)$ is positive definite and radially unbounded.*

Proof. We first show the positive definiteness of $V(X, x)$. From $[X^\top, x]^\top \neq 0_{(m+1)\times 1}$ we have either $X \neq 0_{m\times 1}$ or $X = 0_{m\times 1}$ and $x \neq 0$. If $X \neq 0_{m\times 1}$, then we have $V(X, x) \geq V_1(X) > 0$. If $X = 0_{m\times 1}$ and $x \neq 0$, then by the positiveness of Ξ we have $V(0_{m\times 1}, x) \geq \Xi(0_{m\times 1})V_2(x) > 0$. Clearly, $V(\mathbf{0}_{m\times 1}, 0) = 0$. Thus, $V(X, x)$ is positive definite.

Let us next show the radial unboundedness of $V(X, x)$ by contradiction. Suppose there were a sequence of $\{X_k, x_k, k \in \mathbb{N}\}$ satisfying $\lim_{k\to\infty}(\|X_k\| + \|x_k\|) = \infty$ and a constant $C > 0$ such that $V(X_k, x_k) \leq C < \infty \ \forall k \in \mathbb{N}$. Then, there would be $V_1(X_k) \leq C \ \forall k \in \mathbb{N}$ and $\Xi(X_k)V_2(x_k) \leq C \ \forall k \in \mathbb{N}$. Noticing the positive definiteness and radial unboundedness of V_1, one can show that there is a constant $\delta_1(C) > 0$ such that

$$\|X_k\| \leq \delta_1(C) \quad \forall k \in \mathbb{N}. \tag{B.5}$$

Let $M = \min_{\|X\|\leq\delta_1(C)} \Xi(X)$. Then, by the positiveness and continuity of Ξ we have $M > 0$. Thus, $V_2(x_k) \leq \frac{C}{M} \ \forall k \in \mathbb{N}$. This together with the positive definiteness and radial unboundedness of V_2 in turn implies that there exists a constant $\delta_2(C/M) > 0$ such that

$$\|x_k\| \leq \delta_2(C/M) \quad \forall k \in \mathbb{N}. \tag{B.6}$$

From this and (B.5) we have $\|X_k\| + \|x_k\| \leq \delta_1(C) + \delta_2(C/M) < \infty \ \forall k \in \mathbb{N}$, which contradicts $\lim_{k\to\infty}(\|X_k\| + \|x_k\|) = \infty$. Thus, $V(X, x)$ is radially unbounded. □

Acknowledgments. The authors wish to thank Professor Zigang Pan for helpful discussions on this topic. The authors are also indebted to the anonymous referees for their valuable comments.

REFERENCES

[1] G. ARSLAN AND T. BAŞAR, *Risk-sensitive adaptive trackers for strict-feedback systems with output measurements*, IEEE Trans. Automat. Control, 47 (2002), pp. 1754–1758.

[2] A. BENSOUSSAN, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.

[3] A. BENSOUSSAN, *General finite-dimensional risk-sensitive problems and small noise limits*, IEEE Trans. Automat. Control, 41 (1996), pp. 210–215.

[4] A. BENSOUSSAN AND J. H. VAN SCHUPPEN, *Optimal control of partially observable stochastic systems with an exponential-of-integral performance index*, SIAM J. Control Optim., 23 (1985), pp. 599–613.

[5] H. DENG AND M. KRSTIĆ, *Output-feedback stochastic nonlinear stabilization*, IEEE Trans. Automat. Control, 44 (1999), pp. 328–333.

[6] H. DENG AND M. KRSTIĆ, *Output-feedback stabilization of stochastic nonlinear systems driven by noise of unknown covariance*, Systems Control Lett., 39 (2000), pp. 173–182.

[7] H. DENG, M. KRSTIĆ, AND R. WILLIAMS, *Stabilization of stochastic nonlinear systems driven by noise of unknown covariance*, IEEE Trans. Automat. Control, 46 (2001), pp. 1237–1253.

[8] K. EZAL, Z. PAN, AND P. V. KOKOTOVIĆ, *Locally optimal backstepping design*, IEEE Trans. Automat. Control, 45 (2000), pp. 260–271.

[9] W. H. Fleming and W. M. McEneaney, *Risk-sensitive control and differential games*, in Stochastic Theory and Adaptive Control, Lecture Notes in Control and Inform. Sci. 184, Springer-Verlag, New York, 1992, pp. 185–197.
[10] W. H. Fleming and W. M. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.
[11] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.
[12] J. W. Helton and M. R. James, *Extending H^∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*, Adv. Des. Control 1, SIAM, Philadelphia, 1999.
[13] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, London, 1995.
[14] D. H. Jacobson, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1973), pp. 167–172.
[15] S. Jain and F. Khorrami, *Application of a decentralized adaptive output feedback based on backstepping to power systems*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, IEEE Press, Piscataway, NJ, pp. 1585–1590.
[16] M. R. James and J. S. Baras, *Partially observed differential games, infinite-dimensional Hamilton–Jacobi–Isaacs equations, and nonlinear H_∞ control*, SIAM J. Control Optim., 34 (1996), pp. 1342–1364.
[17] M. R. James, J. Baras, and R. J. Elliott, *Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 780–792.
[18] Z. P. Jiang and H. Nijmeijer, *A recursive technique for tracking control of nonholonomic systems in chained form*, IEEE Trans. Automat. Control, 44 (1999), pp. 265–279.
[19] Z. P. Jiang and J. B. Pomet, *Backstepping-based adaptive controllers for uncertain nonholonomic systems*, in Proceedings of the 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, IEEE Press, Piscataway, NJ, pp. 1573–1578.
[20] I. Kanellakopoulos, P. V. Kokotović, and A. S. Morse, *Systematic design of adaptive controllers for feedback linearizable systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1241–1253.
[21] R. Z. Kha'minslii, *Stochastic Stability of Differential Equations*, S & N International Publishers, Rockville, MD, 1980.
[22] P. Kokotović and M. Arcak, *Constructive nonlinear control: A historical perspective*, Automatica, 37 (2001), pp. 637–662.
[23] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear design of adaptive controllers for linear systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 738–752.
[24] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, John Wiley & Sons, New York, 1995.
[25] W. Lin and R. Pongvuthithum, *Adaptive output tracking of inherently nonlinear systems with nonlinear parameterizations*, IEEE Trans. Automat. Control, 48 (2003), pp. 1737–1749.
[26] Y. Liu, Z. Pan, and S. Shi, *Output feedback control design for strict-feedback stochastic nonlinear systems under a risk-sensitive cost criterion*, IEEE Trans. Automat. Control, 48 (2003), pp. 509–513.
[27] Y. Liu and J. Zhang, *Reduced-order observer-based control design for stochastic nonlinear systems*, Systems Control Lett., 52 (2004), pp. 123–135.
[28] Y. Liu and J. Zhang, *Minimal-order observer and output-feedback stabilization control design of stochastic nonlinear systems*, Science in China (Series F), 47 (2004), pp. 527–544.
[29] Y. Liu, J. Zhang, and Z. Pan, *Design of satisfaction output feedback controls for stochastic nonlinear systems under quadratic tracking risk-sensitive index*, Sciences in China (Series F), 46 (2003), pp. 126–145.
[30] R. Marino and P. Tomei, *Global adaptive output-feedback control of nonlinear systems, Part* I: *Linear parameterization*, IEEE Trans. Automat. Control, 38 (1993), pp. 17–32.
[31] H. Nagai, *Bellman equations of risk-sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.
[32] Z. Pan and T. Başar, *Adaptive controller design for tracking and disturbance attenuation in parametric-feedback nonlinear systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1066–1083.
[33] Z. Pan and T. Başar, *Backstepping controller design for nonlinear stochastic systems under a risk-sensitive cost criterion*, SIAM J. Control Optim., 37 (1999), pp. 957–995.
[34] Z. Pan, K. Ezal, A. Krener, and P. V. Kokotović, *Backstepping design with local optimality matching*, IEEE Trans. Automat. Control, 46 (2001), pp. 1014–1027.

[35] Z. Pan, Y. Liu, and S. Shi, *Output feedback stabilization for stochastic nonlinear systems in observer canonical form with stable zero-dynamics*, Science in China (Series F), 44 (2001), pp. 292–308.

[36] Z. Pan, Y. Liu, and S. Shi, *Output feedback stabilization for stochastic nonlinear systems in observer canonical form with stable zero-dynamics*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, IEEE Press, Piscataway, NJ, pp. 1392–1397.

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

[38] D. Seto, A. M. Annaswamy, and J. Baillieul, *Adaptive control of nonlinear systems with triangular structure*, IEEE Trans. Automat. Control, 39 (1994), pp. 1411–1428.

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

[40] P. Whittle, *Risk-Sensitive Optimal Control*, John Wiley & Sons, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 927–956

GLOBAL STABILIZATION OF THE GENERALIZED KORTEWEG–DE VRIES EQUATION POSED ON A FINITE DOMAIN*

LIONEL ROSIER† AND BING-YU ZHANG‡

Abstract. This paper is concerned with the internal stabilization of the generalized Korteweg–de Vries equation on a bounded domain. The global well-posedness and the exponential stability are investigated when the exponent in the nonlinear term ranges over the interval $[1, 4)$. The global exponential stability is obtained whatever the location where the damping is active, confirming positively a conjecture of Perla Menzala, Vasconcellos, and Zuazua [*Quart. Appl. Math.*, 60 (2002), pp. 111–129].

Key words. generalized Korteweg–de Vries equation, unique continuation property, smoothing effect, Carleman estimate

AMS subject classifications. 35Q20, 93D20, 93B

DOI. 10.1137/050631409

1. Introduction.

In this paper we study the generalized Korteweg–de Vries (GKdV) equation posed on a finite interval $I = (0, L)$ with a localized damping

$$\partial_t u + \partial_x u + a(u)\partial_x u + \partial_x^3 u + b(x)u = 0, \quad 0 \le x \le L, \quad t \ge 0, \tag{1.1}$$

satisfying the homogeneous boundary conditions

$$u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0, \quad t \ge 0, \tag{1.2}$$

and the initial condition

$$u(x,0) = \phi(x), \qquad 0 \le x \le L. \tag{1.3}$$

Here $b \equiv b(x) \in L^2(I)$ is a given nonnegative function with its support ω being a subset of I, and the function $a \equiv a(\mu)$ is a given smooth function satisfying the growth condition

$$a(0) = 0, \qquad |a^{(j)}(\mu)| \le C(1 + |\mu|^{p-j}) \qquad \forall\ \mu \in \mathbb{R} \tag{1.4}$$

for $j = 0, 1$ if $1 \le p < 2$ and for $j = 0, 1, 2$ if $p \ge 2$.

We are mainly concerned with two issues regarding the initial boundary value problem (IBVP) (1.1)–(1.3). One is the well-posedness of the IBVP (1.1)–(1.3) in the classical Sobolev space $H^s(I)$. That is, we aim to establish existence, uniqueness, and persistence properties of solutions corresponding to the given initial data ϕ, together with continuous dependence of solution upon the initial data ϕ. The other one is the long time behavior of solutions of the IBVP (1.1)–(1.3). More precisely, we will

*Received by the editors May 12, 2005; accepted for publication (in revised form) April 24, 2006; published electronically July 17, 2006. This work was partially supported by the Charles Phelps Taft Memorial Fund at the University of Cincinnati.

http://www.siam.org/journals/sicon/45-3/63140.html

†Institut Élie Cartan, UMR 7502 UHP/CNRS/INRIA, B.P. 239, F-54506 Vandœuvre-lès-Nancy Cedex, France (rosier@iecn.u-nancy.fr).

‡Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221 (bzhang@math.uc.edu).

investigate if the solution of (1.1)–(1.3) tends to zero as $t \to \infty$ and under what rate it decays.

When $a(\mu) = \mu$ and $b \equiv 0$, (1.1) is the celebrated Korteweg–de Vries (KdV) equation

$$\partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = 0, \tag{1.5}$$

which was derived by Korteweg and de Vries [18] in 1895 as a model for propagation of surface water waves along a channel. The equation has been studied extremely intensively from various aspects of both mathematics and physics since the early 1960s when *soliton* was discovered through the KdV equation and the *inverse scattering transform*, a so-called nonlinear Fourier transform, was invented to solve the initial value problem of the KdV equation (cf. [19, 21]). In particular, the pure initial value problem for the KdV equation posed on the whole line $\mathbb{R}$ or on a periodic domain $\mathbb{S}$ has received a lot of attention in the past three decades for its well-posedness problem in the classical Sobolev space $H^s(\mathbb{R})$ or $H^s(\mathbb{S})$ (see [1, 2, 5, 6, 10, 13, 14, 15, 16] and the references therein). So far, the best known results are that the pure initial value problem of the KdV equation, when posed on the real line $\mathbb{R}$, is well-posed in the space $H^s(\mathbb{R})$ for $s > -\frac{3}{4}$ [16, 10] and is, when posed on the periodic domain $\mathbb{S}$, well-posed in the space $H^s(\mathbb{S})$ for $s \geq -1$ [12].

For the KdV equation posed on a finite interval, the study of its IBVP began with Bubnov [7], who investigated the general two-point boundary value problem

$$\begin{cases} \partial_t u + u\partial_x u + \partial_x^3 u = f(x,t), \qquad u(x,0) = 0, \\ \alpha_1 \partial_x^2 u(0,t) + \alpha_2 \partial_x u(0,t) + \alpha_3 u(0,t) = 0, \\ \beta_1 \partial_x^2 u(1,t) + \beta_2 \partial_x u(1,t) + \beta_3 u(1,t) = 0, \\ \xi_1 \partial_x u(1,t) + \xi_2 u(1,t) = 0 \end{cases} \tag{1.6}$$

posed on the interval $(0,1)$. Here $\alpha_j,\ \beta_j,\ \xi_i \in \mathbb{R},\ j = 1,2,3,\ i = 1,2$, are constants, and assumptions are imposed so that the L^2-norm of the solutions of the linear version of (1.6) (obtained by dropping the nonlinear term $u\partial_x u$) is decreasing. It was shown in [7] that for given $T > 0$ and $f \in H^1([0,T]; L^2(0,1))$, there exists a $T^* > 0$ depending on $\|f\|_{H^1([0,T];L^2(0,1))}$ such that (1.6) admits a unique solution

$$u \in L^2([0,T^*]; H^3(0,1)), \qquad u_t \in L^\infty([0,T^*]; L^2(0,1)) \cap L^2([0,T^*]; H^1(0,1)).$$

In [4], Bona, Sun, and Zhang studied the following nonhomogeneous boundary value problem for the KdV equation posed on the interval $(0,1)$:

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = 0, \qquad u(x,0) = \phi(x), \\ u(0,t) = h_1(t), \quad u(1,t) = h_2(t), \quad \partial_x u(1,t) = h_3(t). \end{cases} \tag{1.7}$$

The IBVP (1.7) was shown in [4] to be well-posed in the space $H^s(0,1)$ for $s \geq 0$ with $h_j \in H^{(s+1)/3}_{loc}(\mathbb{R}^+)$, $j = 1,2$, and $h_3 \in H^{s/3}_{loc}(\mathbb{R}^+)$. Earlier, in the case of $h_j = 0$, $j = 1,2,3$, Zhang [38] showed that the IBVP (1.7) is well-posed in the space $H^{3k+1}(0,1)$ for $k = 0,1,2,\ldots$, and Perla Menzala, Vasconcellos, and Zuazua [24] showed that the IBVP (1.7) is well-posed in the space $L^2(0,1)$. The reader is referred

to [3, 9, 11] for more references on the IVBP for the KdV equation posed either on a finite interval or on the half-line $\mathbb{R}^+$.

The study of long time behavior of solutions of the KdV equation posed on a finite interval was also started by Bubnov [8]. In [17], Komornik, Russell, and Zhang considered the stabilization problem for the KdV equation posed on finite interval $(0,\pi)$ with the periodic boundary conditions

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = f, \qquad u(x,0) = \phi(x), \\ \\ \partial_x^j u(0,t) = \partial_x^j u(\pi,t), \qquad j = 0,1,2, \end{cases} \tag{1.8}$$

where $f \equiv f(x,t)$ is considered as a control input acting on the whole domain $(0,\pi)$. A special feedback control law $f = \mathcal{B}u$ is designed to ensure the conservation of mass, and the resulting closed loop system

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = \mathcal{B}u, \qquad u(x,0) = \phi(x), \\ \\ \partial_x^j u(0,t) = \partial_x^j u(\pi,t), \qquad j = 0,1,2, \end{cases} \tag{1.9}$$

is demonstrated to be exponentially stable; its solution converges exponentially to the average value of its initial datum ϕ over the domain $(0,\pi)$ as $t \to \infty$. The same problem was studied by Russell and Zhang [29, 31], assuming that the control input f acted only on an open subdomain of $(0,\pi)$. The resulting closed loop system is shown to be locally exponentially stable in the sense that the initial data ϕ is required to be small in a certain sense. In [30], Russell and Zhang studied boundary stabilization of the KdV equation posed on a finite interval $(0,\pi)$ with the periodic boundary conditions. The resulting closed loop system appears in the following form:

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = 0, \qquad u(x,0) = \phi(x), \\ \\ u(0,t) = u(\pi,t), \quad \partial_x^2 u(0,t) = \partial_x^2 u(\pi,t), \quad \partial_x u(\pi,t) = \alpha \partial_x u(0,t), \end{cases} \tag{1.10}$$

where $-1 < \alpha < 1$ is a constant. The system was shown by Russell and Zhang [30] to be locally exponentially stable when $\alpha \neq -\frac{1}{2}$. For the exceptional case $\alpha = -\frac{1}{2}$, the system (1.10) was also shown to be locally exponentially stable later by Sun [34] via a different approach. A similar boundary stabilization problem was studied by Zhang [38] for the KdV equation posed on a finite interval $(0,1)$ with the Dirichlet boundary condition. The resulting closed loop system is of the form

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = 0, \qquad u(x,0) = \phi(x), \\ \\ u(0,t) = 0, \quad u(1,t) = 0, \quad \partial_x u(1,t) = \gamma \partial_x u(0,t) \end{cases} \tag{1.11}$$

with $0 \leq |\gamma| < 1$. When $0 < |\gamma| < 1$, the system (1.11) was shown to be locally exponentially stable.

For the KdV equation posed on a finite interval $(0,L)$ with the homogeneous boundary conditions,

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u = 0, \quad 0 \leq x \leq L, \quad t \geq 0, \\ \\ u(x,0) = \phi(x), \\ \\ u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0; \end{cases} \tag{1.12}$$

it is easily seen that any smooth solution u of (1.12) satisfies

$$\frac{d}{dt}\int_0^L |u(x,t)|^2 dx = -|\partial_x u(0,t)|^2.$$

This leads one to guess that any solution u of (1.12) may decay to zero as $t \to \infty$. However, this may not always be the case. In [25], Rosier discovered that if the length L of the domain $(0, L)$ belongs to the set

$$\mathcal{E} = \left\{ \frac{2\pi}{\sqrt{3}}\sqrt{k^2+kl+l^2},\ k \text{ and } l \text{ are positive integers} \right\},$$

then the linear system associated with the IBVP (1.12) possesses a solution with a constant L^2-norm. It is thus reasonable to say that not all solutions of (1.12) decay to 0 as $t \to \infty$.

If the length L of the interval I does not belong to the set $\mathcal{E}$, it has been demonstrated in [24] by Perla Menzala, Vasconcellos, and Zuazua that there exist $\delta > 0$ and $\gamma > 0$ such that if $\phi \in L^2(I)$ with

$$\|\phi\|_{L^2(I)} \le \delta,$$

then

$$\|u(\cdot,t)\| \le C\|\phi\|_{L^2(I)} e^{-\gamma t} \qquad \forall\, t \ge 0,$$

where C depends only on $\|\phi\|_{L^2(I)}$.

In order to handle the case of $L \in \mathcal{E}$ and to have the solutions of (1.12) with large amplitude stabilized, Perla Menzala, Vasconcellos, and Zuazua [24] introduced an extra damping term $b(x)u$ to the equation in (1.12) to get the following system:

$$\begin{cases} \partial_t u + \partial_x u + u\partial_x u + \partial_x^3 u + b(x)u = 0, \quad 0 \le x \le L, \quad t \ge 0, \\ u(x,0) = \phi(x), \\ u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0. \end{cases} \tag{1.13}$$

Here $b \in L^\infty(I)$ is assumed to be a nonnegative function satisfying $b(x) \ge b_0 > 0$ a.e. in an open, nonempty subset ω of I. Perla Menzala, Vasconcellos, and Zuazua [24] showed the following theorem.

THEOREM 1.1 (Perla Menzala, Vasconcellos, and Zuazua). *For any given $\phi \in L^2(I)$, the IBVP* (1.13) *admits a unique solution $u \in C(\mathbb{R}^+; L^2(I)) \cap L^2_{loc}(\mathbb{R}^+; H^1(I))$. Moreover, if*

$$\omega \text{ contains two sets of the form } (0,\delta) \text{ and } (L-\delta, L) \quad \text{for some } \delta > 0, \tag{1.14}$$

then, for any $L > 0$ and $N > 0$, there exist $C > 0$ and $\mu > 0$ such that for any $\phi \in L^2(I)$ with $\|\phi\|_{L^2(I)} \le N$, the corresponding solution u of (1.13) *satisfies*

$$\|u(\cdot,t)\|_{L^2(I)} \le C\|\phi\|_{L^2(I)} e^{-\mu t} \qquad \forall\, t \ge 0. \tag{1.15}$$

Remarks.

(a) The result of Perla Menzala, Vasconcellos, and Zuazua presented in Theorem 1.1 represents a significant advance in the subject of stabilization of the KdV equation. Indeed, all the previous results except [17], in which the feedback control acts on the whole domain, are local in the sense that only *small amplitude* solutions have been shown to decay exponentially; they are essentially linear stability results. In contrast, the stability result presented in Theorem 1.1 is global; all solutions of (1.13), large or small, decay exponentially in the space $L^2(I)$. It is truly a nonlinear stability result.

(b) Perla Menzala, Vasconcellos, and Zuazua have conjectured in [24] that Theorem 1.1 still holds without the assumption (1.14). Pazoto [22] has proved that this is indeed true. The idea of the proof is as follows: Pazoto shows that the unique continuation property holds, whatever ω is, in proving by multipliers and compactness arguments that any solution vanishing on a subinterval is necessarily smooth.

(c) In Theorem 1.1, the determination of the decay rate ν depends on the size of the initial value ϕ in the space $L^2(I)$. The system (1.12) is locally uniformly exponentially stable in $L^2(I)$.

In this paper, motivated by Perla Menzala, Vasconcellos, and Zuazua's work, we will consider the IBVP of the GKdV equation as described by (1.1)–(1.3) for its well-posedness and long time behavior of its solutions. We will first extend the well-posedness result established in [4] for the IBVP (1.7) for the KdV equation to the IBVP (1.1)–(1.3) for the GKdV equation.

THEOREM 1.2. *Suppose the growth condition* (1.4) *for* $a = a(\mu)$ *is satisfied with*

$$1 \le p < 2.$$

Let $0 \le s \le 3$ *and* $T > 0$ *be given. Then for any given* $\phi \in H^s(I)$ *satisfying the* s*-compatibility conditions*

$$\begin{cases} \phi(0) = \phi(L) = 0 \quad \textit{when } 1/2 < s \le 3/2, \\ \\ \phi(0) = \phi(L) = \phi'(L) = 0 \quad \textit{when } 3/2 < s \le 3, \end{cases} \tag{1.16}$$

the IBVP (1.1)–(1.3) *admits a unique solution*

$$u \in C([0,T]; H^s(I)) \cap L^2(0,T; H^{s+1}(I)).$$

Moreover, the solution u *depends on its initial value continuously in the corresponding spaces.*[1]

Remarks.

(i) We may extend the theorem to include the nonhomogeneous boundary conditions as in [4]. But we choose not to do so here since our main concern is the long time behavior of solutions of (1.1)–(1.3).

(ii) The property that $\phi \in H^s(I)$ implies that the corresponding solution $u \in L^2(0,T; H^{s+1}(I))$ reveals a type of global Kato smoothing effect of the GKdV equation posed on a finite interval. This global Kato smoothing effect together with the property of persistence of regularity presented in Theorem 1.2 yields the following strong smoothing properties for solutions of (1.1)–(1.3):

$$\phi \in L^2(I) \Longrightarrow \ u \in C([\epsilon,T]; H^3(I)) \cap L^2(\epsilon,T; H^4(I)) \quad \text{for any } \epsilon > 0.$$

[1] If a is a real analytic function, then the solution u depends on its initial value analytically in the corresponding spaces (cf. [4, 39]).

In the case of $p \geq 2$, we have the following local well-posedness result.

THEOREM 1.3. *Suppose the growth condition* (1.4) *for* $a = a(\mu)$ *is satisfied with*

$$p \geq 2.$$

Then for any given $\phi \in H^s(I)$ *with* $s = 3$ *or with* 3 *satisfying the* s*-compatibility condition* (1.16) *there exists a* $T^* > 0$ *depending on* $\|\phi\|_{H^s(I)}$ *such that the IBVP* (1.1)–(1.3) *admits a unique solution* $u \in C([0,T^*]; H^s(I)) \cap L^2(0,T^*; H^{s+1}(I))$ *which depends on its initial value continuously.*

In order to get the global well-posedness results in the space $H^1(I)$ or $H^3(I)$, one needs some a priori H^1- or H^3-global estimate. However, those a priori estimates are not available. In fact, the only available a priori estimate for the IBVP (1.1)–(1.3) is an L^2-a priori estimate which is not sufficient to yield a global well-posedness result because we do not know if the IBVP (1.1)–(1.3) is locally well-posed in the space $L^2(I)$. Whether Theorem 1.3 can be extended to a global well-posedness result at least in the case of $2 \leq p < 4$ remains an interesting open question. On the other hand, when $2 \leq p < 4$, the following a priori estimates are known to hold for solutions of (1.1)–(1.3) (cf. Lemma 2.5 in section 2):

$$\|u(\cdot,T)\|^2_{L^2(I)} + \int_0^T (\partial_x u)^2(0,\tau)d\tau + \int_0^T \int_0^L b(x)u^2(x,\tau)dxd\tau = \|\phi\|^2_{L^2(I)}$$

and

$$\int_0^T \int_0^L |\partial_x u|^2 \, dxdt \leq \frac{L+(C+1)T}{2}||\phi||^2_{L^2(I)} + C_p T||\phi||^{\frac{8+2p}{4-p}}_{L^2(I)}.$$

Taking advantage of these estimates, we are able to establish the global existence of solutions of (1.1)–(1.3) in the space $L^2(I)$.

THEOREM 1.4. *Suppose the growth condition* (1.4) *for* $a = a(\mu)$ *is satisfied with*

$$1 \leq p < 4.$$

Then for any given $\phi \in L^2(I)$*, the IBVP* (1.1)–(1.3) *admits a solution*

$$u \in C_w(\mathbb{R}^+; L^2(I)) \cap L^2_{loc}(\mathbb{R}^+; H^1(I)).$$

Note that when $1 \leq p < 2$, a solution given by Theorem 1.4 is identical to the unique solution provided by Theorem 1.2.

With the well-posedness results in hand, we may then investigate the long time behavior of solutions. The following theorem is the main result of this paper.

THEOREM 1.5. *Under the assumptions of Theorem* 1.4*, assume additionally that the support* ω *of the function* b *in* (1.1) *contains a nonempty open subset of* $(0,L)$*. Then there exists a number* $\nu > 0$ *depending only on* L *and a nondecreasing continuous function* $\beta : \mathbb{R}^+ \to \mathbb{R}^+$ *such that for a given* $\phi \in L^2(I)$*, any solution* u *of* (1.1)–(1.3) *provided by Theorem* 1.4 *satisfies*

$$\|u(\cdot,t)\|_{L^2(I)} \leq \beta\big(\|\phi\|_{L^2(I)}\big)e^{-\nu t} \qquad \forall\ t \geq 0. \tag{1.17}$$

Remarks.

(1) Theorem 1.5 confirms positively the conjecture of Perla Menzala, Vasconcellos, and Zuazua [24] for the GKdV equation.

(2) The decay rate ν in Theorem 1.5 depends only on L and does not, in particular, depend on the size of the initial value ϕ in the space $L^2(I)$. The system (1.1)–(1.3) is globally uniformly exponentially stable.

Theorem 1.5 will be proved by the same compactness-uniqueness argument used by Perla Menzala, Vasconcellos, and Zuazua [24]. The key is to establish the following unique continuation property for the solution of the GKdV equation.

Unique continuation property. Let ω be the support of b in the interval I. If $v \in L^\infty(0,T;L^2(I)) \cap L^2(0,T;H^1(I))$ solves

$$\begin{cases} \partial_t v + \partial_x v + a(v)\partial_x v + \partial_x^3 v = 0, \quad 0 \le x \le L, \quad t \ge 0, \\ \\ v(0,t) = 0, \quad v(L,t) = 0, \quad \partial_x v(L,t) = 0 \end{cases}$$

and, in addition, satisfies

$$\partial_x v(0,t) = 0 \qquad \forall\ t \in (0,T)$$

and

$$v(x,t) = 0 \qquad \forall\ (x,t) \in \omega \times (0,T),$$

then v vanishes on $I \times (0,T)$.

When $1 \le p < 2$, the solution v, in fact, belongs to the space $C([\epsilon,T];H^3(I))$ for any $0 < \epsilon < T$ thanks to the strong smoothing property of the GKdV equation (cf. Remark (ii) of Theorem 1.2). Thus v vanishes on the whole domain $I \times (0,T)$ by the standard unique continuation property of the KdV-type equation (cf. [32, 37]). When $2 \le p < 4$, we cannot use the standard unique continuation property because we do not know if $v \in C([\epsilon,T];H^3(I))$. Instead, we will use a new unique continuation property for the GKdV equation as described below.

THEOREM 1.6. *Let $0 < l < L$ and $T > 0$. If $w \in L^\infty(0,T;H^1(0,l))$ solves*

$$\begin{cases} \partial_t w + \partial_x w + a(w)\partial_x w + \partial_x^3 w = 0 & \text{in } (0,l) \times (0,T), \\ w(0,t) = 0 & \text{for a.e. } t \in (0,T), \\ w \equiv 0 & \text{in } (l',l) \times (0,T) \end{cases}$$

with $a \in C^0(\mathbb{R};\mathbb{R})$ and $0 < l' < l$, then $w \equiv 0$ in $(0,l) \times (0,T)$.

The proof of this unique continuation property is mainly based on a key Carleman estimate for the KdV equation established earlier by Rosier [27] (see also Lemma 3.4 in section 3).

When $1 \le p < 2$, for a given $\phi \in L^2(I)$, (1.1)–(1.3) admits a unique solution u which not only decays exponentially in the space $L^2(I)$ but also in the space $H^3(I)$ as described by the following theorem.

THEOREM 1.7. *Under the assumption of Theorem* 1.2, *assume additionally that the support ω of the function b in* (1.1) *contains a nonempty open subset of $(0,L)$. Then there exists a number $\nu > 0$ depending only on L and a continuous function $\alpha : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ such that for any $\varepsilon > 0$ and any $\phi \in L^2(I)$, the corresponding solution u of* (1.1)–(1.3) *satisfies*

$$\|u(\cdot,t)\|_{H^3(I)} \le \alpha\big(\|\phi\|_{L^2(I)}, \epsilon\big) e^{-\nu t} \qquad \forall\ t \ge \epsilon. \tag{1.18}$$

The paper is organized as follows. In section 2, we study the well-posedness problem of the IBVP (1.1)–(1.3) and provide the proofs of Theorems 1.2, 1.3, and 1.4. In section 3, the long time behavior of solutions of (1.1)–(1.3) is investigated.

2. Well-posedness. In this section, attention will be given to the nonlinear IBVP

$$\begin{cases} \partial_t u + \partial_x u + a(u)\partial_x u + \partial_x^3 u + b(x)u = 0, \quad u(x,0) = \phi(x), \quad x \in (0,L), \\ \\ u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0 \end{cases} \tag{2.1}$$

for its well-posedness in the classical Sobolev space $H^s(I)$.

Considered first is the linear problem

$$\begin{cases} \partial_t u + \partial_x u + \partial_x^3 u = 0, \quad u(x,0) = \phi(x)\,, \quad x \in (0,L), \\ \\ u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0 \end{cases} \tag{2.2}$$

with homogeneous boundary conditions and no forcing. Let A be the linear operator defined by

$$Af = -f''' - f'$$

with the domain

$$\mathcal{D}(A) = \{f \in H^3(I),\ f(0) = f(L) = f'(L) = 0\}.$$

The IBVP (2.2) can be written as the initial value problem of an abstract evolution equation in the space $L^2(I)$, namely

$$\frac{du}{dt} = Au, \qquad u(0) = \phi, \tag{2.3}$$

where the spatial variable is suppressed. It is easily verified that both A and its adjoint A^* are dissipative, i.e.,

$$\Big\langle Af, f \Big\rangle_{L^2(I)} \le 0, \qquad \Big\langle A^* g, g \Big\rangle_{L^2(I)} \le 0$$

for any $f \in \mathcal{D}(A)$ and $g \in \mathcal{D}(A^*)$, where $A^* g = g''' + g'$ and

$$\mathcal{D}(A^*) = \{f \in H^3(0,L);\ f(0) = f'(0) = f(L) = 0\}.$$

Thus the operator A is the infinitesimal generator of a C_0-semigroup $W(t)$ in the space $L^2(I)$ (see [23]). By semigroup theory (see [23]), in this situation in the overlying space $L^2(I)$, for any $\phi \in L^2(I)$,

$$u(t) = W(t)\phi$$

belongs to the space $C_b(\mathbb{R}^+; L^2(I))$. The function u thus defined is called a *mild solution* of (2.2). Such solutions certainly solve (2.2) in the sense of distributions. If $\phi \in \mathcal{D}(A)$, then $u(t) = W(t)\phi$ belongs to the space $C(\mathbb{R}^+; H^3(I)) \cap C^1(\mathbb{R}^+; L^2(I))$ and $u(t) \in \mathcal{D}(A)$ for all $t \ge 0$. Moreover, the equation is satisfied in the sense of $C(\mathbb{R}^+; L^2(I))$ and, in particular, pointwise a.e. Such solutions are called *strong solutions*.

LEMMA 2.1. *For any $\phi \in L^2(I)$, $u(t) = W(t)\phi$ satisfies*

$$\|u(\cdot,t)\|^2_{L^2(I)} + \int_0^t (\partial_x u)^2(0,\tau)d\tau = \|\phi\|^2_{L^2(I)}, \tag{2.4}$$

$$\int_0^L xu^2(x,t)dx + 3\int_0^t\int_0^L (\partial_x u)^2(x,\tau)dxd\tau \le (L+t)\int_0^L \phi^2(x)dx \tag{2.5}$$

for any $t \ge 0$. Moreover, u has the property

$$\partial_x u \in C_b([0,L]; L_t^2(\mathbb{R}^+)),$$

and there exists a constant C such that

$$\sup_{x\in I} \|\partial_x u(x,\cdot)\|_{L^2(\mathbb{R}^+)} \le C\|\phi\|_{L^2(I)}. \tag{2.6}$$

Proof. See [4]. □

Note that estimate (2.6) for $W(t)$ reveals a sharp Kato smoothing effect of the system described by the IBVP (2.2). Next, attention is turned to the inhomogeneous linear problem

$$\begin{cases} \partial_t u + \partial_x u + \partial_x^3 u = f(x,t), \qquad u(x,0) = 0\,, \\ \\ u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0\,. \end{cases} \tag{2.7}$$

In terms of the operator A defined above, one may write (2.7) as an initial value problem for an abstract nonhomogeneous evolution equation, namely

$$\frac{du}{dt} = Au + f, \qquad u(0) = 0. \tag{2.8}$$

By standard semigroup theory (see again [23]), for any $f \in L^1_{loc}(\mathbb{R}^+; L^2(I))$,

$$u(t) = \int_0^t W(t-\tau)f(\tau)d\tau \tag{2.9}$$

belongs to the space $C(\mathbb{R}^+; L^2(I))$ and is called a mild solution of (2.8). It is a weak solution of (2.7) in the sense of distributions. In addition, if $f(t) \in \mathcal{D}(A)$ for $t > 0$ and $Af \in L^1_{loc}(\mathbb{R}^+; L^2(I))$, then $u(t)$ given by (2.9) solves (2.8) a.e. on $[0,T)$ and is called a strong solution of (2.8).

LEMMA 2.2. *There exists a constant C such that for any $f \in L^1_{loc}(\mathbb{R}^+; L^2(I))$, the solution u of* (2.7) *satisfies*

$$\|u(\cdot,t)\|_{L^2(I)} + \|\partial_x u(0,\cdot)\|_{L^2(0,t)} \le C\|f\|_{L^1(0,t;L^2(I))} \tag{2.10}$$

and

$$\int_0^L xu^2(x,t)dx + \int_0^t\int_0^L (\partial_x u)^2(x,\tau)dxd\tau \le C(L+t)\|f\|^2_{L^1(0,t;L^2(I))} \tag{2.11}$$

for any $t \ge 0$. Moreover, the solution u has the property

$$\partial_x u \in C_b([0,L]; L_t^2(\mathbb{R}^+)),$$

and there exists a constant C such that

$$\sup_{x\in I} \|\partial_x u(x,\cdot)\|_{L^2(\mathbb{R}^+)} \le C\|f\|_{L^1(0,t;L^2(I))}. \tag{2.12}$$

Proof. See [4]. □

For any $T > 0$ and $0 \le s \le 3$, let X_s be the collection of all functions ϕ in the space $H^s(I)$ satisfying the compatibility conditions (1.16) with its usual topology, and let $Y_{s,T}$ be the collection of

$$v \in C([0,T]; X_s) \cap L^2([0,T]; H^{s+1}(I))$$

with $\partial_x v \in C([0,L]; L^2(0,T))$. A norm $\|\cdot\|_{Y_{s,T}}$ on the space $Y_{s,T}$ is defined by

$$\|v\|_{Y_{s,T}} := \left(\|v\|^2_{C([0,T];H^s(I))} + \|v\|^2_{L^2([0,T];H^{s+1}(I))} + \|\partial_x v\|^2_{C([0,L];L^2(0,T))} \right)^{1/2}$$

for $v \in Y_{s,T}$.

LEMMA 2.3. *Let a be a C^0 function satisfying*

$$|a(\mu)| \le C(1 + |\mu|^p) \quad \text{for any } \mu \in \mathbb{R}$$

with $0 \le p \le 2$. There exists a constant C such that for any $T > 0$ and u, $v \in Y_{0,T}$,

$$\int_0^T \|a(u(\cdot,t))\partial_x v(\cdot,t)\|_{L^2(I)} dt \le CT^{(2-p)/4} \|u\|^p_{Y_{0,T}} \|v\|_{Y_{0,T}} + CT^{1/2} \left(1 + \|u\|^p_{Y_{0,T}}\right) \|v\|_{Y_{0,T}}.$$

Proof. Using the assumption on the function a and the inequality

$$\|u(\cdot,t)\|_{L^\infty(I)} \le C \left(\|u(\cdot,t)\|_{L^2(I)} + \|u(\cdot,t)\|^{\frac{1}{2}}_{L^2(I)} \|\partial_x u(\cdot,t)\|^{\frac{1}{2}}_{L^2(I)} \right),$$

one obtains

$$\|a(u(\cdot,t))\partial_x v(\cdot,t)\|_{L^2(I)} \le C\|u(\cdot,t)\|^p_{L^\infty(I)} \|\partial_x v(\cdot,t)\|_{L^2(I)} + C\|\partial_x v(\cdot,t)\|_{L^2(I)}$$

$$\le C\|u(\cdot,t)\|^{p/2}_{L^2(I)} \|\partial_x u(\cdot,t)\|^{p/2}_{L^2(I)} \|\partial_x v(\cdot,t)\|_{L^2(I)} + C\left(1 + \|u(\cdot,t)\|^p_{L^2(I)}\right) \|\partial_x v(\cdot,t)\|_{L^2(I)}.$$

The first term, when integrated with respect to t, is bounded as follows:

$$\int_0^T \|u(\cdot,t)\|^{p/2}_{L^2(I)} \|\partial_x u(\cdot,t)\|^{p/2}_{L^2(I)} \|\partial_x v(\cdot,t)\|_{L^2(I)} dt$$

$$\le \sup_{0\le t\le T} \|u(\cdot,t)\|^{p/2}_{L^2(I)} \left(\int_0^T \|\partial_x u(\cdot,t)\|^2_{L^2(I)} dt \right)^{p/4} \left(\int_0^T \|\partial_x v(\cdot,t)\|^{4/(4-p)}_{L^2(I)} dt \right)^{(4-p)/4}$$

$$\le T^{(2-p)/4} \|u\|^p_{Y_{0,T}} \|v\|_{Y_{0,T}}$$

if $0 \le p \le 2$. Thus

$$\int_0^T \|a(u(\cdot,t))\partial_x v(\cdot,t)\|_{L^2(I)} dt \le C2^{p/2} T^{(2-p)/4} \|u\|^p_{Y_{0,T}} \|v\|_{Y_{0,T}} + CT^{1/2} \left(1 + \|u\|^p_{Y_{0,T}}\right) \|v\|_{Y_{0,T}}.$$

The proof is complete. □

LEMMA 2.4. *There exists a constant C depending only on L such that for any $T > 0$, $1 \le p \le 2$, $b \in L^2(I)$, and u, v, $w \in Y_{0,T}$,*

$$\int_0^T \|bu\|_{L_x^2(I)} dt \le CT^{1/2} \|b\|_{L^2(I)} \|u\|_{Y_{0,T}}, \tag{2.13}$$

$$\int_0^T \|u\partial_x w\|_{L_x^2(I)} dt \le CT^{1/4} \|u\|_{Y_{0,T}} \|w\|_{Y_{0,T}}, \tag{2.14}$$

$$\int_0^T \|u|w|^{p-1}\partial_x w\|_{L_x^2(I)} dt \le CT^{(2-p)/4} \|u\|_{Y_{0,T}} \|w\|_{Y_{0,T}}^p, \tag{2.15}$$

and

$$\int_0^T \|u|v|^{p-1}\partial_x w\|_{L_x^2(I)} dt \le CT^{(2-p)/4} \|u\|_{Y_{0,T}} \|w\|_{Y_{0,T}} \|v\|_{Y_{0,T}}^{p-1}. \tag{2.16}$$

Proof. The estimate (2.13) follows from the direct calculation

$$\int_0^T \left(\int_0^L |b(x)|^2 |u(x,t)|^2 dx \right)^{\frac{1}{2}} dt \le \int_0^T \sup_{x\in(0,L)} |u(x,t)| \left(\int_0^L |b(x)|^2 dx \right)^{\frac{1}{2}} dt$$
$$\le CT^{\frac{1}{2}} \|b\|_{L^2(I)} \|u\|_{Y_{0,T}}.$$

The other estimates can be proved by a similar argument as that in the proof of Lemma 2.3 and are therefore omitted. The proof is complete. □

Now we turn to the nonlinear problem (2.1). We first present the following global a priori estimates for solutions of the IBVP (2.1).

LEMMA 2.5. *Let a be a C^0 function satisfying*

$$|a(\mu)| \le C(1 + |\mu|^p) \quad \text{for any } \mu \in \mathbb{R}$$

with $0 \le p < 4$. There exists a constant C_p such that for any smooth solution u of (2.1), the following estimates hold:

$$\|u(\cdot,T)\|_{L^2(I)}^2 + \int_0^T (\partial_x u)^2(0,\tau) d\tau + 2\int_0^T \int_0^L b(x) u^2(x,\tau) dx d\tau = \|\phi\|_{L^2(I)}^2 \tag{2.17}$$

and

$$\int_0^T \int_0^L |\partial_x u|^2 \, dx dt \le \frac{L + (C+1)T}{2} ||\phi||_{L^2(I)}^2 + C_p T ||\phi||_{L^2(I)}^{\frac{8+2p}{4-p}} \tag{2.18}$$

for any $T > 0$.

Proof. Multiply both sides of the equation in (2.1) by u and integrate with respect to x over the interval I and about t over the interval $(0,T)$. An integration by parts leads directly to the equality (2.17). To prove the inequality (2.18), let us first introduce the functions

$$A(u) := \int_0^u a(v)\, dv, \qquad \tilde{A}(u) := \int_0^u v a(v)\, dv.$$

Multiplying both sides of the equation in (2.1) by xu and integrating over $(0,L)\times(0,T)$, we obtain after some integrations by parts

$$\frac{1}{2}\left(\int_0^L x|u(x,T)|^2dx-\int_0^L x|\phi(x)|^2dx\right)+\frac{3}{2}\int_0^T\int_0^L|\partial_x u|^2dxdt-\frac{1}{2}\int_0^T\int_0^L|u|^2dxdt$$
$$-\int_0^T\int_0^L\tilde{A}(u)\,dxdt+\int_0^T\int_0^L xb(x)|u|^2dxdt=0.$$

Combining this equality with (2.17), we obtain

$$\frac{3}{2}\int_0^T\int_0^L|\partial_x u|^2dxdt\le\frac{L+T}{2}||\phi||^2_{L^2(I)}+\int_0^T\int_0^L\tilde{A}(u)\,dxdt. \tag{2.19}$$

We infer from the assumption on a that

$$|\tilde{A}(u)|\le C\left(\frac{u^2}{2}+\frac{|u|^{p+2}}{p+2}\right).$$

Hence

$$\begin{aligned}\left|\int_0^T\int_0^L\tilde{A}(u)\,dxdt\right|&\le\frac{C}{2}\int_0^T\int_0^L|u|^2dxdt+\frac{C}{p+2}\int_0^T\int_0^L|u|^{p+2}dxdt\\&\le\frac{CT}{2}||\phi||^2_{L^2(I)}+\frac{C}{p+2}\int_0^T||u||^2_{L^2(I)}\cdot||u||^p_{L^\infty(I)}dt\\&\le\frac{CT}{2}||\phi||^2_{L^2(I)}+\frac{2^{\frac{p}{2}}C}{p+2}\int_0^T||u||^{2+\frac{p}{2}}_{L^2(I)}\cdot||\partial_x u||^{\frac{p}{2}}_{L^2(I)}\,dt\\&\le\frac{CT}{2}||\phi||^2_{L^2(I)}+\frac{2^{\frac{p}{2}}CT^{1-\frac{p}{4}}}{p+2}||\phi||^{2+\frac{p}{2}}_{L^2(I)}\left(\int_0^T||\partial_x u||^2_{L^2(I)}dt\right)^{\frac{p}{4}}\\&\le\frac{CT}{2}||\phi||^2_{L^2(I)}+C_pT||\phi||^{\frac{8+2p}{4-p}}_{L^2(I)}+\frac{1}{2}\int_0^T\int_0^L|\partial_x u|^2dxdt\end{aligned} \tag{2.20}$$

for some positive constant C_p which depends only on C and p. The inequality (2.18) then follows from (2.19) and (2.20). The proof is complete. □

We now present our first well-posedness result for the IBVP (2.1).

PROPOSITION 2.6. *Let a be a C^1 function satisfying*

$$|a(\mu)|\le C(1+|\mu|^p),\quad |a'(\mu)|\le C(1+|\mu|^{p-1})\quad \text{for any } \mu\in\mathbb{R}$$

with $1\le p<2$, and let $T>0$ be given. For any $\phi\in X_0$, the IBVP (2.1) admits a unique solution $u\in Y_{0,T}$, which also satisfies

$$\|u\|_{Y_{0,T}}\le\beta_0(\|\phi\|_{X_0})\|\phi\|_{X_0}, \tag{2.21}$$

where $\beta_0:\mathbb{R}^+\to\mathbb{R}^+$ is a nondecreasing continuous function. Moreover, the corresponding solution map is locally Lipschitz continuous; for any ϕ, $\psi\in X_0$, the corresponding solutions u and v of (2.1) satisfy

$$\|u-v\|_{Y_{0,T}}\le\beta_0(\|\phi\|_{X_0}+\|\psi\|_{X_0})\|\phi-\psi\|_{X_0}. \tag{2.22}$$

Proof. Write the IBVP (2.1) in its integral equation form

$$u(t) = W(t)\phi - \int_0^t W(t-\tau)\Big[(a(u)\partial_x u)(\tau) + (bu)(\tau)\Big]d\tau, \tag{2.23}$$

where the spatial variable is suppressed throughout. For given $\phi \in X_0$, let $r > 0$ and $\theta > 0$ be constants to be determined. Let

$$S_{\theta,r} = \{v \in Y_{0,\theta},\ \|v\|_{Y_{0,\theta}} \le r\}.$$

The set $S_{\theta,r}$ is a closed, convex, and bounded subset of the space $Y_{0,\theta}$ and therefore is a complete metric space in the topology induced from $Y_{0,\theta}$. Define a map Γ on $S_{\theta,r}$ by

$$\Gamma(v) = W(t)\phi - \int_0^t W(t-\tau)\Big[(a(v)\partial_x v)(\tau) + (bv)(\tau)\Big]d\tau$$

for $v \in S_{\theta,r}$. Applying Lemmas 2.2, 2.3, and 2.4, it is adduced that there are constants C_0, C_1, and C_2 for which

$$\begin{aligned}\|\Gamma(v)\|_{Y_{0,\theta}} &\le C_0\|\phi\|_{X_0} + C_1\int_0^\theta \Big[\|a(v)\partial_x v(\cdot,\tau)\|_{L^2(I)} + \|bv(\cdot,\tau)\|_{L^2(I)}\Big]d\tau \\ &\le C_0\|\phi\|_{X_0} + C_1\theta^{(2-p)/4}\|v\|_{Y_{0,\theta}}^{p+1} + C_2\theta^{\frac{1}{2}}(\|b\|_{L^2(I)}+1)\|v\|_{Y_{0,\theta}} \\ &\le C_0\|\phi\|_{X_0} + C_1\theta^{(2-p)/4}r^p\|v\|_{Y_{0,\theta}} + C_2\theta^{\frac{1}{2}}(\|b\|_{L^2(I)}+1)\|v\|_{Y_{0,\theta}}\end{aligned}$$

for any $v \in S_{r,\theta}$. In addition, for any $v,\ w \in S_{r,\theta}$,

$$\begin{aligned}\Gamma(v) - \Gamma(w) &= \int_0^t W(t-\tau)\big[a(v)\partial_x v - a(w)\partial_x w + b(v-w)\big]d\tau \\ &= \int_0^t W(t-\tau)\big[a(v)\partial_x(v-w) + (a(v)-a(w))\partial_x w + b(v-w)\big]d\tau.\end{aligned}$$

Thus

$$\begin{aligned}\|\Gamma(v) - \Gamma(w)\|_{Y_{0,\theta}} &\le C_1\int_0^\theta \big[\|a(v)\partial_x(v-w)\|_{L^2_x(I)} \\ &\qquad + \|(a(v)-a(w))\partial_x w\|_{L^2_x(I)} + \|b(v-w)\|_{L^2_x(I)}\big]\,d\tau \\ &\le C_1\int_0^\theta \big[\|a(v)\partial_x(v-w)\|_{L^2_x(I)} + \|\left(1+|v|^{p-1}+|w|^{p-1}\right)(v-w)\partial_x w\|_{L^2_x(I)} \\ &\qquad + \|b(v-w)\|_{L^2_x(I)}\big]\,d\tau \\ &\le C_1\theta^{\frac{2-p}{4}}\Big(\|v\|_{Y_{0,\theta}}^p + \|w\|_{Y_{0,\theta}}^p + \|v\|_{Y_{0,\theta}}^{p-1}\|w\|_{Y_{0,\theta}}\Big)\|v-w\|_{Y_{0,\theta}} \\ &\qquad + C_1\theta^{1/4}\|w\|_{Y_{0,\theta}}\|v-w\|_{Y_{0,\theta}} + C_2\theta^{\frac{1}{2}}(\|b\|_{L^2(I)}+1)\|v-w\|_{Y_{0,\theta}} \\ &\le 3C_1\theta^{\frac{2-p}{4}}r^p\|v-w\|_{Y_{0,\theta}} + C_1\theta^{1/4}r\|v-w\|_{Y_{0,\theta}} + C_2\theta^{\frac{1}{2}}(\|b\|_{L^2(I)}+1)\|v-w\|_{Y_{0,\theta}}.\end{aligned}$$

Choosing $r > 0$ and $\theta > 0$ so that

$$\begin{cases} r = 2C_0\|\phi\|_{X_0}, \\ \\ 3C_1\theta^{(2-p)/4}r^p + C_1\theta^{1/4}r + C_2\theta^{1/2}(\|b\|_{L^2(I)} + 1) \le \frac{1}{2}, \end{cases} \tag{2.24}$$

then

$$\|\Gamma(v)\|_{Y_{0,\theta}} \le r, \qquad \|\Gamma(v) - \Gamma(w)\|_{Y_{0,\theta}} \le \frac{1}{2}\|v - w\|_{Y_{0,\theta}}$$

for any v, $w \in S_{\theta,r}$. Thus, with such a choice of r and θ, Γ is a contraction mapping of $S_{r,\theta}$. Its fixed point $u = \Gamma(u)$ is the unique solution of the IBVP (2.1) in $S_{\theta,r}$. Note that θ depends only on $\|\phi\|_{X_0}$ and

$$\sup_{0\le t\le\theta} \|u(\cdot,t)\|_{L^2(I)} \le \|\phi\|_{X_0}.$$

By the standard extension argument, one may extend θ to T. The proof is complete. □

Next, we show that the IBVP (2.1) is well-posed in the space X_3.

PROPOSITION 2.7. *Let $b \in H^1(I)$, let a be a C^1 function satisfying*

$$|a(\mu)| \le C(1 + |\mu|^p), \quad |a'(\mu)| \le C(1 + |\mu|^{p-1}) \quad \textit{for any } \mu \in \mathbb{R}$$

with $1 \le p < 2$, and let $T > 0$ be given. For any $\phi \in X_3$, the IBVP (2.1) admits a unique solution $u \in Y_{3,T}$. Moreover, there exists a nondecreasing continuous function $\beta_3 : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|u\|_{Y_{3,T}} \le \beta_3(\|\phi\|_{X_0})\|\phi\|_{X_3}. \tag{2.25}$$

Proof. By Proposition 2.6, (2.1) admits a unique solution $u \in Y_{0,T}$. We just need to show further that $u \in Y_{3,T}$. For this purpose, let $v = u_t$. Then the function v is a solution of

$$\begin{cases} \partial_t v + \partial_x v + \big(a'(u)\partial_x u\big)v + a(u)\partial_x v + \partial_x^3 v + b(x)v = 0, \\ \\ v(x,0) = \phi^*(x), \\ \\ v(0,t) = 0, \quad v(L,t) = 0, \quad \partial_x v(L,t) = 0 \end{cases} \tag{2.26}$$

with

$$\phi^*(x) = -\phi'''(x) - \phi'(x) - a(\phi(x))\phi'(x) - b(x)\phi(x).$$

Observe that $\phi^* \in X_0$ and that there exists a constant $C = C(\|\phi\|_{X_0})$ such that

$$\|\phi^*\|_{X_0} \le C\|\phi\|_{X_3}.$$

We may write (2.26) in its integral form

$$v(t) = W(t)\phi^* - \int_0^t W(t-\tau)\Big(a'(u(\cdot,\tau))\partial_x u(\cdot,\tau)v(\cdot,\tau) + a(u(\cdot,\tau))\partial_x v(\cdot,\tau) + b(\cdot)v(\cdot,\tau)\Big)d\tau.$$

Using Lemma 2.4 and proceeding as in the proof of Proposition 2.6, we see that (2.26) admits a unique solution $v \in Y_{0,T}$. Notice that $u_t = v$, $b \in H^1(I)$, and $u \in$

$Y_{0,T}$. Since $u \in H^1(0,T;H^1(I)) \subset C([0,T];C(\bar{I}))$, we obviously have that $a(u)\partial_x u \in L^2(0,T;L^2(I))$. It then follows from

$$\partial_x^3 u = -\partial_t u - \partial_x u - b(x)u - a(u)\partial_x u = -v - \partial_x u - b(x)u - a(u)\partial_x u \tag{2.27}$$

that $u \in L^2(0,T;H^3(I))$; hence $u \in C([0,T];H^2(I))$. This implies further that $\partial_x u$, bu, and $a(u)\partial_x u$ all belong to the space $C([0,T];L^2(I)) \cap L^2(0,T;H^1(I))$. Using (2.27) again yields that $u \in Y_{3,T}$. The proof is complete. □

According to Propositions 2.6 and 2.7, the IBVP (2.1) defines a continuous nonlinear map K from the space X_j to $Y_{j,T}$ for $j = 0, 3$. Next, we show that K is a continuous map from X_s to $Y_{s,T}$ for $0 \le s \le 3$ by using the following nonlinear interpolation theory (cf. [35, 1]), which implies the well-posedness of the IBVP (2.1) in the space X_s for $0 \le s \le 3$.

Let B_0 and B_1 be two Banach spaces such that $B_1 \subset B_0$ with the inclusion map continuous. Let $f \in B_0$ and, for $t \ge 0$, define

$$K(f,t) = \inf_{g \in B_1} \{\|f-g\|_{B_0} + t\|g\|_{B_1}\}.$$

For $0 < \theta < 1$ and $1 \le p \le +\infty$, define

$$[B_0, B_1]_{\theta,p} = B_{\theta,p} = \left\{ f \in B_0 : \ \|f\|_{\theta,p} = \left(\int_0^{+\infty} K(f,t)^p t^{-\theta p - 1} dt \right)^{1/p} < +\infty \right\}$$

with the usual modification for the case $p = +\infty$. Then $B_{\theta,p}$ is a Banach space with norm $\| \ \|_{\theta,p}$. Given two pairs of indices (θ_1, p_1) and (θ_2, p_2) as above, then $(\theta_1, p_1) \prec (\theta_2, p_2)$ means

$$\begin{cases} \theta_1 < \theta_2 & \text{or} \\ \\ \theta_1 = \theta_2 & \text{and } p_1 > p_2. \end{cases}$$

If $(\theta_1, p_1) \prec (\theta_2, p_2)$, then $B_{\theta_2,p_2} \subset B_{\theta_1,p_1}$ with the inclusion map continuous.

THEOREM 2.8. *Let B_0^j and B_1^j be Banach spaces such that $B_1^j \subset B_0^j$ with continuous inclusion mappings, $j = 1, 2$. Let λ and q lie in the ranges $0 < \lambda < 1$ and $1 \le q \le +\infty$. Suppose A is a mapping such that*

(i) *$A : \ B^1_{\lambda,q} \to B_0^2$ and for $f, \ g \in B^1_{\lambda,q}$,*

$$\|Af - Ag\|_{B_0^2} \le C_0(\|f\|_{B^1_{\lambda,q}} + \|g\|_{B^1_{\lambda,q}})\|f-g\|_{B_0^1}$$

and

(ii) *$A : \ B_1^1 \to B_1^2$ and for $h \in B_1^1$,*

$$\|Ah\|_{B_1^2} \le C_1(\|h\|_{B^1_{\lambda,q}})\|h\|_{B_1^1},$$

where $C_j : \ \mathbb{R}^+ \to \mathbb{R}^+$ are continuous nondecreasing functions, $j = 0, 1$. Then if $(\theta, p) \ge (\lambda, q)$, A maps $B^1_{\theta,p}$ into $B^2_{\theta,p}$ and for $f \in B^1_{\theta,p}$,

$$\|Af\|_{B^2_{\theta,p}} \le C(\|f\|_{B^1_{\lambda,q}})\|f\|_{B^1_{\theta,p}},$$

where for $r > 0$, $C(r) = 4C_0(4r)^{1-\theta}C_1(3r)^{\theta}$.

Remark. This theorem is identical to Theorem 2 of Tartar [35] except that Tartar makes the more restrictive assumption that the nondecreasing functions C_0 and C_1

depend only on the B_0^1 norms of the functions in question. Theorem 2.8 was used by Bona and Scott [1] to prove the global well-posedness of the pure initial value problem for the KdV equation on the whole line in fractional order Sobolev spaces $H^s(\mathbb{R})$.

Here is the promised well-posedness result for the IBVP (2.1) in X_s.

THEOREM 2.9. *Let a be a C^1 function satisfying*

$$|a(\mu)| \le C(1+|\mu|^p), \quad |a'(\mu)| \le C(1+|\mu|^{p-1}) \quad \text{for any } \mu \in \mathbb{R} \tag{2.28}$$

with $1 \le p < 2$, and let $T > 0$ and $0 \le s \le 3$ be given. In addition, assume $b \in L^2(I)$ when $s = 0$ and $b \in H^1(I)$ when $s > 0$. Then, for any $\phi \in X_s$, the IBVP (2.1) admits a unique solution $u \in Y_{s,T}$. Moreover, there exists a nondecreasing continuous function $\beta_s : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|u\|_{Y_{s,T}} \le \beta_s(\|\phi\|_{X_0})\|\phi\|_{X_s}.$$

Proof. Choose

$$B_0^1 = X_0, \quad B_1^1 = X_3, \quad B_0^2 = Y_{0,T}, \quad B_1^2 = Y_{3,T}.$$

Let A be the solution map of the IBVP (2.1): $u = A(\phi)$. For given s with $0 < s < 3$, choose $p = 2$ and $\theta = s/3$. Then

$$B_{\theta,p}^2 = Y_{s,T}, \qquad B_{\theta,p}^1 = X_s.$$

In this case, assumptions (i) and (ii) of Theorem 2.8 are (2.22) and (2.25), respectively, which we have already proved. The proof is then completed by invoking Theorem 2.8. □

By Theorem 2.9, the condition $\phi \in X_s$ implies that the corresponding solution $u \in L^2(0,T;H^{s+1}(I))$. Thus, for any $0 < \epsilon < T$, one can always find a time $t_1 \in (0,\epsilon)$ such that $u(\cdot,t_1) \in H^{s+1}(I)$. The following corollary, which reveals a strong smoothing property of the system (2.1), follows directly from Theorem 2.9.

COROLLARY 2.10. *Under the assumptions of Theorem 2.9, for any $\phi \in L^2(I)$, the corresponding solution u of (2.1) belongs to the space $C([\epsilon,T];H^3(I)) \cap L^2(\epsilon,T;H^4(I))$ for any $\epsilon > 0$.*

In Theorem 2.9, the nonlinear term $a(u)$ of the equation in (2.1) is required to satisfy (2.28) with $1 \le p < 2$. We consider next the case of $p \ge 2$.

Let $\mathcal{B}$ be the linear operator defined by $\mathcal{B}f = b(x)f$. Consider $A_b = A + \mathcal{B}$ as an unbounded operator on $L^2(I)$ with the domain

$$\mathcal{D}(A_b) = \{f \in H^3(I),\ f(0) = f(L) = f'(L) = 0\}.$$

The operator A_b is the infinitesimal generator of a C_0-semigroup $W_b(t)$ in the space $L^2(I)$. For any $\phi \in L^2(I)$,

$$u(t) = W_b(t)\phi$$

belongs to the space $C_b(\mathbb{R}^+;L^2(I))$ and solves

$$\begin{cases} \partial_t u + \partial_x u + \partial_x^3 u + b(x)u = 0, \qquad u(x,0) = \phi(x)\,, \\ \\ u(0,t) = 0, \quad u(L,t) = 0, \quad \partial_x u(L,t) = 0. \end{cases} \tag{2.29}$$

LEMMA 2.11. *Let $T > 0$ and $0 \le s \le 3$ be given. Assume that $b \in L^2(I)$ if $s = 0$ and $b \in H^1(I)$ if $s > 0$. Then there exists a constant C depending only on T and s such that*

$$\|W_b(t)\phi\|_{Y_{s,T}} \le C\|\phi\|_{X_s} \quad \text{for any } \phi \in X_s \tag{2.30}$$

and

$$\left\|\int_0^t W_b(t-\tau)f(\cdot,\tau)d\tau\right\|_{Y_{s,T}} \le C\|f\|_{L^1(0,T;H^s(I))} \quad \text{for any } f \in L^1(0,T;X_s). \tag{2.31}$$

Proof. The estimate (2.30) is established by using a similar (but simpler, because there is no nonlinear term in the equation) argument as that used in the proof of Theorem 2.9. To see that (2.31) is true, note that

$$\begin{aligned}\left\|\int_0^t W_b(t-\tau)f(\cdot,\tau)d\tau\right\|_{C([0,T];H^s(I))} &\le \int_0^T \|W_b(t-\tau)f(\cdot,\tau)\|_{C([\tau,T]_t;H^s(I))}\,d\tau \\ &\le C\int_0^T \|f(\cdot,\tau)\|_{H^s(I)}d\tau\end{aligned}$$

and

$$\begin{aligned}\left\|\int_0^t W_b(t-\tau)f(\cdot,\tau)d\tau\right\|_{L^2(0,T;H^{s+1}(I))} &\le \int_0^T \|W_b(t-\tau)f(\cdot,\tau)\|_{L^2([\tau,T]_t;H^{s+1}(I))}\,d\tau \\ &\le C\int_0^T \|f(\cdot,\tau)\|_{H^s(I)}d\tau.\end{aligned}$$

Moreover, using the same argument as that used in the proof of Proposition 2.17 in [4] yields

$$\sup_{x\in I}\left\|\partial_x\int_0^t W_b(t-\tau)f(\cdot,\tau)d\tau\right\|_{L^2(0,T)} \le C\int_0^T \|f(\cdot,\tau)\|_{H^s(I)}d\tau.$$

Inequality (2.31) then follows by combining the above three estimates. The proof is complete. □

We have the following estimate by direct calculation.

LEMMA 2.12. *Let a be a C^1 function satisfying*

$$|a(\mu)| \le C(1+|\mu|^p), \quad |a'(\mu)| \le C(1+|\mu|^{p-1}) \quad \text{for any } \mu \in \mathbb{R}$$

with $p \ge 1$. There exists a constant C such that

$$\|a(u)\partial_x v\|_{L^1(0,T;H^1(I))} \le CT^{1/2}\left(1+\|u\|_{Y_{1,T}}+\|u\|_{Y_{1,T}}^p\right)\|v\|_{Y_{1,T}} \quad \text{for any } u,\ v \in Y_{1,T}.$$

Using the notation of the semigroup $W_b(t)$, the nonlinear IBVP (2.1) may be written in the following integral form:

$$u(t) = W_b(t)\phi - \int_0^t W_b(t-\tau)a(u(\cdot,\tau))\partial_x u(\cdot,\tau)d\tau.$$

The same arguments as those in the proof of Propositions 2.6 and 2.7 lead to the following (local) well-posedness result for the IBVP (2.1).

THEOREM 2.13. *Let $b \in H^1(I)$, let a be a C^2 function satisfying*

$$|a(\mu)| \le C(1+|\mu|^p), \quad |a'(\mu)| \le C(1+|\mu|^{p-1}), \quad |a''(\mu)| \le C(1+|\mu|^{p-2}) \quad \text{for any } \mu \in \mathbb{R}$$

with $p \ge 2$, and let $j = 1$ or 3 be given. Then for any $\phi \in X_j$ there exists a $T^* > 0$ depending only on $\|\phi\|_{X_1}$ such that the IBVP (2.1) admits a unique solution $u \in Y_{j,T^*}$. Moreover, there exists a nondecreasing continuous function $\alpha_j : \mathbb{R}^+ \to \mathbb{R}^+$ such that

$$\|u\|_{Y_{j,T^*}} \le \alpha_j(\|\phi\|_{X_1})\|\phi\|_{X_j}.$$

Remark. The well-posedness result presented in Theorem 2.13 is local in the sense that the length of the time interval $(0, T^*)$ in which the solution exists depends on the norm of the initial value ϕ in the space X_1. In order to obtain the global well-posedness result, one needs to establish the corresponding global a priori estimate for the IBVP (2.1) in the space X_1, which is not available. In fact, the only available global a priori estimate for the IBVP (2.1) is the L^2-estimate presented in Lemma 2.5. On the other hand, whether the IBVP (2.1) is well-posed in the space $L^2(I)$ in the case of $p \ge 2$ is still an open question.

The next theorem shows that if $\phi \in X_0$, then the IBVP (2.1) admits one solution $u \in C_w(R^+; L^2(I)) \cap L^2_{loc}(R^+; H^1(I))$.

THEOREM 2.14. *Let a be a C^2 function satisfying*

$$|a(\mu)| \le C(1+|\mu|^p), \quad |a'(\mu)| \le C(1+|\mu|^{p-1}), \quad |a''(\mu)| \le C(1+|\mu|^{p-2}) \quad \text{for any } \mu \in \mathbb{R}$$

with $2 \le p < 4$. Then, for any given $\phi \in X_0$, the IBVP (2.1) admits at least one solution $u \in C_w(\mathbb{R}^+; L^2(I)) \cap L^2_{loc}(\mathbb{R}^+; H^1(I))$.

Proof. Let $\{a_n\}$ denote a sequence of functions in $C_0^\infty(\mathbb{R}; \mathbb{R})$ such that

$$|a_n^{(j)}(\mu)| \le C(2 + |\mu|^{p-j}) \qquad \forall n \ge 0,\ \forall \mu \in \mathbb{R},\ j = 0, 1, 2; \tag{2.32}$$

$$a_n(\mu) \to a(\mu) \quad \text{uniformly on each compact set in } R. \tag{2.33}$$

Observe that $|a_n(\mu)| \le C(n)(1 + |\mu|)$ and $|a_n'(\mu)| \le C(n)(1 + |\mu|)$. According to Proposition 2.6, there exists a (unique) solution $u_n \in C(\mathbb{R}^+; L^2(I)) \cap L^2(\mathbb{R}^+; H_0^1(I))$ of the modified boundary initial value problem

$$\begin{cases} \partial_t u_n + \partial_x^3 u_n + \partial_x u_n + a_n(u_n)\partial_x u_n + b u_n = 0, \\ u_n(0,t) = u_n(L,t) = \partial_x u_n(L,t) = 0, \\ u_n(x,0) = \phi(x). \end{cases} \tag{2.34}$$

Let us introduce the functions

$$A_n(u) := \int_0^u a_n(v)\, dv, \qquad \tilde{A}_n(u) := \int_0^u v a_n(v)\, dv.$$

According to Lemma 2.5,

$$||u_n(T)||^2_{L^2(I)} = ||\phi||^2_{L^2(I)} - \int_0^T |\partial_x u_n(0,t)|^2 dt - 2\int_0^T \int_0^L b(x)|u_n|^2 dx dt$$

and

$$\int_0^T \int_0^L |\partial_x u_n|^2\, dx dt \le \frac{L + (C+1)T}{2} ||\phi||^2_{L^2(I)} + C_p T ||\phi||^{\frac{8+2p}{4-p}}_{L^2(0,L)}$$

for any $T \geq 0$ and $n \geq 0$. Thus $\{u_n\}$ is bounded in $L^\infty(\mathbb{R}^+; L^2(I)) \cap L^2_{loc}(\mathbb{R}^+; H^1_0(I))$, and there exist a function $u \in L^\infty(\mathbb{R}^+; L^2(I)) \cap L^2_{loc}(\mathbb{R}^+; H^1(I))$ and a subsequence of $\{u_n\}$, again denoted by $\{u_n\}$, such that

$$u_n \rightharpoonup u \quad \text{in } L^\infty(\mathbb{R}^+; L^2(I)) \text{ weak}*, \tag{2.35}$$

$$u_n \rightharpoonup u \quad \text{in } L^2_{loc}(\mathbb{R}^+; H^1(I)) \text{ weak}. \tag{2.36}$$

To pass to the limit in (2.34) we have to pay some attention to the nonlinear term $a_n(u_n)\partial_x u_n = \partial_x[A_n(u_n)]$. Let us introduce the function $A(u) := \int_0^u a(v)\,dv$. We aim to prove that

$$A_n(u_n) \to A(u) \quad \text{in } \mathcal{D}'(I_x \times (0, +\infty)_t).$$

We first prove the following claim.

Claim 1. Let $T > 0$ and $\alpha \in (1, \frac{6}{p+1}]$. Then $A_n(u_n)$ is bounded in $L^\alpha(I \times (0,T))$.

Proof of Claim 1. It follows from (2.32) that

$$|A_n(u)| \leq C\left(2|u| + \frac{|u|^{p+1}}{p+1}\right) \leq C'(1 + |u|^{p+1})$$

and

$$|A_n(u)|^\alpha \leq C''(1 + |u|^{\alpha(p+1)}),$$

where C' and C'' denote some positive constants which depend only on C, p, and α. Therefore

$$\begin{aligned}\int_0^T \int_0^L |A_n(u_n)|^\alpha dxdt &\leq C''\left(TL + \int_0^T \int_0^L |u_n|^{\alpha(p+1)} dxdt\right)\\ &\leq C''\left(TL + \int_0^T ||u_n||^2_{L^2(I)} \cdot ||u_n||^{\alpha(p+1)-2}_{L^\infty(I)} dt\right)\\ &\leq C''\left(TL + ||\phi||^{1+\frac{\alpha(p+1)}{2}}_{L^2(I)} \int_0^T ||\partial_x u_n||^{\frac{\alpha(p+1)-2}{2}}_{L^2(I)} dt\right).\end{aligned}$$

As $\frac{\alpha(p+1)-2}{2} \leq 2$, the result in Claim 1 follows from Lemma 2.5. □

Next we prove the following claim.

Claim 2. Let T and α be as in Claim 1. Then the sequences $\{a_n(u_n)\partial_x u_n\}$ and $\{\partial_t u_n\}$ are bounded in $L^\alpha(0,T; H^{-2}(I))$.

Proof of Claim 2. Notice first that $L^\alpha(I) \subset H^{-1}(I)$ (as $H^1_0(I) \subset L^\infty(I)$); hence $a_n(u_n)\partial_x u_n = \partial_x[A_n(u_n)]$ is bounded in $L^\alpha(0,T; H^{-2}(I))$. Clearly, $\partial_x^3 u_n$ is bounded in $L^2(0,T; H^{-2}(I))$ and $\partial_x u_n + b u_n$ is bounded in $L^2(0,T; L^2(I))$, so we conclude that $\partial_t u_n = -(\partial_x^3 u_n + \partial_x u_n + a_n(u_n)\partial_x u_n + b u_n)$ is bounded in $L^\alpha(0,T; H^{-2}(I))$. □

As the first embedding in $H^1_0(I) \subset L^2(I) \subset H^{-2}(I)$ is compact, we infer from [33, Cor. 4] that the sequence $\{u_n\}$ is relatively compact in $L^2(0,T; L^2(I))$ for each $T > 0$. Therefore, extracting a subsequence of $\{u_n\}$ again denoted by $\{u_n\}$, we have that

$$u_n \to u \text{ (strongly)} \quad \text{in } L^2_{loc}(\mathbb{R}^+; L^2(I)) \text{ and a.e.} \tag{2.37}$$

Using (2.33) and (2.37), it is easy to see that

$$A_n(u_n(x,t)) \to A(u(x,t)) \quad \text{for a.e. } (x,t) \in (0,L) \times \mathbb{R}^+.$$

On the other hand, picking any $\alpha \in (1, \frac{6}{p+1}]$, we deduce from Claim 1 that there exists a function $g \in L^\alpha_{loc}(\mathbb{R}^+; L^\alpha(I))$ such that (for a subsequence)

$$A_n(u_n) \rightharpoonup g \quad \text{in } L^\alpha((0,L)\times(0,T))$$

for any $T > 0$. The following lemma is an easy consequence of the Egoroff theorem.

LEMMA 2.15. *Let Ω be an open set in R^N, and let $\{f_n\}$ be a sequence of functions in $L^p(\Omega)$ (with $1 < p < \infty$) such that $f_n \rightharpoonup f$ in $L^p(\Omega)$ and $f_n(x) \to g(x)$ a.e. Then $f(x) = g(x)$ a.e.*

We infer from Lemma 2.15 that $g(x,t) = A(u(x,t))$ a.e. Therefore,

$$A_n(u_n) \to A(u) \quad \text{in } \mathcal{D}'((0,L)\times R^+),$$

and taking the spatial derivative, we obtain

$$a_n(u_n)\partial_x u_n \to a(u)\partial_x u \quad \text{in } \mathcal{D}'((0,L)\times R^+). \tag{2.38}$$

Gathering (2.37) and (2.38) together, and taking the limit in (2.34), we conclude that u solves the equation in (2.1) in the sense of distributions.

Since u_n is bounded in $L^\infty(0,T;L^2(I))$ (see (2.17)) and $\partial_t u_n$ is bounded in $L^\alpha(0,T;\, H^{-2}(I))$ (with $\alpha > 1$), we infer from [33, Cor. 4] that for a subsequence, again denoted by $\{u_n\}$,

$$u_n \to u \quad \text{in } C([0,T], H^{-1}(I)) \quad \text{for any } T > 0. \tag{2.39}$$

In particular,

$$u(0) = \lim_{n\to+\infty} u_n(0) = \phi.$$

As $u \in L^\infty(0,T;L^2(I)) \cap C_w([0,T],H^{-1}(I))$, we deduce from [36, Lem. 1.4, p. 263] that $u \in C_w([0,T],L^2(I))$. □

DEFINITION 2.16. *A function $u \in C_w(\mathbb{R}^+;L^2(I)) \cap L^2_{loc}(\mathbb{R}^+;H^1_0(I))$ is said to be a* weak solution *of* (2.1) *if there exist a sequence $\{a_n\}$ of functions in $C^\infty_0(\mathbb{R};\mathbb{R})$ satisfying* (2.32) *and* (2.33) *and a sequence of strong solutions u_n to* (2.34) *such that* (2.35), (2.36), (2.37), *and* (2.39) *hold true.*

3. Exponential stability. In this section we study the long time behavior of weak solutions of (2.1). The goal is to show that any weak solution of (2.1) decays exponentially in the space $L^2(I)$. As a weak solution of (2.1) may fail to be unique, the concept of exponential stability has to be generalized in the following way.

DEFINITION 3.1. *System* (2.1) *is said to be* locally uniformly exponentially stable *in $L^2(I)$ if for any $r > 0$ there exist two constants $C > 0$ and $\nu > 0$ such that for any $\phi \in L^2(I)$ with $||\phi||_{L^2(I)} < r$ and for any weak solution $u = u(x,t)$ of* (2.1), *it holds that*

$$||u(\cdot,t)||^2_{L^2(I)} \le C\,||\phi||^2_{L^2(I)} e^{-\nu t} \qquad \forall t \ge 0. \tag{3.1}$$

If the constant ν in (3.1) *is independent of r, the system* (2.1) *is said to be* globally uniformly exponentially stable *in $L^2(I)$.*

We first show that the system (2.1) is locally uniformly exponentially stable in $L^2(I)$.

PROPOSITION 3.2. *Assume that $a = a(\mu)$ is a C^2 function which satisfies*

$$|a(\mu)| \le C(1+|\mu|^p), \quad |a'(\mu)| \le C(1+|\mu|^{p-1}), \quad |a''(\mu)| \le C(1+|\mu|^{p-2}) \quad \text{for any } \mu \in \mathbb{R}$$

with $1 \le p < 4$ and $b \in L^2(I)$. Then the system (2.1) is locally uniformly exponentially stable in $L^2(I)$.

The following Carleman estimate [27, Prop. 2.3] will play a great role in establishing the unique continuation property of the GKdV equation as described by Lemma 3.4.

LEMMA 3.3. *Let T and l be positive numbers. Then there exist a smooth positive function ψ on $[0,l]$ and two constants $C_0 > 0$, $s_0 > 0$ such that for any*

$$q \in L^2(0,T;H^3(0,l)) \cap H^1(0,T;L^2(0,l)) \tag{3.2}$$

fulfilling

$$q(0,t) = q(l,t) = \partial_x q(l,t) = \partial_x^2 q(l,t) = 0$$

and for any $s \ge s_0$ we have

$$\begin{aligned} &\int_0^T\int_0^l \left\{ \frac{s^5}{t^5(T-t)^5}|q|^2 + \frac{s^3}{t^3(T-t)^3}|\partial_x q|^2 + \frac{s}{t(T-t)}|\partial_x^2 q|^2 \right\} \exp\left(-\frac{2s\psi(x)}{t(T-t)}\right) dxdt \\ &\le C_0 \int_0^T\int_0^l |\partial_t q + \partial_x^3 q + \partial_x q|^2 \exp\left(-\frac{2s\psi(x)}{t(T-t)}\right) dxdt. \end{aligned} \tag{3.3}$$

(Notice that Lemma 3.3 is stated in [27] under the extra assumption that $q \in C^3([0,l]\times[0,T])$, but (3.2) is what is really needed to perform the computations.)

LEMMA 3.4. *Let $l > 0$ and $T > 0$. If $v \in L^\infty(0,T;H^1(0,l))$ solves*

$$\begin{cases} \partial_t v + \partial_x v + a(v)\partial_x v + \partial_x^3 v = 0 & \text{in } (0,l)\times(0,T), \\ v(0,t) = 0 & \text{for a.e. } t \in (0,T), \\ v \equiv 0 & \text{in } (l',l)\times(0,T) \end{cases}$$

with $a \in C^0(\mathbb{R};\mathbb{R})$ and $0 < l' < l$, then $v \equiv 0$ in $(0,l)\times(0,T)$.

Proof. As the function v is not expected to fulfill (3.2), we have to smooth it. For any function $u = u(x,t)$ and any $h > 0$, we set

$$u^{[h]}(x,t) = \frac{1}{h}\int_t^{t+h} u(x,s)\,ds.$$

Recall that if $u \in L^p(0,T;V)$, where $1 \le p \le +\infty$ and V denotes any Banach space, then $u^{[h]} \in W^{1,p}(0,T-h;V)$, $||u^{[h]}||_{L^p(0,T-h;V)} \le ||u||_{L^p(0,T;V)}$, and for $p < \infty$ and $T' < T$

$$u^{[h]} \to u \quad \text{in } L^p(0,T';V) \text{ as } h \to 0.$$

Pick any $T' < T$. Then for any small enough number h, $v^{[h]} \in W^{1,\infty}(0,T';H_0^1(0,l))$, and $v^{[h]}$ solves

$$\partial_t v^{[h]} + \partial_x v^{[h]} + \partial_x^3 v^{[h]} + (a(v)\partial_x v)^{[h]} = 0 \quad \text{in } (0,l)\times(0,T'), \tag{3.4}$$

$$v^{[h]}(0,t) = 0 \quad \text{in } (0,T'), \tag{3.5}$$

$$v^{[h]} \equiv 0 \quad \text{in } (l',l)\times(0,T'). \tag{3.6}$$

As $v \in L^\infty(0,T;H^1(0,l))$, $a(v)\partial_x v \in L^\infty(0,T;L^2(0,l))$; hence it follows from (3.4) that $v^{[h]}_{xxx} \in L^\infty(0,T';L^2(0,l))$ and so $v^{[h]} \in L^\infty(0,T';H^3(0,l))$. It follows then from Lemma 3.3 that for any $s \geq s_0$ we have

$$
\begin{aligned}
(3.7)\quad & \int_0^{T'}\int_0^l \left\{ \frac{s^5}{t^5(T'-t)^5}|v^{[h]}|^2 + \frac{s^3}{t^3(T'-t)^3}|\partial_x v^{[h]}|^2 \right. \\
& \qquad \left. + \frac{s}{t(T'-t)}|\partial_x^2 v^{[h]}|^2 \right\} \exp\left(-\frac{2s\psi(x)}{t(T'-t)}\right) dxdt \\
& \leq C_0 \int_0^{T'}\int_0^l |(a(v)\partial_x v)^{[h]}|^2 \exp\left(-\frac{2s\psi(x)}{t(T'-t)}\right) dxdt \\
& \leq 2C_0 \int_0^{T'}\int_0^l |a(v)\partial_x v^{[h]}|^2 \exp\left(-\frac{2s\psi(x)}{t(T'-t)}\right) dxdt \\
& \quad + 2C_0 \int_0^{T'}\int_0^l |(a(v)\partial_x v)^{[h]} - a(v)\partial_x v^{[h]}|^2 \exp\left(-\frac{2s\psi(x)}{t(T'-t)}\right) dxdt \\
& \equiv I_1 + I_2.
\end{aligned}
$$

We first estimate I_1. Since $a(v) \in L^\infty(0,T;L^\infty(0,l))$, we have

$$
(3.8)\qquad I_1 \leq C \int_0^{T'}\int_0^l |\partial_x v^{[h]}|^2 \exp\left(-\frac{2s\psi(x)}{t(T'-t)}\right) dxdt
$$

for some constant C which does not depend on h. Comparing the powers of s in the right-hand side of (3.8) with those in the left-hand side of (3.7), we deduce that the term I_1 in (3.7) may be dropped by increasing the constants C_0 and s_0 in a convenient way. From now on, we fix s, say, to the value s_0. We claim that $I_2 \to 0$ as $h \to 0$. As $\exp(-\frac{2s_0\psi(x)}{t(T'-t)}) \leq 1$, it is sufficient to prove that

$$
(3.9)\qquad (a(v)\partial_x v)^{[h]} \to a(v)\partial_x v \quad \text{in } L^2(0,T';L^2(0,l)),
$$

$$
(3.10)\qquad a(v)\partial_x v^{[h]} \to a(v)\partial_x v \quad \text{in } L^2(0,T';L^2(0,l)).
$$

Property (3.9) is obvious for $a(v)\partial_x v \in L^2(0,T;L^2(0,l))$, and (3.10) follows from the fact that $a(v) \in L^\infty(0,T;L^\infty(0,l))$ and that $v \in L^2(0,T;H^1(0,l))$. We conclude that as $h \to 0$ the integral term

$$
\begin{aligned}
\int_0^{T'}\int_0^l & \left\{ \frac{s_0^5}{t^5(T'-t)^5}|v^{[h]}|^2 + \frac{s_0^3}{t^3(T'-t)^3}|\partial_x v^{[h]}|^2 \right. \\
& \left. + \frac{s_0}{t(T'-t)}|\partial_x^2 v^{[h]}|^2 \right\} \exp\left(-\frac{2s\psi(x)}{t(T'-t)}\right) dxdt
\end{aligned}
$$

tends to 0. On the other hand, $v^{[h]} \to v$ in $L^2(0,T';L^2(0,l))$. It follows that $v \equiv 0$ in $(0,l)\times(0,T')$. As T' may be taken arbitrarily close to T, we infer that $v \equiv 0$ in $(0,l)\times(0,T)$. This completes the proof of Lemma 3.4. □

The following unique continuation property is a direct consequence of Lemma 3.4.

LEMMA 3.5. *Let $L > 0$ and $T > 0$ be two real numbers, and let $\omega \subset (0,L)$ be a nonempty open set. If $v \in L^\infty(0,T;H^1(I))$ solves*

$$
\begin{cases}
\partial_t v + \partial_x v + a(v)\partial_x v + \partial_x^3 v = 0 & \text{in } (0,L)\times(0,T), \\
v(0,t) = v(L,t) = 0 & \text{for a.e. } t \in (0,T), \\
v \equiv 0 & \text{in } \omega \times (0,T)
\end{cases}
$$

with $a \in C^0(\mathbb{R};\mathbb{R})$, then $v \equiv 0$ in $(0,L)\times(0,T)$.

Proof. Without loss of generality we may assume that $\omega = (l_1, l_2)$ with $0 \leq l_1 < l_2 \leq L$. Pick $l = (l_1 + l_2)/2$. Applying Lemma 3.4 to the function $v(x,t)$ on $(0,l)\times(0,T)$ and then to the function $v(L-x,T-t)$ on $(0,L-l)\times(0,T)$, we conclude that $v \equiv 0$ on $(0,L)\times(0,T)$. □

Now we are in a position to present the proof of Proposition 3.2.

Proof of Proposition 3.2. First, notice that if u is a weak solution of (2.1) emanating from a given initial state $\phi \in L^2(I)$, then by Lemma 2.5 there exists a constant C_p depending only on p such that for any $T > 0$,

$$||u(\cdot,T)||^2_{L^2(I)} = ||\phi||^2_{L^2(I)} - \int_0^T |\partial_x u(0,t)|^2\, dt - 2\int_0^T \int_0^L b|u|^2\, dxdt \tag{3.11}$$

and

$$||\partial_x u||^2_{L^2(0,T;L^2(I))} \leq \frac{L+(C+1)T}{2}||\phi||^2_{L^2(I)} + C_p T||\phi||^{\frac{8+2p}{4-p}}_{L^2(I)}. \tag{3.12}$$

On the other hand, multiplying (2.1) by $(T-t)u$ yields

$$\begin{aligned} T||\phi||^2_{L^2(I)} = &\int_0^T \int_0^L |u(x,t)|^2\, dxdt + \int_0^T (T-t)|\partial_x u(0,t)|^2 dt \\ &+ 2\int_0^T \int_0^L (T-t)b|u(x,t)|^2 dxdt. \end{aligned} \tag{3.13}$$

We now proceed as in the proof of [24, Thm. 3.1]. Let $r > 0$ be given. The proof would be complete if the following claim is true.

Claim 3. For any $T > 0$ and any $r > 0$ there exists a positive constant $C = C(r,T)$ such that for any weak solution u issuing from a state $\phi \in L^2(I)$ with $||\phi||_{L^2(I)} < r$, it holds that

$$||\phi||^2_{L^2(I)} \leq C\left(\int_0^T |\partial_x u(0,t)|^2 dt + 2\int_0^T \int_0^L b(x)|u(x,t)|^2 dxdt\right). \tag{3.14}$$

Indeed, if Claim 3 is proved, then it follows from (3.11) and (3.14) that there exists a constant $\gamma \in (0,1)$ such that

$$||u(\cdot,kT)||^2_{L^2(I)} \leq \gamma^k ||\phi||^2_{L^2(I)} \qquad \forall k \geq 0.$$

Since $||u(\cdot,t)||_{L^2(I)} \leq ||u(\cdot,kT)||_{L^2(I)}$ for $kT \leq t < (k+1)T$, we readily obtain

$$||u(\cdot,t)||^2_{L^2(I)} \leq \frac{1}{\gamma}||\phi||^2_{L^2(I)} e^{\frac{\log\gamma}{T}t}$$

from which the result stated in Proposition 3.2 follows.

To prove Claim 3, because of the identity (3.13), it is sufficient to show that there exists some constant $C_1 > 0$ such that

$$\int_0^T \int_0^L |u|^2 dxdt \leq C_1\left(\int_0^T |\partial_x u(0,t)|^2 dt + 2\int_0^T \int_0^L b(x)|u|^2 dxdt\right), \tag{3.15}$$

provided that $||u(.,0)||_{L^2(I)} < r$. To this end, we argue by contradiction.

Suppose that (3.15) fails to be true. Then there exists a sequence of weak solutions $u_n \in C_w([0,T];L^2(I)) \cap L^2(0,T;H_0^1(I))$ of (2.1) with

$$||u_n(\cdot,0)||_{L^2(I)} < r \tag{3.16}$$

and such that

$$\lim_{n\to+\infty} \frac{||u_n||^2_{L^2(0,T;L^2(I))}}{\int_0^T |\partial_x u_n(0,t)|^2 dt + 2\int_0^T \int_0^L b(x)|u_n|^2 dxdt} = +\infty.$$

Let $\lambda_n := ||u_n||_{L^2(0,T;L^2(I))}$ and $v_n(x,t) := u_n(x,t)/\lambda_n$. Notice that λ_n is bounded from above, according to (3.11) and (3.16). Hence, extracting a subsequence if needed, we may assume that

$$\lambda_n \to \lambda \geq 0.$$

Then v_n fulfills

$$\begin{aligned}
&\partial_t v_n + \partial_x^3 v_n + \partial_x v_n + a(\lambda_n v_n)\partial_x v_n + b v_n = 0,\\
&v_n(0,t) = v_n(L,t) = \partial_x v_n(L,t) = 0,\\
&||v_n||_{L^2(0,T;L^2(I))} = 1,
\end{aligned}$$

and

$$\int_0^T |\partial_x v_n(0,t)|^2 dt + 2\int_0^T \int_0^L b(x)|v_n|^2 dxdt \to 0$$

as $n \to +\infty$. It follows from (3.13) that $||v_n(.,0)||_{L^2(I)} = ||u_n(.,0)/\lambda_n||_{L^2(I)}$ is bounded. Noticing that

$$|a(\lambda_n \mu)| \leq C(1 + |\lambda_n|^p |\mu|^p) \leq C'(1 + |\mu|^p),$$

where C' denotes a positive constant which does not depend on μ and n, we obtain from (3.12) applied to v_n that v_n is bounded in $L^2(0,T;H_0^1(I))$. Proceeding as in the proof of Claim 2, it is easily seen that $a(\lambda_n v_n)\partial_x v_n$ and $\partial_t v_n$ are bounded in $L^\alpha(0,T;H^{-2}(I))$ with $\alpha > 1$. Extracting a subsequence, we may assume that

$$v_n \rightharpoonup v \quad \text{in } L^\infty(0,T;L^2(I)) \text{ weak}*; \tag{3.17}$$

$$v_n \rightharpoonup v \quad \text{in } L^2(0,T;H_0^1(I)) \text{ weak}; \tag{3.18}$$

$$v_n \to v \quad \text{in } L^2(0,T;L^2(I)) \text{ and a.e.}; \tag{3.19}$$

$$v_n \to v \quad \text{in } C([0,T],H^{-1}(I)); \tag{3.20}$$

$$a(\lambda_n v_n)\partial_x v_n \to a(\lambda v)\partial_x v \text{ in } \mathcal{D}'(I\times(0,T)). \tag{3.21}$$

It follows that v solves

$$\partial_t v + \partial_x^3 v + \partial_x v + a(\lambda v)\partial_x v + bv = 0 \quad \text{in } \mathcal{D}'(I\times(0,T)), \tag{3.22}$$

$$v(0,t) = v(L,t) = \partial_x v(L,t) = 0. \tag{3.23}$$

Moreover, $||v||_{L^2(0,T;L^2(I))} = 1$ and $0 = \int_0^T \int_0^L b(x)|v|^2\, dxdt$; hence

$$v \equiv 0 \quad \text{on } \omega \times (0,T).$$

To conclude, the following result is needed.

LEMMA 3.6. *Let $0 < t_1 < t_2 < T$. Then there exists a subinterval $(t_1', t_2') \subset (t_1, t_2)$ such that $v \in L^\infty(t_1', t_2'; H^1(I))$.*

Proof of Lemma 3.6. C will denote here a constant which may vary from line to line. Using (2.39) for each weak solution w_n, we may pick a sequence $\{a_n\}$ in $C_0^\infty(\mathbb{R};\mathbb{R})$ fulfilling (2.32) and such that, if w_n denotes the solution of

$$\begin{cases} \partial_t w_n + \partial_x^3 w_n + \partial_x w_n + a_n(\lambda_n w_n)\partial_x w_n + b w_n = 0, \quad x \in (0,L), \quad t \geq 0, \\ \\ w_n(x,0) = v_n(x,0), \\ \\ w_n(0,t) = w_n(L,t) = \partial_x w_n(L,t) = 0, \end{cases}$$

we have as $n \to \infty$

$$v_n - w_n \to 0 \quad \text{in } C([0,T], H^{-1}(I)). \tag{3.24}$$

As

$$||w_n||_{L^2(0,T;H^1(I))} \leq C,$$

we may pick a sequence $\{\tau_n\}$ in $(t_1, (t_1+t_2)/2)$ such that $\tau_n \to \tau$ and $||w_n(\tau_n)||_{H^1(I)} \leq C$. Using (3.20) and (3.24) we obtain that

$$w_n(\tau_n + \cdot) \to v(\tau + \cdot) \quad \text{in } C([0,\varepsilon], H^{-1}(I))$$

for any $\varepsilon < (t_2 - t_1)/2$. According to Theorem 2.13, for ε sufficiently small,

$$||w_n(\tau_n + \cdot)||_{L^\infty(0,\varepsilon;H^1(I))} \leq C;$$

hence $v \in L^\infty(\tau, \tau+\varepsilon; H^1(I))$. □

Let $t_1 \in (0,T)$, and let $t_2 \in (t_1, T)$. According to Lemma 3.6, $v \in L^\infty(t_1', t_2'; H^1(I))$ for some interval $(t_1', t_2') \subset (t_1, t_2)$. It follows then from Lemma 3.5 that $v \equiv 0$ on $(0,L) \times (t_1', t_2')$. As t_2 is arbitrarily close to t_1, we obtain by continuity of v in $H^{-1}(I)$ that $v(.,t) = 0$. Thus $v \equiv 0$, which contradicts the fact that $||v||_{L^2(0,T;L^2(I))} = 1$. The proof is complete. □

Now we present the main result of this paper, asserting that the system (2.1) is globally uniformly exponentially stable in the space $L^2(I)$.

THEOREM 3.7. *Assume that $a = a(\mu)$ is a C^2 function which satisfies*

$$|a(\mu)| \leq C(1+|\mu|^p), \quad |a'(\mu)| \leq C(1+|\mu|^{p-1}), \quad |a''(\mu)| \leq C(1+|\mu|^{p-2}) \quad \textit{for any } \mu \in \mathbb{R}$$

with $1 \leq p < 4$ and $b \in L^2(I)$. Then the system (2.1) *is globally uniformly exponentially stable in the space $L^2(I)$; i.e., there exist a $\nu^* > 0$ and a continuous nonnegative function $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ such that for a given $\phi \in L^2(I)$, any weak solution u of* (2.1) *satisfies*

$$\|u(\cdot,t)\|_{L^2(I)} \leq \alpha_0(\|\phi\|_{L^2(I)}) e^{-\nu^* t} \qquad \forall\ t \geq 0.$$

Proof. By Proposition 3.2, there exists a $\nu^* > 0$ such that if

$$\|\phi\|_{L^2(I)} \leq 1,$$

then the corresponding solution u of (2.1) satisfies

$$\|u(\cdot,t)\|_{L^2(I)} \leq C\|\phi\|_{L^2(I)} e^{-\nu^* t} \qquad \forall\ t \geq 0 \tag{3.25}$$

for some constant $C \geq 1$ which depends only on $\|\phi\|_{L^2(I)}$. In addition, for a given $r > 0$, there exist two constants $C_r > 0$ and $\nu_r > 0$ such that if $\|\phi\|_{L^2(I)} \leq r$, then any weak solution u of (2.1) satisfies

$$\|u(\cdot,t)\|_{L^2(I)} \leq C_r||\phi||_{L^2(I)}e^{-\nu_r t} \qquad \forall\ t \geq 0.$$

Consequently, setting $T_r := \nu_r^{-1}\ln(rC_r)$, we have that

$$\begin{aligned}\|u(\cdot,t)\|_{L^2(I)} &\leq C\|u(\cdot,T_r)\|_{L^2(I)}e^{-\nu^*(t-T_r)} \qquad \forall\ t \geq T_r\\ &\leq CC_r||\phi||_{L^2(I)}e^{\nu^* T_r}e^{-\nu^* t} \qquad \forall\ t \geq 0.\end{aligned}$$

The proof is complete. □

For a given function $u = u(x,t)$, a real number $s \geq 0$, and an interval (a,b), define

$$\|v\|^s_{Y_{(a,b)}} := \Big(\|v\|^2_{C([a,b];H^s(I))} + \|v\|^2_{L^2([a,b];H^{s+1}(I))} + \|\partial_x v\|^2_{C([0,L];L^2(a,b))}\Big)^{1/2}.$$

As a direct corollary of Theorem 3.7 and Lemma 2.5, we have the following result.

COROLLARY 3.8. *Under the assumptions of Theorem* 3.7*, there exists a* $\nu^* > 0$ *such that for any* $T > 0$ *and* $\phi \in L^2(I)$*, there exists a nonnegative continuous function* $\alpha_0 :\ \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ *such that any weak solution* u *of* (2.1) *satisfies*

$$\|u\|_{Y^0_{(t,t+T)}} \leq \alpha_0(\|\phi\|_{L^2(I)}, T)e^{-\nu^* t}$$

for any $t \geq 0$.

Under the assumptions of Theorem 2.9, if $\phi \in H^s(I)$ with $0 \leq s \leq 3$ satisfies the compatibility conditions (1.16), then the corresponding solution u of (2.1) belongs to the space $C(\mathbb{R}^+; H^s(I))$. Our next theorem shows that the system (2.1) is globally uniformly exponentially stable in the space $H^3(I)$. First, we show that the system (2.1) is globally uniformly exponentially stable in the space $H^3(I)$ if $\phi \in H^3(I)$ as described below.

PROPOSITION 3.9. *Under the assumptions of Theorem* 2.9 *with* $a(0) = 0$ *additionally, there exist a* $\nu^* > 0$ *and a continuous nonnegative function* $\alpha_3 :\ \mathbb{R}^+ \to \mathbb{R}^+$ *such that for any* $\phi \in H^3(I)$ *satisfying the compatibility conditions, the corresponding solution* u *satisfies*

$$\|u(\cdot,t)\|_{H^3(I)} \leq \alpha_3(\|\phi\|_{L^2(I)})\|\phi\|_{H^3(I)}e^{-\nu^* t} \qquad \forall\ t \geq 0.$$

Proof. Let $v = u_t$. Then v solves the following IBVP linearized KdV equation:

$$\begin{cases} \partial_t v + \partial_x v + \partial_x(a(u)v) + bv + \partial_x^3 v = 0, \quad x \in (0,L), \quad t \geq 0,\\ v(x,0) = \phi^*(x),\\ v(0,t) = 0, \quad v(L,t) = 0, \quad v_x(L,t) = 0, \end{cases} \tag{3.26}$$

where

$$\phi^* = -\phi''' - a(\phi)\phi' - \phi' - b\phi.$$

Note that $\phi \in H^3(I)$ implies that $\phi^* \in L^2(I)$ and that there exists a constant $C = C(||\phi||_{L^2(I)})$ such that $\|\phi^*\|_{L^2(I)} \leq C(\|\phi\|_{H^3(I)})$ for any $\phi \in H^3(I)$. The same argument as that in the proof of Proposition 2.6 shows that

$$\|v\|_{Y_{0,T}} \leq \sigma(\|u\|_{Y_{0,T}})\|\phi^*\|_{L^2(I)}, \tag{3.27}$$

where $\sigma : \mathbb{R}^+ \to \mathbb{R}^+$ is a nondecreasing continuous function. By the semigroup property of the system (3.26), we can rewrite (3.27) as

$$\|v\|_{Y_{0,[t,t+T]}} \leq \sigma(\|u\|_{Y_{0,[t,t+T]}})\|v(\cdot,t)\|_{L^2(I)} \tag{3.28}$$

for any $t \geq 0$, where $Y_{0,[t,t+T]}$ is defined as $Y_{0,T} =: Y_{0,[0,T]}$. Using the notation of the semigroup $W_b(t)$, the solution v of (3.26) may be written as

$$v(t) = W_b(t)\phi^* - \int_0^t W_b(t-\tau)(a(u)v)_x(\tau)d\tau.$$

According to Proposition 3.2 (with $a \equiv 0$), $u_1(x,t) = W_b(t)\phi^*(x)$ satisfies

$$\|u_1(\cdot,t)\|_{L^2(I)} \leq Ce^{-\alpha t}\|\phi^*\|_{L^2(I)} \qquad \forall t \geq 0,$$

where C and α are some positive constants independent of ϕ^*. In addition, by Lemma 2.3, $u_2(x,t) = \int_0^t W_b(t-\tau)(\partial_x(a(u)v))(\tau)d\tau$ satisfies

$$\|u_2(\cdot,T)\| \leq C_T\|u\|^p_{Y_{0,T}}\|v\|_{Y_{0,T}}.$$

Thus,

$$\|v(\cdot,T)\|_{L^2(I)} \leq C_1e^{-\alpha T}\|\phi^*\|_{L^2(I)} + C_2\|u\|^p_{Y_{0,T}}\sigma(\|u\|_{Y_{0,T}})\|\phi^*\|_{L^2(I)}.$$

Let $y_n = v(\cdot,nT)$ for $n = 0,1,2,\ldots$, and let w be the solution of the IBVP

$$\begin{cases} \partial_t w + \partial_x w + \partial_x(qw) + bw + \partial_x^3 w = 0, \quad x \in (0,L), \quad 0 \leq t, \\ w(x,0) = y_n, \\ w(0,t) = 0, \quad w(L,t) = 0, \quad w_x(L,t) = 0 \end{cases} \tag{3.29}$$

with $q(x,t) = a(u(x,t+nT))$. Then $y_{n+1}(x) = w(x,T)$ by the semigroup properties of the system (3.26). Consequently, we have the following estimate for y_{n+1}:

$$\|y_{n+1}\|_{L^2(I)} \leq C_1e^{-\alpha T}\|y_n\|_{L^2(I)} + C_2\|u\|^p_{Y_{0,[nT,(n+1)T]}}\sigma(\|u\|_{Y_{0,[nT,(n+1)T]}})\|y_n\|_{L^2(I)}$$

for any $n \geq 0$. Choose T and β such that

$$C_1e^{-\alpha T} = \gamma < 1$$

and

$$\gamma + C_2\beta^p\sigma(\beta) = r < 1.$$

In addition, according to Corollary 3.8, one can choose $N > 0$ such that

$$\|u\|_{Y_{0,[nT,(n+1)T]}} < \beta \qquad \forall\, n \geq N.$$

Consequently, for such chosen T and N,

$$\|y_{n+1}\|_{L^2(I)} \leq r\|y_n\|_{L^2(I)} \qquad \forall\, n \geq N.$$

This inequality implies that there exists a $\nu^* > 0$ such that

$$\|v(\cdot,t)\|_{L^2(I)} = \|\partial_t u(\cdot,t)\|_{L^2(I)} \leq C\alpha_3\left(\|\phi\|_{L^2(I)}\right)\|\phi\|_{H^3(I)}e^{-\nu^* t} \qquad \forall\, t \geq 0. \tag{3.30}$$

Then, using (3.28) and Corollary 3.8 yields that

$$\|v\|_{Y_{0,[t,t+T]}} \le \alpha_4\left(\|\phi\|_{L^2(I)}, T\right)\|\phi\|_{H^3(I)}e^{-\nu^* t} \tag{3.31}$$

for any $t \ge 0$. In particular,

$$\|u_t\|_{L^2(t,t+T;H^1(I))} \le \alpha_4\left(\|\phi\|_{L^2(I)}, T\right)\|\phi\|_{H^3(I)}e^{-\nu^* t}$$

and

$$\|u\|_{L^2(t,t+T;H^1(I))} \le \alpha_0\left(\|\phi\|_{L^2(I)}, T\right)e^{-\nu^* t}$$

for any $t \ge 0$. Consequently,

$$\|a(u(\cdot,t))\partial_x u(\cdot,t)\|_{L^2(I)} + \|\partial_x u(\cdot,t)\|_{L^2(I)} \le \alpha_5\left(\|\phi\|_{L^2(I)}, T\right)\|\phi\|_{H^3(I)}e^{-\nu^* t}$$

for any $t \ge 0$.

The conclusion of Proposition 3.9 follows from the above estimates, since

$$\partial_x^3 u = -\partial_t u - a(u)\partial_x u - \partial_x u - bu.$$

The proof is complete. □

THEOREM 3.10. *Let a be a C^1 function satisfying $a(0) = 0$ and*

$$|a(\mu)| \le C(1+|\mu|^p), \quad |a'(\mu)| \le C(1+|\mu|^{p-1}) \quad \text{for any } \mu \in \mathbb{R}$$

with $1 \le p < 2$. Then there exists a $\nu > 0$ depending only on L such that for any $\phi \in L^2(I)$ and any $\varepsilon > 0$, the corresponding solution u of (2.1) belongs to the space $C(\mathbb{R}^{+*}; H^3(I))$ and satisfies

$$\|u(\cdot,t)\|_{H^3(I)} \le C_\varepsilon e^{-\nu t} \qquad \forall\ t \ge \varepsilon$$

for some constant $C_\varepsilon > 0$.

Proof. It follows directly from Corollary 2.10 and Proposition 3.9. □

Acknowledgments. This work was begun while the second author was visiting the Université Henri Poincaré Nancy 1 in 2003 and completed while the first author was visiting the University of Cincinnati in 2004. The authors thank both institutions for their hospitality, Enrique Zuazua, who brought the paper [22] to their attention, and the anonymous referees for their very useful comments.

REFERENCES

[1] J. L. BONA AND L. R. SCOTT, *Solutions of the Korteweg-de Vries equation in fractional order Sobolev spaces*, Duke Math. J., 43 (1976), pp. 87–99.

[2] J. L. BONA AND R. SMITH, *The initial-value problem for the Korteweg-de Vries equation*, Philos. Trans. Roy. Soc. London Ser. A, 278 (1975), pp. 555–601.

[3] J. L. BONA, S. M. SUN, AND B.-Y. ZHANG, *A non-homogeneous boundary-value problem for the Korteweg-de Vries Equation in a quarter plane*, Trans. Amer. Math. Soc., 354 (2001), pp. 427–490.

[4] J. L. BONA, S. M. SUN, AND B.-Y. ZHANG, *A non-homogeneous boundary-value problem for the Korteweg-de Vries equation posed on a finite domain*, Comm. Partial Differential Equations, 28 (2003), pp. 1391–1436.

[5] J. BOURGAIN, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, part* I: *Schrödinger equations*, Geom. Funct. Anal., 3 (1993), pp. 107–156.

[6] J. Bourgain, *Fourier transform restriction phenomena for certain lattice subsets and applications to non-linear evolution equations, part* II: *the KdV equation*, Geom. Funct. Anal., 3 (1993), pp. 209–262.
[7] B. A. Bubnov, *Generalized boundary value problems for the Korteweg-de Vries equation in bounded domain*, Differential Equations, 15 (1979), pp. 17–21.
[8] B. A. Bubnov, *Solvability in the large of nonlinear boundary-value problems for the Korteweg-de Vries equations*, Differential Equations, 16 (1980), pp. 24–30.
[9] T. Colin and J.-M. Ghidaglia, *An initial-boundary-value problem for the Korteweg-de Vries equation posed on a finite interval*, Adv. Differential Equations, 6 (2001), pp. 1463–1492.
[10] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp global well-posedness for KdV and modified KdV on R and T*, J. Amer. Math. Soc., 16 (2003), pp. 705–749.
[11] A. V. Faminskii, *On an initial boundary value problem in a bounded domain for the generalized Korteweg-de Vries equation*, Funct. Differ. Equ., 8 (2001), pp. 183–194.
[12] T. Kappeler and P. Topalov, *Global Well-Posedness of KdV in* $H^{-1}(T, R)$, Preprint 12-2003, Universität Zürich, Zürich, Switzerland.
[13] T. Kato, *Quasilinear equations of evolution, with applications to partial differential equations*, in Spectral Theory and Differential Equations, Lecture Notes in Math. 448, Springer-Verlag, Berlin, 1975, pp. 27–50.
[14] T. Kato, *On the Korteweg-de Vries equation*, Manuscripta Math., 29 (1979), pp. 89–99.
[15] T. Kato, *On the Cauchy problem for the (generalized) Korteweg-de Vries equations*, Stud. Appl. Math. 8, in Studies in Applied Mathematics, Academic Press, New York, 1983, pp. 93–128.
[16] C. E. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the KdV equation*, J. Amer. Math. Soc., 9 (1996), pp. 573–603.
[17] V. Komornik, D. L. Russell, and B.-Y. Zhang, *Stabilization de l'équation de Korteweg-de Vries*, C. R. Acad. Sci. Paris Sér. I Math., 312 (1991), pp. 841–843.
[18] D. J. Korteweg and G. de Vries, *On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves*, Philos. Mag., 39 (1895), pp. 422–443.
[19] P. D. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.
[20] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes Aux Limites non Linéaires*, Dunod, Paris, 1969.
[21] R. M. Miura, *The Korteweg–de Vries equation: A survey of results*, SIAM Rev., 18 (1976), pp. 412–459.
[22] A. F. Pazoto, *Unique continuation and decay for the Korteweg-de Vries equation with localized damping*, ESAIM Control Optim. Calc. Var., 11 (2005), pp. 473–486.
[23] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1983.
[24] G. Perla Menzala, C. F. Vasconcellos, and E. Zuazua, *Stabilization of the Korteweg-de Vries equation with localized damping*, Quart. Appl. Math., 60 (2002), pp. 111–129.
[25] L. Rosier, *Exact boundary controllability for the Korteweg-de Vries equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 33–55.
[26] L. Rosier, *Exact boundary controllability for the linear Korteweg–de Vries equation on the half-line*, SIAM J. Control Optim., 39 (2000), pp. 331–351.
[27] L. Rosier, *Control of the surface of a fluid by a wavemaker*, ESAIM Control Optim. Cal. Var., 10 (2004), pp. 346–380.
[28] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[29] D. L. Russell and B.-Y. Zhang, *Controllability and stabilizability of the third-order linear dispersion equation on a periodic domain*, SIAM J. Control Optim., 31 (1993), pp. 659–676.
[30] D. L. Russell and B.-Y. Zhang, *Smoothing properties of solutions of the Korteweg-de Vries equation on a periodic domain with point dissipation*, J. Math. Anal. Appl., 190 (1995), pp. 449–488.
[31] D. L. Russell and B.-Y. Zhang, *Exact controllability and stabilizability of the Korteweg-de Vries equation*, Trans. Amer. Math. Soc., 348 (1996), pp. 3643–3672.
[32] J. C. Saut and B. Scheurer, *Unique continuation for some evolution equations*, J. Differential Equations, 66 (1987), pp. 118–139.
[33] J. Simon, *Compact sets in the space* $L^p(0, T; B)$, Ann. Mat. Pura Appl. 4, (1987), pp. 65–96.
[34] S. M. Sun, *The Korteweg–de Vries equation on a periodic domain with singular-point dissipation*, SIAM J. Control Optim., 34 (1996), pp. 892–912.

[35] L. Tartar, *Interpolation non linéaire et régularité*, J. Funct. Anal., 9 (1972), pp. 469–489.
[36] R. Temam, *Navier-Stokes Equations,* North–Holland, Amsterdam, 1984.
[37] B.-Y. Zhang, *Unique continuation for the Korteweg–de Vries equation*, SIAM J. Math. Anal., 23 (1992), pp. 55–71.
[38] B.-Y. Zhang, *Boundary stabilization of the Korteweg-de Vries equation*, in Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, Internat. Ser. Numer. Math. 118, Birkhauser, Basel, 1994, pp. 371–389.
[39] B.-Y. Zhang, *Analyticity of solutions for the generalized Korteweg–de Vries equation with respect to their initial values*, SIAM J. Math. Anal., 26 (1995), pp. 1488–1513.
[40] B.-Y. Zhang, *Exact boundary controllability of the Korteweg–de Vries equation*, SIAM J. Control Optim., 37 (1999), pp. 543–565.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 957–971

COMPLETE STABILITY ANALYSIS OF NEUTRAL-TYPE FIRST ORDER TWO-TIME-DELAY SYSTEMS WITH CROSS-TALKING DELAYS*

RIFAT SIPAHI† AND NEJAT OLGAC†

Abstract. The stability robustness of first order linear time invariant dynamics of *neutral type* with multiple time delays against delay uncertainties is taken into consideration. We depart from a simpler investigation of Hale and Huang [*J. Math. Anal. Appl.*, 178 (1993), pp. 344–362], which studies the same problem for retarded-type systems. On this basis we further introduce two challenging features by including (a) terms that add neutral dynamics and (b) an additional term that introduces cross-talk between the multiple delays. To the best of the authors' knowledge, the stability posture of this class of systems can be treated only by a unique procedure. It is known as cluster treatment of characteristic roots (CTCR), which was recently developed for *retarded*-type dynamics. We first show the applicability of CTCR to the stability analysis of *neutral*-type multiple-delay dynamics. Next, we prove the well-known "small-delay" phenomenon for the dynamics at hand, interestingly, as a natural by-product of the CTCR paradigm. Finally, we present several case studies to display the steps and the strengths of CTCR. This deployment is scalable to treat similar problems with higher order dynamics, which have direct ramifications to some practical control applications.

Key words. neutral time-delayed dynamics, stability, multiple time delays, cluster treatment of characteristic roots (CTCR)

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/050633810

1. Introduction.

We consider a general class of delayed differential equations (DDEs) from the stability robustness perspective, which has not been successfully investigated in the literature. The class is of first order *neutral type*, linear time invariant (LTI) time-delay systems with two cross-talking delays. The dynamics is written in conventional form [6], [7], [8] as

$$\begin{aligned} &\frac{d}{dt}[x(t) - ax(t-\tau_1) - bx(t-\tau_2)] \\ &= cx(t-\tau_1) + dx(t-\tau_2) + fx(t-\tau_1-\tau_2) + gx(t), \end{aligned} \tag{1.1}$$

where a, b, c, d, f, g are all real scalars as well as the dependent variable $x(t)$, $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$. We highlight the $fx(t-\tau_1-\tau_2)$ term as the "delay cross-talk" feature in the problem. The problem is to analyze the stability robustness of this system against time-delay uncertainties in the semi-infinite first quadrant of $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$. The characteristic equation of these dynamics is transcendental,

$$CE = s(1 - ae^{-\tau_1 s} - be^{-\tau_2 s}) - ce^{-\tau_1 s} - de^{-\tau_2 s} - fe^{-(\tau_1+\tau_2)s} - g = 0, \tag{1.2}$$

where arguments of CE, (s, τ_1, τ_2) are suppressed. The stability robustness question of this system reduces to finding (τ_1, τ_2) regions where all the characteristic roots

*Received by the editors June 16, 2005; accepted for publication (in revised form) January 28, 2006; published electronically July 31, 2006. This work was supported partially by research funds from the DoE (DE-FG02-04ER25656) and the NSF (CMS-0439980, CMS-0539980, DMI 0522910).
http://www.siam.org/journals/sicon/45-3/63381.html

†Mechanical Engineering Department, University of Connecticut, Storrs, CT 06269-3139 (rsipahi @hds.utc.fr, olgac@engr.uconn.edu). The first author is currently at Université de Technologie de Compiègne as a postdoc under Chateaubriand Scholarship 2005–2006.

of (1.2) remain in $\mathbf{C}^-$, the open left half of the complex plane. For a notational selection we represent the right half open plane by $\mathbf{C}^+$ and the imaginary axis by $\mathbf{C}^0$ in the rest of the text. The entire complex plane becomes the union of all three, $\mathbf{C} = \mathbf{C}^- \cup \mathbf{C}^0 \cup \mathbf{C}^+$.

It is clear from (1.2) that the highest power of s is accompanied by exponential terms (a and b nonzero) with time-delay signatures, qualifying it as a *neutral type* [2], [14], [6], [7], [8], [3], [1], [15], [20]. Furthermore, the term multiplying f introduces the cross-talk between the delays. That is, the delays τ_1 and τ_2 appear in summation form. We wish to make a clarifying statement here. Although the dynamics in (1.1) are selected as scalar (for ease of conveyance only), the procedure described in the rest of the paper is scalable to higher-dimensional systems (rather than scalar), for which such cross-talk among the delays appears quite naturally [19], [20]. For instance, if one considers the two-dimensional state-space representation of a retarded system as

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{B}_1\mathbf{y}(t-\tau_1) + \mathbf{B}_2\mathbf{y}(t-\tau_2), \qquad y \in \mathbf{R}^2,\ \mathbf{A},\ \mathbf{B}_1,\ \mathbf{B}_2 \in \mathbf{R}^{2\times 2}, \tag{1.3}$$

it is trivial to see that the respective characteristic equation typically carries the signature of $e^{-(\tau_1+\tau_2)s}$ (i.e., cross-talking time delays) beyond the individual delay terms $e^{-\tau_1 s}$ and $e^{-\tau_2 s}$. The system is a first order vector equation, which can be decoupled into two second order scalar equations.

To the best of the authors' knowledge, no methodology exists in the literature today to resolve the stability robustness of such dynamics. For simplicity of the treatment (without loss of generality), we select scalar dynamics in this paper. In this regard, (1.2) is an extended version of the retarded-type characteristic equation treated in a cornerstone study [5] in "multiple-delay systems" literature, where a, b, f are all taken to be zero.

There are some practical ramifications of the class of dynamics we consider here. For instance, multiple time-delay feedback control is used to tune a vibration absorber (called the "delayed resonator") to two distinct excitation frequencies, in (3) in [22], with $n = 2$. The cited investigation, however, does not address the stability question of such systems. Furthermore, in (2) in [16] the authors experimentally demonstrate the viability of using acceleration feedback with time delay that renders neutral dynamics for the *delayed resonator* vibration absorber. As such, neutral dynamics with multiple delays becomes a practical phenomenon, of which the stability repercussions need further investigation. And the scalar form of these dynamics, as given in (1.1), is a meaningful starting platform to present the conceptual process we wish to follow.

It is well known that neutral-type DDEs exhibit drastic differences in their characteristic behavior compared to the retarded-type dynamics [14], [6], [7], [8], [15]. This difference may occur when either one or both components of the delay vector $\{\boldsymbol{\tau}\} = (\tau_1, \tau_2)$ transit from 0 to 0^+. We denote this transition as $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$ in the rest of the text. During this transition, the root continuity argument may collapse and infinitely many unbounded characteristic roots may appear in $\mathbf{C}^+$. Such a root discontinuity, however, never takes place anywhere else in the $\boldsymbol{\tau} \in \mathbf{R}^{2+}$ domain. Clearly, if it occurs, this behavior precludes stability for any $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$ even if the initial stability posture for $\boldsymbol{\tau} = \mathbf{0}$ is stable. This "small-delay phenomenon," or "strong stability condition," has been extensively studied in the literature [2], [14], [6], [7], [8], [3], [1], [15], and conditions guaranteeing root continuity in the transition of $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$ have been investigated with respect to the system properties. It is also known that one should first guarantee the root continuity for $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$ before performing the stability analysis on neutral DDEs. It is shown in [2] that to prevent

the small-delay phenomenon from occurring, the difference equation (also known as discrete kernel operator [14], [15]) consisting of the left-hand side of (1.1),

$$L(\tau_1, \tau_2) = 1 - ae^{-\tau_1 s} - be^{-\tau_2 s} = 0, \tag{1.4}$$

has to be *stable* for $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$. It is further proven [2] that this requires the parametric condition

$$|a| + |b| < 1. \tag{1.5}$$

This inequality constitutes the *necessary and sufficient condition* for neutral dynamics (1.1) to possess root continuity during the transition of $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$, but it is only a *necessary condition* for the system to exhibit stability at a point or in a region within $\boldsymbol{\tau} \in \mathbf{R}^{2+}$, $\boldsymbol{\tau} \neq \mathbf{0}$. This is a very critical nuance, which will be revisited throughout the text.

Dynamics that possess stable discrete kernel operator are called τ-*stabilizable* [14], [15], and we use this terminology in this text as well. Note that τ-stabilizability does not guarantee that for some values of $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$ this system will be stable. But conversely, if the system is not τ-stabilizable (i.e., it is τ-nonstabilizable), no $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$ can impart stability even if nondelayed dynamics, $\boldsymbol{\tau} = \mathbf{0}$, are stable. In other words, τ-stabilizability is a *necessary* but not *sufficient* condition for a stabilizing delay set $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$ to exist, as the name τ-stabilizability implies.

As the main motivation of this study, we state that there is no prior literature which declares the aimed "stability map" of this system in $\{\boldsymbol{\tau}\} \in \mathbf{R}^{2+}$ exhaustively and exactly (i.e., nonconservatively). Exhaustiveness means *all* the stable regions in any given segment of interest in (τ_1, τ_2) space attached to or detached from the origin $(0, 0)$. There is considerable literature treating this very problem, primarily using the Lyapunov stability perspective, which results in very conservative stability bounds in the domain of the time delays; see, for example, [10] and the references therein. Furthering this conservatism, these methods almost exclusively search for a stability region including the origin, $\{\boldsymbol{\tau}\} = \mathbf{0}$ (also called the *delay margin problem*). We distinguished the objective in this paper from the mentioned investigations, as we aim to determine all the stability regions exactly and exhaustively. To resolve this problem, we resort to a very recent conceptual development of the authors: the cluster treatment of characteristic roots (CTCR) [19], [18], [15], [20]. This procedure was originally introduced for retarded systems. In this work we implement the CTCR technique on multiple-delay neutral systems for the first time and obtain the corresponding stability picture.

The CTCR reveals the complete stability robustness against delay uncertainties. We also show that the t-stabilizability condition (i.e., the discrete kernel operator (1.4) to be stable or the parametric inequality (1.5) to hold) is obtained as a by-product of the CTCR. Interestingly, the CTCR approach also offers a very practical way of proving this necessary condition.

One can encounter a small number of investigations in the literature on DDEs with multiple delays. The existing reports consider simpler dynamics than (1.1) where a, b, f are all taken to be zero. These selections reduce the dynamics to a "retarded" class of time-delayed systems (TDS). For example, [5] addresses this question to determine the only stability region attached to the origin (a feature which arises from the simplified dynamics). [12] proposes some interesting lemmas for the location of the characteristic roots when additional conditions are imposed: $f = 0$, $\tau_1 = 1$, $1 < \tau_2 < 2$. The study indicates the overwhelming difficulties which arise for cases

with $\tau_2 > 2$. [11], [21] treat the stability question on second order systems, which are resonant when delays are zero, a different problem from the one treated here. All of these simplifications impose some strong restrictions for the respective methods to be expanded for the problem here. In [19], [15], [20] the treatment entails a novel concept (CTCR) for a very general class of retarded TDS cases overcoming such restrictions. The present text is motivated by the deployment of this new development to a new problem with the inclusion of $a\dot{\mathbf{x}}(t-\tau_1)$ and $b\dot{\mathbf{x}}(t-\tau_2)$ terms (which introduce the neutral behavior) and $f\dot{\mathbf{x}}(t-\tau_1-\tau_2)$ (i.e., delay cross-talk term). These additions complicate the exercise considerably. Furthermore, we will show that this deployment will reveal the τ-stabilizability condition (1.5) as an imbedded by-product.

In the following section we present a review of the enabling paradigm, CTCR. Section 3 discusses the small-delay phenomenon and re-proves the τ-stabilizability condition. In section 4 we display the stability analysis results for some case studies with different selections of a, b, c, d, f, g. Section 5 contains some observations and conclusions based on the new paradigm, CTCR.

2. Review of CTCR. The highlights of the CTCR framework, borrowed from [19], [20], are presented next. We restate that these previous investigations are for *retarded*-type multiple-delay systems. Neutral systems, however, with single delay only are addressed in [18], [14], [15]. We have shown earlier that for single-delay treatments [18], [13], [14], [15] CTCR can be deployed independently from the formation of the dynamics being of retarded or neutral type. The natural pursuit of research brings us to the topic of this paper: neutral-type dynamics with multiple time delays.

We first present some important features and definitions of the general class of systems represented by (1.1):

1. There is a countably infinite number of characteristic roots of (1.2) for a given delay set (τ_1, τ_2). It is impracticable to evaluate their distribution (i.e., their topology for all variations of the delay set $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$) in the quest of assessing the asymptotic stability. It is also known that for the dynamics to switch the stability posture, the characteristic roots should cross the imaginary axis, say at $s = \omega i$. Thus, for a successful stability analysis, one needs to detect exhaustively all the possible imaginary axis crossings. Let us denote the complete set of such crossing frequencies as $\mathbf{\Omega}$, and the corresponding root set as $\mathbf{S}_\Omega$:

$$\mathbf{\Omega} = \{\omega \,|\, CE(s = \omega i, \boldsymbol{\tau}) = 0,\ \boldsymbol{\tau} \in \mathbf{R}^{2+},\ \omega \in \mathbf{R}\}, \tag{2.1a}$$

$$\mathbf{S}_\Omega = \{s = \omega i \,|\, \omega \in \Omega\}. \tag{2.1b}$$

We also assume that this set, $\mathbf{\Omega}$, along with the generating delays, $\boldsymbol{\tau}$, is known completely at this stage.

2. Let us denote the correspondence between the delays (τ_1, τ_2) and ω with the notation $\langle \boldsymbol{\tau}, \omega \rangle$. Notice that, typically, there exist infinitely many characteristic roots ($s = \omega i$, $\omega \in \mathbf{\Omega}$, as $\mathbf{\Omega}$ represents a set of continuum) and the respective $\langle \boldsymbol{\tau}, \omega \rangle$ correspondence. Hence, one needs a treatment of these infinitely many roots simply by grouping them into sets, which show similar characteristics. We call this grouping the *root clustering* operation. To achieve this, however, those identifier characteristics, which we name the *clustering features*, need to be determined.

3. It is known that an imaginary root $s = \omega i$ with $\langle \boldsymbol{\tau}, \omega \rangle$, $\boldsymbol{\tau} = (\tau_1, \tau_2)$, correspondence will be repeated infinitely many times for the mesh points with equidistant grid size:

$$(\tau_{1j}, \tau_{2k}) = \left(\tau_1 + \frac{2\pi}{\omega} j,\ \tau_2 + \frac{2\pi}{\omega} k\right), \qquad j = 0, 1, 2, \ldots,\ k = 0, 1, 2, \ldots. \tag{2.2}$$

Considering the existence of root continuity, we state that *small* perturbations on τ yield *small* perturbations on ω, i.e.,

$$\langle \boldsymbol{\tau} + \boldsymbol{\epsilon}, \omega + \epsilon_c \rangle, \qquad 0 < |\boldsymbol{\epsilon}| \ll 1,\ 0 < |\epsilon_c| \ll 1. \tag{2.3}$$

Clearly, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)$ and ϵ_c are interdependent through the characteristic equation (1.2). Further elaboration on this point is kept outside the document to better streamline the discussions. One can detect infinitely many curves in the $\boldsymbol{\tau}$ domain, which traverse through these infinitely many mesh points (2.2) earmarked by ω. According to the D-subdivision method [4], these trajectories continuously partition the t domain into encapsulated regions in which the number of unstable roots, NU, remains fixed.

4. The *first clustering feature* follows the above comment: It is proven in [19], [15], [20] for much more general forms of the dynamics than (1.1) that there is a manageably small number of curves in (τ_1, τ_2) space which we call the "kernel curves," $\wp_0(\tau_1, \tau_2)$, where *all* the characteristic root sets can be found, containing at least one pair of imaginary roots. These imaginary roots, when determined for the entire domain of delays $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$, constitute the set $\mathbf{S}_\Omega$ as identified earlier (2.1). In summary, the kernel curves correspond to the complete set, $\mathbf{S}_\Omega$, in the sense of $\langle \boldsymbol{\tau}, \omega \rangle$.

The points on the kernel curves satisfy an important condition:

$$0 < \tau_k < \frac{2\pi}{\omega}, \qquad k = 1, 2. \tag{2.4}$$

This condition implies that the points on the kernel curves exhibit the smallest positive delay value for each delay complying with $\langle \boldsymbol{\tau}, \omega \rangle$. Assume that all such points on the kernel curves, call them generically "kernel points," $\boldsymbol{\tau}_{ker}$, are already known for *all* possible $\omega \in \boldsymbol{\Omega}$. By definition, the kernel curve formation is unique for a given system (1.1). The following notation encapsulates the complete representation of the kernel curve:

$$\wp_0 = \left\{ \boldsymbol{\tau} | \langle \boldsymbol{\tau}, \omega \rangle, \boldsymbol{\tau} \in \mathbf{R}^{2+}, \omega \in \boldsymbol{\Omega}, 0 \le \tau_k \le \frac{2\pi}{\omega}, k = 1, 2 \right\}. \tag{2.5}$$

For an exclusive stability analysis, this kernel curve set must be determined *completely* and *exhaustively*. The crucial feature of the kernel curves is that they represent *all* the possible imaginary roots of (1.2) for the entire $\boldsymbol{\tau} \in \mathbf{R}^{2+}$ domain. If there is a $\langle \boldsymbol{\tau}, \omega \rangle$ occurrence, either $\boldsymbol{\tau}$ is on the kernel curves or it can be reduced to a point on the kernel curves using j and k counters defined by (2.2) and (2.4) (one may even call it the "projection on the kernel"). That is, any and every delay set, $\boldsymbol{\tau}$, which results in a root $s \in \mathbf{S}_\Omega$ has to be represented in this "kernel set." The *first clustering feature*, therefore, appears very naturally; it is $\boldsymbol{\Omega}$, which represents the complete loci of $\boldsymbol{\tau}_{ker} \in \mathbf{R}^{2+}$ for which there exists at least a pair of imaginary characteristic roots of (1.2).

5. Kernel curves, $\wp_0(\tau_1, \tau_2)$, are, in fact, the loci of $\boldsymbol{\tau}$ described in (2.2) with the points represented by $j = k = 0$. Notice that there are ∞^2 (two-dimensional infinite) candidate points defined by (2.2) in the $\boldsymbol{\tau} \in \mathbf{R}^{2+}$ domain resulting in the same imaginary root, $\omega i \in \mathbf{S}_\Omega$. Those curves corresponding to $j > 0$ and $k > 0$ are called the "offspring curves," or "offspring" for short, and denoted by $\wp_{jk}(\tau_1, \tau_2)$,

where j and k identify the jth and kth generation offspring in τ_1 and τ_2, respectively. Consequently, the complete set containing the kernel curves and the offspring becomes $\wp(\tau_1, \tau_2)$:

$$\wp(\tau_1, \tau_2) = \wp_0(\tau_1, \tau_2) \cup \bigcup_{j=1}^{\infty} \bigcup_{k=1}^{\infty} \wp_{jk}(\tau_1, \tau_2). \tag{2.6}$$

6. It is critical to note that all the infinitely many trajectories of $\wp_{jk}(\tau_1, \tau_2)$ are created from the kernel curves, $\wp_0(\tau_1, \tau_2)$, via a nonlinear shifting term $2\pi/\omega$, as per (2.2). In addition to this, any kernel point on the trajectories of $\wp_0(\tau_1, \tau_2)$ defined by $j = k = 0$ imposes its ω signature identically onto its offspring ($j > 0$ and $k > 0$). Thus, $\mathbf{\Omega}$ remains *invariant* from kernel curves to offspring curves.

7. The kernel curves and the offspring constitute the complete (and exhaustive) distribution of (τ_1, τ_2) points where the characteristic equation $CE(s, \tau_1, \tau_2)$ has root sets containing at least one pair of imaginary roots. And outside the set $\wp(\tau_1, \tau_2)$ there is not a single point, which renders imaginary characteristic roots. These are the only locations in the (τ_1, τ_2) domain where the system (1.1) could transit from stable to unstable posture (or vice versa). These contours $\wp(\tau_1, \tau_2)$ must be determined exhaustively. Since $\wp(\tau_1, \tau_2)$ is completely generated from the kernel curves $\wp_0(\tau_1, \tau_2)$, it is sufficient for our purposes to determine these kernel curves exhaustively. We present a procedure for achieving this later in this section.

8. *Root tendency invariance property.* The root tendency along the τ_j, $j = 1, 2$, axis at the crossing of $s = \omega i$ is defined by $RT^{\tau_j}_{s=\omega i} = sgn(\Re(\frac{\partial s}{\partial \tau_j}|_{s=\omega i}))$. The root tendencies at $s \in \mathbf{S}_\Omega$ along τ_1 (or τ_2) across the corresponding points on a kernel curve and its offspring remain unchanged so long as τ_2 (or τ_1) is kept fixed. This is a proven feature for much more general constructs than (1.1) known as the "root tendency invariance" property; see [19, Proposition II], [15], [20]. This very strong property constitutes the *second clustering feature* along the contours defined by (10). Through this feature one can mark stabilizing (or destabilizing) transitions along the curves $\wp(\tau_1, \tau_2)$.

2.1. Steps of the CTCR procedure. Based on the above properties one can establish the complete stability robustness picture of the system by performing the following steps of the CTCR procedure:

1. Determine exhaustively the kernel and offspring curves, $\wp(\tau_1, \tau_2)$.

2. Start from a nondelayed system, $\boldsymbol{\tau} = \mathbf{0}$, evaluating $NU(0)$ via a trivial application of the Routh–Hurwitz method.

3. Following line segments in $\boldsymbol{\tau} \in \mathbf{R}^{2+}$, which are parallel to the individual coordinates τ_1 and τ_2, connect the origin ($\boldsymbol{\tau} = \mathbf{0}$) to a point of interest (τ_{1t}, τ_{2t}), where subscript t is used for target.

4. As this path crosses the kernel and offspring curves, increase NU by $+2$ (or -2) for the $RT = +1$ (-1), according to the D-subdivision method of El'sgol'ts [4].

5. Exhaustively identify the regions in $\boldsymbol{\tau} \in \mathbf{R}^{2+}$ where (τ_{1t}, τ_{2t}) results in $NU = 0$ as "stable" and the others ($NU > 0$) "unstable."

This completes the CTCR procedure and the stability robustness picture of this system.

Remember that the delays $\{\boldsymbol{\tau}\}$'s corresponding to the $s \in \mathbf{S}_\Omega$ set were assumed to be known in the first step of the above procedure. However, a methodology is still needed to create the topology in (2.1) and the corresponding kernel curves in (2.5) exhaustively. We discuss a *holographic mapping* procedure next to achieve this. For a

better understanding we will guide the reader through the details of the above steps in section 4 for the special dynamics at hand.

2.2. Exhaustive and complete determination of the kernel curves. This step is very critical to CTCR for obvious reasons. The complete contour of (τ_1, τ_2) which results in at least one imaginary root $s = \omega i$, $\omega \in \mathbf{R}$, must be determined. For this we use the Rekasius substitution, which was first introduced in [17] and also utilized in [19], [13], [15], [20]. It suggests the following representation for exponential terms:

$$e^{-\tau_j s} = \frac{1 - T_j s}{1 + T_j s}, \qquad T_j \in \mathbf{R},\ j = 1, 2. \tag{2.7}$$

This substitution becomes exact for $s = \omega i$, provided that the relation between T_i and τ_i,

$$\tau_{jk} = \frac{2}{\omega}[\tan^{-1}(\omega T_j) + k\pi], \qquad k = 0, 1, 2, \ldots, \tag{2.8}$$

holds. It represents an asymmetric mapping. A given ω and a T_j value correspond to infinitely many equidistant delays, τ_{jk}, $j = 1, 2$, $k = 0, 1, 2, \ldots$, where $\tau_{j,k+1} - \tau_{jk} = \frac{2\pi}{\omega}$ is the fixed grid size in the τ_j distribution. Notice that this grid size is identical for both τ_1 and τ_2 sets. In the other direction, however, the mapping is one-to-one, that is, each τ_j value maps into a single T_j value.

This asymmetric mapping between $\tau_j \leftrightarrow T_j$ transforms the transcendental CE of (1.2) into a new characteristic equation $CE_T(s, T_1, T_2)$, which is of fractional polynomial type. Multiplying this equation by $(1 + T_1 s)(1 + T_2 s)$, one obtains

$$\overline{CE}(s, T_1, T_2) = \sum_{j=0}^{3} b_j(T_1, T_2) s^j = 0. \tag{2.9}$$

Notice that $b_j(T_1, T_2)$ are some multinomials in two parameters T_1, T_2 except b_0, which is a constant. An interesting relation between the infinite-dimensional equation (1.2) and the third degree equation (2.9) is that they share the *same imaginary root sets completely*. The claim is that

$$\begin{aligned} \boldsymbol{\Omega} &= \boldsymbol{\Omega}[s | CE(s, \tau_1, \tau_2) = 0, (\tau_1, \tau_2) \in \mathbf{R}^{2+}] \cap \mathbf{C}^0 \\ &\equiv \overline{\boldsymbol{\Omega}}[s | \overline{CE}(s, T_1, T_2) = 0, (T_1, T_2) \in \mathbf{R}^2] \cap \mathbf{C}^0, \end{aligned} \tag{2.10}$$

where $\overline{\boldsymbol{\Omega}}$ represents the complete topology of root sets of $\overline{CE}$ in the $(T_1, T_2) \in \mathbf{R}^2$. The left-hand side of identity (2.10) consists of all the imaginary characteristic roots of CE for some $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$, and the right-hand side is the same imaginary root set for $\overline{CE}$ at some $(T_1, T_2) \in \mathbf{R}^2$. Every (T_1, T_2) pair, which results in an imaginary root $s = \omega i$ of $\overline{CE}(s, T_1, T_2)$, represents infinitely many (τ_{1j}, τ_{2k}) pairs as per (2.8). In other words, a unique $\boldsymbol{\tau}_{ker}$ and infinitely many offspring arise. For further mathematical details on this property, the reader is referred to [19], [15], [20].

The most beneficial point in transforming $CE(s, \tau_1, \tau_2)$ to $\overline{CE}(s, T_1, T_2)$ is obvious: The parametric equation (2.9) is much easier to study than (1.2). This is a tremendous reduction in the complexity of determining the imaginary characteristic roots of (1.2), as we will observe later. $\overline{\boldsymbol{\Omega}}$ can be determined completely for the entire space of $(T_1, T_2) \in \mathbf{R}^2$ as $(s = \omega i,\ \omega \in \overline{\boldsymbol{\Omega}})$. As per the claim in (2.10) this set of imaginary roots is identical to $\mathbf{S}_\Omega$. Based on these observations the following structured

TABLE 1
The Routh array for $\overline{CE}(s, T_1, T_2)$, arguments suppressed.

s^3	b_3	b_1
s^2	b_2	b_0
s^1	$\frac{b_2 b_1 - b_3 b_0}{b_2} = R_1$	
s^0	b_0	

steps are most natural. One should first find the projection of the kernel curves in (T_1, T_2) space. We will name it the *core curve* just to discriminate its domain (T_1, T_2) from the kernel curves in (τ_1, τ_2). The correspondence between the two sets of curves is discussed in the following segments. The kernel and offspring curves constitute the complete set of contours in (τ_1, τ_2) where the possible stability transition can occur according to the D-subdivision method of [4].

Let us now show how to determine the core curve of $\overline{CE}(s, T_1, T_2)$, i.e., those (T_1, T_2) that give rise to $s = \omega i$ of $\overline{CE}(s, T_1, T_2)$. The easiest procedure for finding all the imaginary roots of such characteristic polynomials is the classical Routh–Hurwitz method [9]. From the well-known rules of the Routh array of Table 1, the imaginary roots of the cubic equation (2.9) are encountered at any point (T_1, T_2) where

$$R_1(T_1, T_2) = 0 \tag{2.11}$$

is satisfied with an additional condition,

$$b_2(T_1, T_2) b_0 > 0. \tag{2.12}$$

At these points one can determine a crossing frequency $\omega = \sqrt{b_0 / b_2(T_1, T_2)} \in \mathbf{R}$.

The implicit expressions (2.11) and (2.12) represent the core curve completely. The full expansion of the Routh array for the system (1.1) is given in the appendix. Notice the formation of $R_1(T_1, T_2)$, which is a multinomial in T_1 and T_2 of degrees 2 and 2, respectively. Therefore, for a given value of T_2 (T_1), there can be at most 2 (2) real T_1 (T_2). That is, the core curve has at most two separate segments. Since there is one-to-one mapping between the core curves in (T_1, T_2) and the kernel curves in (τ_1, τ_2) as proven in [19], [15], [20], this system can possess at most two separate contours for the kernel curves in (τ_1, τ_2), although these two contours may be split into separate segments according to the positivity condition in (2.4). The next step in CTCR is to numerically map these core curves into kernel curves via (2.8) and further to offspring in (τ_1, τ_2) space via (2.2). This completes the determination process of the entire set of kernel and offspring curves, $\wp(\tau_1, \tau_2)$.

With $\wp(\tau_1, \tau_2)$ at hand the second clustering feature can now be utilized. Namely, some segments of the kernel curves are identified as stabilizing transitions as opposed to destabilizing transitions. Ultimately, the D-subdivision methodology is implemented to declare a unique stability map of the system in (τ_1, τ_2) space. This map, in fact, constitutes an exact (nonconservative) declaration of the stability robustness picture against uncertain delays (τ_1, τ_2). And to the best of our knowledge, this is unique.

3. An interesting result on τ-stabilizability. In this section, we further discuss τ-stabilizability and show that it is, in fact, a natural by-product of the CTCR procedure. Several rigorous and cumbersome proofs of the τ-stabilizability condition (1.5) can be found in the literature [6], [7], [8]. As was stated earlier, if this necessary condition is not satisfied, even an infinitesimally small time delay $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$

destabilizes the system, although the nondelayed ($\boldsymbol{\tau} = \mathbf{0}$) system may start asymptotically stable. Infinitely many unbounded unstable roots appear, the root continuity argument collapses in the vicinity, and consequently the D-subdivision method [4], which is really the basis of CTCR, does not apply anymore. The CTCR procedure is, however, still valid except for its final step involving D-subdivision deployment. When applied for small delays, the CTCR procedure indeed gives rise to the τ-stabilizability property. This is the highlight of the discussions in this section.

LEMMA 3.1. *The necessary condition for the system in* (1.1) *to be stable for any* $\boldsymbol{\tau} \in \mathbf{R}^{2+}$ *is the* τ*-stabilizability condition given in* (1.5).

Proof. In order to examine the stability transition of (1.1) as τ varies $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$, we first check the stability posture of the nondelayed dynamics, which is written as

$$CE(s, \boldsymbol{\tau} = \mathbf{0}) = s(1 - a - b) - (c + d + f + g) = 0. \tag{3.1}$$

For this system to represent any dynamics, $1 - a - b$ must be nonzero. Furthermore, if $(1 - a - b)(c + d + f + g)$ is negative (positive), the nondelayed system has one stable (unstable) pole. We next investigate the $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$ transition. During this transition, we wish to observe no change in the stability posture of the system. The contrary occurrence would mean $\omega \to \infty$ (i.e., the Riemann sphere crossing; see [15]), which invites infinitely many unbounded unstable roots. This instability posture is impossible to recover from within the $\tau_1 > 0$, $\tau_2 > 0$ domain. The reason is simply that even for some $RT = -1$ points in the quadrant, $\tau_1 > 0$, $\tau_2 > 0$, there will only be a pair (or finite number of pairs in some degenerate cases) of roots returning to $\mathbf{C}^-$, since the root continuity is now in effect. We need infinitely many such crossings to take place in order to recover from the destabilizing effect caused by the small delay. Thus, $\tau_j : 0 \to \epsilon_j$, $0 \ll \epsilon_j < 1$, has to exert no Riemann sphere crossings for any stable behavior to exist.

If there is such a crossing for small delays, we can use the Rekasius transformed equation (2.9) along with $\boldsymbol{\tau} \to \mathbf{T}$ correspondence given in (2.8). Notice that for small delays, (2.8) becomes

$$\tan\left(\frac{\tau_j \omega}{2}\right) = T_j \omega, \qquad j = 1, 2, \tag{3.2}$$

which implies that $\tau_j : 0 \to \epsilon_j$ corresponds to $T_j : 0 \to \epsilon_j/2$. In short, the τ_j and T_j variations are of the same order of magnitude. In the two-dimensional delay domain such a transition can be represented selecting a directional slope m, which is defined as

$$m = \tau_2/\tau_1 = T_2/T_1, \qquad m \in [0, \infty). \tag{3.3}$$

The "no crossing" requirement should be independent of m. Using the relation (3.3), equation (2.9) becomes

$$\overline{CE}(s, \mathbf{T}) = \overline{CE}(s, T_1, mT1) = 0. \tag{3.4}$$

Taking the limit for $0 < T_1 \ll 1$ on (3.4), we arrive at the following by dropping higher powers of T_1 and favoring the lowest power only in the coefficients of s:

$$\begin{aligned} &\lim_{0<T_1\ll 1} \overline{CE}(s, T_1, mT_1) = (1 + a + b) m T_1^2 s^3 \\ &+ (1 + a - b + m\{1 - a + b\}) T_1 s^2 + (1 - a - b) s + (-c - d - f - g) = 0. \end{aligned} \tag{3.5}$$

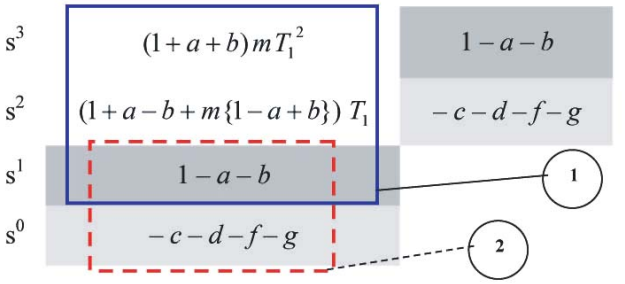


FIG. 1. *The Routh array on* (3.5) *with termwise limit* $T_1 \to \epsilon_1/2$.

We wish to find the conditions on the system parameters a, b, c, d, f, g so that (3.5) does not exhibit an imaginary root. If one forms the respective Routh array and incorporates the small T_1 approximation (i.e., favoring the lowest degree terms) for each term during the formation of this array, Figure 1 is obtained. The first column should exhibit no sign change. It consists of two blocks:

1. The coefficients of the quadratic expression for the discrete kernel operator (1.4) after Rekasius substitution. That is,

$$\begin{aligned}\overline{L}(s, T_1, T_2) &= (1 + T_1 s)(1 + T_2 s) - a(1 - T_1 s) - b(1 - T_2 s) \\ &= (1 + a + b)mT_1^2 s^2 + (1 + a - b + m\{1 - a + b\})T_1 s + (1 - a - b) = 0.\end{aligned} \tag{3.6}$$

2. The block, which is found by the coefficients of s in the nondelayed characteristic equation (3.1). If these two coefficients agree (disagree) in sign, the nondelayed dynamics are stable (unstable), respectively. In turn, it implies that the nondelayed system has no (one) root in $\mathbf{C}^+$.

If we wish to observe no change of stability posture from $T_1 : 0 \to \epsilon_1/2$ transition, independent of m, the terms in block 1 should agree in sign. This proposition yields the conditions

$$c_1 = 1 + a + b > 0, \tag{3.7a}$$

$$c_2 = 1 + a - b + (1 - a + b)m > 0, \tag{3.7b}$$

$$c_1 = 1 - a - b > 0. \tag{3.7c}$$

It is trivial to show that these conditions reduce to a simple form $|a| + |b| < 1$, which is the τ-stabilizability condition (1.5) as shown in the literature [6], [7], [8]. The admissible region in the space of (a, b) is given in Figure 2. If (a, b) satisfy (3.7), it simply guarantees that there will not be a crossing while $\boldsymbol{\tau}_{(1,2)} : 0 \to 0^+$. This property also implies that the system may be stable for some values of $(\tau_1, \tau_2) \in \mathbf{R}^{2+}$. □

Summary observations:

1. Using CTCR, we arrive at the same necessary condition for τ-stabilizability (1.5) as declared in [6], [7], [8] via a simple mathematical manipulation (Figure 1). This shows that τ-stabilizability is a natural by-product of the CTCR procedure.

2. As can be seen from Figure 2, τ-stabilizability is independent of τ_1 and τ_2 and is dictated only by the parametric selection of (a, b) pairs.

4. Example case studies. We present three case studies here for some combinations of parametric possibilities of (1.2). The first case is handled following a step-by-step pursuit of CTCR to assist the reader. The others are offered in their final outcome with some observations.

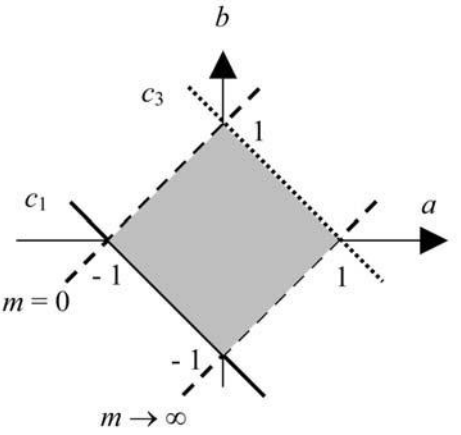


FIG. 2. *Depiction of conditions in* (3.7) *and* τ*-stabilizability region.*

4.1. Case I. Let us take $a = -0.8$, $b = -0.15$, $c = -1.5$, $d = 2$, $f = -5$, $g = 1$ in (1.2). It is clear that these dynamics are τ-stabilizable since the condition in (1.5) is satisfied, i.e., $|a| + |b| < 1$; therefore we proceed to the stability analysis. The characteristic equation is

$$s(1 + 0.8\,e^{-\tau_1 s} + 0.15\,e^{-\tau_2 s}) + 1.5\,e^{-\tau_1 s} - 2\,e^{-\tau_2 s} + 5\,e^{-(\tau_1+\tau_2)s} - 1 = 0, \tag{4.1}$$

which is stable with a pole at $s = -1.7949$ for $\{\boldsymbol{\tau}\} = \mathbf{0}$. After Rekasius substitution, we obtain the corresponding equation to (2.9) as

$$T_1T_2s^3 + (90T_1T_2 + 33T_2 + 7T_1)s^2 + (-50T_2 - 190T_1 + 39)s + 70 = 0. \tag{4.2}$$

The Routh array conditions for the complete core curve and the corresponding imaginary roots, following (2.11), (2.12) and the descriptions in the appendix, are

$$\frac{(-171T_2 - 13.3)T_1^2 + (-31.8T_2 - 45T_2^2 + 2.73)T_1 + 12.87T_2 - 16.5T_2^2}{(90T_2 + 7)T_1 + 33T_2} = 0, \tag{4.3}$$

$$b_2(T_1, T_2)b_0 = (90T_2 + 7)T_1 + 33T_2 > 0. \tag{4.4}$$

Figure 3 displays the viable curves satisfying (4.3) and (4.4) in (T_1, T_2) space. The core curve is those sections of the thick black curve which lie in the background region (shaded). Its determination, however, is exhaustive. Notice that the R_1 expression is quadratic in both T_1 and T_2; see also the appendix, (A.1). Therefore, for one value of T_1 there are at most two real T_2's, and vice versa. That is, at most two separate contours may constitute the core curve. The corresponding kernel curve in (τ_1, τ_2) space is generated numerically using (2.8), as seen in Figure 4 (thick line, shown in red in the electronic version). This figure also contains the complete range of crossing frequencies, $\boldsymbol{\Omega}$. Accordingly the system under consideration can only possess imaginary roots within the range of $1.6 \leq \omega \leq 51.12$ *rad/s* which defines the set $\boldsymbol{\Omega}$. The kernel curves contain all (τ_1, τ_2) points, which create the imaginary root set of the system entirely. That is, these dynamics can have only those imaginary characteristic roots generated by this kernel curve, and *no point in* $\boldsymbol{\tau} \in \mathbf{R}^{2+}$ *space* can render an imaginary root at ω not belonging to $\boldsymbol{\Omega}$. Corresponding offspring curves to the kernel

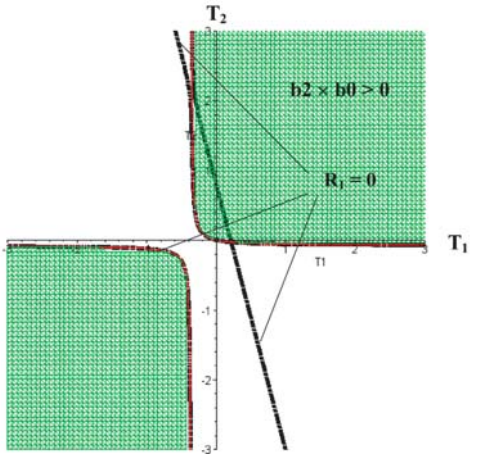


FIG. 3. $\{\mathbf{T}\}$ *domain representation, equality condition* (4.3)*, and the inequality condition in* (4.4).

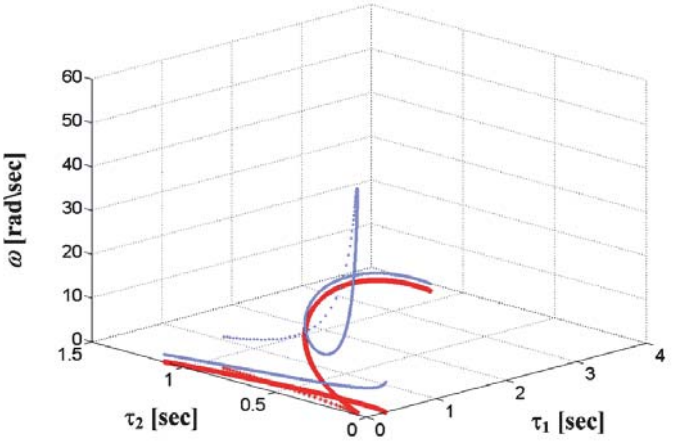


FIG. 4. *The complete kernel formation (thick line, shown in red in the electronic version) of Case* I *and the only possible* ω*'s (thin line, shown in blue in the electronic version).*

are obtained using (2.2) and are displayed in Figure 5 (thin line, shown in blue in the electronic version). They also generate exactly the same set of imaginary roots as the kernel curves, nothing more and nothing less. This concludes the formation of the first clustering feature. All the points (τ_1, τ_2), which render imaginary characteristic roots, are in Figure 5 as kernel and offspring curves.

We now look at the kernel curves and determine the root tendencies. For instance, at points A_1, A_2, and A_3, $RT^{\tau_2} = -1$ (i.e., stabilizing) and it is invariant from kernel curves to all the offspring curves. Using the D-subdivision method of [4] one can determine the number of unstable roots NU in each region in Figure 5. They are sparingly shown in the figure. Obviously wherever $NU = 0$, the stable region appears, which is shaded in the figure. In Figure 5, we also recognize that the dynamics are τ_2-independent stable for $\tau_1 < 0.0595$ s. Just to give an idea of the efficiency of the CTCR procedure, we state the CPU time of creating Figure 5 (on a 3.2 MHz Pentium 4 with 512 MB RAM) to be 25 seconds.

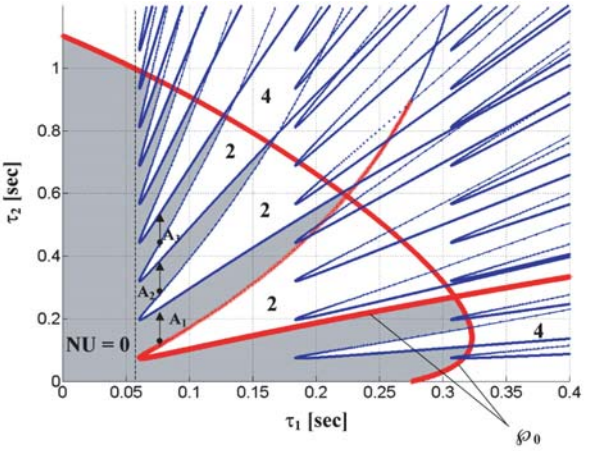


FIG. 5. *Complete stability picture of Case* I.

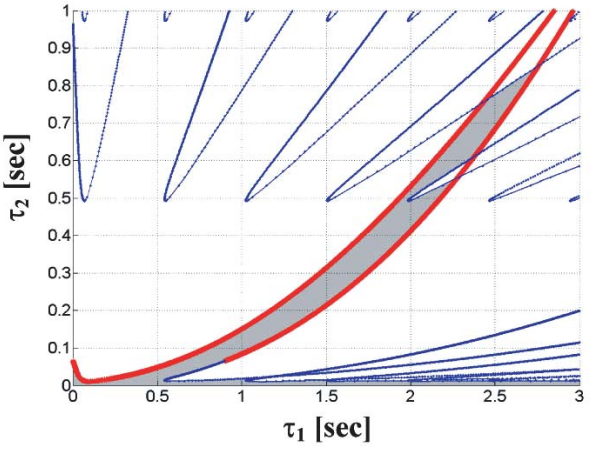


FIG. 6. *Complete stability picture of Case* II.

4.2. Case II. The system with parameters $a = 0.05$, $b = 0.9$, $c = 0.25$, $d = -1$, $f = -1.25$, $g = -1$ is again τ-stabilizable, i.e., $|a| + |b| < 1$. In Figure 6, we suppress the intermediary steps and give the stability picture only. The stable region is shaded ($NU = 0$). We observe that for $\tau_2 < 0.0109$ s the system becomes τ_1-independent stable. Also, we find (following the same logic as in Case I) that the only resonance frequencies of these dynamics are within the range $0.58 \leq \omega \leq 12.97$ rad/s, which is the set $\mathbf{\Omega}$.

4.3. Case III. Take $a = -0.35$, $b = 0.4$, $c = -12$, $d = -18$, $f = 4$, $g = -6$ as the set of parameters. The system is again τ-stabilizable. Without the intermediary steps, we show in Figure 7 the stability picture only. The stable region is shaded ($NU = 0$). We observe that for $\tau_2 < 0.0139$ s the system becomes τ_1-independent stable, and for $\tau_1 < 0.0525$ s the system is τ_2-independent stable. For this case, the only possible resonance frequencies are within the set $\mathbf{\Omega} : \{1.62 \leq \omega \leq 48.06$ $rad/s\}$.

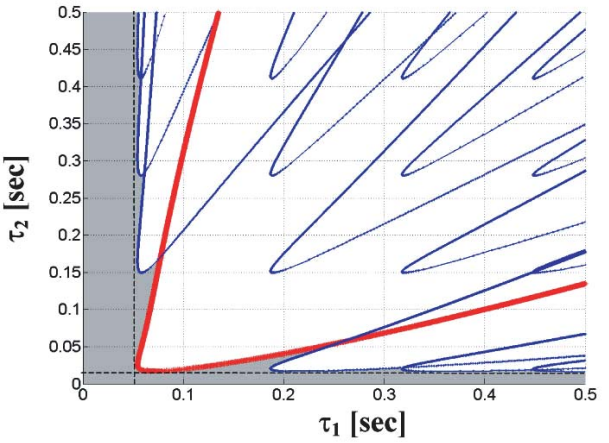


FIG. 7. *Complete stability picture of Case* III.

5. Conclusion. A general class of first order LTI scalar, two-time-delay systems of *neutral type* is studied for their stability posture. The unique features in this work are twofold. First, we have the *additional terms* that convert a cornerstone study on retarded-type dynamics by Hale and Huang [5] into a neutral type. Second, the dynamics are exacerbated by a *cross-talk term in the delays*. Both these added dynamics make the stability analysis considerably more complex. A new framework, the cluster treatment of characteristic roots (CTCR), is deployed to overcome these added complexities. CTCR offers a systematic and numerically very efficient procedure for the creation of complete stability maps of such systems. Moreover, CTCR is also deployed to re-prove the well-known necessary condition of the small delay phenomenon of neutral dynamics. It turns out that the CTCR procedure automatically generates this condition as a by-product. The end results are demonstrated in numerical case studies.

Appendix. Take the most general form of (1.2). Making the substitution of (2.7), creating (2.9), and deploying the Routh array as in Table 1, one obtains the corresponding terms to (2.11) and (2.12) as

$$
\begin{aligned}
R_1(T_1,T_2) = [&(\{2gc - 2fd - c^2 + d^2 - g^2 + f^2\}T_2 + ad - af + ag + bf + g \\
&- bg + bc - bd - ac - c + d - f)T_1^2 + (\{-d^2 - g^2 + c^2 + 2gd - 2fc + f^2\}T_2^2 \\
&+ \{2bd - 2af + 2ac - 2f - 2bf - 2bg + 2g - 2c - 2bc - 2ad - 2ag - 2d\}T_2 \\
&- 1 - b^2 + 2b + a^2)T_1 + (-d + af - ac + g + bc - f + c - bf + ad \\
&+ bg - bd - ag)T_2^2 + (b^2 + 2a - a^2 - 1)T_2]/b_2(T_1,T_2),
\end{aligned}
\tag{A.1}
$$

$$
\begin{aligned}
b_2(T_1,T_2) &= (\{c + d - f - g\}T_2 + 1 + a - b)T_1 + (1 - a + b)T_2, \\
b_0 &= -(c + d + f + g).
\end{aligned}
\tag{A.2}
$$

REFERENCES

[1] R. BELLMAN AND K. COOKE, *Differential Difference Equations*, Academic Press, New York, 1963.

[2] L. Cesari, J. K. Hale, and J. P. LaSalle, *Dynamical Systems: An International Symposium*, Vol. 1, Academic Press, New York, 1976.
[3] J. Chen, *On computing the maximal delay intervals for stability of linear delay systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 1087–1092.
[4] L. E. El'sgol'ts and S. B. Norkin, *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*, Academic Press, New York, 1973.
[5] J. K. Hale and W. Huang, *Global geometry of the stable regions for two delay differential equations*, J. Math. Anal. Appl., 178 (1993), pp. 344–362.
[6] J. K. Hale and S. M. Verduyn Lunel, *An Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.
[7] J. K. Hale and S. M. Verduyn Lunel, *Effects of small delays on stability and control*, in Operator Theory and Analysis, Oper. Theory Adv. Appl. 122, Birkhäuser, Basel, 2001, pp. 275–301.
[8] J. K. Hale and S. M. Verduyn Lunel, *Strong stabilization of neutral functional differential equations*, IMA J. Math. Control Inform., 19 (2001), pp. 1–19.
[9] B. C. Kuo, *Automatic Control Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
[10] C.-H. Lien, K.-W. Yu, and J.-G. Hsieh, *Stability conditions for a class of neutral systems with multiple time delays*, J. Math. Anal. Appl., 245 (2000), pp. 20–27.
[11] S.-I. Niculescu, *On delay robustness analysis of a simple control algorithm in high-speed networks*, Automatica J. IFAC, 38 (2002), pp. 885–889.
[12] R. Nussbaum, *Differential delay equations with two time lags*, Mem. Amer. Math. Soc., 205 (1978).
[13] N. Olgac and R. Sipahi, *An exact method for the stability analysis of time delayed LTI systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 793–797.
[14] N. Olgac and R. Sipahi, *Direct method for analyzing the stability of neutral type LTI-time delayed systems*, Automatica J. IFAC, 40 (2004), pp. 847–853.
[15] N. Olgac and R. Sipahi, *The cluster treatment of characteristic roots and the neutral type time-delayed systems*, ASME J. Dynamic Systems, Measurement and Control, 127 (2005), pp. 88–97.
[16] N. Olgac, H. Elmali, M. Renzulli, and M. Hosek, *Active vibration control of distributed systems using delayed resonator with acceleration feedback*, ASME J. Dynamic Systems, Measurement and Control, 119 (1997), pp. 380–389.
[17] Z. V. Rekasius, *A stability test for systems with delays*, in Proceedings of the Joint Automatic Control Conference, 1980, Paper TP9-A.
[18] R. Sipahi and N. Olgac, *Degenerate cases in using the direct method*, ASME J. Dynamics Systems, Measurement and Control, Special Issue on Time Delayed Systems, 125 (2003), pp. 194–201.
[19] R. Sipahi and N. Olgac, *A novel stability study on multiple time-delay systems (MTDS) using the root clustering paradigm*, in Proceedings of the American Control Conference, Boston, MA, 2004.
[20] R. Sipahi and N. Olgac, *Complete stability robustness of third order LTI multiple time-delay systems*, Automatica J. IFAC, 41 (2005), pp. 1413–1422.
[21] G. Stepan, *Retarded Dynamical Systems: Stability and Characteristic Function*, Longman Scientific & Technical, John Wiley & Sons, New York, 1989.
[22] M. Valasek and N. Olgac, *New concept of active multiple frequency vibration suppression technique*, in Smart Structures, NATO Sci. Ser. 65, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 373–382.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 972–997

SECOND-ORDER OPTIMALITY CONDITIONS FOR SCALAR AND VECTOR OPTIMIZATION PROBLEMS IN BANACH SPACES*

HELMUT GFRERER†

Abstract. In this paper we present a very general and unified theory of second-order optimality conditions for general optimization problems subject to abstract constraints in Banach spaces. Our results apply both to the scalar case and the multicriteria case. Our approach rests essentially on the use of a signed distance function for characterizing metric regularity of a certain multifunction associated with the problem. We prove variational results which show that, in a certain sense, our results are the best possible that one can obtain by using second-order analysis. We demonstrate how recently devised optimality conditions can be derived from our general framework, how they can be extended under weakened assumptions from the scalar case to the multiobjective case, and even how some new results also can be obtained.

Key words. second-order optimality conditions, metric regularity, multipliers, abstract constraints, vector optimization

AMS subject classifications. 49K27, 90C29, 90C30

DOI. 10.1137/040612713

1. Introduction.

In this paper we study second-order optimality conditions, both necessary and sufficient, for the following optimization problem:

$$\text{(P)} \qquad \text{L-minimize } f(x) \quad \text{subject to} \quad g(x) \in K,$$

where $f : X \to U$ and $g : X \to V$ are mappings from the Banach space X into other Banach spaces U and V, $K \subset V$ is a closed convex set, and where $L \subset U$ is a closed convex cone with nonempty interior, $\operatorname{int} L \neq \emptyset$. We define two kinds of local L-minimizers as follows.

DEFINITION 1.1. *An element $\bar{x} \in X$ is called a* local weak minimizer *for* (P) *if $g(\bar{x}) \in K$ and if there exists a neighborhood N of $\bar{x}$ such that for each $x \in N$ with $g(x) \in K$, one has $f(x) - f(\bar{x}) \notin -\operatorname{int} L$. A local weak minimizer $\bar{x}$ is called a* strict local minimizer *for* (P) *if for each $x \in N \setminus \{\bar{x}\}$ with $g(x) \in K$, one has $f(x) - f(\bar{x}) \notin -L$.*

Note that (P) includes the very common problem of constrained scalar minimization, for which $U = \mathbb{R}$ and $L = \mathbb{R}_+$. Local weak minimizers for (P) then amount to usual local minimizers. The case of scalar minimization problems is the one we are most interested in, but by our approach it is also possible to treat vector optimization problems without considerable additional effort in the case of abstract constraints.

A number of well-known optimization problems can be formulated in the form (P) in a natural way. For instance, when $V = \hat{V} \times \mathbb{R}^m$, $K = \{0\} \times \mathbb{R}^m_-$, and $g = (G, g_1, \ldots, g_m)$, the feasible set of (P) is defined by an equality constraint $G(x) = 0$ and finitely many inequality constraints $g_i(x) \leq 0$, and the scalar version of (P) becomes a possibly infinite-dimensional *mathematical programming problem.*

Second-order optimality conditions have been studied in numerous publications; for scalar problems, see [4], [6], [9], [16], [18], [23], [30], [31], and for vector optimization

*Received by the editors August 2, 2004; accepted for publication (in revised form) March 1, 2006; published electronically July 31, 2006.
http://www.siam.org/journals/sicon/45-3/61271.html
†Institute of Computational Mathematics, Johannes Kepler University Linz, A-4040 Linz, Austria (gfrerer@numa.uni-linz.ac.at).

problems, see [5], [8], [20], [21], [22]. Our approach has been strongly inspired by the results of Levitin, Miljutin, and Osmolovskii [26] and Ioffe [16] for the mathematical programming problem. The aim of this paper is to generalize the results from [16], [26] to the problem (P). We present second-order conditions, sufficient (respectively, necessary) for local strict (respectively, weak) minimizers, which differ only by the replacement of an inequality with a strict inequality. Besides this desirable feature, we prove the following variational principle: If the second-order sufficient conditions are not fulfilled at a strict local minimizer $\bar{x}$, then we can find twice continuously differentiable perturbations $\delta f : X \to U$, $\delta g : X \to V$ having vanishing function values and first- and second-order derivatives at $\bar{x}$, such that $\bar{x}$ is not a local weak minimizer for (P), with f and g replaced by $f+\delta f$ and $g+\delta g$, respectively. A similar result also holds for the necessary conditions, provided the following suitable condition on the constraints is fulfilled: If the second-order necessary conditions hold at a point $\bar{x}$, then $\bar{x}$ is a strict local minimizer for some perturbed problem. Let us mention that the condition on the constraints we use is weaker than the very common constraint qualification of Robinson [32].

These variational principles indicate that among those second-order conditions which can be formulated solely in terms of the two convex sets L and K and function values and derivatives up to second order of f and g at a given point $\bar{x}$, our results are optimal. However, specific structures like linearity or convexity are present in many practical problems, and exploiting these structures could of course lead to better optimality conditions.

Given a point $\bar{x} \in X$, fixed throughout this paper, our results can be expressed in terms of a certain multifunction associated with (P) and $\bar{x}$. Let $h : X \to U \times V$ be defined by

$$h(x) := (f(x) - f(\bar{x}),\ g(x)), \qquad C := (-L) \times K, \qquad Y := U \times V. \tag{1.1}$$

Then the multifunction $\Gamma : X \rightrightarrows Y$, given by

$$\Gamma(x) := h(x) - C, \tag{1.2}$$

forms the basis of our investigations. In what follows we use the norm $\|(u,v)\| := \max\{\|u\|, \|v\|\}$ on the product space $Y = U \times V$.

Our notation is fairly standard. In a normed space Z, $B_Z := \{z \in Z : \|z\| \le 1\}$ denotes the closed unit ball and $S_Z := \{z \in Z : \|z\| = 1\}$ denotes the unit sphere. The topological dual space is denoted by Z^*. By $\langle z^*, z\rangle$ we denote the value $z^*(z)$ of the linear functional $z^* \in Z^*$ at $z \in Z$. For a set $D \subset Z$, we denote by $\sigma_D(\cdot)$ its support function $\sigma_D(z^*) := \sup_{z\in D}\langle z^*, z\rangle$ and by $d(\cdot, D)$ the distance function $d(z,D) := \inf_{y\in D}\|z-y\|$. For a set $D \subset Z$ the tangent cone (contingent or Bouligand cone) to D at $z \in D$ is

$$T_D(z) := \Big\{s \in Z : \exists t_n \to 0_+,\ \exists s_n \to s \text{ such that } z + t_n s_n \in D\Big\},$$

equivalent to $T_D(z) = \big\{s : \liminf_{t\to 0_+} d(z+ts, D)/t = 0\big\}$. If D is a convex set, we have that $T_D(z) = \big\{s : \limsup_{t\to 0_+} d(z+ts, D)/t = 0\big\}$.

If W is another normed space, we denote by $L(Z,W)$ the space of all continuous linear operators from Z into W. If $A \in L(Z,W)$, then $A^* : W^* \to Z^*$ denotes the adjoint operator of A.

2. Basic ideas and preliminaries. The aim of this section is to give some general background on the results we will prove. Let us recall the notion of (local) metric regularity as follows.

DEFINITION 2.1. *Let $\Psi : X \rightrightarrows Y$ be a set-valued map, $\bar{y} \in \Psi(\bar{x})$. The multifunction Ψ is called* metrically regular near $(\bar{x}, \bar{y})$ *if there are neighborhoods $N_{\bar{x}}$, $N_{\bar{y}}$ of $\bar{x}$, $\bar{y}$, respectively, and some $k > 0$ such that $d(x, \Psi^{-1}(y)) \leq k\, d(y, \Psi(x))$ for each $(x, y) \in N_{\bar{x}} \times N_{\bar{y}}$. The lower bound of such k will be called the* constant *of metric regularity of Ψ near $(\bar{x}, \bar{y})$ and denoted by* $\operatorname{Reg} \Psi(\bar{x}, \bar{y})$.

For a survey on the theory of metric regularity and also on the equivalent notions of *pseudo-Lipschitz continuity* and *openness with a linear rate*, we refer to [19] and to the monographs [24], [36]. In the context of optimization, in particular for the theory of necessary optimality conditions, the notion of metric regularity is widely used in connection with constraint qualification conditions. Indeed, the usual constraint qualification conditions of Slater, Mangasarian and Fromowitz, and Robinson are equivalent to metric regularity in an appropriate setting.

However, we will use the notion of metric regularity in another context. It was pointed out by Robinson [34] (see [10] for a similar approach) that there is a very close connection between local weak minimizers and a certain singular behavior of the multifunction Γ at $\bar{x}$.

DEFINITION 2.2. *Let $\Psi : X \rightrightarrows Y$ be a set-valued map, $\bar{y} \in \Psi(\bar{x})$. We call $\bar{y}$ an* extreme value *of Ψ at $\bar{x}$ if for some neighborhood N of $\bar{x}$, $\Psi(N)$ is not a neighborhood of $\bar{y}$.*

The following result is due to Robinson [34].

PROPOSITION 2.3. *If $\bar{x}$ is a local weak minimizer of* (P)*, then* 0 *is an extreme value of Γ at $\bar{x}$.*

Of course, if 0 is an extreme value of Γ at $\bar{x}$, then Γ is not metrically regular near $(\bar{x}, 0)$. This fact was used in [34] to obtain first-order necessary conditions. The notion of extremals is common with first-order optimality conditions (see [28] for a recent survey); however, it seems that Robinson [34] was the first to state that the usual (Fritz John or Kuhn–Tucker) conditions can be interpreted as negations of sufficient conditions for metric regularity of Γ near $(\bar{x}, 0)$. The second-order necessary conditions we present in this paper rely on the following observation.

PROPOSITION 2.4. *Let (x_n) be a sequence converging to $\bar{x}$ such that for each n, the multifunction Γ is metrically regular near $(x_n, 0)$. Then $\bar{x}$ is not a local weak minimizer for* (P).

Proof. We show that 0 is not an extreme value of Γ at $\bar{x}$. Then the assertion follows immediately from Proposition 2.3. Indeed, for any neighborhood N of $\bar{x}$, we can fix some x_n and some $r > 0$ such that $x_n + rB_X \subset N$. Since Γ is metrically regular near $(x_n, 0)$, 0 is not an extreme value of Γ at x_n. Hence, $\Gamma(N) \supset \Gamma(x_n + rB_X)$ is a neighborhood of 0 and we may conclude that 0 is not an extreme value of Γ at $\bar{x}$. ∎

It is clear from the proof of Proposition 2.4 that it would be sufficient to assume that Γ is not open at $(x_n, 0)$. However, we assumed the stronger property of metric regularity because there exist fairly accurate characterizations of metric regularity of Γ. It was shown by Ioffe [19] that metric regularity of Γ near $(\tilde{x}, 0)$ can be characterized by the slope of the functions $\psi_y(x) := d(y, \Gamma(x))$ for y near 0 and x near $\tilde{x}$. A quite similar result was derived by Rockafellar [35]. A link to sufficient conditions is given in terms of the function $\psi_0(x)$ by the following simple observation: The element $\bar{x}$ is a strict local minimizer for (P) if and only if for some neighborhood N of $\bar{x}$ we have

$$\psi_0(x) = d(0, \Gamma(x)) = \max\{d(f(x) - f(\bar{x}), -L),\ d(g(x), K)\} > 0, \quad \bar{x} \neq x \in N.$$

In our special case, $\psi_y(x)$ is simply the distance from y to the closed convex set $\Gamma(x) = h(x) - C$. Hence, instead of the very general theory for characterizing metric regularity by subdifferential calculus given in [19], we can use another approach providing more geometric insight. The core of our investigations will be the notion of a signed distance function, which seems to have been first used in the context of mathematical programming by Hiriart-Urruty [14], [15].

LEMMA 2.5. *Let D be a closed convex subset of a normed space Z and let $\hat{d}_D : Z \to \mathbb{R} \cup \{\pm\infty\}$ be given by*

$$\hat{d}_D(z) := \sup_{z^* \in S_{Z^*}} \{\langle z^*, z\rangle - \sigma_D(z^*)\}.$$

Then, for all $z \in Z$ we have

$$\hat{d}_D(z) = \begin{cases} d(z, D) & \text{if } z \notin D, \\ -\max\{\rho : z + \rho B_Z \subset D\} & \text{if } z \in D. \end{cases}$$

Proof. If $D = \emptyset$ or $D = Z$, then the assertion follows easily from the relations $\sigma_\emptyset(\cdot) \equiv -\infty$, $d(\cdot, \emptyset) \equiv +\infty$, $\sigma_Z(\cdot) \equiv +\infty$, $\sup\{\rho : z + \rho B_Z \subset Z\} = \infty$.

Now assume $\emptyset \neq D \neq Z$. If $z \notin D$, we obtain by the minimum norm duality theorem (see, e.g., [27]) $0 < d(z, D) = \max\{\langle z^*, z\rangle - \sigma_D(z^*) : z^* \in B_{Z^*}\}$. Since $z^* \to \langle z^*, z\rangle - \sigma_D(z^*)$ is positively homogenous and $d(z, D) > 0$, the maximum must be attained at some point $z^* \in S_{Z^*}$. Hence, $\hat{d}_D(z) = d(z, D)$ for all $z \notin D$. Now, let $z \in D$ and set $\bar{\rho} := \sup\{\rho : z + \rho B_Z \subset D\}$. Since D is closed, we have $z + \bar{\rho} B_Z \subset D$. Then, for each $z^* \in S_{Z^*}$ we have $\sigma_D(z^*) \geq \sigma_{z+\bar{\rho}B_Z}(z^*) = \langle z^*, z\rangle + \bar{\rho}$, implying $\hat{d}_D(z) \leq -\bar{\rho}$. On the other hand, for each $\rho > \bar{\rho}$ there exists some $z_\rho \in (z + \rho B_Z) \setminus D$, and thus

$$\begin{aligned} 0 < d(z_\rho, D) = \hat{d}_D(z_\rho) &= \sup_{z^* \in S_{Z^*}} \{\langle z^*, z_\rho\rangle - \sigma_D(z^*)\} \\ &\leq \sup_{z^* \in S_{Z^*}} \{\langle z^*, z\rangle - \sigma_D(z^*)\} + \sup_{z^* \in S_{Z^*}} \langle z^*, z_\rho - z\rangle = \hat{d}_D(z) + \|z_\rho - z\| \leq \hat{d}_D(z) + \rho, \end{aligned}$$

showing $\hat{d}_D(z) > -\rho$ for each $\rho > \bar{\rho}$. Hence, $\hat{d}_D(z) \geq -\bar{\rho}$ also holds. ∎

For $z \in D$, $-\hat{d}_D(z)$ is the radius of the largest ball around z contained in D and acts as a measure for "openness" of D around z. Note that for $z \notin D$ the distance $d(z, D) = \hat{d}_D(z)$ may be interpreted as a measure for "openness" of the complementary set $Z \setminus D$ around z.

In the rest of this section we reformulate some well-known results on metric regularity for a set-valued mapping $\Gamma : X \rightrightarrows Y$ of the form $\Gamma(x) = h(x) - C$ in connection with Lemma 2.5, where X and Y are Banach spaces, $h : X \to Y$ is an arbitrary mapping, and $C \subset Y$ is a closed convex subset of Y.

Assume $h(\bar{x}) \in C$. If h is strictly differentiable at $\bar{x}$, i.e., h is Fréchet differentiable at $\bar{x}$ and the derivative $h'(\bar{x})$ satisfies

$$\lim_{x, x' \to \bar{x}} \frac{\|h(x) - h(x') - h'(\bar{x})(x - x')\|}{\|x - x'\|} = 0,$$

then the following statements are equivalent:

(1) The multifunction Γ is metrically regular near $(\bar{x}, 0)$.

(2) The linearization $\Psi(x) := h(\bar{x}) + h'(\bar{x})(x - \bar{x}) - C$ is metrically regular near $(\bar{x}, 0)$.

(3) There holds $0 \in \operatorname{int}(h(\bar{x}) + h'(\bar{x})X - C)$.

This is basically Cominetti's Corollary 2.2 in [9]. It follows further from [9, Thm. 2.1] that $\operatorname{Reg}\Gamma(\bar{x},0) = \operatorname{Reg}\Psi(\bar{x},0)$. By the generalized open mapping theorem [33, Thm. 1] we can equivalently replace condition (3) with one of the following conditions:

(3′) There is some $\tilde{\kappa} \geq 0$ such that $0 \in \operatorname{int}(h(\bar{x}) + \tilde{\kappa}h'(\bar{x})B_X - C)$.

(3″) For all $\kappa > 0$ one has $0 \in \operatorname{int}(h(\bar{x}) + \kappa h'(\bar{x})B_X - C)$.

Though condition (3) is more convenient to verify in applications, we work with condition (3′) since it also provides some bound on $\operatorname{Reg}\Psi(\bar{x},0)$ (see Lemma 2.9 below).

In what follows we use extensively the function $\hat{d}_C(y,A,\kappa) : Y \times L(X,Y) \times \mathbb{R} \to \mathbb{R}$ given by

$$\hat{d}_C(y,A,\kappa) := \sup_{y^* \in S_{Y^*}} \{\langle y^*, y\rangle - \sigma_C(y^*) - \kappa\|A^*y^*\|\}.$$

The following proposition will be the core of the duality theory used in this paper.

PROPOSITION 2.6. *For each $A \in L(X,Y)$, each $y \in Y$, and each $\kappa \geq 0$ let $D_C(y,A,\kappa)$ be given by $D_C(y,A,\kappa) := y + \kappa AB_X - C$. Then one has*

$$\hat{d}_C(y,A,\kappa) = \begin{cases} d(0, D_C(y,A,\kappa)) & \text{if } 0 \notin \operatorname{cl} D_C(y,A,\kappa), \\ 0 & \text{if } 0 \in \operatorname{bd} D_C(y,A,\kappa), \\ -\sup\ \rho\ :\ \rho B_Y \subset D_C(y,A,\kappa) & \text{if } 0 \in \operatorname{int} D_C(y,A,\kappa). \end{cases}$$

Proof. We apply Lemma 2.5 to the set $D := \operatorname{cl} D_C(y,A,\kappa)$. Since

$$\sigma_D(y^*) = \sigma_{D_C(y,A,\kappa)}(y^*) = \sup_{x \in \kappa B_X, c \in C} \langle y^*, y + Ax - c\rangle = \langle y^*, y\rangle + \kappa\|A^*y^*\| + \sigma_C(-y^*),$$

we obtain

$$\begin{aligned} \hat{d}_D(0) &= \sup_{y^* \in S_{Y^*}} \{\langle y^*, 0\rangle - \sigma_D(y^*)\} = \sup_{y^* \in S_{Y^*}} \{\langle -y^*, y\rangle - \kappa\|A^*y^*\| - \sigma_C(-y^*)\} \\ &= \sup_{y^* \in S_{Y^*}} \{\langle y^*, y\rangle - \kappa\|A^*y^*\| - \sigma_C(y^*)\} = \hat{d}_C(y,A,\kappa). \end{aligned}$$

Next consider the closed convex set $M := \{(x, y+Ax-c) : x \in \kappa B_X,\ c \in C\} \subset X \times Y$. Since X is complete, the projection of M onto X is the bounded set κB_X, and the projection of M onto Y amounts to $D_C(y,A,\kappa)$, we may apply [33, Lem. 1] to obtain $\operatorname{int} D = \operatorname{int} D_C(y,A,\kappa)$. Since we also have $d(0,D) = d(0, D_C(y,A,\kappa))$, the assertion of the theorem follows from Lemma 2.5. ∎

It follows that Γ is metrically regular near $(\bar{x},0)$ if and only if $\hat{d}_C(h(\bar{x}), h'(\bar{x}), \tilde{\kappa}) < 0$ for some $\tilde{\kappa} \geq 0$. Using a slightly different but equivalent formula, this was already observed by Sach [37].

We now state some basic properties of the function $\hat{d}_C(y,A,\kappa)$.

LEMMA 2.7. *Let $A \in L(X,Y)$ be fixed. Then the function $\hat{d}_C(\cdot,A,\cdot)$ is convex on $Y \times \mathbb{R}$. For each fixed y, the function $\hat{d}_C(y,A,\cdot)$ is monotonically decreasing on $\mathbb{R}$, and for each fixed $\kappa \in \mathbb{R}$, the function $\hat{d}_C(\cdot,A,\kappa)$ is Lipschitz on Y with constant 1.*

Proof. Since $\hat{d}_C(\cdot,A,\cdot)$ is the pointwise supremum of a family of affine linear functions, convexity easily follows. The other assertions also follow immediately. ∎

The following two lemmas basically are the results of the equivalence between metric regularity and openness with a linear rate.

LEMMA 2.8. *Let $A \in L(X,Y)$, $y \in Y$, and $\kappa \geq 0$ be such that $\hat{d}_C(y,A,\kappa) < 0$. Then*

$$d(0, A^{-1}(C-y)) \leq \frac{\kappa}{d(y,C) - \hat{d}_C(y,A,\kappa)} d(y,C).$$

Proof. If $y \in C$, then obviously $d(y,C) = d(0,A^{-1}(C-y)) = 0$ holds. Now suppose $y \notin C$. Then $d(y,C) = d(0,y-C) = \hat{d}_C(y,A,0) > 0$ and $\bar{\alpha} := d(y,C)/(d(y,C) - \hat{d}_C(y,A,\kappa)) \in (0,1)$. Convexity of $\hat{d}_C(y,A,\cdot)$ implies that for each $\alpha \in (\bar{\alpha},1]$,

$$\hat{d}_C(y,A,\alpha\kappa) \leq \hat{d}_C(y,A,0) + \alpha(\hat{d}_C(y,A,\kappa) - \hat{d}_C(y,A,0)) < 0,$$

yielding $0 \in \operatorname{int}(y + \alpha\kappa AB_X - C)$ by Proposition 2.6. Hence, for each $\alpha \in (\bar{\alpha},1]$ there exists some $x_\alpha \in \alpha\kappa B_X$ such that $x_\alpha \in A^{-1}(C-y)$. Therefore,

$$d(0,A^{-1}(C-y)) \leq \inf_{\alpha \in (\bar{\alpha},1]} \|x_\alpha\| \leq \inf_{\alpha \in (\bar{\alpha},1]} \alpha\kappa = \bar{\alpha}\kappa = \frac{\kappa}{d(y,C) - \hat{d}_C(y,A,\kappa)} d(y,C). \quad \square$$

LEMMA 2.9. *For given $A \in L(X,Y)$, $\tilde{y} \in C$, and $\tilde{x} \in X$ consider the set-valued map $\Psi(x) := \tilde{y} + A(x-\tilde{x}) - C$. If Ψ is metrically regular near $(\tilde{x},0)$, then*

$$\operatorname{Reg}\Psi(\tilde{x},0) = \lim_{\kappa \to 0_+} \frac{\kappa}{-\hat{d}_C(\tilde{y},A,\kappa)} = \inf\left\{ \frac{\kappa}{-\hat{d}_C(\tilde{y},A,\kappa)} \,:\, \kappa > 0 \right\}.$$

Proof. As already stated above, Ψ is metrically regular near $(\tilde{x},0)$ if and only if $0 \in \operatorname{int}(\tilde{y} + \kappa AB_X - C)$ for all $\kappa > 0$, i.e., $\hat{d}_C(\tilde{y},A,\kappa) < 0$. Convexity of $\hat{d}_C(\tilde{y},A,\cdot)$, together with $\hat{d}_C(\tilde{y},A,0) \leq 0$, implies that the function $\eta(\kappa) := \kappa/(-\hat{d}_C(\tilde{y},A,\kappa))$ is increasing for positive κ. Thus, $\bar{\eta} := \lim_{\kappa \to 0_+} \eta(\kappa)$ exists and $\bar{\eta} = \inf\{\eta(\kappa) \,:\, \kappa > 0\}$. Since Ψ is assumed to be metrically regular near $(\tilde{x},0)$, for each $\epsilon > 0$ there are positive scalars k and ρ with $k < \operatorname{Reg}\Psi(\tilde{x},0) + \epsilon$ and $\operatorname{int}(\rho B_Y) \subset \Psi(\tilde{x} + k\rho B_X)$. Hence, $\hat{d}_C(\tilde{y},A,k\rho) \leq -\rho < 0$ by Proposition 2.6 and $\operatorname{Reg}\Psi(\tilde{x},0) + \epsilon \geq k \geq \eta(k\rho) \geq \bar{\eta}$. On the other hand, for arbitrarily fixed $\tilde{\kappa} > 0$ and $t \in (0,1)$, let $\epsilon := -t\hat{d}_C(\tilde{y},A,\tilde{\kappa})$ and $k := \eta(\tilde{\kappa})/(1-t)$. Next we choose neighborhoods $N_{\tilde{x}}$ of $\tilde{x}$ and N_0 of 0 such that $\|A(x-\tilde{x}) - y\| \leq \epsilon$ for all $x \in N_{\tilde{x}}$ and $y \in N_0$. Then, for any $x \in N_{\tilde{x}}$ and $y \in N_0$ we have $\hat{d}_C(w,A,\tilde{\kappa}) \leq \hat{d}_C(\tilde{y},A,\tilde{\kappa}) + \epsilon = -\kappa/k$, where $w := \tilde{y} - y + A(x-\tilde{x})$. Since $d(y,\Psi(x)) = d(w,C) \geq 0$, by using Lemma 2.8 we obtain

$$d(x,\Psi^{-1}(y)) = d(0,A^{-1}(C-w)) \leq \frac{\kappa d(w,C)}{\kappa/k + d(w,C)} \leq k d(y,\Psi(x)),$$

showing $\operatorname{Reg}\Psi(\tilde{x},0) \leq k = \eta(\tilde{\kappa})/(1-t)$. Since we may choose $t > 0$ and $\tilde{\kappa}$ arbitrarily small, $\operatorname{Reg}\Psi(\tilde{x},0) \leq \bar{\eta}$ follows. $\square$

We now state the main result for deriving the second-order necessary optimality conditions. It concerns the solvability of inclusions of the form $h(x) \in C$ near a given point x^0. This question has been extensively studied in varying degrees of generality by several authors and is ultimately related to the notion of metric regularity and the equivalent notions of pseudo-Lipschitz continuity and openness with a linear rate. We refer here only to [3], [9], [11], [12], [25], [29], [32]. The following proposition is an adaptation of such results by using the signed distance function $\hat{d}_C$ to express the constant of metric regularity.

PROPOSITION 2.10. *Let $\hat{x} \in X$ be given and suppose that there exist a continuous linear mapping $A \in L(X,Y)$, a vector $x^0 \in X$, and scalars $R > 0$, $\hat{\kappa} \geq 0$, and $\gamma > 0$ such that the following conditions are fulfilled:*

$$\|h(x') - h(x) - A(x' - x)\| \leq \gamma \|x' - x\| \ \forall x, x' \in \hat{x} + RB_X, \tag{2.1}$$

$$h(\hat{x}) + A(x^0 - \hat{x}) \in C, \ r := \|x^0 - \hat{x}\| < R/2, \tag{2.2}$$

$$2\gamma(\hat{\kappa} + 3r) + \hat{d}_C(h(\hat{x}), A, \hat{\kappa}) < 0. \tag{2.3}$$

Then there exists some $\tilde{x} \in x^0 + rB_X$ such that $h(\tilde{x}) \in C$ and Γ is metrically regular near $(\tilde{x}, 0)$. Moreover, $\hat{d}_C(h(\tilde{x}), A, \hat{\kappa} + \|\tilde{x} - \hat{x}\|) \leq \hat{d}_C(h(\tilde{x}), A, \hat{\kappa}) + \gamma\|\tilde{x} - \hat{x}\|$.

Proof. Though the statement also follows from [25, Thm. 5.1] (with some different constants appearing in (2.3)), we prefer to give a proof based on the "basic lemma" of [19]. Let $x \in x^0 + rB_X \subset \hat{x} + RB_X$ be arbitrarily fixed. Using (2.1), (2.3), and the inequality $-\|x - \hat{x}\|\|A^*y^*\| \leq -\langle y^*, A(x - \hat{x})\rangle$ we obtain

$$\begin{aligned}
&\hat{d}_C(h(x), A, \hat{\kappa} + \|x - \hat{x}\|) = \sup_{y^* \in S_{Y^*}} \{\langle y^*, h(x)\rangle - \sigma_C(y^*) - (\hat{\kappa} + \|x - \hat{x}\|)\|A^*y^*\|\} \\
&\leq \sup_{y^* \in S_{Y^*}} \{\langle y^*, h(\hat{x})\rangle - \sigma_C(y^*) - \hat{\kappa}\|A^*y^*\| + \langle y^*, h(x) - h(\hat{x})\rangle - \langle y^*, A(x - \hat{x})\rangle\} \\
&\leq \sup_{y^* \in S_{Y^*}} \{\langle y^*, h(\hat{x})\rangle - \sigma_C(y^*) - \hat{\kappa}\|A^*y^*\|\} + \sup_{y^* \in S_{Y^*}} \langle y^*, h(x) - h(\hat{x}) - A(x - \hat{x})\rangle \\
&\leq \hat{d}_C(h(\hat{x}), A, \hat{\kappa}) + \gamma\|x - \hat{x}\| \leq \hat{d}_C(h(\hat{x}), A, \hat{\kappa}) + 2\gamma r < 0.
\end{aligned}$$

Hence, by Lemma 2.8 there exists $\triangle x \in X$ with $h(x) + A\triangle x \in C$ and

$$\|\triangle x\| \leq \frac{\hat{\kappa} + \|x - \hat{x}\|}{-\hat{d}_C(h(\hat{x}), A, \hat{\kappa}) - \gamma\|x - \hat{x}\|} d(h(x), C) \leq \frac{\hat{\kappa} + 2r}{-\hat{d}_C(h(\hat{x}), A, \hat{\kappa}) - 2\gamma r} d(h(x), C).$$

Now choose $\alpha \in (0, 1]$ such that $u := x + \alpha\triangle x \in \hat{x} + RB_X$. Then we have

$$\begin{aligned}
d(h(u), C) &\leq d(h(x) + A\alpha\triangle x, C) + \|h(u) - h(x) - \alpha A\triangle x\| \\
&\leq (1 - \alpha) d(h(x), C) + \gamma\alpha\|\triangle x\| \\
&\leq d(h(x), C) - \alpha\|\triangle x\| \left(\frac{-\hat{d}_C(h(\hat{x}), A, \hat{\kappa}) - 2\gamma r}{\hat{\kappa} + 2r} - \gamma \right) \\
&= d(h(x), C) - K^{-1}\|x - u\|,
\end{aligned}$$

where $K := (\hat{\kappa} + 2r)/(-\hat{d}_C(h(\hat{x}), A, \hat{\kappa}) - \gamma(\hat{\kappa} + 4r))$. Using (2.1) and (2.2) we obtain

$$d(h(x^0), C) \leq d(h(\hat{x}) + A(x^0 - \hat{x}), C) + \|h(x^0) - h(\hat{x}) - A(x^0 - \hat{x})\| \leq \gamma r.$$

Rearranging (2.3) yields $Kd(h(x^0), C) \leq K\gamma r < r$ and by [19, basic lem.] there is some $\tilde{x}$ with $d(h(\tilde{x}), C) = 0$ and $\|\tilde{x} - x^0\| \leq Kd(h(x^0), C) < r$.

In a second step we show that Γ is metrically regular near $(\tilde{x}, 0)$. Since $\hat{d}_C(h(\tilde{x}), A, \hat{\kappa} + \|\tilde{x} - \hat{x}\|) \leq \hat{d}_C(h(\hat{x}), A, \hat{\kappa}) + \gamma\|\tilde{x} - \hat{x}\| < 0$, we obtain from Lemma 2.9 that $\Psi(x) := h(\tilde{x}) + A(x - \tilde{x}) - C$ is metrically regular near $(\tilde{x}, 0)$ and

$$\operatorname{Reg} \Psi(\tilde{x}, 0) \leq \frac{\hat{\kappa} + \|\tilde{x} - \hat{x}\|}{-d_C(h(\hat{x}), A, \hat{\kappa}) - \gamma\|\tilde{x} - \hat{x}\|} \leq \frac{\hat{\kappa} + 2r}{-d_C(h(\hat{x}), A, \hat{\kappa}) - 2\gamma r} < \frac{1}{2\gamma}.$$

Since $\Gamma(x) = \Psi(x) + (h(x) - h(\tilde{x}) - A(x - \tilde{x}))$ and $h(\cdot) - h(\tilde{x}) - A(\cdot - \tilde{x})$ is Lipschitz near $\tilde{x}$ with constant γ, we may conclude from [9, Thm. 2.1] that Γ is metrically regular near $(\tilde{x}, 0)$. ∎

3. The main optimality conditions. In what follows let $\bar{x} \in X$ be fixed, let the function $h : X \to Y$ and the closed convex set $C \subset Y$ be given by (1.1), and let Γ be given by (1.2). The second-order sufficient conditions given in this paper are an equivalent characterization of essential local minimizers of second order given by the following definition of Penot [31].

DEFINITION 3.1. *We say that $\bar{x}$ is an* essential local minimizer of second order *for problem* (P) *if $\bar{x}$ is feasible and there exists some neighborhood N of $\bar{x}$ and some real $\beta > 0$ such that*

$$d(0, \Gamma(x)) \geq \beta \|x - \bar{x}\|^2 \quad \forall x \in N. \tag{3.1}$$

Note that for scalar f, Definition 3.1 is equivalent to Penot's definition [31] of essential local minimizers of second order.

Since $d(0, \Gamma(x)) = d(h(x), C) = \max\{d(f(x) - f(\bar{x}), -L), d(g(x), K)\}$, an essential local minimizer of second order also fulfills

$$d(f(x) - f(\bar{x}), -L) \geq \beta \|x - \bar{x}\|^2 \quad \forall x \in g^{-1}(K) \cap N, \tag{3.2}$$

which clearly implies that $\bar{x}$ is a strict local minimizer. A point $\bar{x} \in g^{-1}(K)$ satisfying (3.2) is called by Jiménez and Novo [20] a *strict local minimizer of second order* for problem (P). Note that for scalar f, (3.2) is nothing else than the so-called *quadratic growth condition.* Equation (3.2) is also sufficient for $\bar{x}$ to be an essential local minimizer, provided the mapping $g(x) - K$ associated with the constraints is metrically regular near $(\bar{x}, 0)$ and f is Lipschitz near $\bar{x}$. To see this, note that by Definition 2.1 there are $k > 0$ and a neighborhood $\hat{N}$ of $\bar{x}$ such that for each $x \in \hat{N}$ there is some $\hat{x} \in g^{-1}(K)$ with $\|\hat{x} - x\| \leq k d(0, g(x) - K) = k d(g(x), K)$. We can also assume that f is Lipschitz on $\hat{N}$ with constant $\lambda > 0$. Now let us fix $\hat{\beta} > 0$ such that $\hat{\beta}(1 + k\lambda) \leq \beta/4$ and choose a positive radius $r \leq 1/(2k\hat{\beta})$ such that $\bar{x} + \frac{3}{2} r B_X \subset \hat{N} \cap N$. If $d(g(x), K) \geq \hat{\beta}\|x - \bar{x}\|^2$, then obviously $d(0, \Gamma(x)) \geq \hat{\beta}\|x - \bar{x}\|^2$ follows. On the other hand, for all $x \in \bar{x} + r B_X$ with $d(g(x), K) \leq \hat{\beta}\|x - \bar{x}\|^2$ we have $\|\hat{x} - x\| \leq k\hat{\beta}\|x - \bar{x}\|^2 \leq \frac{1}{2}\|x - \bar{x}\|$, implying $\hat{x} \in N$ and $\|\hat{x} - \bar{x}\| \geq \frac{1}{2}\|x - \bar{x}\|$. Using (3.2) we obtain

$$\begin{aligned} d(f(x) - f(\bar{x}), -L) &\geq d(f(\hat{x}) - f(\bar{x}), -L) - \lambda\|x - \hat{x}\| \geq \beta\|\hat{x} - \bar{x}\|^2 - \lambda k d(g(x), K) \\ &\geq \left(\frac{\beta}{4} - \lambda k \hat{\beta}\right) \|\hat{x} - \bar{x}\|^2 \geq \hat{\beta}\|x - \bar{x}\|^2. \end{aligned}$$

It follows that $d(0, \Gamma(x)) \geq \hat{\beta}\|x - \bar{x}\|^2$ holds for all $x \in \bar{x} + r B_X$.

Our optimality conditions are valid under the following assumption.

ASSUMPTION 1. *The mapping h is Fréchet differentiable at $\bar{x}$, and for some radius $\bar{r} > 0$ and some scalar $\eta \geq 0$ we have*

$$\|h(x_1) - h(x_2) - h'(\bar{x})(x_1 - x_2)\| \leq \eta \max\{\|x_1 - \bar{x}\|, \|x_2 - \bar{x}\|\}\|x_1 - x_2\|$$

for all $x_1, x_2 \in \bar{x} + \bar{r} B_X$.

For example, Assumption 1 is satisfied if h is continuously differentiable at $\bar{x}$ and h' is Lipschitz near $\bar{x}$ or if h is twice Fréchet differentiable at $\bar{x}$.

THEOREM 3.2. *Suppose that Assumption* 1 *is satisfied at $\bar{x}$. If $\bar{x}$ is a local weak minimizer for* (P)*, then*

$$\liminf_{\substack{x \to \bar{x} \\ \tau \to 0_+}} \frac{\hat{d}_C(h(x), h'(\bar{x}), \tau\|x - \bar{x}\|)}{\|x - \bar{x}\|^2} \geq 0. \tag{3.3}$$

Moreover, a feasible point $\bar{x}$ is an essential local minimizer of second order for (P) *if and only if*

$$\liminf_{\substack{x\to\bar{x}\\ \tau\to 0_+}} \frac{\hat{d}_C(h(x), h'(\bar{x}), \tau\|x-\bar{x}\|)}{\|x-\bar{x}\|^2} > 0. \tag{3.4}$$

Proof. We prove the first assertion by contraposition. Assume that (3.3) does not hold. We will show that $\bar{x}$ is not a local weak minimizer. In fact, if (3.3) fails to hold, there are sequences $x_n \to \bar{x}$, $\tau_n \to 0_+$ and a positive scalar $\beta > 0$ such that $\hat{d}_C(h(x_n), h'(\bar{x}), \tau_n\|x_n-\bar{x}\|) \le -\beta\|x_n-\bar{x}\|^2 < 0$ for all n. By Proposition 2.6, for each n we can find some $s_n \in \kappa_n B_X$ with $h(x_n) + h'(\bar{x})s_n \in C$, where $\kappa_n := \tau_n\|\bar{x}-x_n\|$. Let $\bar{r}$ and η be given by Assumption 1. Then for all n sufficiently large we have $\|x_n-\bar{x}\| + 3\kappa_n < \bar{r}$, $8\eta(\|x_n-\bar{x}\| + 3\kappa_n)\kappa_n - \beta\|x_n-\bar{x}\|^2 < 0$, and we may apply Proposition 2.10 with data $\hat{x} := x_n$, $A := h'(\bar{x})$, $x^0 := x_n + s_n$, $R = 3\kappa_n$, $\hat{\kappa} := \kappa_n$, and $\gamma := \eta(\|x_n-\bar{x}\| + 3\kappa_n)$ to find some $\tilde{x}_n \in X$ with $h(\tilde{x}_n) \in C$ and $\|x_n-\tilde{x}_n\| \le 2\kappa_n$ such that Γ is metrically regular at $\tilde{x}_n$. Since $\tilde{x}_n \to \bar{x}$, it follows from Proposition 2.4 that $\bar{x}$ is not a local weak minimizer and the first assertion is proved. To prove the only if part of the second assertion, we use quite similar arguments. Suppose that (3.4) does not hold. We will show that $\bar{x}$ is not an essential local minimizer of second order. If (3.4) fails to hold, we can find sequences $x_n \to \bar{x}$ and $\tau_n \to 0_+$ such that $\hat{d}_C(h(x_n), h'(\bar{x}), \tau_n\|x_n-\bar{x}\|) \le \|x_n-\bar{x}\|^2/n$ for all n. Using Proposition 2.6, we may also assume that $d(0, h(x_n) + \tau_n\|x_n-\bar{x}\|h'(\bar{x})B_X - C) \le \|x_n-\bar{x}\|^2/n$. Hence, for each n there exist some $s_n \in \tau_n\|x_n-\bar{x}\|B_X$ and some $c_n \in C$ such that $\|h(x_n) + h'(\bar{x})s_n - c_n\| \le (2/n)\|x_n-\bar{x}\|^2$. Using Assumption 1 we obtain

$$\begin{aligned} d(0, \Gamma(x_n+s_n)) &\le \|h(x_n+s_n) - c_n\| \\ &\le \|h(x_n) + h'(\bar{x})s_n - c_n\| + \|h(x_n+s_n) - h(x_n) - h'(\bar{x})s_n\| \\ &\le (2/n)\|x_n-\bar{x}\|^2 + \eta(\|x_n-\bar{x}\| + \|s_n\|)\|s_n\| = o(\|x_n-\bar{x}\|^2). \end{aligned}$$

Since $\|x_n+s_n\| \ge \|x_n-\bar{x}\|(1-\tau_n)$, it follows that $\bar{x}$ is not an essential local minimizer of second order. To complete the proof of the theorem we must show the if part of the second assertion. This follows immediately from the inequality

$$d(0, \Gamma(x)) = d(h(x), C) \ge \hat{d}_C(h(x), h'(\bar{x}), \tau\|x-\bar{x}\|),$$

valid for all $x \in X$ and $\tau > 0$. ∎

In what follows we use the following definition to refer to points satisfying the necessary condition (3.3).

Definition 3.3. *We say that $\bar{x}$ is a* critical point of second order *for* (P) *if $\bar{x}$ is feasible for* (P)*, h is Fréchet differentiable at $\bar{x}$, and condition* (3.3) *is fulfilled.*

Now we derive second-order conditions from Theorem 3.2 by means of a second-order approximation of h. To simplify the presentation, we restrict ourselves to the case that some uniform second-order expansion of h at $\bar{x}$ is possible.

Assumption 2. *h is second-order directionally differentiable at $\bar{x}$ in the sense that h is Fréchet differentiable at $\bar{x}$, the limit*

$$h''(\bar{x}; z) := \lim_{t\to 0_+} \frac{h(\bar{x}+tz) - h(\bar{x}) - th'(\bar{x})z}{t^2/2}$$

exists for all $z \in X$, and convergence is uniform with respect to z in bounded sets.

If Assumption 2 is satisfied, we obviously have $h''(\bar{x};\alpha z) = \alpha^2 h''(\bar{x};z)$ for all $\alpha \geq 0$ and all $z \in X$. Further, a uniform approximation of the form $h(x) = h(\bar{x}) + h'(\bar{x})(x-\bar{x}) + \frac{1}{2}h''(\bar{x};x-\bar{x}) + o(\|x-\bar{x}\|^2)$ holds. Hence, the following corollary follows immediately.

COROLLARY 3.4. *Suppose $\bar{x}$ is feasible for* (P) *and suppose that Assumptions* 1 *and* 2 *are satisfied. Then, $\bar{x}$ is a critical point (respectively, essential local minimizer) of second order for* (P) *if and only if there exists some scalar $\beta \geq 0$ (respectively, $\beta > 0$) such that for each sequence $(z_n, t_n, \tau_n) \subset S_X \times \mathbb{R}_+ \times \mathbb{R}_+$ with $t_n \to 0_+$ and $\tau_n \to 0_+$ one has*

$$\liminf_{n\to\infty} \sup_{y^* \in S_{Y^*}} \left\{ \frac{\langle y^*, h(\bar{x}) + t_n h'(\bar{x}) z_n \rangle - \sigma_C(y^*)}{t_n^2} + \frac{1}{2}\langle y^*, h''(\bar{x};z_n)\rangle - \frac{\tau_n}{t_n}\|h'(\bar{x})^* y^*\| \right\} \geq \beta. \tag{3.5}$$

The following two theorems give further details on the optimality conditions of Theorem 3.2 and Corollary 3.4.

THEOREM 3.5. *Suppose that a feasible point $\bar{x}$ is not an essential local minimizer of second order for* (P). *Then there exists a twice continuously differentiable mapping $\delta h := (\delta f, \delta g)$ satisfying $\delta h(\bar{x}) = 0$, $\delta h'(\bar{x}) = 0$, and $\delta h''(\bar{x}) = 0$, such that $\bar{x}$ is not a local weak minimizer for* (P) *with f and g replaced by $f + \delta f$ and $g + \delta g$, respectively.*

Proof. Since $\bar{x}$ is not an essential local minimizer, there exists a sequence $x_n \to \bar{x}$ satisfying $d_n := \max\{d(f(x_n) - f(\bar{x}), -L),\ d(g(x_n), K)\} = o(\|x_n - \bar{x}\|^2)$. Hence, we can find a sequence $y_n = (u_n, v_n)$ in $(-\operatorname{int} L) \times K$ such that $\|h(x_n) - y_n\| = \max\{\|f(x_n) - f(\bar{x}) - u_n\|, \|g(x_n) - v_n\|\} \leq d_n + \|x_n - \bar{x}\|^3 =: \bar{d}_n$. Taking a subsequence if necessary, we may suppose that

$$\sum_{n=1}^{\infty} \frac{\bar{d}_n}{\|x_n - \bar{x}\|^2} < \infty. \tag{3.6}$$

We shall now construct a subsequence of (x_n), still denoted by (x_n), satisfying

$$\tau_n := \|x_n - \bar{x}\| \leq \frac{d(\bar{x}, A_{n-1})}{4} \leq \frac{\tau_{n-1}}{4} \quad \forall n \geq 2, \tag{3.7}$$

where A_n denotes the convex hull of $x_1, \ldots, x_n$. Let m_n denote the dimension of the linear hull of $x_1 - \bar{x}, \ldots, x_n - \bar{x}$. If the sequence (m_n) is bounded, $x_1 - \bar{x}, x_2 - \bar{x}, \ldots$ are contained in a finite-dimensional subspace $E \subset X$. Hence, we can find a subsequence of (x_n), still denoted by (x_n), and $z \in E$, $\|z\| = 1$ such that $(x_n - \bar{x})/\tau_n \to z$. By passing once more to a subsequence if necessary, we may also assume that $\|(x_n - \bar{x})/\tau_n - z\| \leq 1/2$ for all n and $\tau_n \leq \tau_{n-1}/8$ for all $n \geq 2$. Now, let $n \geq 2$ and $x \in A_{n-1}$ be arbitrarily fixed. Then x can be expressed as a convex combination $\sum_{i=1}^{n-1} \alpha_i x_i$, and we obtain

$$\begin{aligned} \|x - \bar{x}\| &= \left\| \sum_{i=1}^{n-1} \alpha_i (x_i - \bar{x}) \right\| \geq \left\| \sum_{i=1}^{n-1} \alpha_i \tau_i z \right\| - \left\| \sum_{i=1}^{n-1} \alpha_i \tau_i \left(\frac{x_i - \bar{x}}{\tau_i} - z \right) \right\| \\ &\geq \frac{1}{2} \sum_{i=1}^{n-1} \alpha_i \tau_i \geq \frac{\tau_{n-1}}{2}. \end{aligned}$$

Since $x \in A_{n-1}$ was arbitrary, we have $d(\bar{x}, A_{n-1}) \geq \tau_{n-1}/2 \geq 4\tau_n$, and (3.7) follows. On the other hand, if the sequence (m_n) is unbounded, we can find a subsequence of (x_n), still denoted by (x_n), such that $m_n = n$ holds for all n. Then we have $\bar{x} \notin A_n$ for all n. Since A_n is the convex hull of finitely many points, A_n is closed. Therefore, we have $d(\bar{x}, A_n) > 0$ for all n, and by taking again a suitable subsequence, (3.7) holds.

By the minimum norm duality theorem (see, e.g., [27]), we can find for each $n \geq 2$ some $p_n^* \in B_{X^*}$ such that $d(x_n, A_{n-1}) = \langle p_n^*, x_n \rangle - \sigma_{A_{n-1}}(p_n^*)$. Then, for each $n' < n$ we have $x_{n'} \in A_{n-1}$, and hence

$$\langle p_n^*, x_n - x_{n'} \rangle \geq \langle p_n^*, x_n \rangle - \sigma_{A_{n-1}}(p_n^*) = d(x_n, A_{n-1}) \geq d(\bar{x}, A_{n-1}) - \tau_n \geq 3\tau_n.$$

On the other hand, for each n there exists also $q_n^* \in S_{X^*}$ such that $\langle q_n^*, x_n - \bar{x} \rangle = \tau_n$, implying

$$\langle q_n^*, x_n - x_{n'} \rangle = \langle q_n^*, x_n - \bar{x} \rangle + \langle q_n^*, \bar{x} - x_{n'} \rangle \geq \tau_n - \tau_{n'} \geq \frac{3}{4}\tau_n$$

for each $n' > n$. Now, for each n let

$$\delta h_n(x) := \max\left\{1 - 2\frac{\langle p_n^*, x_n - x \rangle^2 + \langle q_n^*, x_n - x \rangle^2}{\tau_n^2}, 0\right\}^3 (y_n - h(x_n)) \quad \forall x \in X,$$

where $p_1^* := 0$. It follows that

$$\delta h_n(x_{n'}) = \begin{cases} y_n - h(x_n) & \text{for } n = n', \\ 0 & \text{for } n \neq n'. \end{cases} \tag{3.8}$$

Further, δh_n is twice continuously differentiable and

$$\begin{aligned} \delta h_n'(x)s &= -12c_n^2(x)B_n(x,s)r_n, \\ \delta h_n''(x)st &= 12c_n(x)r_n\left(8B_n(x,s)B_n(x,t) - c_n(x)\frac{\langle p_n^*, s\rangle\langle p_n^*, t\rangle + \langle q_n^*, s\rangle\langle q_n^*, t\rangle}{\tau_n^2}\right), \end{aligned}$$

where $c_n(x) := \max\{1 - 2(\langle p_n^*, x_n - x\rangle^2 + \langle q_n^*, x_n - x\rangle^2)/\tau_n^2, 0\}$, $r_n := y_n - h(x_n)$, and $B_n(x,s) := (\langle p_n^*, x - x_n\rangle\langle p_n^*, s\rangle + \langle q_n^*, x - x_n\rangle\langle q_n, s\rangle)/\tau_n^2$. Obviously, $0 \leq c_n(x) \leq 1$ holds for all $x \in X$. If $c_n(x) > 0$, we have $|\langle p_n^*, x - x_n\rangle| \leq \tau_n/\sqrt{2}$ and $|\langle q_n^*, x - x_n\rangle| \leq \tau_n/\sqrt{2}$, implying $|B_n(x,s)| \leq \sqrt{2}\|s\|/\tau_n$. Hence, by these facts and (3.6), we obtain

$$\sum_{n=1}^{\infty} \|\delta h_n(x)\| \leq \sum_{n=1}^{\infty} \bar{d}_n < \infty, \quad \sum_{n=1}^{\infty} \|\delta h_n'(x)\| \leq \sum_{n=1}^{\infty} 12\sqrt{2}\frac{\bar{d}_n}{\tau_n} < \infty,$$

$$\sum_{n=1}^{\infty} \|\delta h_n''(x)\| \leq \sum_{n=1}^{\infty} 216\frac{\bar{d}_n}{\tau_n^2} < \infty$$

for all $x \in X$. Since Y is a Banach space, we may conclude that $\delta h := \sum_{n=1}^{\infty} \delta h_n$ is well defined and twice continuously differentiable due to the uniform convergence with respect to x of $\sum_n \delta h_n$, $\sum_n \delta h_n'$, and $\sum_n \delta h_n''$. Because $\langle q_n^*, x_n - \bar{x}\rangle = \tau_n$, we have $c_n(\bar{x}) = 0$ for all n. Therefore, $\delta h(\bar{x}) = 0$, $\delta h'(\bar{x}) = 0$, $\delta h''(\bar{x}) = 0$. Splitting δh into the two components δf and δg, it remains to show that $\bar{x}$ is not a local weak

minimizer for (P), with f and g replaced by $f+\delta f$ and $g+\delta g$. Since $\delta h(\bar{x})=0$, $\bar{x}$ is feasible also for the perturbed problem, and the value of the objective function at $\bar{x}$ remains unchanged. However, since $\delta h(x_n)=y_n-h(x_n)$ for each n by (3.8), we obtain

$$(g+\delta g)(x_n)=g(x_n)+(v_n-g(x_n))=v_n\in K,$$

$$(f+\delta f)(x_n)-(f+\delta f)(\bar{x})=f(x_n)+u_n-(f(x_n)-f(\bar{x}))-f(\bar{x})=u_n\in -\operatorname{int} L.$$

Since $x_n\to\bar{x}$, this completes the proof. □

THEOREM 3.6. *Assume that $\bar{x}$ is a critical point of second order for problem* (P). *Further assume that*

$$\operatorname{int}(h'(\bar{x})B_X-C)\neq\emptyset. \tag{3.9}$$

Then there exist a mapping $\delta h=(\delta f,\delta g)$ with $\delta h(x)=\psi(\|x-\bar{x}\|)y$, where $y\in Y$ and $\psi:\mathbb{R}_+\to\mathbb{R}_+$ is a twice continuously differentiable function satisfying $\psi(0)=\psi'(0)=\psi''(0)=0$, such that $\bar{x}$ is a strict local minimizer for (P) *with f and g replaced by $f+\delta f$ and $g+\delta g$, respectively.*

Proof. Define

$$\eta(r):=-\min\left\{0,\inf_{\bar{x}\neq x\in\bar{x}+rB_X}\frac{\hat{d}_C(h(x),h'(\bar{x}),\|x-\bar{x}\|^{3/2})}{\|x-\bar{x}\|^2}\right\},\quad r>0.$$

Then, η is monotonically increasing and $\lim_{r\to 0_+}\eta(r)=0=:\eta(0)$. We shall now construct a twice continuously differentiable function $\psi:\mathbb{R}_+\to\mathbb{R}_+$ so that $\psi(0)=\psi'(0)=\psi''(0)=0$ and $\psi(r)>r^2\eta(r)$ for all $r\in(0,1)$. At first we define recursively a sequence (φ_n) of positive reals by $\varphi_0:=\eta(1)+1$ and $\varphi_n:=\max\{\varphi_{n-1}/2,\eta(2^{-n})\}$. If $\varphi_n=\varphi_{n-1}/2$, we obviously have $\varphi_n<\varphi_{n-1}$. On the other hand, if $\varphi_n=\eta(2^{-n})$, then by the monotonicity of η we have $\varphi_n\le\eta(2^{-(n-1)})\le\varphi_{n-1}$. Hence, the sequence (φ_n) is decreasing. Moreover, it is easy to see that $\lim_n\varphi_n=0$. Hence, we can construct a continuously increasing function $\varphi:\mathbb{R}_+\to\mathbb{R}_+$ in such a way that $\varphi(0)=0$, φ is linear on every interval $[2^{-(n+1)},2^{-n}]$, $\varphi(2^{-n})=\varphi_n$, and $\varphi(t)=\varphi_0+(t-1)\varphi_0$ for $t>1$. Define $\psi'(r):=256\int_0^r\varphi(\xi)d\xi$ for all $r\ge 0$. Since $\varphi(\xi)$ is positive and increasing for $\xi>0$ and since $\varphi_{n+1}\ge\varphi_n/2$, we get

$$\psi'(2^{-n})>256\int_{2^{-(n+1)}}^{2^{-n}}\varphi(2^{-(n+1)})d\xi=256\cdot 2^{-(n+1)}\varphi_{n+1}\ge 64\cdot 2^{-n}\varphi_n$$

for all n. Now define $\psi(r):=\int_0^r\psi'(\xi)d\xi$ for $r\ge 0$. Let $r\in(0,1)$ be arbitrarily fixed and let n denote the unique number such that $r\in[2^{-(n+1)},2^{-n})$. Using similar arguments as above we obtain

$$\begin{aligned}\psi(r)&>\int_{2^{-(n+2)}}^{2^{-(n+1)}}\psi'(2^{-(n+2)})d\xi=2^{-(n+2)}\psi'(2^{-(n+2)})\\&>64\cdot(2^{-(n+2)})^2\varphi_{n+2}\ge(2^{-n})^2\varphi_n\ge(2^{-n})^2\eta(2^{-n})\ge r^2\eta(r).\end{aligned}$$

Further, it is easy to see that ψ has the claimed differentiability properties.

By virtue of (3.9) we can find some $\tilde{y}\in Y$ so that $0\in\operatorname{int}(\tilde{y}+h(\bar{x})+h'(\bar{x})B_X-C)$. By Proposition 2.6 we have $-2\rho:=\hat{d}_C(\tilde{y}+h(\bar{x}),h'(\bar{x}),1)<0$. Now let $\delta h(x):=$

$\psi(\|x-\bar{x}\|)y$, where $y := -\tilde{y}/\rho$. Now we show that $\bar{x}$ is a strict local minimizer for the perturbed problem. By Lemma 2.7, $\hat{d}_C(\cdot, h'(\bar{x}), \cdot)$ is convex. Thus

$$\begin{aligned}\hat{d}_C((1+\alpha)h(x) - \alpha(\tilde{y}+h(\bar{x})), h'(\bar{x}), (1+\alpha)\|x-\bar{x}\|^{3/2} - \alpha) \\ \geq (1+\alpha)\hat{d}_C(h(x), h'(\bar{x}), \|x-\bar{x}\|^{3/2}) - \alpha\hat{d}_C(\tilde{y}+h(\bar{x}), h'(\bar{x}), 1) \\ \geq -(1+\alpha)\|x-\bar{x}\|^2\eta(\|x-\bar{x}\|) + 2\alpha\rho\end{aligned}$$

for all $\alpha \geq 0$ and all x. By Lipschitz continuity of $\hat{d}_C(\cdot, h'(\bar{x}), \kappa)$ we obtain

$$\begin{aligned}&\hat{d}_C(h(x) - \alpha\tilde{y}, h'(\bar{x}), (1+\alpha)\|x-\bar{x}\|^{3/2} - \alpha) \\ &\geq -\|x-\bar{x}\|^2\eta(\|x-\bar{x}\|) + \alpha[2\rho - \|x-\bar{x}\|^2\eta(\|x-\bar{x}\|) - \|h(x)-h(\bar{x})\|].\end{aligned} \tag{3.10}$$

Finally, we can find some $R \in (0,1)$ such that for all $x \in \bar{x} + RB_X$ we have

$$\|x-\bar{x}\|^2\eta(\|x-\bar{x}\|) + \|h(x)-h(\bar{x})\| < \rho, \ \frac{\psi(\|x-\bar{x}\|)}{\rho} \leq \|x-\bar{x}\|^{3/2}.$$

Hence, for all $x \in \bar{x} + RB_X$, $x \neq \bar{x}$, we have

$$\kappa(x) := \left(1 + \frac{\psi(\|x-\bar{x}\|)}{\rho}\right)\|x-\bar{x}\|^{3/2} - \frac{\psi(\|x-\bar{x}\|)}{\rho} \geq \frac{\psi(\|x-\bar{x}\|)}{\rho}\|x-\bar{x}\|^{3/2} > 0,$$

and by using (3.10) with $\alpha = \psi(\|x-\bar{x}\|)/\rho$,

$$\hat{d}_C((h+\delta h)(x), h'(\bar{x}), \kappa(x)) \geq -\|x-\bar{x}\|^2\eta(\|x-\bar{x}\|) + \psi(\|x-\bar{x}\|) > 0,$$

which together with Proposition 2.6 shows that $d((h+\delta h)(x), C) > 0$. Using $\delta f(\bar{x}) = 0$, we obtain that $\bar{x}$ is a strict local minimizer for the perturbed problem. □

Now, the smoothness of the perturbation function δh given by Theorem 3.6 depends on geometric properties of the underlying space X. However, in any case, δh fulfills Assumptions 1 and 2: It is easy to see that δh is Fréchet differentiable at $\bar{x}$ with $\delta h'(\bar{x}) = 0$ and second-order directionally differentiable with

$$\delta h''(\bar{x}; z) = \lim_{t \to 0_+} \frac{\psi(t\|z\|) - \psi(0) - t\psi'(0)}{t^2/2} y = 0 \quad \forall z \in X, \tag{3.11}$$

and that the convergence in (3.11) is uniform with respect to z in bounded sets. Further, δh fulfills Assumption 1 with arbitrarily small η. In fact, let $\eta > 0$ be arbitrarily fixed. Since ψ is twice continuously differentiable with $\psi''(0) = \psi'(0) = 0$, we can find some $r_\eta > 0$ so that $\|y\||\psi'(r)| \leq \eta r$ for all $r \in [0, r_\eta]$. By the mean-value theorem, for every $x_1, x_2 \in \bar{x} + r_\eta B_X$ we can find some r in the interval $[\|x_1-\bar{x}\|, \|x_2-\bar{x}\|] \subset [0, r_\eta]$ so that $\psi(\|x_1-\bar{x}\|) - \psi(\|x_2-\bar{x}\|) = \psi'(r)(\|x_1-\bar{x}\| - \|x_2-\bar{x}\|)$. Since $\|y\||\psi'(r)| \leq \eta r \leq \eta \max\{\|x_1-\bar{x}\|, \|x_2-\bar{x}\|\}$ and $\delta h'(\bar{x}) = 0$, we obtain

$$\begin{aligned}\|\delta h(x_1) - \delta h(x_2) - \delta h'(\bar{x})(x_1-x_2)\| &= \|\psi'(r)(\|x_1-\bar{x}\| - \|x_2-\bar{x}\|)y\| \\ &\leq |\psi'(r)|\|y\|\|x_1-x_2\| \leq \eta\max\{\|x_1-\bar{x}\|, \|x_2-\bar{x}\|\}\|x_1-x_2\|.\end{aligned}$$

Hence, Assumption 1 is fulfilled for δh.

Moreover, it is easy to see that δh is twice continuously differentiable if X is a Hilbert space. Similarly, if X is Hilbertizable (i.e., there exists a norm $|||\cdot|||$ that is equivalent to the one of X and for which X is a Hilbert space), some straightforward

modifications of the proof of Theorem 3.6 yield that the assertion remains valid with some twice continuously differentiable perturbation function of the form $\delta h(x) = \psi(|\|x - \bar{x}\||)y$.

Now let us discuss condition (3.9). If (3.9) is not fulfilled, as a consequence of Proposition 2.6 we have that the necessary conditions (3.3) and (3.5) with $\beta = 0$ automatically hold, regardless of whether $\bar{x}$ is a local weak minimizer or not. In this case, the necessary conditions only reflect some degeneration of the problem.

DEFINITION 3.7. *Let h be differentiable at $\bar{x}$. Then $\bar{x}$ is called* nondegenerate *if* (3.9) *holds, i.e.,* $\operatorname{int}(h'(\bar{x})B_X - C) \neq \emptyset$.

Note that condition (3.9) is actually a condition on the constraints: since we assume $\operatorname{int} L \neq \emptyset$, (3.9) holds if and only if

$$\operatorname{int}(g'(\bar{x})B_X - K) \neq \emptyset. \tag{3.12}$$

As a consequence of the generalized open mapping theorem [33, Thm. 1], this can also be equivalently written as $\operatorname{int}(g'(\bar{x})X - K) \neq \emptyset$. This shows that condition (3.12) is weaker than Robinson's constraint qualification

$$0 \in \operatorname{int}(g(\bar{x}) + g'(\bar{x})X - K), \tag{3.13}$$

stating the metric regularity of the constraint mapping $g(x) - K$ at $(\bar{x}, 0)$.

If $\operatorname{int} K \neq \emptyset$, then (3.12) is trivially fulfilled. If the constraints can be subdivided into general inequality constraints of the form $\tilde{g}(x) \in \tilde{K}$ with $\operatorname{int} \tilde{K} \neq \emptyset$ and "equality" constraints $G(x) \in Q$, where the interior of the closed convex set Q is possibly empty, $\operatorname{int} Q = \emptyset$, then condition (3.12) holds if and only if the condition on the "equality" constraints

$$\operatorname{int}(G'(\bar{x})B_X - Q) \neq \emptyset \tag{3.14}$$

holds. For instance, in the case of the mathematical programming problem, we have $Q = \{0\}$, and then conditions (3.12) and (3.14), respectively, are equivalent to the well-known Lyusternik condition $G'(\bar{x})X = \hat{V}$. It has been already observed by several authors (see, e.g., [4], [5], [18], [26]) that only some regularity assumptions on the equality constraints are required for the development of second-order necessary conditions. Another special case of interest appears when the "equality" constraints are explicit constraints of the form $x \in Q$, i.e., $G(x) = x$. Then condition (3.14), and hence also condition (3.12), is fulfilled for such structured constraints, and consequently $\bar{x}$ is nondegenerate.

Note that condition (3.9) is essential in Theorem 3.6: For the scalar problem

$$\min_{x \in \mathbb{R}^2} -x_1^2 - x_2^2 \quad \text{subject to} \quad x_1 x_2 = 0,$$

the point $\bar{x} = (0, 0)$ is degenerate (i.e., (3.9) is not fulfilled) and by using the second-order necessary conditions [1, Thm. 3.2] one can see that the conclusion of Theorem 3.6 does not hold. For a survey on second-order conditions in the degenerate case for a special class of problems, we refer to the monograph [2]. Note that our definition of degenerate points corresponds to that of *abnormal points* in [1] and [2].

In this paper we restrict ourselves to necessary optimality conditions for nondegenerate local minimizers only. However, in another paper [13] we present new second-order necessary conditions which are meaningful also in the degenerate case. These conditions rest also upon Propositions 2.4 and 2.10; however, second-order derivatives are used to build first-order approximations of first derivatives, whereas in this paper second-order derivatives are used only to build second-order expansions of our problem functions.

4. First-order necessary conditions revisited. Up to now we have said nothing about first-order necessary conditions. One reason is that some kind of first-order necessary conditions are implicitly contained in the second-order conditions; another reason is that at an essential local minimizer of second order the classical multiplier conditions need not hold, as can be seen from an example that will be presented in the next section.

Throughout this section we assume that h is strictly differentiable at $\bar{x}$. Then, Fritz John–type optimality conditions can be written in the form

$$f'(\bar{x})^*u^* + g'(\bar{x})^*v^* = 0, \quad 0 \neq (u^*, v^*) \in L^* \times N_K(g(\bar{x})) \subset U^* \times V^*, \tag{4.1}$$

where $L^* := \{u^* \in U^* : \langle u^*, u\rangle \geq 0 \text{ for all } u \in L\}$ is the dual cone of the cone L and $N_K(z) := \{v^* \in V^* : \langle v^*, v - z\rangle \leq 0 \text{ for all } v \in K\}$ is the normal cone to K at z. Setting $y^* := (u^*, v^*)$ and using the notation of h and C, condition (4.1) can also be written in a shorter form as

$$h'(\bar{x})^*y^* = 0, \quad y^* \in N_C(h(\bar{x})), \quad y^* \neq 0. \tag{4.2}$$

In what follows we denote the set of multipliers y^* satisfying the Fritz John conditions (4.2) by Λ_{FJ}. It should be noted that in general Λ_{FJ} may be empty. An additional condition has to be imposed to ensure the existence of a nontrivial multiplier y^* at a local weak minimizer for (P). For instance, condition (3.9) can be used, as easily follows from Theorem 3 in [34].

We say that the Kuhn–Tucker conditions hold if the Fritz John conditions (4.1) hold with $u^* \neq 0$. It is well known that at a local weak minimizer the Kuhn–Tucker conditions hold under Robinson's constraint qualification (3.13); see [34, Thm. 4]. We denote by Λ_{KT} the set of multipliers y^* satisfying the Kuhn–Tucker conditions.

We now discuss some aspects of first-order conditions in terms of the material presented in section 2. If $\bar{x}$ is a local weak minimizer of problem (P), then by Proposition 2.3 $h(\cdot) - C$ is not metrically regular near $(\bar{x}, 0)$, or equivalently, since h is assumed to be strictly differentiable at $\bar{x}$,

$$\hat{d}_C(h(\bar{x}), h'(\bar{x}), 1) = \sup_{y^* \in S_{Y^*}} \{\langle y^*, h(\bar{x})\rangle - \sigma_C(y^*) - \|h'(\bar{x})^*y^*\|\} \geq 0.$$

Since $h(\bar{x}) \in C$ we have $\langle y^*, h(\bar{x})\rangle - \sigma_C(y^*) \leq 0$ for all y^*. Hence, $h(\cdot) - C$ is not metrically regular near $(\bar{x}, 0)$ if and only if there exists a sequence $(y_n^*) \subset S_{Y^*}$ with

$$\lim_{n\to\infty} \{\sigma_C(y_n^*) - \langle y_n^*, h(\bar{x})\rangle\} = 0, \quad \lim_{n\to\infty} \|h'(\bar{x})^*y_n^*\| = 0. \tag{4.3}$$

LEMMA 4.1. *If $\bar{x}$ is a critical point of second order for problem* (P)*, then $h(\cdot) - C$ is not metrically regular at $(\bar{x}, 0)$.*

Proof. We argue by contraposition. Assume on the contrary that $h(\cdot) - C$ is metrically regular near $(\bar{x}, 0)$. Then $\bar{x}$ is not a local weak minimizer; i.e., there exist a sequence $(x_n) \to \bar{x}$ such that $x_n \neq \bar{x}$ and $h(x_n) \in C$ for all n. Define $\tau_n := \|x_n - \bar{x}\|$. Then, by the definition of a critical point of second order, we obtain a sequence $(y_n^*) \subset S_{Y^*}$ such that

$$\liminf_{n\to\infty} \left\{ \frac{\langle y_n^*, h(x_n)\rangle - \sigma_C(y_n^*)}{\|x_n - \bar{x}\|^2} - \|h'(\bar{x})^*y_n^*\| \right\} \geq 0.$$

Since $h(x_n) \in C$ we have $\langle y_n^*, h(x_n)\rangle - \sigma_C(y_n^*) \leq 0$, implying $\lim_{n\to\infty} \|h'(\bar{x})^*y_n^*\| = \lim_{n\to\infty} \langle y_n^*, h(x_n)\rangle - \sigma_C(y_n^*) = 0$. By continuity of h at $\bar{x}$, we obtain that the sequence

(y_n^*) satisfies (4.3). Hence, $h(\cdot) - C$ is not metrically regular at $(\bar{x}, 0)$, which is a contradiction. ∎

Now let $(y_n^*) \subset S_{Y^*}$ be an arbitrary sequence satisfying (4.3). By the Alaoglu–Bourbaki theorem, B_{Y^*} is weakly* compact, and hence the sequence (y_n^*) has at least one weak* accumulation point. Let $\bar{y}^* \in B_{Y^*}$ denote an arbitrary weak* accumulation point of (y_n^*). Since both $\sigma_C(\cdot) - \langle \cdot, h(\bar{x})\rangle$ and $\|h'(\bar{x})^* \cdot \|$ are weakly* lower semicontinuous, we have $h'(\bar{x})^*\bar{y}^* = 0$ and $\sigma_C(\bar{y}^*) - \langle \bar{y}^*, h(\bar{x})\rangle = 0$, where the latter can also be written as $\bar{y}^* \in N_C(h(\bar{x}))$. This implies $\bar{y}^* \in \Lambda_{FJ}$, provided $\bar{y}^* \neq 0$. Now assume that $\bar{x}$ is nondegenerate and let $\bar{y} \in h(\bar{x}) + \operatorname{int}(h'(\bar{x})B_X - C)$ and $-\rho := \hat{d}_C(h(\bar{x}) - \bar{y}, h'(\bar{x}), 1) < 0$. Then we have $\langle y_n^*, h(\bar{x}) - \bar{y}\rangle - \sigma_C(y_n^*) - \|h'(\bar{x})^* y_n^*\| \leq -\rho$, implying $\langle \bar{y}^*, \bar{y}\rangle \geq \rho > 0$. Therefore, $0 \neq \bar{y}^*$ and consequently $\bar{y}^* \in \Lambda_{FJ}$. Thus, we have shown the following result, which will be useful in what follows.

PROPOSITION 4.2. *Assume that $\bar{x}$ is nondegenerate and let $(y_n^*) \subset S_{Y^*}$ be a sequence satisfying* (4.3). *Then, for each weak* accumulation point $\bar{y}^*$ of (y_n^*) one has $\bar{y}^* \in \Lambda_{FJ} \cap B_{Y^*}$.*

In the case of scalar minimization the commonly used set of Lagrange multipliers is given by $\{v^* \in V^* : (1, v^*) \in \Lambda_{FJ}\}$. It is well known that the set of Lagrange multipliers is convex and weakly* compact if Robinson's constraint qualification (3.13) is fulfilled (see, e.g., [7, Thm. 3.9]). Now we generalize this result to problem (P), where we also replace the assumption of Robinson's constraint qualification with the weaker condition (3.9).

For arbitrary $y \in Y$ let $\Lambda_{FJ}^y := \{y^* \in \Lambda_{FJ} : \langle y^*, y\rangle = 1\}$. Note that in the case of scalar minimization the set of Lagrange multipliers is related to Λ_{FJ}^y with $y = (1, 0) \in \mathbb{R} \times V$ and that Robinson's constraint qualification holds if and only if $(1, 0) \in \operatorname{int}(h(\bar{x}) + h'(\bar{x})X - C)$. The following theorem appears to be new even in the scalar case.

THEOREM 4.3. *Let $\Lambda_{FJ} \neq \emptyset$ and assume that $\bar{x}$ is nondegenerate. Then for all $y \in \operatorname{int}(h(\bar{x}) + h'(\bar{x})B_X - C)$ the set Λ_{FJ}^y is a nonempty, convex, and weakly* compact subset of Y^* and Λ_{FJ} is generated by Λ_{FJ}^y in the sense that $\Lambda_{FJ} = (0, +\infty)\Lambda_{FJ}^y$.*

Proof. It is clear that Λ_{FJ}^y is convex and weakly* closed for all $y \in Y$, but possibly empty. Now let $y \in \operatorname{int}(h(\bar{x}) + h'(\bar{x})B_X - C)$ be arbitrarily chosen and let $-\rho := \hat{d}_C(h(\bar{x}) - y, h'(\bar{x}), 1) < 0$. Then, for all $y^* \in \Lambda_{FJ}$ we have

$$\left\langle \frac{y^*}{\|y\|^*}, -y \right\rangle = \left\langle \frac{y^*}{\|y\|^*}, h(\bar{x}) - y \right\rangle - \sigma_C\left(\frac{y^*}{\|y\|^*}\right) - \left\| h'(\bar{x})^* \frac{y^*}{\|y\|^*} \right\| \leq -\rho,$$

implying $\langle y^*, y\rangle \geq \rho\|y^*\|$. Hence, Λ_{FJ}^y is also nonempty and Λ_{FJ} is generated by Λ_{FJ}^y. Moreover, for all $y^* \in \Lambda_{FJ}^y \subset \Lambda_{FJ}$ we have $1 = \langle y^*, y\rangle \geq \rho\|y^*\|$ showing that Λ_{FJ}^y is bounded. Hence, together with the Alaoglu–Bourbaki theorem, the weak* compactness follows. ∎

For the sake of completeness we state the following corollary (see [34, Thm. 4]).

COROLLARY 4.4. *Let $\Lambda_{FJ} \neq \emptyset$ and assume that Robinson's constraint qualification $0 \in \operatorname{int}(g(\bar{x}) + g'(\bar{x})X - K)$ is fulfilled. Then one has $\Lambda_{FJ} = \Lambda_{KT}$.*

5. Simplification of the second-order conditions. Needless to say, the second-order optimality conditions given by Corollary 3.4 can be verified in practice only with some great effort. We derive now second-order conditions which are easier to handle. However, we restrict ourselves to those conditions which are equivalent to those of Corollary 3.4, at least under some additional assumptions on the problem. We devote this section also to comparisons with other works.

Throughout this section we assume that Assumptions 1 and 2 are fulfilled. It follows that $h''(\bar{x};\cdot)$ is Lipschitzian on bounded sets. Indeed, let $z_1, z_2 \in X$ be fixed. For any $\epsilon > 0$ there exists some $\delta > 0$ such that for all $t \in (0,\delta)$ one has

$$\left\| \frac{h(\bar{x}+tz_i) - h(\bar{x}) - th'(\bar{x})z_i}{t^2/2} - h''(\bar{x};z_i) \right\| \le \epsilon, \quad i = 1,2.$$

Hence, using Assumption 1, it follows for all $t > 0$ sufficiently small that

$$\begin{aligned} \|h''(\bar{x};z_1) - h''(\bar{x};z_2)\| &\le \left\| \frac{h(\bar{x}+tz_1) - h(\bar{x}+tz_2) - th'(\bar{x})(z_1-z_2)}{t^2/2} \right\| + 2\epsilon \\ &\le 2\eta \max\{\|z_1\|, \|z_2\|\}\|z_1 - z_2\| + 2\epsilon. \end{aligned}$$

Since ϵ was arbitrary, the claimed Lipschitz continuity of $h''(\bar{x};\cdot)$ on bounded sets follows.

LEMMA 5.1. *Let $(z_n, t_n, \tau_n) \subset S_X \times \mathbb{R}_+ \times \mathbb{R}_+$ be a sequence such that $t_n \to 0_+$, $\tau_n \to 0_+$. If there is some $\beta > 0$ such that $d(h(\bar{x}) + t_n h'(\bar{x})z_n, C) \ge \beta t_n$ for all n, then $\liminf_{n\to\infty} t_n^{-2} \hat{d}_C(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{t_n^2}{2}h''(\bar{x};z_n), h'(\bar{x}), \tau_n t_n) = \infty$.*

Proof. For each n let $y_n^* \in S_{Y^*}$ be with $\langle y_n^*, h(\bar{x}) + t_n h'(\bar{x})z_n\rangle - \sigma_C(y_n^*) = d(h(\bar{x}) + t_n h'(\bar{x})z_n, C)$ according to the minimum norm duality theorem. Then, by the definition of $\hat{d}_C$ we have $t_n^{-2}\hat{d}_C(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{t_n^2}{2}h''(\bar{x};z_n), h'(\bar{x}), \tau_n t_n) \ge \frac{1}{2}\langle y_n^*, h''(\bar{x};z_n)\rangle + t_n^{-1}(\beta - \tau_n \|h'(\bar{x})^* y_n^*\|) \to \infty$. ∎

Therefore, in order to verify the second-order conditions (3.5) we need only consider sequences (z_n, t_n, τ_n) with $d(h(\bar{x}) + t_n h'(\bar{x})z_n, C) = o(t_n)$, which can also be written as $d(h'(\bar{x})z_n, T_C(h(\bar{x}))) \to 0$.

LEMMA 5.2. *If $\bar{x}$ is not a critical point (respectively, essential local minimizer) of second order, then there are a scalar $\rho > 0$ (respectively, $\rho \ge 0$) and a sequence $(z_n, t_n, \tau_n) \subset S_X \times \mathbb{R}_+ \times \mathbb{R}_+$ with $t_n \to 0_+$, $\tau_n \to 0_+$ such that $d(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{t_n^2}{2}h''(\bar{x};z_n), C) = o(t_n^2)$ and*

$$\liminf_{n\to\infty} t_n^{-2}\hat{d}_C\left(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{t_n^2}{2}h''(\bar{x};z_n), h'(\bar{x}), C\right) \le -\rho.$$

Proof. If $\bar{x}$ is not a critical point of second order, consider the sequences $(x_n) \to \bar{x}$, $(\tau_n) \to 0_+$, (κ_n), $(\tilde{x}_n)$ and the scalars $\beta > 0$, η used in the proof of Theorem 3.2. By using the last assertion of Proposition 2.10 we also obtain

$$\begin{aligned} \hat{d}_C(h(\tilde{x}_n), h'(\bar{x}), 3\kappa_n) &\le \hat{d}_C(h(\tilde{x}_n), h'(\bar{x}), \kappa_n + \|\tilde{x}_n - x_n\|) \\ &\le -\beta\|x_n - \bar{x}\|^2 + 2\eta(\|x_n - \bar{x}\| + \kappa_n)\kappa_n \le -\frac{3}{4}\beta\|x_n - \bar{x}\|^2. \end{aligned} \tag{5.1}$$

By Assumption 2 the assertion follows with $\tilde{t}_n := \|\tilde{x}_n - \bar{x}\|$, $\tilde{z}_n := t_n^{-1}(\tilde{x}_n - \bar{x})$, $\tilde{\tau}_n := 3\kappa_n/t_n$, and $\rho = 3\beta/4$.

The assertion about essential local minimizers follows easily from Definition 3.1. ∎

Next we consider the optimality conditions (3.5) for sequences $(z_n, t_n, \tau_n) \subset S_X \times \mathbb{R}_+ \times \mathbb{R}_+$, where (z_n) is convergent to some $z \in S_X$. By Lemma 5.1 we have only to consider the case $z \in \mathcal{C}(\bar{x})$, where the set

$$\mathcal{C}(\bar{x}) := \{z \in X : h'(\bar{x})z \in T_C(h(\bar{x}))\}$$

is called the *critical cone* of problem (P) at $\bar{x}$. We also need a tool to describe the possible curvature of a set. As a consequence of Lemma 5.2, the use of the *second-order compound tangent set* introduced by Penot [30] seems to be an appropriate approach.

DEFINITION 5.3. *Let E, F be normed spaces, let D be a subset of F, let $A \in L(E,F)$ be a continuous linear operator, and let $\bar{f} \in D$. Then for $z \in E$ the* second-order compound tangent set *to D at $(\bar{f}, z)$ (with respect to A) is the set*

$$D''_A(\bar{f}, z) := \{w \in F : \exists z_n \to z,\ t_n \to 0_+ :\ d(\bar{f} + t_n A z_n + t_n^2 w/2, D) = o(t_n^2)\}.$$

THEOREM 5.4. *If $\bar{x}$ is an essential local minimizer (respectively, nondegenerate critical point) of second order for problem* (P)*, then there exists a scalar $\beta > 0$ (respectively, $\beta \geq 0$) such that for each $z \in \mathcal{C}(\bar{x}) \cap S_X$ either we have $C''_{h'(\bar{x})}(h(\bar{x}), z) = \emptyset$ or for each $w \in C''_{h'(\bar{x})}(h(\bar{x}), z)$ there exists a multiplier $y^* \in \Lambda_{FJ} \cap B_{Y^*}$ with*

$$\langle y^*, h''(\bar{x}; z) - w\rangle \geq \beta. \tag{5.2}$$

Proof. Assume that $\bar{x}$ is a critical point (respectively, essential local minimizer) of second order, let $\beta \geq 0$ (respectively, $\beta > 0$) be as in Corollary 3.4, and let $z \in \mathcal{C}(\bar{x}) \cap S_X$ be arbitrarily chosen. In the case of $C''_{h'(\bar{x})}(h(\bar{x}), z) = \emptyset$ there is nothing to prove. Thus, let $w \in C''_{h'(\bar{x})}(h(\bar{x}), z)$. By the definition of $C''_{h'(\bar{x})}(h(\bar{x}), z)$, there is a sequence (z_n, t_n, w_n) with limit $(z, 0_+, w)$ such that $h(\bar{x}) + t_n h'(\bar{x}) z_n + \frac{1}{2} t_n^2 w_n \in C$. Since $z \in S_X$ and $z_n \to z$, we may also assume that $(z_n) \subset S_X$. By Corollary 3.4, taking $\tau_n = \sqrt{t_n}$, there exists a sequence $(y_n^*) \subset S_{Y^*}$, approaching the supremum in (3.5) sufficiently accurately, such that

$$\begin{aligned} \liminf_{n\to\infty} \Bigg\{ & \frac{\langle y_n^*, h(\bar{x}) + t_n h'(\bar{x}) z_n\rangle - \sigma_C(y_n^*)}{t_n^2} \\ & + \frac{1}{2}\langle y_n^*, h''(\bar{x}; z_n)\rangle - \frac{\|h'(\bar{x})^* y_n^*\|}{\sqrt{t_n}} \Bigg\} \geq \beta. \end{aligned} \tag{5.3}$$

Since $h''(\bar{x}, \cdot)$ is continuous, $\langle y_n^*, h''(\bar{x}; z_n)\rangle$ is uniformly bounded. Therefore, we have $\langle y_n^*, h(\bar{x}) + t_n h'(\bar{x}) z_n\rangle - \sigma_C(y_n^*) \geq (\beta - k) t_n^2$ for some constant $k > 0$ and all n sufficiently large, and together with $\langle y_n^*, h(\bar{x})\rangle \leq \sigma_C(y_n^*)$ we obtain $\langle y_n^*, h(\bar{x})\rangle - \sigma_C(y_n^*) \to 0$. We also have $\langle y_n^*, h(\bar{x}) + t_n h'(\bar{x}) z_n + \frac{1}{2} t_n^2 w_n\rangle \leq \sigma_C(y_n^*)$, and by (5.3) we obtain

$$\liminf_{n\to\infty} \left\{ \frac{1}{2}\langle y_n^*, -w_n + h''(\bar{x}; z_n)\rangle - \frac{1}{\sqrt{t_n}} \|h'(\bar{x})^* y_n^*\| \right\} \geq \beta.$$

Hence, $h'(\bar{x})^* y_n^* \to 0$ and $\liminf_{n\to\infty} \langle y_n^*, h''(\bar{x}; z) - w\rangle \geq \beta$. By the Alaoglu–Bourbaki theorem, the sequence (y_n^*) has at least one weak* accumulation point $y^* \in B_{Y^*}$. If $\bar{x}$ is nondegenerate, every accumulation point y^* belongs to Λ_{FJ} by Proposition 4.2 and also fulfills (5.2). Hence, the assertion about nondegenerate critical points is proved. If $\bar{x}$ is an essential local minimizer, it follows that every weak* accumulation point y^* of (y_n^*) satisfies $\langle y^*, h''(\bar{x}; z) - w\rangle \geq \beta > 0$, showing $y^* \neq 0$ and hence $y^* \in \Lambda_{FJ}$. This completes the proof of the theorem. □

Note that the case $C''_{h'(\bar{x})}(h(\bar{x}), z) = \emptyset$ may be treated as a kind of second-order condition where no multipliers are involved.

COROLLARY 5.5. *Let $\bar{x}$ be an essential local minimizer (respectively, nondegenerate critical point) of second order for* (P)*. Then there exists a scalar $\beta > 0$ (respectively, $\beta \geq 0$) such that for every $z \in \mathcal{C}(\bar{x}) \cap S_X$ satisfying $C''_{h'(\bar{x})}(h(\bar{x}), z) \neq \emptyset$ and*

every convex subset W of $C''_{h'(\bar{x})}(h(\bar{x}),z)$ there exists a multiplier $y^* \in \Lambda_{FJ}$ such that $\langle y^*, h''(\bar{x};z)\rangle - \sigma_W(y^*) \geq \beta$.

Proof. Let $\tilde{\Lambda} := (\Lambda_{FJ} \cup \{0\}) \cap B_{Y^*}$ (respectively, $\tilde{\Lambda} := \Lambda^y_{FJ}$ with $y \in \operatorname{int}(h(\bar{x}) + h'(\bar{x})B_X - C)$). Then, by Theorem 5.4 we have $\inf_{w\in W} \sup_{y^*\in\tilde{\Lambda}} \langle y^*, h''(\bar{x};z) - w\rangle \geq \beta$. Since $\tilde{\Lambda}$ is weakly* compact, the assertion follows from the classical minsup theorem of Moreau. ∎

Theorem 5.4 and Corollary 5.5 generalize the necessary optimality conditions [30] to vector optimization problems. Moreover, the assumption on metric regularity of the constraints used in [30] is replaced with condition (3.9). As a consequence, we have to formulate the necessary conditions in terms of Fritz John multipliers $y^* \in \Lambda_{FJ}$ instead of multipliers $y^* \in \Lambda_{KT}$.

THEOREM 5.6. *Let X be finite dimensional. Suppose that for each $z \in \mathcal{C}(\bar{x}) \cap S_X$ satisfying $C''_{h'(\bar{x})}(h(\bar{x}),z) \neq \emptyset$ there is a scalar $\beta_z \geq 0$ such that for each $w \in C''_{h'(\bar{x})}(h(\bar{x}),z)$ there exists a multiplier $y^* \in \Lambda_{FJ}$ with*

$$\langle y^*, h''(\bar{x};z) - w\rangle \geq \beta_z. \tag{5.4}$$

Then $\bar{x}$ is a critical point of second order for (P)*. If $\beta_z > 0$ for all $z \in \mathcal{C}(\bar{x}) \cap S_X$ satisfying $C''_{h'(\bar{x})}(h(\bar{x}),z) \neq \emptyset$, then $\bar{x}$ is even an essential local minimizer of second order.*

Proof. The proof is by contraposition. Suppose that $\bar{x}$ is not an essential local minimizer of second order. Let the scalar $\rho \geq 0$ and the sequence $(z_n, t_n, \tau_n) \subset S_X \times \mathbb{R}_+ \times \mathbb{R}_+$ be chosen so that the conclusion of Lemma 5.2 holds. Since X is finite dimensional we may assume that (z_n) converges to some $z \in S_X$, where we have eventually passed to a subsequence. Since $d(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{1}{2}t_n^2 h''(\bar{x},z_n), C) = o(t_n^2)$ we have $z \in \mathcal{C}(\bar{x})$ and $h''(\bar{x};z) \in C''_{h'(\bar{x})}(h(\bar{x}),z)$. By taking $w = h''(\bar{x};z)$ we see that (5.4) cannot hold with $\beta_z > 0$, which is a contradiction. Now assume that $\bar{x}$ is not even a critical point of second order. Then $\rho > 0$, and we may also assume that $\hat{d}_C(h(\bar{x}) + t_n h'(\bar{x})z_n + (1/2)t_n^2 h''(\bar{x};z_n), h'(\bar{x}), \tau_n t_n) < -\rho t_n^2/2$ for all n. Then, for each $w \in \frac{\rho}{2}B_Y$ and each n there is some $\xi_{n,w} \in B_X$ with $h(\bar{x}) + t_n h'(\bar{x})(z_n + \tau_n\xi_{n,w}) + \frac{1}{2}t_n^2(h''(\bar{x};z_n) + w) \in C$. Thus, $h''(\bar{x};z) + w \in C''_{h'(\bar{x})}(h(\bar{x}),z)$. Let $\epsilon > 0$ be chosen so small such that $w := -\epsilon h''(\bar{x};z) \in \frac{\rho}{2}B_Y$. By our hypothesis there is some $\bar{y}^* \in \Lambda_{FJ}$ such that $\langle \bar{y}^*, -w\rangle = \epsilon\langle \bar{y}^*, h''(\bar{x};z)\rangle \geq 0$. Since $\bar{y}^* \neq 0$ and Λ_{FJ} is a cone we may assume that $\bar{y}^* \in \Lambda_{FJ} \cap S_{Y^*}$. Since $\langle \bar{y}^*, h(\bar{x})\rangle - \sigma(\bar{y}^*) = \|h'(\bar{x})^*\bar{y}^*\| = 0$ we obtain $\langle \bar{y}^*, h''(\bar{x};z_n)\rangle = 2t_n^{-2}(\langle \bar{y}^*, h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{1}{2}t_n^2 h''(\bar{x};z_n)\rangle - \sigma_C(\bar{y}^*) - \tau_n t_n \|h'(\bar{x})^*\bar{y}^*\|) < -\rho$ for all n. Passing to the limit, we obtain $\langle \bar{y}^*, h''(\bar{x};z)\rangle \leq -\rho < 0$, which clearly contradicts $\epsilon\langle \bar{y}^*, h''(\bar{x};z)\rangle \geq 0$. ∎

Theorem 5.6 generalizes the sufficient conditions of [31, Prop. 5.1] to vector optimization problems. Moreover, we do not assume under any circumstances that the set of Fritz John multipliers Λ_{FJ} is nonempty: Theorem 5.6 states that $\bar{x}$ is an essential local minimizer of second order if $C''_{h'(\bar{x})}(h(\bar{x}),z) = \emptyset$ for every $z \in \mathcal{C}(\bar{x}) \cap S_X$ and X is finite dimensional, even if $\Lambda_{FJ} = \emptyset$. On the other hand, from Theorem 5.4 it follows that at an essential local minimizer of second order, the set Λ_{FJ} of multipliers is not empty if there is some $z \in \mathcal{C}(\bar{x}) \cap S_X$ with $C''_{h'(\bar{x})}(h(\bar{x}),z) \neq \emptyset$.

COROLLARY 5.7. *Let X be finite dimensional and assume that the set of Fritz John multipliers Λ_{FJ} at $\bar{x}$ is empty $\Lambda_{FJ} = \emptyset$. Then the element $\bar{x}$ is an essential local minimizer of second order for problem* (P) *if and only if for all $z \in \mathcal{C}(\bar{x}) \cap S_X$ one has $C''_{h'(\bar{x})}(h(\bar{x}),z) = \emptyset$.*

The following example illustrates the situation of Corollary 5.7.

Example 5.1. Let $X = \mathbb{R}$ and choose a closed convex set $K \subset V$ of an infinite-dimensional space V such that K does not have a supporting functional at some boundary point $\bar{v} \in \operatorname{bd} K$, i.e., $N_K(\bar{v}) = \{0\}$. Let $s \in V$ be a direction such that $\bar{v} + xs \in K$ if $x \geq 0$ and $\bar{v} + xs \notin K$ if $x < 0$. Consider the scalar optimization problem

$$\min_{x \in \mathbb{R}} f(x) := x \quad \text{subject to} \quad g(x) := \bar{v} + xs \in K.$$

Clearly, $\bar{x} = 0$ is a strictly local minimizer; however, the set of multipliers Λ_{FJ} is empty. Indeed, the Fritz John conditions read as

$$u^* + \langle v^*, s \rangle = 0, \quad u^* \geq 0, \quad v^* \in N_K(\bar{v})$$

and can be fulfilled only by the trivial solution $(u^*, v^*) = 0$. Since $N_K(\bar{v}) = \{0\}$, we have $T_K(\bar{v}) = V$. Hence, $-s \in T_K(\bar{v})$ and therefore $\mathcal{C}(0) = \mathbb{R}_-$. We obtain from the definition that $\bar{x} = 0$ is an essential local minimizer of second order if and only if s is so that $\liminf_{x \to 0-} x^{-2} d(g(x), K) > 0$. By Corollary 5.7, this is also equivalent to the condition that the second-order compound tangent set of $C = \mathbb{R}_- \times K$ at $(0, \bar{v})$ in direction -1 is empty. To be more concrete, let $V = L_1([0,1])$ be the Banach space of measurable, absolutely integrable functions on $[0,1]$, $K = \{v \in L_1([0,1]) : v(\xi) \geq 0 \text{ a.e.}\}$, $\bar{v}(\xi) = \xi$, $s(\xi) = 1$. Then $g(x)(\xi) = \xi + x$ and we obtain $d(g(x), K) = \int_0^1 -\min\{\xi + x, 0\} d\xi = \frac{1}{2}x^2$ for $x < 0$. Hence, $\bar{x} = 0$ is an essential local minimizer in this case. Now let us show that $C''_{h'(\bar{x})}(h(\bar{x}), -1) = \emptyset$. To do this we show that the second-order compound tangent set $K''_s(\bar{v}, -1)$ is empty. Assume that there is a sequence $(z_n, t_n, v_n) \in \mathbb{R} \times \mathbb{R}^+ \times L_1([0,1])$ with limit $(-1, 0_+, v)$ such that $\bar{v} + t_n z_n s + \frac{1}{2} t_n^2 v_n \in K$. It follows that $v_n(\xi) \geq -2t_n^{-2}(\xi + t_n z_n)$ for almost all $\xi \in [0, -t_n z_n]$ and therefore $\int_0^{-t_n z_n} v_n(\xi) d\xi \geq z_n^2 \to 1$. On the other hand, we have $\int_0^{-t_n z_n} v(\xi) d\xi \to 0$, and therefore $v_n \not\to v$, which is a contradiction. This shows that the compound tangent set $K''_s(\bar{v}, -1)$ (and hence also $C''_{h'(\bar{x})}(h(\bar{x}), -1)$) is empty.

The computation of second-order compound tangent sets is a rather complex task. In some situations the compound tangent set can be equivalently replaced with another set which is easier to handle. Let us consider the following notion of outer second-order regularity as introduced in [6].

DEFINITION 5.8. *Let D be a subset of the normed space F and let $\bar{f} \in D$, $s \in F$.*

1. *The* outer second-order tangent set *to D at $\bar{f}$ in direction s is the set*

$$T_D^2(\bar{f}, s) := \{r \in F : \exists t_n \to 0_+ \text{ such that } d(\bar{f} + t_n s + t_n^2 r/2, D) = o(t_n^2)\}.$$

2. *Let E be another normed space and let $A \in L(E, F)$ be a continuous linear mapping. We say that a set $\mathcal{A}_{D,A}(\bar{f}, s)$ is an* upper second-order approximation set *for D at the point $\bar{f}$ in direction s with respect to A if for any sequence $(f_n) \subset D$ of the form $f_n = \bar{f} + t_n s + \frac{1}{2} t_n^2 r_n$, where $t_n \to 0_+$ and $r_n = Ae_n + w_n$ with (w_n) being a convergent sequence in F and (e_n) being a sequence in E satisfying $t_n e_n \to 0$, the following condition holds:*

$$\lim_{n \to \infty} d(r_n, \mathcal{A}_{D,A}(\bar{f}, s)) = 0.$$

We say that D is outer second-order regular *at the point $\bar{f}$ in direction s with respect to A if $T_D^2(\bar{f}, s)$ is an upper second-order approximation set at the point $\bar{f}$ in direction s with respect to A.*

Obviously, $T_D^2(\bar f, Az) \subset D_A''(\bar f, z)$ always hold. Hence, Theorem 5.4 remains true if we replace $C''_{h'(\bar x)}(h(\bar x), z)$ with the smaller set $T_C^2(h(\bar x), h'(\bar x)z)$. Further, $D_A''(\bar f, z)$ is always an upper second-order approximation set for D at $\bar f$ in direction Az with respect to A. Indeed, let $f_n = \bar f + t_n Az + \frac{1}{2}t_n^2 r_n \in D$ be a sequence, where $t_n \to 0_+$, $r_n = Ae_n + w_n$ with $w_n \to w \in F$, and $t_n e_n \to 0$. Then, since $f_n = \bar f + t_n A(z + \frac{1}{2}t_n e_n) + \frac{1}{2}t_n^2 w_n$, we obtain $w \in D_A''(\bar f, z)$. Hence, $d(r_n, D_A''(\bar f, z) + \mathrm{Range}(A)) \to 0$. Noting that $D_A''(\bar f, z) + \mathrm{Range}(A) = D_A''(\bar f, z)$, we see that $D_A''(\bar f, z)$ is an upper second-order approximation set.

Using similar arguments we can also show that any upper second-order approximation set $\mathcal{A}_{D,A}(\bar f, Az)$ satisfies $D_A''(\bar f, z) \subset \mathrm{cl}\,(\mathcal{A}_{D,A}(\bar f, Az) + \mathrm{Range}(A))$.

Use of the concepts of outer second-order tangent sets and upper second-order approximation sets has the advantage that the constraints and the objective can be treated separately.

THEOREM 5.9.

1. *Let $\bar x$ be an essential local minimizer (respectively, nondegenerate critical point) of second order for problem* (P). *Then there is a real $\beta > 0$ (respectively, $\beta \ge 0$) such that for every critical direction $z \in \mathcal{C}(\bar x) \cap S_X$ satisfying $T_K^2(g(\bar x), g'(\bar x)z) \ne \emptyset$ and for every $v \in T_K^2(g(\bar x), g'(\bar x)z)$ there exists some multiplier $y^* = (u^*, v^*) \in \Lambda_{FJ} \cap B_{Y^*}$ such that*

$$\langle u^*, f''(\bar x; z)\rangle + \langle v^*, g''(\bar x; z) - v\rangle \ge \beta.$$

2. *Assume that X is finite dimensional. To every $z \in \mathcal{C}(\bar x) \cap S_X$ let there correspond an upper second-order approximation set $\mathcal{A}_K(z) := \mathcal{A}_{K,g'(\bar x)}(g(\bar x), g'(\bar x)z)$ for the set K at the point $g(\bar x)$ in the direction $g'(\bar x)z$ with respect to $g'(\bar x)$. Assume that for each $z \in \mathcal{C}(\bar x) \cap S_X$ there is some $\beta_z \ge 0$ such that for each $v \in \mathcal{A}_K(z)$ there exists some $y^* = (u^*, v^*) \in \Lambda_{FJ} \cap B_{Y^*}$ with*

$$\langle u^*, f''(\bar x; z)\rangle + \langle v^*, g''(\bar x; z) - v\rangle \ge \beta_z. \tag{5.5}$$

Then, $\bar x$ is a critical point of second order. In addition, if $\beta_z > 0$ for every $z \in \mathcal{C}(\bar x) \cap S_X$, then $\bar x$ is even an essential local minimizer of second order.

Proof. Since for every $v \in T_K^2(g(\bar x), g'(\bar x)z)$ we have $(0, v) \in T_C^2(h(\bar x), h'(\bar x)z) \subset C''_{h'(\bar x)}(h(\bar x), z)$, the first part follows immediately from Theorem 5.4. To show the second part, we use Theorem 5.6. Let $z \in \mathcal{C}(\bar x) \cap S_X$, satisfying $C''_{h'(\bar x)}(h(\bar x), z) \ne \emptyset$ and $w \in C''_{h'(\bar x)}(h(\bar x), z)$, be arbitrarily fixed. By the definition of compound tangent sets there is a sequence (z_n, t_n, w_n) with limit $(z, 0_+, w)$ such that $h(\bar x) + t_n h'(\bar x)z_n + \frac{1}{2}t_n^2 w_n \in C$. Setting $w_n = (u_n, v_n)$ we have $g(\bar x) + t_n g'(\bar x)z + \frac{1}{2}t_n^2(2t_n^{-1}g'(\bar x)(z_n - z) + v_n) \in K$. Hence, there is a sequence $(\tilde v_n) \subset \mathcal{A}_K(z)$ with $\nu_n := 2t_n^{-1}g'(\bar x)(z_n - z) + v_n - \tilde v_n \to 0$. Now let $y_n^* := (u_n^*, v_n^*) \in \Lambda_{FJ} \cap B_{Y^*}$ with $\langle u_n^*, f''(\bar x; z)\rangle + \langle v_n^*, g''(\bar x; z) - \tilde v_n\rangle \ge \beta_z$. Since Λ_{FJ} is a cone we may assume $y_n^* \in S_{Y^*}$. Because of $t_n f'(\bar x)z + \frac{1}{2}t_n^2(2t_n^{-1}f'(\bar x)(z_n - z) + u_n) \in -L$, $u_n^* \in L^*$, and because L is a cone, we have $\langle u_n^*, 2t_n^{-1}f'(\bar x)(z_n - z) + u_n\rangle \le -2t_n^{-1}\langle u_n^*, f'(\bar x)z\rangle$. Together with $h'(\bar x)^* y_n^* = 0$ we obtain

$$\begin{aligned}\langle y_n^*, w_n\rangle &= \langle y_n^*, 2t_n^{-1}h'(\bar x)(z_n - z) + w_n\rangle \\ &= \langle u_n^*, 2t_n^{-1}f'(\bar x)(z_n - z) + u_n\rangle + \langle v_n^*, \tilde v_n + \nu_n\rangle \\ &\le -2t_n^{-1}\langle u_n^*, f'(\bar x)z\rangle - \beta_z + \langle u_n^*, f''(\bar x; z)\rangle + \langle v_n^*, g''(\bar x; z)\rangle + \langle v_n^*, \nu_n\rangle.\end{aligned}$$

Rearranging yields $\langle y_n^*, h''(\bar x; z) - w_n\rangle \ge \beta_z + 2t_n^{-1}\langle u_n^*, f'(\bar x)z\rangle - \langle v_n^*, \nu_n\rangle$. Since $z \in \mathcal{C}(\bar x)$ we have $(f'(\bar x)z, g'(\bar x)z) \in T_{-L}(0) \times T_K(g(\bar x))$ and, because of $(u_n^*, v_n^*) \in L^* \times$

$N_K(g(\bar{x}))$ we obtain $\langle u_n^*, f'(\bar{x})z\rangle \le 0$ and $\langle v_n^*, g'(\bar{x})z\rangle \le 0$. Thus, $0 = \langle y_n^*, h'(\bar{x})z\rangle = \langle u_n^*, f'(\bar{x})z\rangle + \langle v_n^*, g'(\bar{x})z\rangle \le 0$ and consequently $\langle u_n^*, f'(\bar{x})z\rangle = \langle v_n^*, g'(\bar{x})z\rangle = 0$. Together with $(w_n, \nu_n) \to (w, 0)$ we obtain $\liminf_{n\to\infty}\langle y_n^*, h''(\bar{x}; z) - w\rangle \ge \beta_z$. By the Alaoglu–Bourbaki theorem the sequence y_n^* has a weak* accumulation point y^*, and by definition of the weak* topology, we also have $\langle y^*, h''(\bar{x}; z) - w\rangle \ge \beta_z$. If $\beta_z > 0$, we obviously have $y^* \ne 0$ and thus $y^* \in \Lambda_{FJ}$. If $\beta_z = 0$, and $\bar{x}$ is nondegenerate, then we have $y^* \in \Lambda_{FJ}$ by Proposition 4.2. Finally, if $\bar{x}$ is degenerate, then $\bar{x}$ is always a critical point of second order for problem (P). Now the assertion follows from Theorem 5.6. □

For ease of presentation we have not included in the sufficient conditions of Theorem 5.9 the case that for some directions $z \in \mathcal{C}(\bar{x}) \cap S_X$ no upper second-order approximation set $\mathcal{A}_K(z) := \mathcal{A}_{K,g'(\bar{x})}(g(\bar{x}), g'(\bar{x})z)$ whatsoever exists. We also dispense with formulating a corollary analogous to Corollary 5.5. Theorem 5.9 improves the results of [6] even for scalar optimization problems. Our necessary conditions for a local minimum require only $\bar{x}$ to be nondegenerate, whereas Robinson's constraint qualification (3.13) is used in [6]. However, second-order necessary conditions are stated in [7, Thm. 3.50] for a special case of nondegenerate minimizers $\bar{x}$ without presuming Robinson's constraint qualification, namely when the constraints can be written in the form $g(x) = (\tilde{g}(x), x)$, $K = \tilde{K} \times Q$, where $\operatorname{int}\tilde{K} \ne \emptyset$, i.e., general inequality constraints together with explicit constraints.

The second-order sufficient conditions in [6] state that if for each $z \in \mathcal{C}(\bar{x}) \cap S_X$ there exists an upper second-order approximation set $\mathcal{A}_K(z) := \mathcal{A}_{K,g'(\bar{x})}(g(\bar{x}), g'(\bar{x})z)$, and if there exists $y^* = (u^*, v^*) \in \Lambda_{FJ}$ such that

$$\langle u^*, f''(\bar{x}; z)\rangle + \langle v^*, g''(\bar{x}; z)\rangle > \sigma_{\mathcal{A}_K(z)}(v^*), \tag{5.6}$$

then $\bar{x}$ is a strict locally optimal solution fulfilling the quadratic growth condition (3.2). Note that (5.6) always implies (5.5), and condition (5.5) is actually weaker for nonconvex approximation sets $\mathcal{A}_K(z)$.

Example 5.2. Let $\mathbb{S}^m$ denote the Hilbert space of symmetric $m \times m$ matrices equipped with the scalar product $\langle x, y\rangle = \operatorname{trace}(xy)$, and denote by $\mathbb{S}^m_+ \subset \mathbb{S}^m$ the cone of positive semidefinite $m \times m$ matrices. Consider for arbitrarily fixed γ the scalar problem

$$\begin{aligned} &\min_{x\in\mathbb{S}^2} f(x) := \frac{1}{2}(x_{11}^2 - x_{12}^2) \\ \text{subject to}\quad &g_1(x) := \frac{1}{2}(\det x + \langle \bar{s}, x\rangle) + \gamma(x_{12} - 1)^2 = 0, \\ &x \in \mathbb{S}^2_+, \end{aligned}$$

at $\bar{x} := \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, where $\bar{s} := \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$. In terms of problem (P) we have $g(x) = (g_1(x), x)$ and $K = \{0\} \times \mathbb{S}^2_+$. There holds $g_1'(\bar{x}) = \bar{s} \ne 0$, and together with $\operatorname{int}\mathbb{S}^2_+ \ne \emptyset$ it easily follows that $\bar{x}$ is nondegenerate. Now one can show that $N_{\mathbb{S}^2_+}(\bar{x}) = \{t\bar{s} : t \le 0\}$, and hence we have $g'(\bar{x})^* v^* = 0$ with $v^* := (1, -\bar{s}) \in N_K(0, \bar{x})$, and it follows that Robinson's constraint qualification (3.13) is not fulfilled. It also follows that the second-order qualification condition (TR) used in [22], [31], $g'(\bar{x})X - T_{T_K(g(\bar{x}))}(g'(\bar{x})z) = V$, is not fulfilled for any $z \in \mathcal{C}(\bar{x})$. Some straightforward calculations show that $\bar{x}$ is the only feasible point, and hence is a strict minimizer when $\gamma > 0$. When $\gamma = 0$ the feasible set amounts to the ray $t\bar{x}, t \ge 0$, and hence every feasible point is a local nonstrict minimizer. Finally, when $\gamma < 0$, $\bar{x}$ is no longer a local minimizer. Now let us consider

the second-order conditions of Theorem 5.9. It is possible to show (see section 5.3.1 in [7]) that the critical cone evaluates as $\mathcal{C}(\bar{x}) = \{z \in \mathbb{S}^2 : z_{11} \le z_{12},\ \langle \bar{s}, z\rangle = 0\}$ and that $T^2_K((0,\bar{x}),(0,z)) = \{(0,v) : \langle \bar{s}, v\rangle + 2\det z \ge 0\}$ holds for all $z \in \mathcal{C}(\bar{x})$. The set Λ_{FJ} of multipliers is given by all $y^* := (\alpha,\mu,\eta) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{S}^2 \neq (0,0,0)$ so that

$$\alpha f'(\bar{x}) + \mu g_1'(\bar{x}) + \eta = \alpha \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix} + \mu\bar{s} + \eta = 0, \quad \eta \in N_{\mathbb{S}^2_+(\bar{x})} = \{t\bar{s} : t \le 0\}.$$

Thus, $\Lambda_{FJ} = \{(0,\mu,-\mu\bar{s}) : \mu > 0\}$. Now, for arbitrary $z \in \mathcal{C}(\bar{x})$, $v \in T^2_K((0,\bar{x}),(0,z))$, and $y^* = (0,\mu,-\mu\bar{s})$, where $\mu > 0$, the left-hand side of relation (5.5) amounts to

$$\kappa(z,v,\mu) := \mu(2\gamma z_{12}^2 + \det z + \langle \bar{s}, v\rangle).$$

By setting $z := \frac{1}{\sqrt{6}}\left(\begin{smallmatrix} 0 & 1 \\ 1 & 2 \end{smallmatrix}\right) \in \mathcal{C}(\bar{x}) \cap S_X$, $v := \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) \in T^2_K((0,\bar{x}),(0,z))$ we obtain $\kappa(z,v,\mu) = \mu\gamma/3$, and from the first part of Theorem 5.9 we conclude that $\bar{x}$ is not a critical point of second order, and therefore also not a local minimizer when $\gamma < 0$. Note that the second-order necessary conditions known from the literature cannot be applied to this problem. In the case of $\gamma = 0$ it follows that the point $\bar{x}$ is not an essential local minimizer of second order.

Now let us turn to sufficient conditions. For $z \in \mathcal{C}(\bar{x})$ we have $z_{11} \le z_{12} \le z_{22}$, $0 \le z_{12} - z_{11} = z_{22} - z_{12} =: \epsilon$, and $0 \le \epsilon^2 = (z_{12} - z_{11})(z_{22} - z_{12}) = -\det z - 2z_{12}^2 + z_{12}(z_{11} + z_{22}) = -\det z$. Thus, $\|z\|^2 = z_{11}^2 + 2z_{12}^2 + z_{22}^2 = 4z_{12}^2 + 2\epsilon^2 = 4z_{12}^2 - 2\det z$, and since $\langle \bar{s}, v\rangle \ge -2\det z$ holds for $v \in T^2_K((0,\bar{x}),(0,z))$ we obtain

$$\kappa(z,v,\mu) \ge \mu(2\gamma z_{12}^2 - \det z) \ge \frac{1}{2}\mu \min\{\gamma, 1\}\|z\|^2$$

for $\gamma \ge 0$. Using [7, Prop. 3.89] and [6, Cor. 4.6] we obtain that $K = \{0\} \times \mathbb{S}^2_+$ is outer second-order regular at $(0,\bar{x})$ in any direction $(0,z)$, $z \in T_{\mathbb{S}^2_+}(\bar{x})$, and with respect to any mapping A. By using the second part of Theorem 5.9 with $\mathcal{A}_K(z) = T^2_K((0,\bar{x}),(0,z))$ we obtain that $\bar{x}$ is an essential local minimizer (respectively, critical point) of second order when $\gamma > 0$ (respectively, $\gamma \ge 0$). Note that for $\gamma = 0$ the point $\bar{x}$ is a nondegenerate critical point but not an essential local minimizer of second order. Hence, from Theorems 3.5 and 3.6 it follows that, on the one hand, there exist twice continuously differentiable perturbations of the objective and the constraints with vanishing function values and first- and second-order derivatives such that $\bar{x}$ is a strict local minimum for the perturbed problem, and on the other hand, there exist perturbations of this kind such that $\bar{x}$ is not a local minimum. Further note that the second-order necessary conditions make sense even though the multiplier associated with the objective is always zero; i.e., the Kuhn–Tucker conditions are not fulfilled.

The necessary conditions of Theorem 5.4 can be considerably simplified if

$$0 \in C''_{h'(\bar{x})}(h(\bar{x}), z) \quad \forall z \in \mathcal{C}(\bar{x}). \tag{5.7}$$

This condition is fulfilled, for instance, when K is a generalized polyhedral set (i.e., K can be represented as the intersection of a closed affine subspace and a finite number of closed half spaces), or more generally, if $0 \in T^2_K(g(\bar{x}), g'(\bar{x})z)$ for all z in a dense subset of $\mathcal{C}(\bar{x})$ (this is the so-called *extended polyhedricity condition* in [7]). It is clear from the definition of compound tangent sets that $C''_{h'(\bar{x})}(h(\bar{x}),z) \subset \mathrm{cl}\,(T_C(h(\bar{x})) + \mathrm{Range}(h'(\bar{x})))$. Hence, for each $w \in C''_{h'(\bar{x})}(h(\bar{x}),z)$ and every $y^* \in \Lambda_{FJ}$

we have $\langle y^*, w\rangle \leq 0$. Therefore, if (5.7) holds, the necessary conditions (5.2) of Theorem 5.4 can be rewritten as

$$\forall z \in \mathcal{C}(\bar{x}) \cap S_X \ \exists y^* \in \Lambda_{FJ} : \ \langle y^*, h''(\bar{x}; z)\rangle \geq \beta.$$

In the case of the mathematical programming problem, with $\beta = 0$, we obtain exactly the common second-order necessary conditions [16].

In order to prove sufficiency we need the following regularity condition on the cone of critical directions.

CONDITION 5.10. *For every sequence* $(z_n) \subset S_X$ *with* $d(h'(\bar{x})z_n, T_C(h(\bar{x}))) \to 0$ *there holds* $d(z_n, \mathcal{C}(\bar{x})) \to 0$.

THEOREM 5.11. *Assume that Condition* 5.10 *holds at* $\bar{x}$ *and assume that there exists some* $\beta \geq 0$ *such that for each* $z \in \mathcal{C}(\bar{x}) \cap S_X$ *we can find some multiplier* $y^* \in \Lambda_{FJ} \cap B_{Y^*}$ *satisfying* $\langle y^*, h''(\bar{x}; z)\rangle \geq \beta$. *Then* $\bar{x}$ *is a critical point of second order. If* $\beta > 0$, *then* $\bar{x}$ *is even an essential local minimizer of second order.*

Proof. The proof is by contraposition. Assume that $\bar{x}$ is not a critical point (respectively, essential local minimizer). Let the scalar $\rho > 0$ (respectively, $\rho \geq 0$) and the sequence $(z_n, t_n, \tau_n) \subset S_X \times \mathbb{R}_+ \times \mathbb{R}_+$ be chosen so that the conclusion of Lemma 5.2 holds. From $d(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{1}{2}t_n^2 h''(\bar{x}; z_n), C) = o(t_n^2)$ we obtain $d(h'(\bar{x})z_n, T_C(h(\bar{x}))) = O(t_n)$ and by Condition 5.10 we can find another sequence, say (z_n'), with $(z_n') \subset \mathcal{C}(\bar{x}) \cap S_X$ and $z_n - z_n' \to 0$. By the assumptions of the theorem there is some $\beta \geq 0$ (respectively, $\beta > 0$) so that for each n we can find some $(y_n^*) \subset \Lambda_{FJ} \cap B_{Y^*}$ with $\langle y_n^*, h''(\bar{x}; z_n')\rangle \geq \beta$. Without loss of generality we have $y_n^* \in S_{Y^*}$. Since $h''(\bar{x}; \cdot)$ is Lipschitzian on bounded sets we get

$$\begin{aligned} -\rho &\geq \liminf_{n\to\infty} \frac{\hat{d}_C(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{1}{2}t_n^2 h''(\bar{x}; z_n), h'(\bar{x}), \tau_n t_n)}{t_n^2} \\ &\geq \liminf_{n\to\infty} \frac{1}{2}\langle y_n^*, h''(\bar{x}; z_n)\rangle \geq \frac{1}{2}(\beta + \liminf_{n\to\infty} \langle y_n^*, h''(\bar{x}; z_n) - h''(\bar{x}; z_n')\rangle) = \frac{1}{2}\beta, \end{aligned}$$

which is a contradiction. □

Apart from the case when X is finite dimensional, Condition 5.10 is fulfilled also in other situations. For instance, for the mathematical programming problem, when Range$(h'(\bar{x}))$ is closed, we can apply Hoffman's lemma [17] to verify Condition 5.10, and we obtain the standard second-order sufficient conditions [16] from the above theorem. Hence, the common second-order conditions are equivalent characterizations of critical points and essential local minimizers of second order, respectively.

The following result is new even in the case of scalar minimization and even when Robinson's constraint qualification condition (3.13) is fulfilled.

THEOREM 5.12. *Assume that* $\bar{x}$ *is a nondegenerate critical point (respectively, essential local minimizer) of second order for problem* (P). *Then there is a real* $\beta \geq 0$ *(respectively,* $\beta > 0$) *such that the following holds: For any pair* $(\bar{y}, w) \in Y \times Y$ *such that there is a sequence* $(z_n, t_n) \subset S_X \times \mathbb{R}_+$ *with* $t_n \to 0_+$, $d(h(\bar{x}) + t_n h'(\bar{x})z_n + \frac{1}{2}t_n^2 w, C) = o(t_n^2)$, *and* $h''(\bar{x}; z_n) \to \bar{y}$, *we can find some* $y^* \in \Lambda_{FJ} \cap B_{Y^*}$ *such that* $\langle y^*, \bar{y} - w\rangle \geq \beta$.

Proof. The proof follows along the same lines as the proof of Theorem 5.4 and is therefore omitted. □

Using a proof similar to that of Theorem 5.6 one can show that the second-order conditions of Theorem 5.12 are also sufficient for critical points (respectively, essential local minimizers) if Y is finite dimensional.

Example 5.3 (cf. [7, Ex. 3.25]). Consider the scalar minimization problem

$$\min_{x\in\ell_2} \sum_{i=1}^{\infty} \frac{1}{i} x_i - x_i^2 \quad \text{subject to} \quad x \in K,$$

where $X = \ell_2$ is the Hilbert space of square summable sequences and $K := \{(x_i) : x_i \geq 0 \text{ for all } i\}$. It follows that $Y = \mathbb{R} \times \ell_2$ and $C = \mathbb{R}_- \times K$. Since $g(x) = x$, $\bar{x} = 0$ is nondegenerate and the cone of critical directions and the set of multipliers, respectively, are given by $\mathcal{C}(0) = \{0\}$ and $\Lambda_{FJ} = \{(u^*, v^*) \in \mathbb{R} \times \ell_2 : u^* > 0, v_i^* = u^*/i \; for \; all \; i\}$. Now consider the sequence of directions $z^n = (z_k^n) \in \ell_2$, $n = 1, \ldots$, with $z_k^n = 1$ if $k = n$ and $z_k^n = 0$ if $k \neq n$, and the sequence of scalars $t_n = 2n^{-1}$. Clearly, $h''(\bar{x}; z_n) \to \bar{y} := (-2, 0) \in \mathbb{R} \times \ell_2$. For $w := (-1, 0) \in \mathbb{R} \times \ell_2$ we obtain $d(h(\bar{x}) + t_n h'(\bar{x}) z_n + \frac{1}{2} t_n^2 w, C) = 0$. Hence, we have $\langle y^*, \bar{y} - w \rangle = u^*(-2 - (-1)) = -u^* < 0$ for all $y^* := (u^*, v^*) \in \Lambda_{FJ}$, and from Theorem 5.12 it follows that $\bar{x} = 0$ is not a critical point.

Acknowledgments. The author thanks the anonymous referees for their helpful comments and suggestions.

REFERENCES

[1] A. V. Arutyunov, *Second-order conditions in extremal problems. The abnormal points*, Trans. Amer. Math. Soc., 350 (1998), pp. 4341–4365.
[2] A. V. Arutyunov, *Optimality Conditions: Abnormal and Degenerate Problems*, Math. Appl. 526, Kluwer Academic Publishers, Dordrect, Boston, London, 2000.
[3] J. P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.
[4] A. Ben-Tal, *Second order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
[5] A. Ben-Tal and J. Zowe, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.
[6] J. F. Bonnans, R. Cominetti, and A. Shapiro, *Second order optimality conditions based on parabolic second order tangent sets*, SIAM J. Optim., 9 (1999), pp. 466–492.
[7] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.
[8] A. Cambini, L. Martein, and R. Cambini, *A new approach to second order optimality conditions in vector optimization*, in Advances in Multiple Objective and Goal Programming, Lecture Notes in Econom. and Math. Systems 455, R. Caballero, F. Ruiz, and R. E. Steur, eds., Springer, Berlin, pp. 219–227.
[9] R. Cominetti, *Metric regularity, tangent sets, and second-order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.
[10] S. Dolecki, *Abstract study of optimality conditions*, J. Math. Anal. Appl., 73 (1980), pp. 24–48.
[11] A. Dontchev, *Local convergence of the Newton method for generalized equations*, C. R. Acad. Sci. Paris Ser. 1, 332 (1996), pp. 327–331.
[12] P. C. Duong and H. Tuy, *Stability, surjectivity, and local invertibility of nondifferentiable mappings*, Acta Math. Vietnam., 3 (1978), pp. 13–27.
[13] H. Gfrerer, *Second-order Necessary Conditions for General Nonlinear Optimization Problems: The Degenerate Case*, preprint, University of Linz, Linz, Austria, 2005.
[14] J. B. Hiriart-Urruty, *Tangent cones, generalized gradients, and mathematical programming in Banach spaces*, Math. Oper. Res., 4 (1979), pp. 79–97.
[15] J. B. Hiriart-Urruty, *New concepts in nondifferentiable programming*, Bull. Soc. Math. France Mém., 60 (1979), pp. 57–85.
[16] A. D. Ioffe, *Necessary and sufficient conditions for a local minimum.* 3: *Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.
[17] A. D. Ioffe, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.

[18] A. D. Ioffe, *On some recent developments in the theory of second order optimality conditions*, in Optimization—Fifth French-German Conference Lecture Notes in Math. 1405, S. Dolecki, ed., Springer-Verlag, Berlin, 1989, pp. 55–68.
[19] A. D. Ioffe, *Metric regularity and subdifferential calculus*, Russian Math. Surveys, 55 (2000), pp. 501–558.
[20] B. Jiménez and V. Novo, *First- and second-order sufficient conditions for strict minimality in multiobjective programming*, Numer. Funct. Anal. Optim., 23 (2002), pp. 303–322.
[21] B. Jiménez and V. Novo, *Second-order necessary conditions in set constrained differentiable vector optimization*, Math. Methods Oper. Res., 58 (2003), pp. 299–317.
[22] B. Jiménez and V. Novo, *Optimality conditions in differentiable vector optimization via second-order tangent sets*, Appl. Math. Optim., 49 (2004), pp. 123–144.
[23] H. Kawasaki, *An envelope-like effect of infinitely many inequality constraints on second-order necessary conditions for minimization problems*, Math. Program., 41 (1988), pp. 73–96.
[24] D. Klatte and B. Kummer, *Nonsmooth Equations in Optimization. Regularity, Calculus, Methods, and Applications*, Nonconvex Optim. Appl. 60, Kluwer Academic Publishers, Dordrecht, Boston, London, 2002.
[25] B. Kummer, *Metric regularity: Characterizations, nonsmooth variations, and successive approximation*, Optimization, 46 (1999), pp. 247–281.
[26] E. S. Levitin, A. A. Milyutin, and N. P. Osmolovskii, *Conditions of high order for a local minimum in problems with constraints*, Russian Math. Surveys, 33 (1978), pp. 97–168.
[27] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[28] B. S. Mordukhovich, J. S. Treiman, and Q. J. Zhu, *An extended extremal principle with applications to multiobjective programming*, SIAM J. Optim., 14 (2003), pp. 359–379.
[29] J.-P. Penot, *On regularity conditions in mathematical programming*, Math. Program. Stud., 19 (1982), pp. 167–199.
[30] J.-P. Penot, *Optimality conditions in mathematical programming and composite optimization*, Math. Program., 67 (1994), pp. 225–245.
[31] J.-P. Penot, *Second-order conditions for optimization problems with constraints*, SIAM J. Control Optim., 37 (1998), pp. 303–318.
[32] S. M. Robinson, *Stability theory for systems of inequalities, part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[33] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.
[34] S. M. Robinson, *First order conditions for general nonlinear optimization*, SIAM J. Appl. Math., 30 (1976), pp. 597–607.
[35] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
[36] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
[37] P. H. Sach, *Regularity, calmness, and support principles*, Optimization, 19 (1988), pp. 13–27.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 998–1016

DUAL KALMAN–BUCY FILTERS AND INTERACTIVE ENTROPY PRODUCTION*

NIGEL J. NEWTON†

Abstract. This paper investigates partially observed linear Gaussian systems in continuous time from an information-theoretic viewpoint. It identifies information *supply*, *storage*, and *dissipation* processes that quantify various dependencies between the unobservable and observable components of the state variables of such systems, which we call the *signal* and *observation* processes. The storage is that information in the past of the observation useful for estimating the present and future of the signal. It is so called because it is the information "stored" in the state variable of the Kalman–Bucy filter. The optimal properties of the Kalman–Bucy filter bring about a particular symmetry in its joint dynamics with the signal. This enables a *dual* system, with the same structure as the original, to be derived by time reversal. The information supply and dissipation processes of this dual problem are those of the original, with roles interchanged. The same symmetry enables us to define an entropic measure of the local time-asymmetry of the *interaction* between the signals and filters, which we call the *rate of interactive entropy production*. This is shown to be directly connected with the rates of information supply and dissipation. It is also intimately connected with the entropy production of an abstract statistical mechanical system at or near a stationary nonequilibrium state.

Key words. Bayesian inference, dual filters, entropic derivative, information theory, nonequilibrium statistical mechanics, statistical filtering

AMS subject classifications. 93E11, 93E20, 94A15, 62F15, 60E10, 60G35, 82C05

DOI. 10.1137/050623267

1. Introduction.

This article investigates continuous-time, partially observed linear systems from an information-theoretic viewpoint. The systems considered have state variables comprising unobservable and observable components, which we call the *signal* and the *observation*, respectively. The key property of the state variable here is that it is Markov. This allows the future and the past of certain estimation problems to be separated by the present. More specifically, we define information *supply*, *storage*, and *dissipation* processes, which quantify the information contained in the past and present values of the observation on, respectively, the entire path of the signal, the present and future values of the signal, and the past of the signal. In the context of filtering, one is not concerned with the improvement of estimates of past values of the signal as new information supply becomes available, but rather with the optimal estimation of the current signal value from the current supply. For this reason it is not necessary for the filter state variable to hold the entire current information supply, but only that part useful for estimating the present and the future of the signal, i.e., the storage. The storage is so called because it is "stored" in the state variable of the Kalman–Bucy filter.

*Received by the editors January 25, 2005; accepted for publication (in revised form) March 11, 2006; published electronically July 31, 2006. This work was partially supported by Leverhulme Trust Research Fellowship 2003/0426, by MURI grant F49620-02-1-0325 (Complex Adaptive Networks for Cooperative Control), by ARO-MURI grant DAAD19-00-1-0466 (Data Fusion in Large Arrays of Microsensors (SensorWeb)) and by NSF-ITR grant CCR-0325774 Collaborative Research: New Approaches to Experimental Design and Statistical Analysis of Genomic and Structural Biological Data from Multiple Sources.

http://www.siam.org/journals/sicon/45-3/62326.html

†Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (njn@essex.ac.uk, njnewton@mit.edu).

The signal and filter processes are jointly *and* marginally Markov, and because of this the processes have a particular symmetry that yields a *dual* filtering problem when time reversed. This dual problem has a structure identical to that of the original. Furthermore, its information supply, storage, and dissipation processes are intimately related to those of the original problem; the information *flows* occurring in the dual problems are time-reversed versions of each other.

We build on earlier work appearing in [12], which is centered on a statistical mechanical analogy for stable, time-homogeneous, nondegenerate, Gaussian processes. In this analogy, an abstract statistical mechanical system characterized by the process interacts with a heat bath in such a way that the "universe" comprising the system and the heat bath obeys a law of nondecrease of entropy analogous to the second law of thermodynamics. Under the special conditions of [12] (stability, time-homogeneity, and nondegeneracy) the process has a unique invariant distribution to which other distributions of the process converge. This invariant distribution is the maximum entropy "state" of the abstract universe. In the theory of nonequilibrium statistical mechanics, invariant distributions are called *stationary equilibrium states* or *stationary nonequilibrium states* according to the flow of some quantity (often energy) in the state, an equilibrium state being one for which this flow is zero. (See, for example, [1], [5], or [8].)

This statistical mechanical analogy is applied in [12] to the *joint* process comprising the signal and its Kalman–Bucy filter. It is shown there that, whatever the nature of the invariant distribution of the signal process alone, that of the joint process is a stationary nonequilibrium state. Energy is shown to flow, in the stationary state, around a loop comprising the heat bath and two subsystems termed the "conditional signal" and the "filter." This energy flow is accompanied by an entropy flow, which can be isolated from any flows of entropy occurring *within* either of the subsystems. This is done by what we define as *interactive entropy production*. This quantity can be interpreted, independently of the statistical mechanical analogies in [12], as an entropic measure of the time-asymmetry of the interaction between the signal and filter processes. This interpretation is meaningful without the special conditions of [12]. The definition of interactive entropy production is extended in this paper to the more general case, and its properties are investigated. It is intimately related to the information flows that occur in the original and dual filtering problems.

Information-theoretic aspects of filters have received much attention in the literature. For example, a variety of information-theoretic measures of optimality, such as the mutual information between the observation and the estimation error, are considered in [4], [7], and [16]. In particular, these papers show that many of these measures are optimized by the Kalman–Bucy filter in the linear Gaussian case. More interesting is the fact that they provide a framework for deriving suboptimal filters in the nonlinear case. The measure of interest in this paper is the mutual information between the observation path and the signal value. The results can be extended to nonlinear systems (see [13] and [14]), but then involve the infinite-dimensional nonlinear filter. The results most closely connected with the information flow aspects of this paper are to be found in [3] and [15].

2. Information flow in the Kalman–Bucy filter. This section investigates the information flows that occur between signal and observation processes in a partially observed linear Gaussian system. Up to a point it interprets classical results; what we call here the information *supply* was first evaluated for partially observed diffusions in [3], and a study of the mutual information for the filter and path estimators

was made in [15]. (See also [10].) Results of the type presented in this section, only for the time-homogeneous, nondegenerate case, were developed in [12].

All random quantities are defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t,\ t \in [-T,T]), \mathbb{P})$. (The use of a *symmetric* time interval simplifies the discussion of the dual filter in section 3.) The signal, X, is an $\mathbb{R}^n$-valued continuous-time process satisfying

$$X_t = \xi + \int_{-T}^{t} A(s)X_s\,ds + V_t \quad \text{for } t \in [-T,T], \tag{2.1}$$

where A is a continuous $\mathbb{R}^{n\times n}$-valued function on $[-T,T]$, ξ is an $\mathcal{F}_{-T}$-measurable $\mathbb{R}^n$-valued Gaussian random variable with mean zero and nonsingular covariance matrix P_i, and

$$V_t = \int_{-T}^{t} \Psi(s)\,d\tilde{V}_s \quad \text{for } t \in [-T,T].$$

Here, Ψ is a Lipschitz continuous $\mathbb{R}^{n\times r}$-valued function on $[-T,T]$, and $(\tilde{V}_t, \mathcal{F}_t)$ is an r-vector Brownian motion, left-shifted in time by T, for some $r \le n$. With this definition, X is a zero-mean Gaussian process with covariance matrix process P satisfying the following equation:

$$P(t) = P_i + \int_{-T}^{t} \left(A(s)P(s) + P(s)A(s)' + \Sigma_V(s)\right)\,ds \quad \text{for } t \in [-T,T],$$

where $\Sigma_V = \Psi\Psi'$.

The observation, Y, is an $\mathbb{R}^n$-valued continuous-time process satisfying

$$Y_t = \xi + \zeta + \int_{-T}^{t} \Sigma_W(s)X_s\,ds + W_t \quad \text{for } t \in [-T,T], \tag{2.2}$$

where ζ is an $\mathcal{F}_{-T}$-measurable Gaussian random variable, independent of ξ, with mean zero and nonsingular covariance matrix M,

$$W_t = \int_{-T}^{t} \Gamma(s)\,d\tilde{W}_s \quad \text{for } t \in [-T,T],$$

Γ is a Lipschitz continuous $\mathbb{R}^{n\times d}$-valued function on $[-T,T]$ for some $d \le n$, $(\tilde{W}_t, \mathcal{F}_t)$ is a d-vector left-shifted Brownian motion, independent of V, and $\Sigma_W = \Gamma\Gamma'$.

Remark 2.1. Y comprises an "initial" observation Y^i $(= \xi + \zeta)$, and a "running" observation $(Y_t - Y^i,\ t \in [-T,T])$. The former can be thought of as summarizing partial observations of an extended version of X that solves an equation of type (2.1) over times prior to $-T$.

Remark 2.2. The running observation is essentially of dimension d since it can be replaced with the d-dimensional process

$$\tilde{Y}^r_t = \int_{-T}^{t} \Gamma(s)'X_s\,ds + \tilde{W}_t \quad \text{for } t \in [-T,T],$$

with no loss in information content regarding the signal.

Remark 2.3. The signal and observation noise processes, V and W, have "built-in" nonstandard quadratic covariations

$$d[V,V']_t = \Sigma_V(t)\,dt \quad \text{and} \quad d[W,W']_t = \Sigma_W(t)\,dt.$$

This avoids any further need for specific square roots of Σ_V and Σ_W and clarifies the discussion of duality in section 3.

Remark 2.4. The more general model, in which ξ and ζ have nonzero means and the linear drift coefficients in (2.1) and (2.2) are replaced with affine functions, can easily be re-expressed in the form above. It is then an equivalent filtering problem to estimate the centered signal from a centered version of the observations.

The Kalman–Bucy filter generates the conditional distribution of X_t given $(Y_s, s \in [-T,t])$, which is Gaussian. The *covariance* form of the filter propagates the mean vector, $\hat{X}$, and covariance matrix, Q, of the conditional distribution. These evolve as follows:

$$\begin{aligned} \hat{X}_{-T} &= P_i(P_i+M)^{-1}Y^i, \\ d\hat{X}_t &= (A(t)-Q(t)\Sigma_W(t))\,\hat{X}_t\,dt + Q(t)\,dY_t, \\ Q(-T) &= \left(P_i^{-1}+M^{-1}\right)^{-1}, \\ \dot{Q}(t) &= A(t)Q(t)+Q(t)A(t)'+\Sigma_V(t)-Q(t)\Sigma_W(t)Q(t). \end{aligned} \tag{2.3}$$

(See, for example, [2].) Since P_i and M are positive definite, the inverse matrices here are well defined, and $P(t)$, $Q(t)$, and $P(t)-Q(t)$ are all positive definite for all t.

Let $\mathcal{F}^Y$ be the filtration generated by Y, and let ν be the associated innovations process:

$$\begin{aligned} \mathcal{F}_t^Y &:= \sigma(Y_s,\ s\in[-T,t]) \\ \nu_t &:= Y_t - \hat{X}_{-T} - \int_{-T}^t \Sigma_W(s)\hat{X}_s\,ds \qquad \text{for } t\in[-T,T]. \end{aligned} \tag{2.4}$$

Then $(\nu_t, \mathcal{F}_t^Y)$ is a continuous n-vector Gaussian martingale with the same quadratic covariation as W, but with a nonzero initial value having the n-variate Gaussian distribution $N(0, M-Q(-T))$. (Here, and in what follows, we denote the n-variate Gaussian distribution with mean vector μ and covariance matrix Σ by $N(\mu,\Sigma)$ and denote its density (when $|\Sigma|\neq 0$) by $n(\mu,\Sigma)$.) Of course the filter can be expressed entirely in terms of the innovations process as follows:

$$\hat{X}_t = P_iM^{-1}\nu_{-T} + \int_{-T}^t A(s)\hat{X}_s\,ds + \int_{-T}^t Q(s)\,d\nu_s. \tag{2.5}$$

(This, together with the representation (2.4) for Y, and the fact that (2.5) has a strong solution, confirms that the filtration generated by ν coincides with $\mathcal{F}^Y$.)

We define information *supply*, *storage*, and *dissipation* processes as follows:

$$\begin{aligned} S(t) &:= I(X\,;\,(Y_s, s\in[-T,t])), \\ C(t) &:= I((X_s, s\in[t,T])\,;\,(Y_s, s\in[-T,t])) \quad \text{for } t\in[-T,T], \\ D(t) &:= S(t)-C(t), \end{aligned} \tag{2.6}$$

where, for random variables Θ and Φ taking values in Borel spaces and having joint and marginal distributions $\mathbb{P}_{\Theta\Phi}$, $\mathbb{P}_\Theta$, and $\mathbb{P}_\Phi$, $I(\Theta\,;\,\Phi)$ is the *mutual information*:

$$I(\Theta\,;\,\Phi) = \begin{cases} \displaystyle\int \log\left(\frac{d\mathbb{P}_{\Theta\Phi}}{d(\mathbb{P}_\Theta\otimes\mathbb{P}_\Phi)}\right) d\mathbb{P}_{\Theta\Phi} & \text{if the integral exists,} \\ +\infty & \text{otherwise.} \end{cases} \tag{2.7}$$

Since X is Markov, and the increments of W are independent of X and of each other, $\hat{X}_t$ is a sufficient statistic for the problem of estimating the future of X (X_s, $s \in [t,T]$), from the past of Y (Y_s, $s \in [-T,t]$), and so

$$\begin{aligned} C(t) &= I((X_s, s \in [t,T])\,;\, \hat{X}_t) \\ &= I(X_t\,;\, \hat{X}_t) \\ &= \tfrac{1}{2}\log|P_i + M| - \tfrac{1}{2}\log|M| + \tfrac{1}{2}\int_{-T}^{t} \operatorname{tr}(\Sigma_W(s)Q(s))\,ds \\ &\quad - \tfrac{1}{2}\int_{-T}^{t} \operatorname{tr}\left(\Sigma_V(s)(Q(s)^{-1} - P(s)^{-1})\right)\,ds. \end{aligned} \tag{2.8}$$

The information supply can be evaluated by means of a simple generalization of Theorem 3 in [3]. This involves a measure transformation on $(\Omega, \mathcal{F}_T)$. Let $\mathbb{P}^M$ be the measure on $\mathcal{F}_T$ whose Radon–Nikodym derivative with respect to $\mathbb{P}$ is L_T, where L is the continuous process defined as follows:

$$\begin{aligned} L_t := \frac{n(0, P_i + M)(Y^i)}{n(\xi, M)(Y^i)} \exp\Bigg(&- \int_{-T}^{t} (X_s - \hat{X}_s)'\,dW_s \\ &- \tfrac{1}{2}\int_{-T}^{t} (X_s - \hat{X}_s)'\Sigma_W(s)(X_s - \hat{X}_s)\,ds\Bigg). \end{aligned} \tag{2.9}$$

Lemma 2.1.

(i) *$\mathbb{P}$ and $\mathbb{P}^M$ are mutually absolutely continuous probability measures.*

(ii) *Under $\mathbb{P}^M$, (ξ, V) and ν are independent, but retain the marginal distributions they have under $\mathbb{P}$.*

(iii) *Under $\mathbb{P}^M$, the processes X and $\hat{X}$ are independent, but retain the marginal distributions they have under $\mathbb{P}$.*

Proof. It is a standard result that

$$\mathbf{E}\left(\exp\left(\tfrac{1}{2}\int_{-T}^{T} (X_t - \hat{X}_t)'\Sigma_W(t)(X_t - \hat{X}_t)\,dt\right) \,\middle|\, \mathcal{F}_{-T}\right) < \infty \quad \text{a.s.}$$

(See, for example, [9].) Thus it follows from the Girsanov theorem that, under $\mathbb{P}$, $(L_t, \mathcal{F}_t)$ is a martingale with

$$\mathbf{E}\,(L_t \,|\, \mathcal{F}_{-T}) = \frac{n(0, P_i + M)(Y^i)}{n(\xi, M)(Y^i)} \quad \text{for all } t \in [-T,T]$$

so that $\mathbf{E}L_T = 1$, and this establishes part (i). Now, for each $t \in [-T,T]$,

$$\begin{bmatrix} V_t \\ \nu_t \end{bmatrix} = \begin{bmatrix} 0 \\ \nu_{-T} \end{bmatrix} + \int_{-T}^{t} \begin{bmatrix} 0 \\ \Sigma_W(s)(X_s - \hat{X}_s) \end{bmatrix} ds + \begin{bmatrix} V_t \\ W_t \end{bmatrix},$$

and so it follows from Girsanov's theorem that, under $\mathbb{P}^M$, $((V_t, \nu_t), \mathcal{F}_t)$ is a $2n$-dimensional Gaussian martingale with quadratic covariation

$$d\left[\begin{bmatrix} V \\ \nu \end{bmatrix}, \begin{bmatrix} V \\ \nu \end{bmatrix}'\right]_t = \begin{bmatrix} \Sigma_V(t) & 0 \\ 0 & \Sigma_W(t) \end{bmatrix} dt. \tag{2.10}$$

The independence of ξ and ν_{-T} under $\mathbb{P}^M$ follows from elementary manipulations of the n-variate Gaussian densities in (2.9), and this establishes part (ii). Part (iii) follows from part (ii) and the innovations representation of $\hat{X}$ in (2.5). □

In what follows, we shall call $\mathbb{P}^M$ the *product-of-marginals* measure.

The information supply can now be evaluated as follows:

$$\begin{aligned} S(t) &= \mathbf{E}\log\left(L_t^{-1}\right) \\ &= \frac{1}{2}\log|P_i + M| - \frac{1}{2}\log|M| + \frac{1}{2}\int_{-T}^{t} \mathrm{tr}(\Sigma_W(s)Q(s))\,ds, \end{aligned} \tag{2.11}$$

and so the dissipation has the form

$$D(t) = \frac{1}{2}\int_{-T}^{t} \mathrm{tr}\left(\Sigma_V(s)(Q(s)^{-1} - P(s)^{-1})\right)\,ds. \tag{2.12}$$

It also easily follows that, for any $s \le t$,

$$\begin{aligned} S(t) - D(s) &= C(s) + \mathbf{E}\log\left(L_t^{-1}L_s\right) \\ &= I((X_r, r \in [s,T])\,;\,(Y_r, r \in [-T,t])), \end{aligned}$$

so that the dissipation, $D(s)$, is the information on X derived from $(Y_r,\ r \in [-T,t])$ that is of no use in estimating the values that X takes from time s onward.

In one respect, the Kalman–Bucy filter can be thought of as being a data *encoder*. At time t, it partially encodes the observation path $(Y_s,\ s \in [-T,t])$ as the statistic $\hat{X}_t$. Clearly this encoding is *lossy* because the observation path cannot be recovered from $\hat{X}_t$. Since both $(Y_s,\ s \in [-T,t])$ and $\hat{X}_t$ take values in uncountably infinite spaces, we cannot define absolute information quantities for them, and so we cannot define a meaningful measure of the *total* information loss arising from the encoding. However, we *can* define a meaningful measure of information loss in the context of a specific estimation problem; this is simply the amount by which the *mutual* information between the estimand and $\hat{X}_t$ is less than that between the estimand and $(Y_s,\ s \in [-T,t])$. For the problem of estimating the whole process X, the encoding of the Kalman–Bucy filter at time t loses information, but for the problem of estimating the future of X, $(X_s,\ s \in [t,T])$, it is lossless. We note that S, C, and D are all *mutual* information quantities and, as such, have absolute meaning.

In a dynamic sense, the Kalman–Bucy filter can be thought of as an *information processor* that receives new observation-derived information at the rate $\dot{S}(t)$, and dissipates historical information at the rate $\dot{D}(t)$. It stores an amount of observation-derived information $C(t)$, which is that part of the supply to date useful for estimating the present and future of X.

3. The dual filter. The duality ideas for discrete-time systems in [6] are extended here to the continuous-time system (2.1), (2.2). These ideas use the concept of *dual bases* for finite-dimensional, linear spaces. When the state and observation processes of a finite-dimensional, discrete-time system, evolving over a finite time interval, are taken as a set of basis vectors, the vectors of a dual basis can be shown to be the state and observation processes of another system, which is naturally termed the *dual system*. In the continuous-time setting of this article, the dual system is derived through time reversal.

It turns out that the dual filters are most conveniently expressed in *information form*. For the Kalman–Bucy filter of section 2, this propagates the vector process Z

$(:= Q^{-1}\hat{X})$ instead of $\hat{X}$. The signal and information filter processes, X and Z, are *coupled* Markov processes evolving *jointly* according to the following equations:

$$\begin{aligned} \begin{bmatrix} X_{-T} \\ Z_{-T} \end{bmatrix} &= \begin{bmatrix} I_n & 0 \\ M^{-1} & M^{-1} \end{bmatrix} \begin{bmatrix} \xi \\ \zeta \end{bmatrix}, \\ d\begin{bmatrix} X_t \\ Z_t \end{bmatrix} &= A_J(t) \begin{bmatrix} X_t \\ Z_t \end{bmatrix} dt + d\begin{bmatrix} V_t \\ W_t \end{bmatrix}, \end{aligned} \tag{3.1}$$

where I_n is the $n \times n$ identity matrix, and

$$A_J(t) := \begin{bmatrix} A(t) & 0 \\ \Sigma_W(t) & -\big(A(t) + \Sigma_V(t)Q(t)^{-1}\big)' \end{bmatrix}. \tag{3.2}$$

Under the product-of-marginals measure, $\mathbb{P}^M$, X, and Z are *independent* Markov processes whose evolutions are conveniently expressed by the "marginal" equations:

$$\begin{aligned} \begin{bmatrix} X_{-T} \\ Z_{-T} \end{bmatrix} &= \begin{bmatrix} I_n & 0 \\ 0 & M^{-1}(P_i + M)M^{-1} \end{bmatrix} \begin{bmatrix} \xi \\ \nu_{-T} \end{bmatrix}, \\ d\begin{bmatrix} X_t \\ Z_t \end{bmatrix} &= A_M(t) \begin{bmatrix} X_t \\ Z_t \end{bmatrix} dt + d\begin{bmatrix} V_t \\ \nu_t \end{bmatrix}, \end{aligned} \tag{3.3}$$

where

$$A_M(t) := \begin{bmatrix} A(t) & 0 \\ 0 & -\big(A(t) + \Sigma_V(t)Q(t)^{-1}\big)' + \Sigma_W(t)Q(t) \end{bmatrix}. \tag{3.4}$$

The dual system is obtained by time-reversing (X, Z). For each $t \in [-T, T]$, let

$$\begin{aligned} X_t^* &:= Z_{-t}, \\ Z_t^* &:= X_{-t}, \\ \mathcal{F}_t^* &:= \sigma(X_s^*, Z_s^*,\ s \in [-T, t]), \\ A^*(t) &:= A(-t)' + Q(-t)^{-1}\Sigma_V(-t) - \Sigma_W(-t)\big(Q(-t)^{-1} - P(-t)^{-1}\big)^{-1}, \\ \Sigma_V^*(t) &:= \Sigma_W(-t), \\ \Sigma_W^*(t) &:= \Sigma_V(-t), \\ Q^*(t) &:= Q(-t)^{-1} - P(-t)^{-1}, \\ V_t^* &:= X_t^* - X_{-T}^* - \int_{-T}^{t} A^*(s)X_s^*\, ds, \\ W_t^* &:= Z_t^* - Z_{-T}^* - \int_{-T}^{t} \big(\Sigma_W^*(s)X_s^* - (A^*(s) + \Sigma_V^*(s)Q^*(s)^{-1})'Z_s^*\big)\, ds, \\ \nu_t^* &:= P(T)^{-1}Z_{-T}^* + \int_{-T}^{t} \Sigma_W^*(s)\,(X_s^* - Q^*(s)Z_s^*)\, ds + W_t^*. \end{aligned} \tag{3.5}$$

Proposition 3.1.

(i) *Under* $\mathbb{P}$, *the process* $((V_t^*, W_t^*),\ \mathcal{F}_t^*,\ t \in [-T, T])$ *is a* $2n$*-dimensional, Gaussian martingale with initial value zero and quadratic covariation*

$$\left[\begin{bmatrix} V^* \\ W^* \end{bmatrix}, \begin{bmatrix} V^* \\ W^* \end{bmatrix}'\right]_t = \int_{-T}^{t} \begin{bmatrix} \Sigma_V^*(s) & 0 \\ 0 & \Sigma_W^*(s) \end{bmatrix} ds. \tag{3.6}$$

(ii) *Under* $\mathbb{P}^M$*, the process* $((V_t^*, \nu_t^*),\ \mathcal{F}_t^*,\ t \in [-T,T])$ *is a* $2n$*-dimensional, Gaussian martingale with initial value* $(0, P(T)^{-1}Z_{-T}^*)$ *(which is independent of* X_{-T}^**) and the quadratic covariation of* (3.6).

(iii) *For each* $t \in [-T,T]$,

$$\begin{aligned}\mathcal{F}_t^* &= \sigma(X_{-T}^*,\, Z_{-T}^*;\, (V_s^*,\, W_s^*),\, s \in [-T,t]) \\ &= \sigma(X_{-T}^*,\, Z_{-T}^*;\, (V_s^*,\, \nu_s^*),\, s \in [-T,t]).\end{aligned} \tag{3.7}$$

Proof. Let $(\Theta(s,t),\ -T \le s \le t \le T)$ be the matrix fundamental solution of (3.1) without the "driving" process; i.e.,

$$\begin{aligned}&\Theta(s,s) = I_{2n} \quad \text{for } -T \le s \le T, \\ &\frac{\partial \Theta(s,t)}{\partial t} = A_J(t)\Theta(s,t) \quad \text{for } -T \le s \le t \le T.\end{aligned} \tag{3.8}$$

Then it follows that for $-T \le s \le t \le T$,

$$\mathbf{E}\begin{bmatrix} Z_s^* \\ X_s^* \end{bmatrix}\begin{bmatrix} Z_t^* \\ X_t^* \end{bmatrix}' = \Theta(-t,-s)P_J(-t),$$

where $P_J(t)$ is the covariance matrix of (X_t, Z_t):

$$P_J(t) := \begin{bmatrix} P(t) & (P(t)-Q(t))Q(t)^{-1} \\ Q(t)^{-1}(P(t)-Q(t)) & Q(t)^{-1}(P(t)-Q(t))Q(t)^{-1} \end{bmatrix}. \tag{3.9}$$

Straightforward calculations now show that, for any $-T \le r \le s \le t \le T$,

$$\mathbf{E}\begin{bmatrix} Z_r^* \\ X_r^* \end{bmatrix}\begin{bmatrix} V_t^* - V_s^* \\ W_t^* - W_s^* \end{bmatrix}' = 0,$$

and that (3.6) holds, and so (V^*, W^*) is a zero-mean, independent-increments, Gaussian process, independent of (X_{-T}^*, Z_{-T}^*) and with the quadratic covariation of (3.6). It is thus a Gaussian martingale with this quadratic covariation with respect to the filtration it generates. It now follows from the definitions of V^* and W^* that

$$(X^*, Z^*) = \Phi^*\left((X_{-T}^*, Z_{-T}^*),\, (V^*, W^*)\right),$$

where Φ^* is the strong solution of (3.10) below. This proves the first equation in (3.7) and proves part (i) of the proposition. The second equation in (3.7) and part (ii) of the proposition can be proved in a similar manner, but starting from the matrix fundamental solution of (3.3). □

It follows from the definitions of V^*, W^*, and ν^* that (X^*, Z^*) satisfies the following two equations:

$$\begin{aligned}&\begin{bmatrix} X_{-T}^* \\ Z_{-T}^* \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ (M^*)^{-1} & (M^*)^{-1} \end{bmatrix}\begin{bmatrix} \xi^* \\ \zeta^* \end{bmatrix}, \\ &d\begin{bmatrix} X_t^* \\ Z_t^* \end{bmatrix} = A_J^*(t)\begin{bmatrix} X_t^* \\ Z_t^* \end{bmatrix}dt + d\begin{bmatrix} V_t^* \\ W_t^* \end{bmatrix},\end{aligned} \tag{3.10}$$

and

$$\begin{aligned}&\begin{bmatrix} X_{-T}^* \\ Z_{-T}^* \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & (M^*)^{-1}(P_i^* + M^*)(M^*)^{-1} \end{bmatrix}\begin{bmatrix} \xi^* \\ \nu_{-T}^* \end{bmatrix}, \\ &d\begin{bmatrix} X_t^* \\ Z_t^* \end{bmatrix} = A_M^*(t)\begin{bmatrix} X_t^* \\ Z_t^* \end{bmatrix}dt + d\begin{bmatrix} V_t^* \\ \nu_t^* \end{bmatrix},\end{aligned} \tag{3.11}$$

where A_J^* and A_M^* are defined as in (3.2) and (3.4), but with dual quantities, and

$$\begin{aligned} \xi^* &:= X^*_{-T}, \\ \zeta^* &:= M^* Z^*_{-T} - X^*_{-T}, \\ M^* &:= Q(T)^{-1}, \\ P_i^* &:= Q(T)^{-1}(P(T) - Q(T))Q(T)^{-1}. \end{aligned} \tag{3.12}$$

Equations (3.10) and (3.11) are the equivalents of (3.1) and (3.3) for a dual filtering problem: straightforward calculations show that X^* is a zero-mean Gaussian process, that $X^*(t)$ has covariance matrix

$$P^*(t) := Q(-t)^{-1}(P(-t) - Q(-t))Q(-t)^{-1},$$

and that Z^* is its information filter process given the dual observations

$$Y_t^* = \xi^* + \zeta^* + \int_{-T}^t \Sigma_W^*(s) X_s^*\, ds + W_t^* \quad \text{for } t \in [-T, T].$$

These are made up of initial and running components in the same way as was Y. To complete the duality, we note that Q^*, as defined in (3.5), is indeed the conditional covariance matrix for this filter. Thus, in reverse time, the information filter of the original problem, Z, becomes an autonomous signal process, and the signal process of the original problem, X, becomes its Kalman–Bucy filter, in information form, given the dual observations Y^*.

In analogy with the definitions of $S(t)$, $C(t)$, and $D(t)$ of section 2, we can identify the information supply, storage, and dissipation for the dual filter:

$$\begin{aligned} S^*(t) &:= I(X^*\,;\, (Y_s^*, s \in [-T, t])) \\ &= \tfrac{1}{2}\log|P_i^* + M^*| - \tfrac{1}{2}\log|M^*| + + \tfrac{1}{2}\int_{-T}^t \operatorname{tr}(\Sigma_W^*(s)Q^*(s))\, ds \\ &= S^*(-T) + (D(T) - D(-t)), \\ C^*(t) &:= I((X_s^*, s \in [t, T])\,;\, (Y_s^*, s \in [-T, t])) \\ &= C(-t), \\ D^*(t) &:= S^*(t) - C^*(t) \\ &= \tfrac{1}{2}\int_{-T}^t \operatorname{tr}\left(\Sigma_V^*(s)\left(Q^*(s)^{-1} - P^*(s)^{-1}\right)\right) ds \\ &= S(T) - S(-t). \end{aligned} \tag{3.13}$$

This shows that the time-reversed information flows for the original problem can be interpreted as the forward-time information flows for the dual problem, with supply and dissipation swapping roles.

It may seem, at first sight, that $\dot{S}^*(-t)$ should equal $\dot{S}(t)$ since both concern the supply of information between X and Z over the short time interval $[t, t + \delta t]$. That this is not so is because S and S^* are *cumulative* information quantities with derivatives that represent the flow rate of *new* information. What constitutes new information between X and Z depends on the time direction over which information is being accumulated. Over the time interval $[t, t + \delta t]$ an amount $S(t + \delta t) - S(t)$ of new information is supplied in the forward problem. Because this is new, it relates to

a dependency between $X_{t+\delta t}$ and $Z_{t+\delta t}$ that is not present between $(X_s,\ s \in [-T,t])$ and $(Z_s,\ s \in [-T,t])$. For this reason the dual filter dissipates it and

$$D^*(-t) - D^*(-t-\delta t) = S(t+\delta t) - S(t).$$

A central role is played in the information supply rate of the original filter by the "signal-to-noise ratio" matrix $\Sigma_W(t)$. Its dual, $\Sigma_V(t)$, controls the rate of information *dissipation* for the forward filter: the larger it is, the greater is the effect of the signal noise, and this increases the rate at which the signal "forgets its past." We can think of information for the forward filter as being supplied by the observation process, Y, stored temporarily in the filter state, Z_t, and then absorbed by the signal noise process, V. This notion that the signal noise can *erase* past information supply is at the heart of some statistical mechanical analogies developed in [12].

Equation (3.10) provides an "inverse construction" of the pair (X,Z). In (3.1), X is defined in an autonomous fashion, and Z is defined in a way that couples it to X and results in a particular joint distribution, P_{XZ}. In the inverse construction (3.10), this dependency is reversed. Clearly Y can be found from Z (and vice versa) by means of (3.3) and (2.4), and so (3.10) also provides an inverse construction of the pair (X,Y). This construction can be formulated in terms of optimal control, as follows.

The equations for Z^* in (3.11) and (3.10) can be rewritten as

$$dZ_t^* = \bar{A}(t)Z_t^*\,dt + dv_t^* \tag{3.14}$$

and

$$dZ_t^* = \bar{A}(t)Z_t^*\,dt + \Sigma_V(-t)U^*(Z_t^*,t)\,dt + dW_t^*, \tag{3.15}$$

respectively, where

$$\begin{aligned}\bar{A}(t) &:= -\left(A^*(t) + \Sigma_V^*(t)Q^*(t)^{-1}\right)' + \Sigma_W^*(t)Q^*(t)\\ &= -A(-t) - \Sigma_V(-t)P(-t)^{-1}\end{aligned} \tag{3.16}$$

and

$$U^*(z^*,t) = X_t^* - \left(Q(-t)^{-1} - P(-t)^{-1}\right)z^*. \tag{3.17}$$

Equation (3.14) is the autonomous equation for X, (2.1), reversed in time; (3.15) can be thought of as being a *controlled* version of (3.14), in which the control, U^*, is random through its dependence on X^*. Regarding the outcome of Y (and hence X^*) as a fixed parameter, the effect of the control is to change the distribution of the path Z^* into a regular version of its Y-conditional distribution. In fact U^* is the optimal control in an LQG (linear quadratic Gaussian) problem for Z^* with the *apparent information* for estimators of X given Y as cost. (See [11].) The use of a suboptimal control in this time-reversed problem corresponds to the dissipation of information at a rate exceeding $\dot{D}$. This fact is useful in the study of suboptimal filtering. (See [14].) Of course, the same arguments can be applied to the equations for Z in (3.3) and (3.1) to characterize the dissipation of the dual filtering problem in terms of a dual LQG control problem for Z in forward time.

4. Interactive entropy production. This section further investigates the interaction between the signal and filter processes of (3.1). Starting from a concept of

entropy production for the signal, it defines an entropic measure of the local time-asymmetry of this interaction, which we shall call the *rate of interactive entropy production* between the signal and filter processes. This was introduced in the context of a statistical mechanical analogy for the signal and filter under special conditions in [12]. The (pertinent) special conditions are recalled here, as follows:

(S1) Σ_V and Σ_W are not dependent on t;

(S2) Σ_V and Σ_W are positive definite for all $t \in [-T, T]$.

We first consider the signal process alone under these conditions. For each $-T \le t - \epsilon < t + \epsilon \le T$, let $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$ be regular X_t-conditional distributions of the processes $(X_u,\, u \in [t, t+\epsilon])$ and $(Z^*_u,\, u \in [-t, -t+\epsilon])$, respectively. (These are segments of the process X extending forward and backward from time t by an amount ϵ.) It follows from (S1), (S2), and the Girsanov theorem that $\Pi^X_{t+\epsilon|t} \sim \Pi^X_{t-\epsilon|t}$ (a.s.), and that

$$\frac{d\Pi^X_{t+\epsilon|t}}{d\Pi^X_{t-\epsilon|t}}(X) = \exp\biggl(\int_t^{t+\epsilon} X'_s(A(s) - \bar{A}(s-2t))'\Sigma_V^{-1}\, dV_s + \frac{1}{2}\int_t^{t+\epsilon} X'_s(A(s) - \bar{A}(s-2t))'\Sigma_V^{-1}(A(s) - \bar{A}(s-2t))X_s\, ds\biggr),$$

where $\bar{A}$ is as defined in (3.16). The rate of entropy production of X at time $t \in (-T, T)$, when (S1) and (S2) hold, can then be defined as follows:

$$\begin{aligned} R_X(t) &:= \lim_{\epsilon \downarrow 0} \epsilon^{-1} \mathbf{E} h\left(\Pi^X_{t+\epsilon\,|\,t} | \Pi^X_{t-\epsilon|t}\right) \\ &= \tfrac{1}{2}\mathrm{tr}\left((A(t) - \bar{A}(-t))'\Sigma_V^{-1}(A(t) - \bar{A}(-t))P(t)\right), \end{aligned} \tag{4.1}$$

where, for measures α and β on the same measurable space, $h(\alpha\,|\,\beta)$ is the *relative entropy*

$$h(\alpha\,|\,\beta) = \begin{cases} \displaystyle\int \log\left(\frac{d\alpha}{d\beta}\right) d\alpha & \text{if the integral exists,} \\ +\infty & \text{otherwise.} \end{cases} \tag{4.2}$$

This definition of R_X is too strong in the general case since it is then possible for the measures $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$ to be relatively singular. This can occur even if (S2) holds, since a small difference between the quadratic variations of X on either side of time t will result in the relative singularity of $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$. We weaken the definition here by replacing $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$ in (4.1) with their respective $t+\epsilon$ and $t-\epsilon$ marginals. The eventuality that $\Sigma_W(t)$ is singular is handled by a relaxation argument, as follows. For any $\delta \in (0, 1]$, let X^δ be the $\mathbb{R}^n$-valued process defined by the following equation:

$$X^\delta_t = \xi + \int_{-T}^t A(s)X^\delta_s\, ds + V_t + \delta B_t \quad \text{for } t \in [-T, T], \tag{4.3}$$

where $(B_t, \mathcal{F}_t)$ is a left-shifted n-vector Brownian motion independent of V (and W). Let $\mathcal{L}^\delta(t,x)$ and $\bar{\mathcal{L}}^\delta(t,x)$ be the forward and backward generators of the process X^δ:

$$\mathcal{L}^\delta(t,x) := \sum_{i=1}^n (A(t)x)_i \frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^n \left(\Sigma_V(t) + \delta^2 I_n\right)_{i,j} \frac{\partial^2}{\partial x_i \partial x_j},$$

$$\bar{\mathcal{L}}^\delta(t,x) := \sum_{i=1}^n (\bar{A}^\delta(t)x)_i \frac{\partial}{\partial x_i} + \frac{1}{2}\sum_{i,j=1}^n \left(\Sigma_V(t) + \delta^2 I_n\right)_{i,j} \frac{\partial^2}{\partial x_i \partial x_j},$$

where

$$\bar{A}^{\delta}(t) := -A(-t) - \left(\Sigma_V(-t) + \delta^2 I_n\right) P^{\delta}(-t)^{-1}, \tag{4.4}$$

and P^{δ} is the covariance matrix process of X^{δ}. Let $\mathcal{L}$ $(= \mathcal{L}^0)$ and $\bar{\mathcal{L}}$ $(= \bar{\mathcal{L}}^0)$ be the forward and backward generators of the original signal X. The following lemma provides some justification for the definition that follows, and it establishes some facts about the processes $(X^{\delta},\, \delta \in (0,1])$.

LEMMA 4.1.

$$\lim_{\delta \downarrow 0} \sup_{t \in [-T,T]} |X_t - X_t^{\delta}| = 0 \quad a.s., \tag{4.5}$$

$$\lim_{\delta \downarrow 0} \sup_{t \in [-T,T]} \|P(t) - P^{\delta}(t)\| = 0, \tag{4.6}$$

and, for any $f \in C_b^2(\mathbb{R}^n;\, \mathbb{R})$,

$$\begin{aligned} &\lim_{\delta \downarrow 0} \sup_{t,x} \left|(\mathcal{L}f)(t,x) - \left(\mathcal{L}^{\delta} f\right)(t,x)\right| = 0, \\ &\lim_{\delta \downarrow 0} \sup_{t,x} \left|\left(\bar{\mathcal{L}} f\right)(t,x) - \left(\bar{\mathcal{L}}^{\delta} f\right)(t,x)\right| = 0. \end{aligned} \tag{4.7}$$

Proof. The difference, $X - X^{\delta}$, satisfies a simple linear equation and can be represented by the following convolution integral:

$$X_t - X_t^{\delta} = \delta \int_{-T}^{t} \begin{bmatrix} I_n & 0 \end{bmatrix} \Theta(s,t) \begin{bmatrix} I_n \\ 0 \end{bmatrix} dB_s,$$

where Θ is as defined in (3.8). Equations (4.5) and (4.6) readily follow from this.

It now follows from (4.6) that $\sup_t \|\bar{A}(t) - \bar{A}^{\delta}(t)\|$ converges to zero, and this establishes (4.7). □

DEFINITION 4.2. *For each* $t \in (-T, T)$, *the* rate of entropy production *of the process* X *at time* t *is*

$$\begin{aligned} R_X(t) := \lim_{\delta \downarrow 0} \lim_{\epsilon \downarrow 0} \epsilon^{-1} \mathbf{E} h\left(\pi^{X,\delta}_{t+\epsilon|t} \,|\, \pi^{X,\delta}_{t-\epsilon|t}\right) \quad &\text{if the limit in } \delta \text{ exists}, \\ +\infty \quad &\text{otherwise}, \end{aligned} \tag{4.8}$$

where, for each $\delta \in (0,1]$ *and each* $\epsilon \in (0, \max(T-t,\, T+t)]$, $\pi^{X,\delta}_{t+\epsilon|t}$ *and* $\pi^{X,\delta}_{t-\epsilon|t}$ *are regular* X_t^{δ}*-conditional distributions of* $X^{\delta}_{t+\epsilon}$ *and* $X^{\delta}_{t-\epsilon}$, *respectively.*

The following proposition shows that this definition is consistent with that in (4.1), and evaluates $R_X(t)$.

PROPOSITION 4.3.

(i) *For any* $t \in (-T, T)$,

$$\begin{aligned} R_X(t) = \lim_{\delta \downarrow 0} \frac{1}{2} \mathrm{tr}\left((A(t) - \bar{A}^{\delta}(-t))' \left(\Sigma_V(t) + \delta^2 I_n\right)^{-1} (A(t) - \bar{A}^{\delta}(-t)) P^{\delta}(t) \right) \\ \text{if the limit exists}, \\ +\infty \quad \text{otherwise}. \end{aligned} \tag{4.9}$$

(ii) *For any* $t \in (-T, T)$ *such that* $|\Sigma_V(t)| \neq 0$,

$$R_X(t) = \frac{1}{2} \mathrm{tr}\left((A(t) - \bar{A}(-t))' \Sigma_V(t)^{-1} (A(t) - \bar{A}(-t)) P(t) \right). \tag{4.10}$$

Proof. $\pi^{X,\delta}_{t+\epsilon|t}$ and $\pi^{X,\delta}_{t-\epsilon|t}$ are nondegenerate Gaussian distributions on $\mathbb{R}^n$, with means $m^{+\delta}(\epsilon)$ and $m^{-\delta}(\epsilon)$, and covariance matrices $P^{+\delta}(\epsilon)$ and $P^{-\delta}(\epsilon)$, satisfying

$$\begin{aligned}
m^{+\delta}(\epsilon) &= X^\delta_t + \int_t^{t+\epsilon} A(s)m^{+\delta}(s)\,ds,\\
m^{-\delta}(\epsilon) &= X^\delta_t + \int_t^{t+\epsilon} \bar{A}^\delta(s-2t)m^{-\delta}(s)\,ds,\\
P^{+\delta}(\epsilon) &= \int_t^{t+\epsilon} \left(A(s)P^{+\delta}(s) + P^{+\delta}(s)A(s)' + \Sigma_V(s) + \delta^2 I_n\right)\,ds,\\
P^{-\delta}(\epsilon) &= \int_t^{t+\epsilon} \left(\bar{A}^\delta(s-2t)P^{-\delta}(s) + P^{-\delta}(s)\bar{A}^\delta(s-2t)' + \Sigma_V(2t-s) + \delta^2 I_n\right)\,ds.
\end{aligned}$$

Since Σ_V is Lipschitz continuous, it follows that, for each $\delta \in (0,1]$,

$$\begin{aligned}
P^{+\delta}(\epsilon) &= (\Sigma_V(t) + \delta^2 I_n)\epsilon + O(\epsilon^2),\\
P^{-\delta}(\epsilon) &= (\Sigma_V(t) + \delta^2 I_n)\epsilon + O(\epsilon^2),\\
P^{-\delta}(\epsilon)^{-1}P^{+\delta}(\epsilon) &= I_n + O(\epsilon),
\end{aligned}$$

and

$$\log\left(P^{-\delta}(\epsilon)^{-1}P^{+\delta}(\epsilon)\right) = P^{-\delta}(\epsilon)^{-1}P^{+\delta}(\epsilon) - I_n + O(\epsilon^2).$$

From these expressions it follows that

$$\begin{aligned}
h\left(\pi^{X,\delta}_{t+\epsilon\,|\,t}\,|\,\pi^{X,\delta}_{t-\epsilon|t}\right) &= -\frac{1}{2}\log\left(\left|P^{-\delta}(\epsilon)^{-1}P^{+\delta}(\epsilon)\right|\right) + \frac{1}{2}\mathrm{tr}\left(P^{-\delta}(\epsilon)^{-1}P^{+\delta}(\epsilon) - I_n\right)\\
&\quad + \frac{1}{2}\left(m^{+\delta}(\epsilon) - m^{-\delta}(\epsilon)\right)' P^{-\delta}(\epsilon)^{-1}\left(m^{+\delta}(\epsilon) - m^{-\delta}(\epsilon)\right)\\
&= \frac{1}{2}\left(m^{+\delta}(\epsilon) - m^{-\delta}(\epsilon)\right)' P^{-\delta}(\epsilon)^{-1}\left(m^{+\delta}(\epsilon) - m^{-\delta}(\epsilon)\right) + O(\epsilon^2)\\
&= \frac{1}{2}\left(X^\delta_t\right)'\left(A(t) - \bar{A}^\delta(-t)\right)'\left(\Sigma_V(t) + \delta^2 I_n\right)^{-1}\left(A(t) - \bar{A}^\delta(-t)\right)X^\delta_t\epsilon\\
&\quad + r(X^\delta_t, \delta, \epsilon)\epsilon^2,
\end{aligned}$$

where $r(X^\delta_t, \delta, \epsilon)$ is quadratic in the components of X^δ_t with coefficients that are uniformly bounded in (δ, ϵ). So

$$\begin{aligned}
&\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}h\left(\pi^{X,\delta}_{t+\epsilon|t}\,|\,\pi^{X,\delta}_{t-\epsilon|t}\right)\\
&\qquad = \frac{1}{2}\mathrm{tr}\left(\left(A(t) - \bar{A}^\delta(-t)\right)'\left(\Sigma_V(t) + \delta^2 I_n\right)^{-1}\left(A(t) - \bar{A}^\delta(-t)\right)P^\delta(t)\right),
\end{aligned}$$

which proves part (i). Part (ii) follows from this and (4.6). □

Remark 4.1. $R_X(t)$ measures the *degree of time-asymmetry* of the process X at time t. Imagine a game in which one player secretly samples X at the three times $t-\epsilon$, t, and $t+\epsilon$, tosses a coin, reverses the time sequence of the samples if "heads" occurs, and then shows the other player the sequence, asking whether or not it has been reversed. $R_X(t)$ is a measure of the average degree of ease with which the second player could answer correctly.

Remark 4.2. If $\Sigma_V(t)$ is singular, then it is possible (but not necessary) for $R_X(t)$ to be infinite. Consider, for example, the time-homogeneous case in which

$$A = \begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix} \quad \text{and} \quad \Sigma_V = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

This has invariant distribution $N(0, P_{SS})$ with covariance matrix

$$P_{SS} = \frac{1}{24} \begin{bmatrix} 7 & 2 \\ 2 & 1 \end{bmatrix}.$$

It can be shown that if X_t has this distribution (for example), then $R_X(t)$ is infinite. Furthermore, the second player in the game described above could determine whether or not the sequence of samples had been reversed with probability one by comparing the signs of $X_{t,1} - 2X_{t,2}$ and $X_{t+\epsilon,2} - X_{t-\epsilon,2}$ for sufficiently small ϵ. Even though this example is hypoelliptic, this is not enough to make $R_X(t)$ finite.

Remark 4.3. The rate of entropy production of the time-reversed process Z^* at time t is the same as that of the forward process X at time $-t$.

Remark 4.4. R_X is *homeomorphism invariant*. If f is a continuous, one-to-one mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$, then it induces new distributions on $\mathbb{R}^n$ corresponding to $\pi^{X,\delta}_{t+\epsilon|t}$ and $\pi^{X,\delta}_{t-\epsilon|t}$. These can be used to define the rate of entropy production of the process $f(X_t)$ at time t. This is, of course, equal to $R_X(t)$.

Remark 4.5. $R_X(t)$ is nonnegative. It is zero if and only if $\bar{A}(-t) = A(t)$, in which case the dynamics of X are locally time symmetric at time t.

We now use the rates of entropy production of the joint process (X, Z), and those of the signal and filter processes, X and Z, to define an information-theoretic measure of the *interaction* between the latter two, which we shall call the *rate of interactive entropy production*. Crucial to the definition is the fact that X and Z are both *marginally* as well as jointly Markov. Like that of (marginal) entropy production, the definition of interactive entropy production relies on a relaxation argument. Let X^δ be as defined in (4.3), and let Y^δ be the following relaxed observation process:

$$Y^\delta_t = \xi + \zeta + \int_{-T}^{t} (\Sigma_W(s) + \delta^2 I_n) X^\delta_s \, ds + W_t + \delta \tilde{B}_t, \tag{4.11}$$

where $(\tilde{B}_t, \mathcal{F}_t)$ is a left-shifted n-vector Brownian motion, independent of V, W, and B. Let (Z^δ, Q^δ) be the information form of the Kalman–Bucy filter for X^δ given Y^δ, and let $\mathbb{P}^{M,\delta}$ be the associated product-of-marginals measure. (See the discussion preceding Lemma 2.1.)

The following lemma provides some justification for the definition that follows, and it establishes some facts about the processes $(Z^\delta, Q^\delta, \delta \in (0, 1])$.

LEMMA 4.4.

$$\lim_{\delta \downarrow 0} \sup_{t \in [-T,T]} |Z_t - Z^\delta_t| = 0 \quad a.s., \tag{4.12}$$

$$\lim_{\delta \downarrow 0} \sup_{t \in [-T,T]} \|Q(t) - Q^\delta(t)\| = 0, \tag{4.13}$$

and, for any $f \in C_b^2(\mathbb{R}^{2n};\mathbb{R})$,

$$\begin{aligned}\lim_{\delta\downarrow 0}\sup_{t,x}\left|\left(\mathcal{L}^J f\right)(t,x)-\left(\mathcal{L}^{J,\delta} f\right)(t,x)\right|&=0,\\ \lim_{\delta\downarrow 0}\sup_{t,x}\left|\left(\bar{\mathcal{L}}^J f\right)(t,x)-\left(\bar{\mathcal{L}}^{J,\delta} f\right)(t,x)\right|&=0,\end{aligned}\tag{4.14}$$

where $\mathcal{L}^{J,\delta}$ *and* $\bar{\mathcal{L}}^{J,\delta}$ *are the forward and backward generators of the joint process* (X^δ, Z^δ), *and* $\mathcal{L}^J$ $(=\mathcal{L}^{J,0})$ *and* $\bar{\mathcal{L}}^J$ $(=\bar{\mathcal{L}}^{J,0})$ *are the forward and backward generators of* (X,Z).

Proof. The difference $Q-Q^\delta$ satisfies the following equation:

$$\begin{aligned}\frac{d}{dt}\left(Q-Q^\delta\right)(t)=A(t)\left(Q-Q^\delta\right)(t)+\left(Q-Q^\delta\right)(t)A(t)'\\ -\left(Q+Q^\delta\right)(t)\Sigma_W(t)\left(Q-Q^\delta\right)(t)+\delta^2 Q^\delta(t)^2.\end{aligned}$$

Since $\sup_{t,\delta}\|Q^\delta(t)\|$ is bounded, there exist finite positive constants K_1 and K_2 such that

$$\frac{d}{dt}\|Q(t)-Q^\delta(t)\|\le K_1\|Q(t)-Q^\delta(t)\|+K_2\delta^2 \quad \text{for all } t\in[-T,T],$$

and so (4.13) follows from the Gronwall lemma. Equation (4.12) now follows from (3.1) and (4.6) by straightforward approximation. The fact that $P^\delta(t)$, $Q^\delta(t)$, and $\bar{A}^\delta(t)$ converge to $P(t)$, $Q(t)$, and $\bar{A}(t)$, respectively, establishes (4.14). ∎

Since the joint process (X^δ, Z^δ) is Markov under both $\mathbb{P}$ and $\mathbb{P}^{M,\delta}$, it has rates of entropy production ($R_J^\delta(t)$ and $R_M^\delta(t)$) under these measures. The difference $R_J^\delta(t)-R_M^\delta(t)$ isolates that part of $R_J^\delta(t)$ associated with the interaction of X^δ and Z^δ. This is the rationale for the following definition.

Definition 4.5. *For each* $t\in(-T,T)$, *the* rate of interactive entropy production *between the signal and filter processes* X *and* Z *at time* t *is*

$$R_I(t):=\lim_{\delta\downarrow 0}\left(\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}h\left(\pi^{J,\delta}_{t+\epsilon|t}\,|\,\pi^{J,\delta}_{t-\epsilon|t}\right)-\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}h\left(\pi^{M,\delta}_{t+\epsilon|t}\,|\,\pi^{M,\delta}_{t-\epsilon|t}\right)\right),\tag{4.15}$$

where $\pi^{J,\delta}_{t+\epsilon|t}$ *is a regular* (X_t^δ, Z_t^δ)*-conditional distribution of* $(X^\delta_{t+\epsilon}, Z^\delta_{t+\epsilon})$ *under* $\mathbb{P}$, $\pi^{J,\delta}_{t-\epsilon|t}$ *is a regular* (X_t^δ, Z_t^δ)*-conditional distribution of* $(X^\delta_{t-\epsilon}, Z^\delta_{t-\epsilon})$ *under* $\mathbb{P}$, *and* $\pi^{M,\delta}_{t+\epsilon|t}$ *and* $\pi^{M,\delta}_{t-\epsilon|t}$ *are corresponding regular conditional distributions under* $\mathbb{P}^{M,\delta}$.

Proposition 4.6. *For any* $t\in(-T,T)$,

$$R_I(t)=\dot{S}(t)+\dot{D}(t),\tag{4.16}$$

where S *and* D *are the information supply and dissipation processes of* (2.6).

Proof. It follows from Proposition 4.3 that

$$\begin{aligned}R_J^\delta(t)&:=\lim_{\epsilon\downarrow 0}\epsilon^{-1}\mathbf{E}h\left(\pi^{J,\delta}_{t+\epsilon|t}\,|\,\pi^{J,\delta}_{t-\epsilon|t}\right)\\ &=\frac{1}{2}\mathrm{tr}\Big((A_J^\delta(t)-UA_J^{\delta*}(-t)U)'(\Sigma_J(t)+\delta^2 I_{2n})^{-1}\\ &\qquad\times(A_J^\delta(t)-UA_J^{\delta*}(-t)U)P_J^\delta(t)\Big)\end{aligned}\tag{4.17}$$

and

$$R_M^\delta(t) := \lim_{\epsilon\downarrow 0} \epsilon^{-1}\mathbf{E}^M h\left(\pi^{M,\delta}_{t+\epsilon|t} \,|\, \pi^{M,\delta}_{t-\epsilon|t}\right)$$

$$= \frac{1}{2}\mathrm{tr}\Big((A_M^\delta(t) - UA_M^{\delta *}(-t)U)'(\Sigma_J(t) + \delta^2 I_{2n})^{-1} \times (A_M^\delta(t) - UA_M^{\delta *}(-t)U)P_M^\delta(t)\Big), \tag{4.18}$$

where A_J^δ, $A_J^{\delta *}$, A_M^δ, $A_M^{\delta *}$, and P_J^δ are the equivalents of A_J, A_J^*, A_M, A_M^*, and P_J for the relaxed problem, U is the $2n \times 2n$ transposition matrix

$$U = \begin{bmatrix} 0 & I_n \\ I_n & 0 \end{bmatrix}, \tag{4.19}$$

Σ_J is the joint quadratic covariation rate matrix of (2.10), and P_M^δ is the product-of-marginals covariance matrix process for the relaxed problem,

$$P_M^\delta = \begin{bmatrix} P^\delta & 0 \\ 0 & (Q^\delta)^{-1}(P^\delta - Q^\delta)(Q^\delta)^{-1} \end{bmatrix}. \tag{4.20}$$

Straightforward calculations now show that

$$R_J^\delta(t) - R_M^\delta(t) = \frac{1}{2}\mathrm{tr}\left((\Sigma_W(t) + \delta^2 I_n)Q^\delta(t)\right) + \frac{1}{2}\mathrm{tr}\left((\Sigma_V(t) + \delta^2 I_n)\left(Q^\delta(t)^{-1} - P^\delta(t)^{-1}\right)\right),$$

which, together with (2.11), (2.12), (4.6), and (4.13), completes the proof. □

Remark 4.6. R_I inherits the properties of R_X discussed in Remarks 4.1, 4.3, 4.4, and 4.5. It measures the time-asymmetry of the *interaction* between the signal and filter processes, and as shown by Proposition 4.6, this is entirely associated with the flow of information. The rate of interactive entropy production of the dual system at time $-t$ is the same as that of the original system at time t.

Remark 4.7. $R_I(t)$ is finite even if $R_J(t)$ and $R_M(t)$ are not. Consider the example system with signal as in Remark 4.2 and with the time-homogeneous observation covariance matrix $\Sigma_W = \Sigma_V$. When in their joint invariant distribution, the signal and filter processes are strongly locally time asymmetric in the sense that $R_X(t) = R_Z(t) = +\infty$. However, their interaction is not, in the sense that $R_I(t)$ is finite. The strong time-asymmetry of the joint process is associated with the *marginal* dynamics of X and Z and not with their interaction.

Remark 4.8. If $|\Sigma_V(t)| \neq 0$ and $|\Sigma_W(t)| \neq 0$, then

$$R_I(t) = R_J(t) - R_M(t) = R_J(t) - R_X(t) - R_Z(t), \tag{4.21}$$

where R_J and R_M are the (finite) rates of entropy production of the joint process (X, Z) under $\mathbb{P}$ and $\mathbb{P}^M$, respectively, and R_Z is the rate of entropy production of the filter process, Z.

5. Connections with statistical mechanics. In [12] a statistical mechanical analogy for the signal is developed that depends on (S1), (S2), and an additional condition, as follows:

(S3) A does not depend on t, and all of its eigenvalues have negative real parts. Under these conditions X has the unique invariant distribution, $N(0, P_{SS})$, where

$$AP_{SS} + P_{SS}A' + \Sigma_V = 0. \tag{5.1}$$

The analogy involves an abstract statistical mechanical system whose *state* is a probability measure on $(\mathbb{R}^n, \mathcal{B}^n)$. In fact the state at time t is the distribution of X_t, $N(0, P(t))$. In state μ, the system has *energy* $\mathcal{E}(\mu)$, *entropy* $\mathcal{S}(\mu)$, and *free energy* $\mathcal{F}(\mu)$, where

$$\begin{aligned} \mathcal{E}(\mu) &:= \int H(x)\mu(dx), \\ \mathcal{S}(\mu) &:= -h(\mu \,|\, \mathrm{Leb}), \\ \mathcal{F}(\mu) &:= \mathcal{E}(\mu) - \mathcal{S}(\mu); \end{aligned} \tag{5.2}$$

h is the relative entropy of (4.2), Leb is the Lebesgue measure, and H is the following *Hamiltonian*:

$$H(x) := \tfrac{1}{2} x' P_{SS}^{-1} x. \tag{5.3}$$

It is shown in [12] that the free energy $\mathcal{F}$ is minimized by the *stationary state*, $N(0, P_{SS})$, and that, for any state process $(\mu_t,\, t \in [-T, T])$ satisfying the Kolmogorov forward (Fokker–Planck) equation associated with X, the free energy is nonincreasing. In particular this is true of the state process $(N(0, P(t)),\, t \in [-T, T])$. The state evolves toward the minimum free energy state, $N(0, P_{SS})$, in such a way that the free energy is nonincreasing. The introduction of a *heat bath*, with which the abstract statistical mechanical system interacts, connects this property with the second law of thermodynamics. (See [12].)

In the statistical mechanics literature, stationary states are referred to as stationary *equilibrium* states, or stationary *nonequilibrium* states, according to the net *flow* of some quantity (often energy) in the state, an equilibrium state being one for which this flow is zero. (See, for example, [1], [5], or [8].) Stationary nonequilibrium states of physical systems are typically associated with strictly increasing entropy. For example, if a cylinder of gas is heated at one end and cooled at the other, a stationary *nonequilibrium* state is eventually set up in the gas, in which heat flows from the hot end to the cool end—heat flowing from a hot object to a cool object causes entropy to increase in the universe. For this reason, nonequilibrium states have associated nonzero rates of entropy production. For the abstract system associated with the signal process, X, the rate of entropy production is defined in Definition 4.2. This is zero (and $N(0, P_{SS})$ is a stationary *equilibrium* state) if and only if X is *self-adjoint* in the sense that $P_{SS}A' = AP_{SS}$. Otherwise $N(0, P_{SS})$ is a stationary *nonequilibrium* state.

The statistical mechanical analogy of [12] is also applied there to the joint process (X, Z), again under the special conditions (S1)–(S3). This describes a "joint" system with Hamiltonian

$$H_J(x, z) := \frac{1}{2} \begin{bmatrix} x' & z' \end{bmatrix} P_{J,SS}^{-1} \begin{bmatrix} x \\ z \end{bmatrix}, \tag{5.4}$$

where $P_{J,SS}$ is the stationary covariance matrix, to which $P_J(t)$ of (3.9) converges. This joint system is shown to exhibit, when in its stationary state $N(0, P_{J,SS})$, an

energy flow around a loop comprising two subsystems and a heat bath. These subsystems are termed the "conditional signal" and the "filter" systems. They are characterized by the processes $X - QZ$ and Z in the same way that the original system is characterized by X, and the joint system by (X, Z). The subsystems are "physically distinct" in the sense that (i) their characterizing processes, $X - QZ$ and Z, are *marginally* Markov; (ii) the energy of the joint system is the sum of the energies of the subsystems; and (iii) the entropy of the joint system is the sum of the entropies of the subsystems. Property (i), in particular, shows that it is possible to identify rates of entropy production for each of the two subsystems as well as for the joint system. These rates correspond to nonequilibrium statistical mechanics occurring *within* the subsystems. By subtracting them from that of the joint system, a rate of entropy production associated purely with the interaction between the subsystems is isolated. Of course, this is the rate of interactive entropy production, defined in section 4. The energy flow in the stationary nonequilibrium state of this analogy is, therefore, directly connected with the information flows of the Kalman–Bucy filter. The reader is referred to [12] for a complete discussion of this analogy.

These statistical mechanical analogies can be extended to the more general setting of this paper in a number of ways. For example, one can regard the time-varying nature of A and Σ_V as being the effect of external *forces* applied to the abstract system associated with X. Suppose that (S2) holds and that

(S3′) for each $t \in [-T, T]$, all eigenvalues of $A(t)$ have negative real parts.

Then we can define a "stationary" covariance matrix, $P_{SS}(t)$, for each t as the solution of the following equation:

$$A(t)P_{SS}(t) + P_{SS}(t)A(t)' + \Sigma_V(t) = 0. \tag{5.5}$$

The time-varying nature of P_{SS} now gives rise to a *time-varying* Hamiltonian

$$H(t, x) := \tfrac{1}{2}x'P_{SS}(t)^{-1}x. \tag{5.6}$$

If the energy, entropy, and free energy functionals of the abstract system remain as defined in (5.2), but with this time-varying Hamiltonian, then the minimum free energy state, $N(0, P_{SS}(t))$, changes with time. If the external force remains constant on some interval $[s, t]$, then the state of the system will evolve toward $N(0, P_{SS}(s))$ on this interval, and will do so in a manner that gives rise to a nonincreasing free energy. On the other hand, if the external force continues to change, then the state will "attempt to track" the moving target, $N(0, P_{SS}(t))$. In this case, the abstract system will rarely, if ever, be in its minimum free energy state.

If the changes in A and Σ_V occur on a time scale that is large in comparison to the magnitude of the eigenvalues of A, then the system state may be close to the minimum free energy state at all times. The changing nature of $\mathcal{E}(N(0, P(t)))$ is then partly due to interaction with the heat bath (short time scale), and partly due to energy transferred to or from the system by the time-varying force (long time scale). This is analogous to a cylinder of gas (with identical or different end temperatures) whose volume is changed slowly by the movement of a piston.

Of course, the abstract system associated with the joint process (X, Z) can also be thought of in this way. Even though, in general, the energy of this system will be time-varying, it is still possible to identify a circular flow of energy around a loop of subsystems. As in the special case of [12], this is directly connected with the flow of information between the signal and filter processes.

The reader interested in further discussion of statistical mechanical analogies is referred to [12] and [14].

6. Conclusions. This paper has developed dual linear filtering problems and has defined and related information flows for them. The information processing carried out by the dual filter has been shown to be a time-reversed version of that carried out by the original. The information flow rates between signal and filter processes are intimately connected with entropy production in nonequilibrium statistical mechanical systems. This interpretation is developed fully elsewhere (see [12] and [14]).

The derivation of dual filters by the techniques of this paper is not restricted to linear models. The same approach can be taken with nonlinear filters. (See [13].) The dual filters thus obtained are finite-dimensional and typically nonlinear. Of course this does not herald the "El Dorado" of finite-dimensional nonlinear filtering because the finite-dimensionality of these particular filters depends on a very special relation existing between the distribution and dynamics of the signal and observation processes.

Acknowledgment. The author would like to thank Professor Sanjoy Mitter of the Laboratory for Information and Decision Systems (LIDS) at MIT for suggesting this subject area as one fruitful for research, for providing encouragement and support during a number of visits made by the author to LIDS, and for the many engaging discussions that took place during those visits.

REFERENCES

[1] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim, *Macroscopic fluctuation theory for stationary non-equilibrium states*, J. Statist. Phys., 107 (2002), pp. 635–675.

[2] M. H. A. Davis, *Linear Estimation and Stochastic Control*, Chapman and Hall, New York, 1977.

[3] T. E. Duncan, *On the calculation of mutual information*, SIAM J. Appl. Math., 19 (1970), pp. 215–220.

[4] X. Feng, K. A. Loparo, and Y. Fang, *Optimal state estimation for stochastic systems: An information theoretic approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 771–785.

[5] G. Gallavotti, *Statistical Mechanics—A Short Treatise*, Springer-Verlag, Berlin, 1999.

[6] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice–Hall, Upper Saddle River, NJ, 2000.

[7] P. Kalata and R. Priemer, *Linear prediction, filtering and smoothing: An information-theoretic approach*, Inform. Sci., 17 (1979), pp. 1–14.

[8] J. L. Lebowitz and H. Spohn, *A Gallavotti-Cohen type symmetry in the large deviation functional for stochastic dynamics*, J. Statist. Phys., 95 (1999), pp. 333–366.

[9] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes* I. *General Theory*, Springer-Verlag, New York, 1977.

[10] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes* II. *Applications*, Springer-Verlag, New York, 1978.

[11] S. K. Mitter and N. J. Newton, *A variational approach to nonlinear estimation*, SIAM J. Control Optim., 42 (2003), pp. 1813–1833.

[12] S. K. Mitter and N. J. Newton, *Information and entropy flow in the Kalman-Bucy filter*, J. Statist. Phys., 118 (2005), pp. 145–176.

[13] N. J. Newton, *Dual nonlinear filters and interactive entropy production*, SIAM J. Control Optim., submitted.

[14] N. J. Newton, *The interactive statistical mechanics of nonlinear filters*, submitted.

[15] E. Mayer-Wolf and M. Zakai, *On a formula relating the Shannon information to the Fisher information for the filtering problem*, in Filtering and Control of Random Processes, Lecture Notes in Control and Inform. Sci. 61, H. Korezlioglu, G. Mazziotto, and S. Szpirglas, eds., Springer, Berlin, 1984, pp. 164–171.

[16] Y. Tomita, S. Omatu, and T. Soeda, *An application of the information theory to filtering problem*, Inform. Sci., 11 (1976), pp. 13–27.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 1017–1038

REGULARITY OF THE EIKONAL EQUATION WITH NEUMANN BOUNDARY CONDITIONS IN THE PLANE: APPLICATION TO FRONTS WITH NONLOCAL TERMS*

PIERRE CARDALIAGUET† AND CLAUDIO MARCHI‡

Abstract. The first part of the paper is devoted to length estimates of the boundary of the reachable set for the plane and state constrained controlled system $x'(t) = c(t, x(t))b(t)$ (where $|b(t)| \leq 1$ a.e.). This study is motivated in the second part by the analysis of dislocation dynamics, which can be modeled as a curve $\Gamma(t)$ moving in an open set $\Omega \subset \mathbb{R}^2$ according to some nonlocal law with Neumann boundary conditions on $\partial\Omega$. The length estimates of the first part play a crucial role in the proof of the existence and uniqueness of a *viscosity* solution for this model.

Key words. dislocation dynamics, eikonal equation, Hamilton–Jacobi equations, Neumann boundary condition, discontinuous viscosity solutions, nonlocal equations

AMS subject classifications. 35F25, 35D05

DOI. 10.1137/050635894

1. Introduction. In this paper, we study the dynamics of a curve $\Gamma(t) \subset \Omega$ which moves according to the following nonlocal law:

$$V_{t,x} = \bar{c}_1(t,x) + I\left(\mathbf{1}_{K(t)}\right)(t,x), \tag{1}$$

where Ω is an open subset of $\mathbb{R}^2$, $V_{t,x}$ is the normal velocity at the point x at time t, $K(t)$ is the set enclosed by the curve $\Gamma(t)$, the map I is defined by

$$I(\rho)(t,x) := \int_\Omega \bar{c}_2(t,x,y)\rho(y)\,dy \qquad \forall \rho \in \mathbb{L}^1(\Omega),$$

and $\bar{c}_2$ solves the *elasticity equation* (see [2, 3]). Moreover, the curve Γ can touch the boundary $\partial\Omega$ only orthogonally. Our investigation originated from the so-called *phase field* model of dislocation, introduced by Rodney, Le Bouar, and Finel [13]: a dislocation is a line of crystal defect, and it moves on its slip plane with a normal velocity proportional to the Peach–Koehler force acting on the line:

$$V_{t,x} = \bar{c}_1(t,x) + \bar{c}_2 \star_x \mathbf{1}_{K(t)}(t,x), \tag{2}$$

where $\bar{c}_1$ is the contribution given by an exterior field, while $\bar{c}_2 \star_x \mathbf{1}_{K(t)}$ (the convolution is done only w.r.t. the space variable $x \in \mathbb{R}^2$) is a force created by the dislocation itself. Equations of this type have been studied only in the whole space: one of the main issues of this note is that we reduced the slip plane to a open subset Ω and addressed the dynamics' modifications given by the presence of $\partial\Omega$.

*Received by the editors July 13, 2005; accepted for publication (in revised form) April 24, 2006; published electronically July 31, 2006.

http://www.siam.org/journals/sicon/45-3/63589.html

†Université de Bretagne Occidentale, Laboratoire de Mathématiques UMR 6205, 6 Av. Le Gorgeu, BP 809, 29285 Brest, France (Pierre.Cardaliaguet@univ-brest.fr).

‡Dipartimento di Matematica, Università della Calabria, Ponte P. Bucci 30b, Arcavacata di Rende (CS), 87036 Italy (marchi@mat.unical.it). This author was partially supported by the project RTN "Evolution Equations for Deterministic and Stochastic Systems" HPRN-CT-2002-00281, financed by the European Community.

For dynamics in the whole space, let us recall that if we set

$$\rho(t,x) = \mathbf{1}_{K(t)}(x) \equiv \begin{cases} 1 & \text{if} \quad x \in K(t), \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

then ρ is a (discontinuous) viscosity solution to the following Cauchy problem for a nonlocal Hamilton–Jacobi equation

$$\begin{cases} \frac{\partial \rho}{\partial t} = (\bar{c}_1 + \bar{c}_2 \star_x \rho)\,|\nabla \rho| & \text{in} \quad (0,+\infty) \times \mathbb{R}^2, \\ \rho(0,x) = \mathbf{1}_{K_0}(x) & \text{on} \quad \mathbb{R}^2, \end{cases}$$

where K_0 is the starting set. For $\bar{c}_2 \geq 0$, the equation fulfills the *inclusion principle*, so the existence and uniqueness of the solution follow from a result by Cardaliaguet [9]; from a physical point of view, however, this case is not interesting because $\bar{c}_2$ needs to have zero mean. For $\bar{c}_2$ changing sign, Alvarez, Hoch, Le Bouar, and Monneau [2, 3] proved the short time existence and uniqueness of the solution, provided that the dislocation starts from a Lipschitz graph. Finally, in a forthcoming paper, Alvarez, Cardaliaguet, and Monneau [1] addressed the existence and uniqueness of the solution in an arbitrary time interval under the assumption that the starting set K_0 fulfills the *interior sphere condition* of radius $r > 0$ (i.e., for every $x \in \partial K_0$, there exists a unitary vector p such that $B(x - rp, r) \subseteq K_0$) and that $\bar{c}_i$ fulfills (beside some regularity assumptions)

$$\bar{c}_1(t,x) \geq \|\bar{c}_2(t,\cdot)\|_{\mathbb{L}^1(\mathbb{R}^2)} \qquad \forall (t,x) \in (0,+\infty) \times \mathbb{R}^2$$

(which assures that the dislocation is noncontracting: $\bar{c}_1 + \bar{c}_2 \star_x \mathbf{1}_K \geq 0$ for every Borel set K). Actually, these results also apply when the space is $\mathbb{R}^N$, even if the only interesting case for the phase field model is $N = 2$.

In our study of (1), the function ρ defined by (3) is a discontinuous solution to the following problem with Neumann boundary conditions in the viscosity sense:

$$\begin{cases} \frac{\partial \rho}{\partial t} = \big(\bar{c}_1 + I(\rho)\big)|\nabla \rho| & \text{in} \quad (0,+\infty) \times \Omega, \\ \rho(0,x) = \mathbf{1}_{K_0}(x) & \text{on} \quad \Omega, \\ \frac{\partial \rho}{\partial \nu} = 0 & \text{on} \quad (0,+\infty) \times \partial\Omega \end{cases} \tag{4}$$

(ν is the outward unity normal to $\partial\Omega$). One of the main issues of this paper is that we establish the existence and uniqueness (in a sense that will be specified in section 4) of the solution to the above initial boundary value problem under some interior sphere condition on the initial set K_0 and a condition on $\bar{c}_1$ and $\bar{c}_2$ ensuring that the dislocation is expanding.

As in [1], we shall use the Banach fixed point theorem for the mapping Φ on $C^0([0,T], \mathbb{L}^1(\Omega))$ that associates to any $\rho^0 \in C^0([0,T], \mathbb{L}^1(\Omega))$ the unique discontinuous solution to

$$\begin{cases} \frac{\partial \rho}{\partial t} = c_{\rho^0}|\nabla \rho| & \text{in} \quad (0,T) \times \Omega, \\ \rho(0,x) = \mathbf{1}_{K_0}(x) & \text{on} \quad \Omega, \\ \frac{\partial \rho}{\partial \nu} = 0 & \text{on} \quad (0,T) \times \partial\Omega \end{cases}$$

with $c_{\rho^0} := \big(\bar{c}_1 + I(\rho^0)\big)$. In order to prove that Φ is a well-defined contraction, we use the *representation formula*: for $0 \leq \rho^0 \leq 1$, the level set $\{\Phi(\rho^0)(t,\cdot) = 1\}$ coincides with the *reachable set* (for the precise definition, see formula (7) below) of the following control problem with reflecting boundary:

$$\begin{cases} y'(s) = c_{\rho^0}(s, y(s))\big[b(s) - \langle b(s); \nu(y(s))\rangle \mathbf{1}_{\partial\Omega}(y(s))\nu(y(s))\big] & \text{a.e. in } [0,t], \\ y(0) \in K_0, \quad y(s) \in \overline{\Omega} \qquad \forall s \in [0,t], \end{cases} \tag{5}$$

where the control $b : [0,t] \to B(0,1)$ is a measurable map and $\nu(y(s))$ is the outward unitary normal to $\partial\Omega$ in $y(s)$; namely, $z \in \{\Phi(\rho^0)(t,\cdot) = 1\}$ if and only if there exists a measurable map $b : [0,t] \to B(0,1)$ and a solution y to (5) with $y(t) = z$.

For proving that Φ is a contraction, we are led to estimate the Lebesgue measure of the reachable set and the Hausdorff measure of its boundary. In the case $\Omega = \mathbb{R}^N$ treated in [1], these estimates were based on the fact that the interior sphere assumption stated on the initial set K_0 propagates: the reachable set of the (unconstrained) control system satisfies the interior sphere condition for all times (see [7]). This gives a bound from above for the curvature of the reachable set at each time t, from which an estimate of the length of its boundary is derived. In our case, when $\Omega \neq \mathbb{R}^2$, the interior sphere condition does *not* propagate: as we show below, the curvature of the reachable set can blow up near the boundary of Ω. The main contribution of this paper is to explain how to overcome this difficulty, at least in dimension 2. In particular we show that the boundary of the reachable set has bounded length, although not necessarily finite curvature. From this we deduce estimates on the Lebesgue measure of the reachable set and the existence and uniqueness for problem (4).

The paper is organized as follows: In the rest of this introduction we give an example in which many features of our problem appear, and we set some notation. Section 2 is devoted to the estimates of the volume and the boundary of the reachable set of problem (5). The existence and uniqueness of the solution to (4) (and, in particular, the dynamics of Γ) are established in section 4 (using some technical estimates stated in section 3). Finally, the appendix is devoted to the proof of an extension of the Pontryagin maximum principle (stated in section 2).

1.1. Toy example. We give here an elementary example showing that the reachable set of a reflected controlled system of the form (5) can have a curvature which blows up at some point.

Consider $c \equiv 1$, $\Omega = \mathbb{R}^2 \backslash B(0,1)$, and $K_0 = B\left((-4,0),1\right)$. In this case the reachable set $\mathcal{R}(t)$ for the reflected control problem (5) can be explicitly computed. For this we note that $\mathcal{R}(t)$ is just the set of points $x \in \overline{\Omega}$ which is at a *geodesic distance* (in Ω) from K_0 not larger than t.

Hence for $t \in [0,2]$, $\mathcal{R}(t) = \bar{B}\big((-4,0),1+t\big)$ (see Figure 1). For $t \in [2,\sqrt{15}]$, $\mathcal{R}(t) = \bar{B}\big((-4,0),1+t\big)\backslash B(0,1)$ (see Figure 2).

After $\bar{t} = \sqrt{15} - 1$, some geodesics need to bend around $\partial\Omega$ for reaching the points in the "cone of shade" due to the presence of $\mathcal{C}\Omega$ (see Figure 3). In this case, all the points in this cone of shade are reached by geodesics passing through either the point $P_1 = (-1/4, -\sqrt{15}/4)$ or the point $P_2 = (-1/4, \sqrt{15}/4)$. These geodesics are made up of a straight line from $(-4,0)$ to P_1 or P_2, an arc on $\partial\Omega$, and again a straight line tangent to $\partial\Omega$. For the sake of simplicity, let us consider the coordinate system centered in $(0,0)$ with the ξ-axis and η-axis given, respectively, by the vector $(-1/4, -\sqrt{15}/4)$ and $(\sqrt{15}/4, -1/4)$ (i.e., in the new system $P_1 = (1,0)$). For $t \in (\bar{t}, 4)$, the coordinates of the point of $\partial\mathcal{R}(t)$ reached by a geodesic passing through P_1 and remaining on $\partial\Omega$ from the time $\bar{t}$ to the time s (for some $s \in [\bar{t}, t]$) are given by

$$\begin{cases} \xi(s) = \cos(s-\bar{t}) - (t-s)\sin(s-\bar{t}), \\ \eta(s) = \sin(s-\bar{t}) + (t-s)\cos(s-\bar{t}). \end{cases}$$

By standard arguments, one can verify that the curvature of $\partial\mathcal{R}(t)$ at that point is $(t-s)^{-2}$; in particular, as $s \to t$, it tends to $+\infty$.

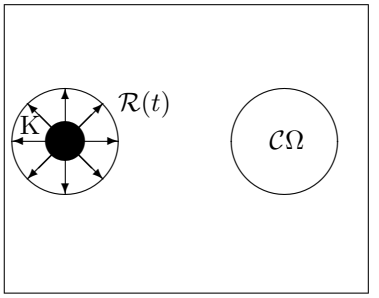


FIG. 1.

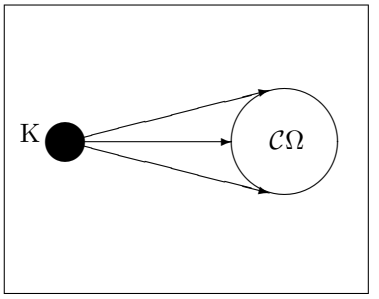


FIG. 2.

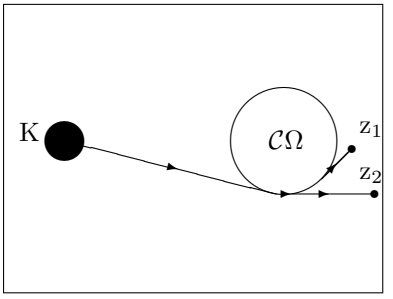


FIG. 3.

Notation. For each $K \subset \mathbb{R}^2$, $\partial_{\mathbb{R}^2} K$ and ∂K stand for its boundary in the natural topology of $\mathbb{R}^2$ and, respectively, in the induced topology of Ω; in other words, we have

$$\partial K := \partial_{\mathbb{R}^2} K \cap \Omega.$$

Moreover, $\mathcal{C}K := \mathbb{R}^2 \backslash K$, and $\mathbf{1}_K$ is the *characteristic* function of K; i.e., $\mathbf{1}_K(x) = 1$ if $x \in K$, and $\mathbf{1}_K(x) = 0$ if $x \in \mathcal{C}K$. For all sets K and H in $\mathbb{R}^2$, $d^{\mathcal{H}}(K, H)$ stands for the Hausdorff distance between the sets K and H:

$$d^{\mathcal{H}}(K, H) := \max\left\{ \max_{s \in K} \operatorname{dist}(s, H),\ \max_{u \in H} \operatorname{dist}(u, K) \right\},$$

where $\operatorname{dist}(s, H) := \min\{|s - u| : u \in H\}$. We denote by $B(p, r)$ the ball centered in $p \in \mathbb{R}^2$ with radius $r > 0$; B is the abridged notation of $B(0, 1)$. For every $x, y \in \Omega$,

$d^\Omega(x,y)$ is their *geodesic distance*; namely,

$$d^\Omega(x,y) := \inf\left\{ l \mid \begin{array}{c} \exists \text{ an AC arc, contained in } \Omega, \text{ with length } l \\ \text{and extremities in } x \text{ and } y \end{array} \right\}.$$

For a function ρ, ρ^* (resp., ρ_*) denotes the upper semicontinuous envelope (resp., the lower semicontinuous envelope) of ρ.

2. Main estimates for the reachable set. In this section we consider the reachable set for the reflected control problem

$$\text{(6)} \quad \begin{cases} y'(t) = c(t,y(t))\big[b(t) - \langle b(t), \nu(y(t))\rangle \mathbf{1}_{\partial\Omega}(y(t))\nu(y(t))\big] & \text{a.e. in } (0,T), \\ y(t) \in \overline{\Omega} \qquad \forall t \geq 0, \\ y(0) = y_0 \in K_0, \end{cases}$$

where $b : (0,T) \to B$ is a time measurable control and $\nu(y(t))$ is the outward unit normal to $\partial\Omega$ at $y(t)$. Let us recall that if Ω has a $\mathcal{C}^2$ boundary and c is sufficiently regular, then, for any measurable control $b : (0,T) \to B$ and any initial position $y_0 \in K_0$, (6) has a unique solution.

The reachable set at time $t > 0$ for this reflected control system is given by

$$\text{(7)} \qquad \mathcal{R}(t) = \{x \in \overline{\Omega} \mid \exists y \text{ solution to (6) on } [0,t] \text{ with } y(t) = x\}.$$

Our aim is to show that the boundary of this set has a finite Hausdorff measure and that its volume depends in a Lipschitz continuous way on the time.

For doing this we will need the following conditions on c, K_0, and Ω:

$$\text{(8)} \quad \begin{cases} \text{(i)} & c \text{ is continuous and derivable w.r.t. the second variable,} \\ \text{(ii)} & 0 < m \leq c(t,x) \leq M \qquad \forall (t,x) \in \mathbb{R}^+ \times \Omega, \\ \text{(iii)} & |c(t,x_1) - c(t,x_2)| \leq L_0|x_1 - x_2| \qquad \forall (t,x_1),(t,x_2) \in \mathbb{R}^+ \times \Omega, \\ \text{(iv)} & |D_x c(t,x_1) - D_x c(t,x_2)| \leq L_1|x_1 - x_2| \qquad \forall (t,x_1),(t,x_2) \in \mathbb{R}^+ \times \Omega \end{cases}$$

for some positive constants m, M, L_0, and L_1.

The initial set $K_0 \subset \Omega$ has to satisfy the following:

$$\text{(9)} \qquad K_0 \text{ is compact and fulfills the interior sphere condition with radius } r;$$

namely, for any $x \in \partial K_0$, there is some $p \in \mathbb{R}^2$, $|p| = 1$, with $B(x - rp, r) \subset K_0$. On Ω we assume that

$$\text{(10)} \qquad \begin{cases} \Omega \text{ is connected, } \partial\Omega \text{ is of class } \mathcal{C}^2, \text{ and } \mathcal{C}\Omega \text{ fulfills the} \\ \qquad \text{interior sphere condition with radius } r_1 > 0. \end{cases}$$

We also require that

$$\text{(11)} \qquad \exists\, k_0 > 0 \quad \text{such that} \quad |x-y| \leq d^\Omega(x,y) \leq k_0|x-y| \qquad \forall x, y \in \Omega$$

(the first inequality being obvious).

We finally introduce the minimal time function τ, which plays a crucial role in our study:

$$\text{(12)} \qquad \tau(x) := \inf\{t : x \in \mathcal{R}(t)\};$$

it is the time needed by the reflected controlled system to reach a point x when starting from K_0.

By standard argument (see [5]), we have the following lemma.

LEMMA 2.1. *The function τ is Lipschitz continuous with a Lipschitz constant less than k_0/m, where m and k_0 are defined in* (8) *and* (11). *Moreover, τ satisfies the Hamilton–Jacobi equation $c(\tau(x),x)|\nabla\tau(x)| = 1$ in the viscosity sense in $\Omega \setminus K_0$.*

We also note that the level sets of τ fulfill

$$\{x : \tau(x) = t\} \supseteq \partial\mathcal{R}(t); \tag{13}$$

in general, the inclusion is strict.

2.1. Extremal solutions and the Pontryagin maximum principle.

DEFINITION 2.2. *An admissible trajectory y is called extremal on $[0,t]$ if $\tau(y(t)) = t$.*

It is well known (for a detailed study of extremal trajectories, see [4, 8, 10, 15] and the references therein) that if y is an extremal trajectory on $[0,t]$, then $\tau(y(s)) = s$ for any $s \in [0,t]$. Throughout this section, $\mathcal{Y}(t)$ denotes the set of extremal trajectories on $[0,t]$:

$$\mathcal{Y}(t) = \{y \text{ extremal trajectory on } [0,t]\}.$$

The following result, proved in the appendix, is an extension of the Pontryagin maximum principle for our reflected control problem. It is inspired by similar results of Frankowska [12] on the regularity of the state and the adjoint for state constraints system. Let us note that the equation on p' is new in this context.

Here we denote by d the signed distance to the boundary of Ω:

$$d(x) = \begin{cases} -d_{\partial\Omega}(x) & \text{if } x \in \Omega, \\ d_\Omega(x) & \text{otherwise.} \end{cases}$$

From assumption (10), the function d is $\mathcal{C}^2$ in a neighborhood of $\partial\Omega$. Finally, for any $s \in \mathbb{R}$, we set $(s)_+ = \max\{s, 0\}$.

LEMMA 2.3. *Let $x(\cdot)$ be an extremal trajectory on the time interval $[0,T]$. Then there is a Lipschitz continuous function $p : [0,T] \to \mathbb{R}^N \setminus \{0\}$ and a measurable map $\lambda : [0,T] \to \mathbb{R}$ such that*

$$\begin{cases} x'(t) = c(t,x(t))\dfrac{p(t)}{|p(t)|} & \forall t \in (0,T), \\ p'(t) = -\left[D_x c(t,x(t)) - \lambda(t) Dd(x(t))\right]|p(t)| & \text{for a.e. } t \in (0,T), \\ \lambda(t) = \Big(-\Big\langle D^2 d(x(t)) c(t,x(t)) \frac{p(t)}{|p(t)|}, \frac{p(t)}{|p(t)|} \Big\rangle \\ \qquad\qquad + \langle Dd(x(t)), D_x c(t,x(t)) \rangle \Big)_+ \mathbf{1}_{\partial\Omega}(x(t)) & \text{for a.e. } t \in (0,T), \\ x(t) \in \overline{\Omega} \qquad \forall t \in (0,T). \end{cases} \tag{14}$$

The map p is called the adjoint of x.

Remark 2.4. Lemma 2.3 states, in particular, that any extremal trajectory is $\mathcal{C}^{1,1}$.

Remark 2.5. Problem (6) can be rewritten as a constraint problem; actually, its extremal trajectories coincide with those of problem

$$\begin{cases} y'(t) = c(t,y(t))b(t) & \text{a.e. in } (0,T), \\ y(t) \in \overline{\Omega} \qquad \forall t \geq 0, & y(0) = y_0 \in K_0, \end{cases}$$

and the function λ is a multiplier. Of course, other types of Neumann boundary conditions give rise to control problems that fail this property.

We now state several consequences of Lemma 2.3.

COROLLARY 2.6. *For any $T > 0$, the set $\mathcal{Y}(T)$ is compact w.r.t. the $\mathcal{C}^1$ norm.*

Proof. From standard arguments the set $\mathcal{Y}(T)$, endowed with the $\mathcal{C}^0$ norm, is compact. Let (y_n) be a converging sequence of $\mathcal{Y}(T)$ for the $\mathcal{C}^0$ norm, and let y be its limit. We claim that y_n converges to y for the $\mathcal{C}^1$ norm.

From Lemma 2.3, for any n, there is some Lipschitz map $p_n : [0,T] \to \mathbb{R}^N \backslash \{0\}$ such that $|p_n(0)| = 1$ and (14) holds. In particular, the (p_n) are Lipschitz continuous with a Lipschitz constant depending only on the regularity of Ω and of c. Hence a subsequence, still denoted (p_n), converges uniformly to some Lipschitz continuous function p such that $|p(0)| = 1$ and p is a solution to

$$\begin{cases} p'(t) = -\left[D_x c(t,x(t)) - \lambda(t) Dd(x(t))\right] |p(t)|, \\ \lambda(t) = \Big(-\left\langle D^2 d(x(t)) c(t,x(t)) \frac{p(t)}{|p(t)|}, \frac{p(t)}{|p(t)|} \right\rangle \\ \qquad\qquad + \langle Dd(x(t)), D_x c(t,x(t)) \rangle \Big)_+ \mathbf{1}_{\partial\Omega}(x(t)) \end{cases}$$

for a.e. $t \in (0,T)$. In particular, we deduce that $p(t) \neq 0$. Therefore, since $y_n'(t) = c(t, y_n(t)) p_n(t)/|p_n(t)|$, (y_n') uniformly converges to $y'(t) = c(t,y(t))p(t)/|p(t)|$.

So we have proved that for any sequence (y_n) of $\mathcal{Y}(T)$, there is a subsequence which converges in the $\mathcal{C}^1$ norm, whence we get the desired result. □

The next statement explains that two extremal trajectories on $[0,T]$ when crossing on $(0,T)$ necessarily have the same velocity.

LEMMA 2.7. *Let y_1 and y_2 be two extremal trajectories on $[0,T]$ for which there is some $t \in (0,T)$ with $y_1(t) = y_2(t) \in \partial\Omega$. Then $y_1'(t) = y_2'(t)$.*

Remark 2.8. Contrary to what happens in the unconstrained case ($\Omega = \mathbb{R}^2$), two different extremal trajectories of our reflected control problem can indeed cross, in particular on the boundary of Ω (as shown in Figure 3).

Proof. Since Ω has a $\mathcal{C}^2$ boundary, $y_1(t) = y_2(t) \in \partial\Omega$, and y_1 and y_2 are $\mathcal{C}^1$ and remain in $\overline{\Omega}$ on $[0,T]$, we deduce that the vectors $y_1'(t)$ and $y_2'(t)$ are necessarily tangent to $\partial\Omega$. Since, moreover, $|y_1'(t)| = c(t, y_1(t)) = |y_2'(t)|$ and we are in the plane, this leads to $y_1'(t) = \pm y_2'(t)$. Suppose for a while that $y_1'(t) = -y_2'(t)$. Then, thanks to Lemma 2.1, we have for any $h > 0$ sufficiently small

$$2h = \tau(y_2(t+h)) - \tau(y_1(t-h)) \leq \frac{k_0}{m} |y_1(t-h) - y_2(t+h)| = \frac{k_0}{m} h\epsilon(h),$$

where $\epsilon(h) \to 0$ as $h \to 0^+$. This is impossible. So $y_1'(t) = y_2'(t)$. □

As a consequence, we have the following technical result which shall be useful in the proof of the main perimeter estimate.

COROLLARY 2.9. *Let $\delta > 0$ and $T > 0$ be fixed. There is some positive σ such that if y_1 and y_2 are two extremal trajectories on some interval $[0, t_1 + \delta]$ and $[0, t_2 + \delta]$, for $t_1, t_2 \in (0,T]$, and if $y_1(t_1) \in \partial\Omega$, $y_2(t_2) \in \partial\Omega$ with $|y_1(t_1) - y_2(t_2)| \leq \sigma$, then*

$$\left| \frac{y_1'(t_1)}{|y_1'(t_1)|} - \frac{y_2'(t_2)}{|y_2'(t_2)|} \right| \leq c_b |y_1(t_1) - y_2(t_2)|,$$

where c_b depends only on the $\mathcal{C}^2$ regularity of $\partial\Omega$.

Proof. Let us first define the constant c_b. Since we are in the plane and Ω has a $\mathcal{C}^2$ boundary, there is a constant c_b (depending on the curvature of $\partial\Omega$ in $K_0 + B_{MT}$)

such that, for any $y_1, y_2 \in \partial\Omega \cap (K_0 + B_{MT})$ and for any v_1, v_2 unit tangent vector to $\partial\Omega$ at y_1 and y_2, we have either

$$|v_1 - v_2| \le c_b|y_1 - y_2| \qquad \text{or} \qquad |v_1 + v_2| \le c_b|y_1 - y_2|.$$

We now argue by contradiction, by assuming that there is a sequence $\sigma_n \to 0^+$ and, for any n, some pair y_1^n and y_2^n of extremal trajectories on some time intervals $[0, t_1^n + \delta]$ and $[0, t_2^n + \delta]$, such that

$$|y_1^n(t_1^n) - y_2^n(t_2^n)| \le \sigma_n \tag{15}$$

and

$$\left| \frac{(y_1^n)'(t_1^n)}{|(y_1^n)'(t_1^n)|} - \frac{(y_2^n)'(t_2^n)}{|(y_2^n)'(t_2^n)|} \right| > c_b|y_1^n(t_1^n) - y_2(t_2^n)|.$$

Let us observe that by conditions (8), we have

$$y_i^n(t) \in K_0 + B_{MT} \qquad \forall n \in \mathbb{N},\, t \in [0,T] \quad (i = 1,\, 2).$$

From the definition of c_b, the previous inequality implies that

$$\left| \frac{(y_1^n)'(t_1^n)}{|(y_1^n)'(t_1^n)|} + \frac{(y_2^n)'(t_2^n)}{|(y_2^n)'(t_2^n)|} \right| \le c_b|y_1^n(t_1^n) - y_2(t_2^n)|, \tag{16}$$

because $(y_1^n)'(t_1^n)/|(y_1^n)'(t_1^n)|$ and $(y_2^n)'(t_2^n)/|(y_2^n)'(t_2^n)|$ are tangent to $\partial\Omega$ at $y_1^n(t_1^n)$ and $y_2^n(t_2^n)$, respectively.

Thanks to Corollary 2.6 we can extract subsequences of (y_1^n), (y_2^n), (t_1^n), and (t_2^n)—still denoted (y_1^n), (y_2^n), (t_1^n), and (t_2^n)—converging to some y_1, y_2, t_1, and t_2, where y_1 and y_2 are extremal, on $[0, t_1 + \delta]$ and $[0, t_2 + \delta]$. From (15), we have $y_1(t_1) = y_2(t_2)$ and therefore $t_1 = t_2$. Then (16) becomes $y_1'(t_1) = -y_2'(t_2)$ with $y'(t_1) \neq 0$ because of Lemma 2.3. This is in contradiction with Lemma 2.7. ∎

2.2. Main estimates. We now state the main results of this section. The first one states, thanks to (13), that the length of the boundary of the reachable set—or, more precisely, its Hausdorff measure $\mathcal{H}^1$ (see [11])—remains bounded. The second one explains that the volume of the reachable set depends in a Lipschitz continuous way on the time.

PROPOSITION 2.10. *For $T > 0$, there exists a constant C such that*

$$\mathcal{H}^1\big(\{\tau = t\}\big) \le C \qquad \forall t \in [0,T].$$

The constant C depends on Ω and on the constants r, m, M, L_0, L_1, and T in assumptions (8)–(11).

As a consequence we have the following corollary.

COROLLARY 2.11. *Under the assumptions of Proposition* 2.10,

$$|\mathcal{R}(t_1)\backslash\mathcal{R}(t_0)| \le MC(t_1 - t_0) \qquad \forall 0 \le t_0 \le t_1 \le T,$$

where C is the constant in Proposition 2.10 *and M is given by* (8).

Remark 2.12. In particular, the map

$$t \to \rho(t) \equiv \mathbf{1}_{\mathcal{R}(t)}$$

is locally Lipschitz continuous in $\mathbb{L}^1(\Omega)$.

The proofs of Proposition 2.10 and Corollary 2.11 are the aim of the rest of the section. For proving Proposition 2.10, we divide the set $\{\tau = t\}$ into two sets: the first one consists of points which can be reached with extremal trajectories remaining in the interior of Ω. The second one is made up of points for which the associated extremal trajectories have to touch $\partial\Omega$. For the first set, the techniques of [1] can be adapted, although they have to be localized. For the second set, the key idea amounts to comparing its length with the length of $\partial\Omega$.

Let us introduce more precise notation: for $t \geq 0$ we set

$$E_t := \{x \in \partial\Omega \mid \exists y \in \mathcal{Y}(t),\ \exists s \in [0,t] \text{ with } y(s) = x\}, \tag{17}$$

$$D_t^{int} := \{x \in \Omega \mid \exists y \in \mathcal{Y}(t) \text{ with } y(t) = x \text{ and } y([0,t]) \cap \partial\Omega = \emptyset\}, \tag{18}$$

$$D_t^{bnd} := \{x \in \Omega \mid \exists y \in \mathcal{Y}(t),\ \exists s \in [0,t) \text{ with } y(t) = x \text{ and } y(s) \in \partial\Omega\}. \tag{19}$$

We note that D_t^{int} is exactly the subset of points of $\{\tau = t\}$ which can be reached with extremal trajectories remaining in the interior of Ω, while D_t^{bnd} is the set of points of $\{\tau = t\}$ reached by extremal trajectories passing by $\partial\Omega$. Furthermore,

$$\{\tau = t\} = D_t^{int} \cup D_t^{bnd}.$$

2.3. Estimates for the set reachable by passing by $\partial\Omega$. In this part we estimate the length of D_t^{bnd}. For this we heavily use the fact that we are dealing with a plane system. We do not know if similar constructions can be done in higher dimensions.

Since we cannot directly estimate the length of D_t^{bnd}, we need to introduce an approximation of this set given, for any $\delta > 0$, by

$$D_t^{bnd}(\delta) = \left\{ x \in \Omega \mid \begin{array}{l} \exists y \in \mathcal{Y}(t),\ \exists s \in [0, t-\delta) \text{ with } y(t) = x, \\ y(s) \in \partial\Omega \text{ and } y((s,t]) \cap \partial\Omega = \emptyset \end{array} \right\}.$$

We note for later use that the set D_t^{bnd} is the increasing limit of the $D_t^{bnd}(\delta)$ as δ decreases to 0^+.

LEMMA 2.13. *Let $T > 0$ be fixed. There are $C_1 > 0$ and, for any $\delta > 0$, a constant $\sigma > 0$ such that if y_1, y_2 belong to $D_t^{bnd}(\delta)$, $y_1(\cdot)$, $y_2(\cdot)$ belong to $\mathcal{Y}(t)$, and $s_1, s_2 \in [0, t-\delta)$ with (for $j = 1, 2$)*

$$y_j(t) = y_j, \quad y_j(s_j) \in \partial\Omega, \quad y_j((s_j,t]) \cap \partial\Omega = \emptyset, \quad \text{and } |y_1(s_1) - y_2(s_2)| \leq \sigma,$$

then

$$|y_1 - y_2| \leq C_1 |y_1(s_1) - y_2(s_2)|.$$

The constant C_1 depends on the various constants of the problem and on T but not on δ.

Proof. From Corollary 2.9, there is some positive σ such that if y_1 and y_2 are two extremal trajectories on some interval $[0, t_1 + \delta]$ and $[0, t_2 + \delta]$ and if $y_1(t_1) \in \partial\Omega$, $y_2(t_2) \in \partial\Omega$ (for some $t_1, t_2 \leq T$) with $|y_1(t_1) - y_2(t_2)| \leq \sigma$, then

$$\left| \frac{y_1'(t_1)}{|y_1'(t_1)|} - \frac{y_2'(t_2)}{|y_2'(t_2)|} \right| \leq c_b |y_1(t_1) - y_2(t_2)|, \tag{20}$$

where c_b depends only on the $\mathcal{C}^2$ regularity of $\partial\Omega$.

Now let y_j, $y_j(\cdot)$ and s_j $(j = 1, 2)$ be as in the statement. Without loss of generality, we assume that $s_1 \leq s_2$. Let p_1 and p_2 be the adjoint of y_1 and y_2 defined in Lemma 2.3. On $[s_j, t]$, the pair (y_j, p_j) is the unique solution of the differential equation

$$\left\{ \begin{array}{ll} y_j'(s) = c(s, y_j(s)) \dfrac{p_j(s)}{|p_j(s)|} & \forall s \in (s_j, t), \\ p_j'(s) = -D_x c(s, y_j(s))\, |p_j(s)| & \text{for a.e. } s \in (s_j, t) \end{array} \right. \tag{21}$$

with initial condition $(y_j(s_j), p_j(s_j))$. Without loss of generality, we can assume that $|p_j(s_j)| = 1$. From these equations, we deduce that

$$\frac{d}{ds}\left(|p_j(s)|\right) = \frac{\langle p_j'(s), p_j(s)\rangle}{|p_j(s)|} = -\langle D_x c(s, y_j(s)), p_j(s)\rangle,$$

whence, for some constant $L > 0$,

$$-L|p_j(s)| \leq \frac{d}{ds}|p_j(s)| \leq L|p_j(s)|$$

and, by integration,

$$e^{-L(s-s_j)} \leq |p_j(s)| \leq e^{L(s-s_j)} \qquad \forall s \in [s_j, t]. \tag{22}$$

In particular, (21) can be written as a Cauchy problem with a Lipschitz right-hand side.

From the Lipschitz continuity of τ given in Lemma 2.1 we have

$$|s_2 - s_1| = \tau(y_2(s_2)) - \tau(y_1(s_1)) \leq \frac{k_0}{m}|y_2(s_2) - y_1(s_1)|. \tag{23}$$

We also note that inequality (20) can be rewritten as

$$|p_1(s_1) - p_2(s_2)| \leq c_b|y_1(s_1) - y_2(s_2)|. \tag{24}$$

Finally, because of (21) and the bounds on c and $D_x c$ (stated in (8)), there is some constant C_0 such that

$$|y_1(s_2) - y_1(s_1)| + |p_1(s_2) - p_1(s_1)| \leq C_0|s_2 - s_1| \leq \frac{C_0 k_0}{m}|y_2(s_2) - y_1(s_1)|. \tag{25}$$

Putting together (23), (24), and (25) and using the Gronwall inequality for (21) then easily gives

$$|y_1(s) - y_2(s)| + |p_1(s) - p_2(s)| \leq C_1|y_1(s_1) - y_2(s_2)|$$

for any $s \in [s_2, t]$ for some constant $C_1 = C_1(T, m, c_b, k_0, M, L_0, L_1)$. Setting $s = t$ gives the desired inequality. □

LEMMA 2.14. *Fix $T > 0$. With the same constant C_1 as in Lemma 2.13, we have*

$$\mathcal{H}^1(D_t^{bnd}) \leq C_1 \mathcal{H}^1(\partial\Omega) \qquad \forall 0 \leq t \leq T.$$

Proof. Let C_1 and, for $\delta > 0$ fixed, σ be as in Lemma 2.13. Let $\varepsilon > 0$ and a family of sets $\{B_i\}_{i\in\mathbb{N}}$ be such that

$$\mathrm{diam} B_i \leq \inf\{\varepsilon, \sigma\}, \quad \bigcup_{i=1}^{+\infty} B_i \supseteq \partial\Omega, \quad \text{and} \quad \mathcal{H}^1(\partial\Omega) \geq \sum_{i=1}^{+\infty} \mathrm{diam} B_i - \varepsilon.$$

We denote by K_i the set of points of $D_t^{bnd}(\delta)$ which can be reached by an extremal trajectory touching $\partial\Omega$ for the last time in B_i; in other words, $x \in K_i$ if and only if there exists an extremal trajectory y with $y(t) = x$ and $y(s_0) \in B_i$ for $s_0 := \max\{s \in [0,t] : y(s) \in \partial\Omega\}$.

Since we have chosen $\mathrm{diam}B_i \leq \sigma$, Lemma 2.13 states that

$$\mathrm{diam}(K_i) \leq C_1 \mathrm{diam}B_i \qquad \forall i \in \mathbb{N}. \tag{26}$$

Moreover, since

$$\bigcup_{i=1}^{+\infty} K_i = D_t^{bnd}(\delta) \qquad \text{and} \qquad \mathrm{diam}(K_i) \leq C_1\epsilon,$$

we have

$$\mathcal{H}^1_{C_1\varepsilon}(D_t^{bnd}(\delta)) \leq \sum_{i=1}^{+\infty} \mathrm{diam}(K_i) \leq C_1 \left(\mathcal{H}^1(\partial\Omega) + \varepsilon\right)$$

(for the precise definition of $\mathcal{H}^1_\delta$, see [11, pp. 60ff.]). As $\varepsilon \to 0$, we get

$$\mathcal{H}^1(D_t^{bnd}(\delta)) \leq C_1 \mathcal{H}^1(\partial\Omega).$$

Recalling that C_1 does not depend on δ, we can let $\delta \to 0^+$ to get the result. □

2.4. Estimates for the set reachable by remaining in the interior of Ω. We now aim at computing the length of D_t^{int}, the part of the boundary consisting of the points which can be reached by remaining in the interior of Ω.

Let us recall that the interior sphere condition does *not* propagate in $\Omega \neq \mathbb{R}^2$ (contrary to what happens in $\mathbb{R}^2$, eventually with a radius depending on time; see [1] and [7]). However, a local version of propagation (established in Lemma 2.15) holds also in Ω. This property and a bound on the measure of ∂K_0 (still due to the interior sphere condition) will be crucial in the proof of our estimate for $\mathcal{H}^1(D_{t_1}^{int})$ (see Lemma 2.16 and Remark 2.17).

The estimates below follow the original ideas of Cannarsa and Frankowska [7] and the computations of [1]. In particular, the two following lemmata are just a localized version of Lemmata 3.1 and 6.1 of [1].

LEMMA 2.15. *Under assumption* (9)*, let $y(\cdot)$ be some extremal trajectory on the time interval $[0,t]$, with adjoint $p(\cdot)$, for which there is some $\eta > 0$ with*

$$B(y(s), \eta) \subset \Omega \qquad \forall s \in [0,t].$$

Then, if we set $\kappa = 3L_0 + L_1$, we have

$$B\left(y(t) - re^{-\kappa t}\frac{p(t)}{|p(t)|}, re^{-\kappa t}\right) \cap B\left(y(t), \eta e^{-L_0 t}\right) \subset \mathcal{R}(t).$$

Proof. We explain only the main differences with the proofs of Lemmata 3.1 and 6.1 of [1]. Without loss of generality, we assume that $|p(0)| = 1$. We define the vector field $f(s,y) = c(s,y)p(s)/|p(s)|$ and set $\bar\kappa = 2L_0 + L_1$. We consider, for any $\theta \in B(0, |p(t)|)$ such that

$$re^{-\bar\kappa t}|p(t) - \theta| \leq \eta e^{-L_0 t}, \tag{27}$$

the solution y_θ of the backward differential equation

$$\begin{cases} y'_\theta(s) = f(s, y_\theta(s)) & \forall s \in [0,t], \\ y_\theta(t) = y(t) - re^{-\bar\kappa t}(p(t) - \theta). \end{cases}$$

Following [1], one can show that $y_\theta(0) \in K_0$. Furthermore, a straightforward application of the Gronwall lemma yields to

$$y_\theta(s) \in B(y(s), \eta e^{-L_0 s}) \subset \Omega \qquad \forall s \in [0,t],$$

because $y_\theta(t) \in B(y(t), \eta e^{-L_0 t})$ thanks to (27). So y_θ is a solution to the control system (6) which remains in Ω on $[0,t]$, and therefore $y_\theta(t) \in \mathcal{R}(t)$. This proves that

$$B\left(y(t) - re^{-\bar\kappa t}p(t), re^{-\bar\kappa t}|p(t)|\right) \cap B\left(y(t), \eta e^{-L_0 t}\right) \subset \mathcal{R}(t).$$

Since $\kappa = \bar\kappa + L_0$ and $|p(s)| \geq e^{-L_0 t}$ (thanks to the differential equation satisfied by p), the proof is complete. □

LEMMA 2.16. *For $0 \leq t_0 \leq t_1 \leq T$, we have*

$$\mathcal{H}^1(D^{int}_{t_1}) \leq e^{C_2(t_1 - t_0)}\mathcal{H}^1(D^{int}_{t_0}),$$

where $C_2 = L_0 + Me^{\kappa T}/r$ with $\kappa = 3L_0 + L_1$.

Remark 2.17. In particular, since K_0 satisfies the interior ball condition, $\mathcal{H}^1(\partial K_0)$ is finite (Lemma 2.2 of [1]), and we have (setting $t_0 = 0$ and $t_1 = t$)

$$\mathcal{H}^1(D^{int}_t) \leq e^{C_2 t}\mathcal{H}^1(\partial K_0) \qquad \forall t \in [0,T].$$

Proof. Without loss of generality, we can assume that $\mathcal{H}^1(D^{int}_{t_0}) < +\infty$. For $\eta > 0$ let us set

$$D^{int}_t(\eta) = \left\{ x \in \Omega \mid \begin{array}{c} \tau(x) = t \text{ and } \exists y \in \mathcal{Y}(t) \text{ with } y(t) = x \\ \text{and } d_{\partial\Omega}(y(s)) \geq \eta \text{ for any } s \in [0,t] \end{array} \right\}.$$

We note that $D^{int}_t(\eta)$ is closed and that

$$D^{int}_t \subset \bigcup_{\eta > 0} D^{int}_t(\eta). \tag{28}$$

Let us fix $\epsilon > 0$ and $\delta > 0$, and let us choose an (at most) countable collection of compact sets (K_i) such that $K_i \subset D^{int}_{t_0}(\eta)$,

$$D^{int}_{t_0}(\eta) = \bigcup_i K_i, \quad \operatorname{diam}(K_i) \leq \min\left\{\eta e^{-(L_0 + C_2)T}, \delta e^{-C_2 T}\right\} \qquad \forall i, \tag{29}$$

and

$$\mathcal{H}^1(D^{int}_{t_0}(\eta)) \geq \sum_i \operatorname{diam}(K_i) - \epsilon.$$

For any $s \in [t_0, t_1]$, let $K'_i(s)$ be the set of points $y \in D^{int}_s(\eta)$ for which there is an extremal trajectory $y(\cdot)$ with $y(s) = y$, $y(t_0) \in K_i$ and $d_{\partial\Omega}(y(t)) \geq \eta$ for any $t \in [0,s]$. We note for later use that

$$D^{int}_s(\eta) \subset \bigcup_i K'_i(s) \qquad \forall s \in [t_0, t_1]. \tag{30}$$

We claim that

$$\operatorname{diam}(K_i'(s)) \le \min\{\eta e^{-L_0 T}, e^{C_2 s}\operatorname{diam}(K_i)\} \qquad \forall s \in [t_0, t_1]. \tag{31}$$

For proving the claim, let

$$\theta = \inf\{s \in [t_0, t_1] \ : \ \operatorname{diam}(K_i'(s)) \ge \eta e^{-L_0 T}\}$$

with convention $\inf \emptyset = t_1$. One readily checks from its definition that $s \to \operatorname{diam}(K_i'(s))$ is right upper semicontinuous and left continuous. Therefore $\theta > t_0$ because

$$\operatorname{diam}(K_i'(t_0)) = \operatorname{diam}(K_i) < \eta e^{-L_0 T}.$$

For $y_1, y_2 \in K_i'(\theta)$, let $y_1(\cdot)$ and $y_2(\cdot)$ be two extremal trajectories on $[0, \theta]$ with $y_j(\theta) = y_j$, $d_{\partial\Omega}(y_j(t)) \ge \eta$ for any $t \in [0, \theta]$ and $y_j(t_0) \in K_i$ for $j = 1, 2$. We denote by $p_j(\cdot)$ the adjoint of $y_j(\cdot)$. Since $y_j(s) \in K_i'(s)$ for $s \in [t_0, \theta)$, and from the definition of θ, we have

$$|y_1(s) - y_2(s)| \le \operatorname{diam}(K_i'(s)) < \eta e^{-L_0 T} \qquad \forall s \in [t_0, \theta).$$

In particular, $y_2(s)$ belongs to $B(y_1(s), \eta e^{-L_0 s})$. Since $d_{\partial\Omega}(y_1(s)) \ge \eta$ for any $s \in [0, t_1]$, and since $y_2(s)$ lies on the boundary of the reachable set, Lemma 2.15 states that $y_2(s)$ does not belong to the interior of the ball centered at $y_1(s) - re^{-\kappa s} p_1(s)/|p_1(s)|$ and of radius $re^{-\kappa s}$. Hence

$$|y_2(s) - y_1(s)|^2 + 2re^{-\kappa s}\langle y_2(s) - y_1(s), p_1(s)/|p_1(s)|\rangle \ \ge \ 0.$$

Reversing the roles of y_1 and y_2 also gives

$$|y_2(s) - y_1(s)|^2 + 2re^{-\kappa s}\langle y_1(s) - y_2(s), p_2(s)/|p_2(s)|\rangle \ \ge \ 0.$$

whence we get

$$\left\langle y_1(s) - y_2(s), \frac{p_1(s)}{|p_1(s)|} - \frac{p_2(s)}{|p_2(s)|} \right\rangle \ \le \ \frac{e^{\kappa T}}{r}|y_1(s) - y_2(s)|^2. \tag{32}$$

Setting $\rho(s) = \frac{1}{2}|y_1(s) - y_2(s)|^2$, we have, for $s \in [t_0, \theta)$,

$$\begin{aligned}
\rho'(s) &= \langle y_1'(s) - y_2'(s), y_1(s) - y_2(s)\rangle \\
&= \left\langle c(s, y_1(s))\tfrac{p_1(s)}{|p_1(s)|} - c(s, y_2(s))\tfrac{p_2(s)}{|p_2(s)|}, y_1(s) - y_2(s) \right\rangle \\
&= \tfrac{1}{2}(c(s, y_1(s)) - c(s, y_2(s)))\left\langle \tfrac{p_2(s)}{|p_2(s)|} + \tfrac{p_1(s)}{|p_1(s)|}, y_1(s) - y_2(s) \right\rangle \\
&\qquad + \tfrac{1}{2}(c(s, y_1(s)) + c(s, y_2(s)))\left\langle \tfrac{p_1(s)}{|p_1(s)|} - \tfrac{p_2(s)}{|p_2(s)|}, y_1(s) - y_2(s) \right\rangle \\
&\le L_0|y_1(s) - y_2(s)|^2 + \tfrac{Me^{\kappa T}}{r}|y_1(s) - y_2(s)|^2
\end{aligned}$$

(in the last inequality, relations (8) and (32) have been used). Therefore, by the definition of C_2, we have $\rho'(s) \le 2C_2\rho(s)$ and deduce that

$$|y_1(s) - y_2(s)| \le |y_1(t_0) - y_2(t_0)|e^{C_2(s-t_0)} \le \operatorname{diam}(K_i)e^{C_2(s-t_0)} \qquad \forall s \in [t_0, \theta],$$

because $y_1(t_0)$ and $y_2(t_0)$ belong to K_i. So we have proved that, for all $s \in [t_0, \theta]$,

$$\operatorname{diam}(K_i'(s)) \leq \operatorname{diam}(K_i) e^{C_2(s-t_0)}.$$

Therefore, because of the definition of θ, we complete the proof of our claim (31) if we show that $\theta = t_1$. To this end, for $s = \theta$, we get

$$\operatorname{diam}(K_i'(\theta)) \leq \operatorname{diam}(K_i) e^{C_2(\theta-t_0)} < \eta e^{-L_0 T}$$

from the choice of $\operatorname{diam}(K_i)$ in (29). In particular, since $s \to \operatorname{diam}(K_i'(s))$ is right upper semicontinuous, we necessarily have $\theta = t_1$, and (31) is proved.

We now complete the proof of the lemma: the family $(K_i'(t_1))$ is a covering of $D_{t_1}^{int}(\eta)$, from (30), with

$$\operatorname{diam}(K_i'(t_1)) \leq \operatorname{diam}(K_i) e^{C_2(t_1-t_0)} \leq \delta$$

from (29). Therefore

$$\mathcal{H}^1_\delta(D_{t_1}^{int}(\eta)) \leq \sum_i \operatorname{diam}(K_i'(t_1)) \leq e^{C_2(t_1-t_0)} (\mathcal{H}^1(D_{t_0}^{int}(\eta)) + \epsilon).$$

Letting $\delta, \epsilon, \eta \to 0^+$ gives the desired result. □

2.5. Proofs of Proposition 2.10 and of Corollary 2.11.

Proof of Proposition 2.10. Taking into account Lemmata 2.14 and 2.16, we obtain

$$\mathcal{H}^1\big(\{\tau = t\}\big) = \mathcal{H}^1\big(D_t^{bnd}\big) + \mathcal{H}^1\big(D_t^{int}\big) \leq C_1 \mathcal{H}^1(\partial\Omega) + e^{C_2 T} \mathcal{H}^1(\partial K_0),$$

where $\mathcal{H}^1(\partial K_0)$ is finite because K_0 is compact and satisfies the interior ball condition (see, for instance, [1]). □

Proof of Corollary 2.11. Let us recall that τ is Lipschitz continuous and satisfies the Hamilton–Jacobi equation $c(\tau(x), x)|\nabla\tau(x)| = 1$ in the viscosity sense and thus a.e. in $\Omega \setminus K_0$. From assumption (8), we therefore have that

$$\frac{1}{M} \leq |\nabla\tau(x)| \leq \frac{k_0}{m} \qquad \text{a.e. in } \ \Omega \setminus K_0. \tag{33}$$

The coarea formula states that

$$\int_{\mathcal{R}(t_1)\setminus\mathcal{R}(t_0)} |\nabla\tau(x)|\, dx = \int_{t_0}^{t_1} \mathcal{H}^1\big(\{\tau = s\}\big)\, ds. \tag{34}$$

By assumption (33) and Proposition 2.10, we deduce, respectively, that

$$\int_{\mathcal{R}(t_1)\setminus\mathcal{R}(t_0)} |\nabla\tau(x)|\, dx \geq \frac{1}{M} |\mathcal{R}(t_1)\setminus\mathcal{R}(t_0)|,$$

$$\int_{t_0}^{t_1} \mathcal{H}^1\big(\{\tau = s\}\big)\, ds \ \leq \ C(t_1 - t_0).$$

Substituting the previous two inequalities into (34) completes the proof. □

3. Estimate for contraction. Let us consider two velocities c_1 and c_2, both satisfying assumptions (8), and the corresponding reflected control problems (for $j = 1, 2$)

$$\text{(35)}\quad \begin{cases} y'(t) = c_j(t, y(t))\big[b(t) - \langle b(t), \nu(y(t))\rangle \mathbf{1}_{\partial\Omega}(y(t))\nu(y(t))\big] & \text{a.e. in } (0,T), \\ y(0) \in K_j, \end{cases}$$

where $b : \mathbb{R}_+ \to B$ denotes the time measurable control. We assume that the sets K_1 and K_2 fulfill the interior sphere condition of radius $r > 0$ and that Ω satisfies (10)–(11). We denote by $\mathcal{R}_j(t)$ the reachable set of (35) at time t.

The main result of this section is some estimate of the volume of the symmetric difference $\mathcal{R}_1(t)\Delta\mathcal{R}_2(t)$.

PROPOSITION 3.1. *Under the above assumptions on c_j, K_j, and Ω, there exists for any $T > 0$ a constant $N = N(\Omega, r, m, M, L_0, L_1, T)$ such that*

$$|\mathcal{R}_1(t)\Delta\mathcal{R}_2(t)| \le N\gamma(t)$$

with

$$\gamma(t) := d^{\mathcal{H}}(K_1, K_2)e^{(L_0+M/r_1)t} + \int_0^t \|c_1(s,\cdot) - c_2(s,\cdot)\|_\infty e^{(L_0+M/r_1)(t-s)}ds,$$

where r_1 is the constant introduced in (10).

The proof is postponed until the end of the section.

Our starting point is the following lemma.

LEMMA 3.2. *Under the previous hypotheses on c_j, K_j, and Ω, we have*

$$\mathcal{R}_2(t) \subseteq \mathcal{R}_1(t) + \gamma(t)B.$$

In particular, we deduce that

$$\mathcal{R}_2(t) \subseteq \mathcal{R}_1(t + k_0\gamma(t)/m),$$

where k_0 is the constant introduced in inequality (11).

Proof. We set $\gamma_0 := d^{\mathcal{H}}(K_1, K_2)$ and, without loss of generality (thanks to Remark 2.5), consider an extremal trajectory z of (35) for $j = 2$ having the form

$$\begin{cases} z'(t) = c_2(t, z(t))b_2(t), \\ z(0) = z_0 \in K_2. \end{cases}$$

We define y as the reflected trajectory with velocity c_1, driven by the control b_2:

$$\begin{cases} y'(t) = c_1(t, y(t))\big[b_2(t) - \langle b_2(t), \nu(y(t))\rangle \mathbf{1}_{\partial\Omega}(y(t))\nu(y(t))\big], \\ y(0) = y_0, \end{cases}$$

where $y_0 \in K_1$ satisfies $|z_0 - y_0| \le \gamma_0$. For the sake of simplicity, we denote

$$b_1(t) := b_2(t) - \langle b_2(t), \nu(y(t))\rangle \mathbf{1}_{\partial\Omega}(y(t))\nu(y(t)).$$

For $g(s) := |z(s) - y(s)|^2/2$, there holds that

$$\begin{aligned} g'(s) &= \langle c_2(s, z(s))b_2(s) - c_1(s, y(s))b_1(s), z(s) - y(s)\rangle \\ &= \langle [c_2(s, z(s)) - c_1(s, z(s))]\, b_2(s), z(s) - y(s)\rangle \\ &\qquad + \langle c_1(s, z(s))b_2(s) - c_1(s, y(s))b_1(s), z(s) - y(s)\rangle \qquad \text{(36)} \\ &\le \|c_1(s,\cdot) - c_2(s,\cdot)\|_\infty |z(s) - y(s)| \\ &\quad + \langle c_1(s, z(s))b_2(s) - c_1(s, y(s))b_1(s), z(s) - y(s)\rangle. \end{aligned}$$

Using assumption (8), we get

$$\begin{aligned}
&\langle c_1(s,z(s))b_2(s) - c_1(s,y(s))b_1(s), z(s)-y(s)\rangle \\
&= \langle [c_1(s,z(s)) - c_1(s,y(s))]\, b_2(s), z(s)-y(s)\rangle \\
&\qquad - c_1(s,y(s))\langle b_1(s)-b_2(s), z(s)-y(s)\rangle \\
&\le L_0|z(s)-y(s)|^2 \\
&\qquad + c_1(s,y(s))\langle b_2(s), \nu(y(s))\rangle\langle \nu(y(s)), z(s)-y(s)\rangle \mathbf{1}_{\partial\Omega}(y(s)).
\end{aligned} \tag{37}$$

Let us recall that $\mathcal{C}\Omega$ fulfills the interior sphere condition with radius r_1. We claim that, for almost every $s \in [0,t]$, there holds that

$$\langle b_2(s), \nu(y(s))\rangle\langle \nu(y(s)), z(s)-y(s)\rangle \mathbf{1}_{\partial\Omega}(y(s)) \le \frac{1}{r_1}|z(s)-y(s)|^2. \tag{38}$$

This inequality is obvious if $y \notin \partial\Omega$. Consider $y \in \partial\Omega$, whence we get

$$|z(s) - y(s) - r_1\nu(y(s))|^2 \ge r_1^2,$$

from which we deduce that

$$\langle \nu(y(s)), z(s)-y(s)\rangle \le \frac{1}{2r_1}|z(s)-y(s)|^2.$$

By the arbitrariness of points $y(s)$ and $z(s)$, we deduce our claim (38).

Substituting relations (37) and (38) into (36), we deduce that

$$g'(s) \le \|c_1(s,\cdot) - c_2(s,\cdot)\|_\infty |z(s)-y(s)| + (L_0 + M/r_1)|z(s)-y(s)|^2;$$

in particular, for $\rho(s) := |z(s)-y(s)|$, the previous inequality gives

$$\rho'(s) \le \|c_1(s,\cdot) - c_2(s,\cdot)\|_\infty + (L_0 + M/r_1)\rho(s).$$

Applying the Gronwall lemma, we get

$$\rho(t) \le \rho(0)e^{(L_0+M/r_1)t} + \int_0^t \|c_1(s,\cdot) - c_2(s,\cdot)\|_\infty e^{(L_0+M/r_1)(t-s)}\, ds.$$

By the arbitrariness of z_0, the first part of the statement is proved.

By assumption (8) and inequality (11), the minimal time for reaching a point $y \in \Omega$ starting from $\bar{y} \in \Omega$ is less than or equal to $k_0|y-\bar{y}|/m$ (this value is the time needed to follow the geodesic between y and $\bar{y}$ at the minimal velocity m). Therefore, the second part of the statement is an immediate consequence of the previous one. □

Proof of Proposition 3.1. We estimate only the measure of $|\mathcal{R}_2(t)\backslash\mathcal{R}_1(t)|$; the other difference can be estimated by similar arguments, so we shall omit it. From Lemma 3.2 and Corollary 2.11 we have

$$|\mathcal{R}_2(t)\backslash\mathcal{R}_1(t)| \le |\mathcal{R}_1(t + k_0\gamma(t)/m)\backslash\mathcal{R}_1(t)| \le MCk_0\gamma(t)/m.$$

Setting $N = MCk_0/m$ completes the proof. □

4. Application to dislocation dynamics in a region. The aim of this section is to investigate the existence and uniqueness of the solution to the following nonlocal problem:

$$\left\{\begin{array}{ll} \frac{\partial \rho}{\partial t} = \big(\bar{c}_1 + I(\rho)\big)|\nabla \rho| & \text{in} \quad (0,+\infty)\times\Omega, \\ \rho(0,x) = \rho_0(x) \equiv \mathbf{1}_{K_0}(x) & \text{on} \quad \Omega, \\ \frac{\partial \rho}{\partial \nu} = 0 & \text{on} \quad (0,+\infty)\times\partial\Omega, \end{array}\right. \tag{39}$$

where ν is the outward unitary normal to $\partial\Omega$ and the map I is defined as follows:

$$I(\rho)(t,x) = \int_\Omega \bar{c}_2(t,x,y)\rho^*(y)\,dy \qquad \forall \rho \in \mathbb{L}^1(\Omega) \tag{40}$$

for any function $\rho : \Omega \to [0,1]$, where ρ^* denotes the upper semicontinuous envelope of ρ.

Besides assumptions (9)–(11) on the sets Ω and K_0, we require throughout this section that the maps $\bar{c}_1 : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R}$ and $\bar{c}_2 : \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ satisfy the following:

$$\left\{\begin{array}{ll} \text{(i)} & \bar{c}_1 \text{ and } \bar{c}_2 \text{ are continuous and derivable w.r.t. the second variable,} \\ \text{(ii)} & |\bar{c}_1(t,x)| + \|\bar{c}_2(t,x,\cdot)\|_1 \le M \qquad \forall (t,x) \in \mathbb{R}^+ \times \Omega, \\ \text{(iii)} & |\bar{c}_1(t,x_1) - \bar{c}_1(t,x_2)| + \|\bar{c}_2(t,x_1,\cdot) - \bar{c}_2(t,x_2,\cdot)\|_1 \le L_0|x_1 - x_2| \\ & \qquad \forall (t,x_1),(t,x_2) \in \mathbb{R}^+ \times \Omega, \\ \text{(iv)} & |D_x\bar{c}_1(t,x_1) - D_x\bar{c}_1(t,x_2)| + \|D_x\bar{c}_2(t,x_1,\cdot) - D_x\bar{c}_2(t,x_2,\cdot)\|_1 \\ & \qquad \le L_1|x_1 - x_2|, \qquad \forall (t,x_1),(t,x_2) \in \mathbb{R}^+ \times \Omega, \\ \text{(v)} & |\bar{c}_2(t,x,y)| \le M \qquad \forall (t,x,y) \in \mathbb{R}^+ \times \Omega \times \Omega \end{array}\right. \tag{41}$$

(where $\|\cdot\|_1$ denotes the $\mathbb{L}^1$ norm in Ω) for some positive constants M, L_0, and L_1, and, for some $m > 0$, the following relation holds:

$$\bar{c}_1(t,x) - \|\bar{c}_2(t,x,\cdot)\|_1 \ge m \qquad \forall (t,x) \in [0,+\infty)\times\Omega. \tag{42}$$

This assumption implies that the dislocation is always expanding. We note that under the assumption (41) and (42), the map $c_\rho := \bar{c}_1 + I(\rho)$ satisfies (8) for any map $\rho \in C^0([0,+\infty), \mathbb{L}^1(\Omega,[0,1]))$. Moreover, the constants entering into relations (8) do not depend on the function ρ.

Let us recall that in the framework of discontinuous *viscosity* solutions (see the monographs by Barles [5] and by Bardi and Capuzzo-Dolcetta [4] for an overview) uniqueness of the solution means that two solutions ρ_1 and ρ_2 have the same upper and lower semicontinuous envelopes: $(\rho_1)^* = (\rho_2)^*$ and $(\rho_1)_* = (\rho_2)_*$. Now we can introduce our definition of viscosity solution.

DEFINITION 4.1. *A function ρ is a viscosity solution to problem* (39) *if* $\rho \in C^0([0,+\infty), \mathbb{L}^1(\Omega))$ *and if the unique discontinuous solution u to the initial-boundary value problem*

$$\left\{\begin{array}{ll} \frac{\partial u}{\partial t} = c_\rho(t,x)|\nabla u| & \textit{in} \quad (0,+\infty)\times\Omega, \\ u(0,x) = \rho_0(x) \equiv \mathbf{1}_{K_0}(x) & \textit{on} \quad \Omega, \\ \frac{\partial u}{\partial \nu} = 0 & \textit{on} \quad (0,+\infty)\times\partial\Omega \end{array}\right. \tag{43}$$

with $c_\rho(t,x) := \bar{c}_1(t,x) + I(\rho)(t,x)$ fulfills $u^* = \rho^*$ and $u_* = \rho_*$ in $[0,+\infty)\times\Omega$.

The well-posedness of system (43) is ensured by the following lemma, which also provides a representation formula for this system.

LEMMA 4.2. *Under the above assumptions on* Ω, K_0, $\bar{c}_1$, *and* $\bar{c}_2$, *for any* $\rho \in C^0([0,+\infty), \mathbb{L}^1(\Omega))$ *there exists a unique discontinuous solution* u *to the initial boundary value problem* (43). *Moreover, for* $t > 0$, *we have*

$$u(t,x) \in \{0,1\} \qquad \forall x \in \overline{\Omega} \qquad \text{and} \qquad \{x \mid u^*(t,x) = 1\} = \mathcal{R}(t), \tag{44}$$

where $\mathcal{R}(t)$ *is the reachable set for the control problem* (6) *introduced in section* 2 *with* $c(t,x) = c_\rho(t,x)$.

Proof. In order to show the relation between the level sets of the solution ρ and the reachable set $\mathcal{R}(t)$, let us introduce the following Mayer problem: the value function $\psi(t,x)$ is defined as

$$\psi(t,x) := \max_b \mathbf{1}_{K_0}(y(0)),$$

where y is the solution to the following *backward* reflected control problem:

$$\begin{cases} y'(s) = c(s,y(s))\big[b(s) - \langle b(s), \nu(y(s))\rangle \mathbf{1}_{\partial\Omega}(y(s))\nu(y(s))\big] & \text{a.e. in } [0,t], \\ y(t) = x. \end{cases}$$

Arguing as in [14], one can easily prove that ψ is a solution to (43).

In order to prove the uniqueness of the solution, let us first show that the *fattening* phenomenon (for the precise definition, see [6]) does not occur in $[0,T]$ for every $T \in \mathbb{R}^+$. To this end, we introduce $u_0 \in BUC(\Omega)$, $u_0 \le 1$, such that $\{x \mid u_0(x) \ge 0\} = K_0$ and $\{x \mid u_0(x) > 0\} = \text{Int } K_0$, and we consider the following problem:

$$\begin{cases} \frac{\partial u}{\partial t} = c_\rho(t,x)|\nabla u| & \text{in} \quad (0,+\infty)\times\Omega, \\ u(0,x) = u_0(x) & \text{on} \quad \Omega, \\ \frac{\partial u}{\partial \nu} = 0 & \text{on} \quad (0,+\infty)\times\partial\Omega. \end{cases} \tag{45}$$

It is well known (see [6]) that there exists exactly one solution $u \in UC$ of this problem. For every $s \in \mathbb{R}$, we set $K_s := \{u_0 \ge s\}$ and denote $\mathcal{R}_s(t)$ the reachable set in time t when starting from K_s. The function $\mathbf{1}_{\mathcal{R}_s(t)}(x)$ is a (discontinuous) viscosity solution to problem (45) with u_0 replaced by $\mathbf{1}_{K_s}$. The comparison principle (which applies because u is sufficiently regular; see [6]) entails

$$u(t,x) - s - \epsilon \le \mathbf{1}_{\mathcal{R}_s(t)}(x) \qquad \forall s,\epsilon \in \mathbb{R},\ x \in \mathbb{R}^2,\ t \in \mathbb{R}^+$$

and, in particular,

$$\{(t,x) \mid u(t,x) \ge s + \epsilon\} \subseteq \mathcal{R}_s(t) \qquad \forall s,\epsilon \in \mathbb{R},\ t \in \mathbb{R}^+. \tag{46}$$

On the other hand, let v_ϵ be the UC solution to problem (45) with u_0 replaced by $v_{0\epsilon} \in UC$ such that $\{v_{0\epsilon} \ge 0\} = K_{s-\epsilon}$, $\{v_{0\epsilon} > 0\} = \text{Int } K_{s-\epsilon}$ and $v_{0\epsilon} \ge 1$ on K_s. Therefore, again by the comparison principle, we have $\mathbf{1}_{\mathcal{R}_s} \le v_\epsilon$ and also

$$\mathcal{R}_s(t) \subseteq \{v_\epsilon(t,x) \ge 0\} = \{u(t,x) - s + \epsilon \ge 0\} \tag{47}$$

(the equality is due to the *geometricity* of the partial differential equation; see [6]).

By the stability theorem, passing to the limit as $\epsilon \to 0$ in relations (46)–(47), we obtain

$$\overline{\{(t,x) \mid u(t,x) > s\}} \subseteq \cup_t \mathcal{R}_s(t) \subseteq \{(t,x) \mid u(t,x) \ge s\} \qquad \forall s \in \mathbb{R}.$$

Taking into account assumptions (8) and (11), we have

$$\lim_{\eta\to 0^+} \cup_{t\in[0,T]}\mathcal{R}_{s+\eta}(t) = \cup_{t\in[0,T]}\mathcal{R}_s(t).$$

From the last two relations, we deduce that

$$\overline{\{(t,x)\in[0,T]\times\Omega \mid u(t,x)>s\}} = \{(t,x)\in[0,T]\times\Omega \mid u(t,x)\geq s\} \qquad \forall s\in\mathbb{R},$$

whence, by the arbitrariness of T, fattening does not occur. Finally, the uniqueness of the solution follows by an adaptation of the arguments used by Barles, Soner, and Souganidis [6]. □

Let us state the main result of this section.

THEOREM 4.3. *Under the assumptions* (9)–(11) *on the sets* Ω *and* K_0 *and* (41)–(42) *on* $\bar{c}_1$ *and* $\bar{c}_2$*, the initial boundary value problem* (39) *has a unique discontinuous viscosity solution* $\rho\in C^0\left([0,+\infty),\mathbb{L}^1(\Omega)\right)$.

Proof. Let us fix $T_0>0$ and prove the existence and uniqueness of the solution in a strip $[0,T_0]$. For this it is enough to show that there is some positive τ with the following property: for any $T\in[0,T_0]$, if the existence and uniqueness of solutions hold on $[0,T]$, then they hold on $[0,T+\tau]$.

Let us first define τ. For this let C, r_1, and N be the constants given in Propositions 2.10 and 3.1 for the time T_0+1. We fix $\tau\in(0,1]$ and $0<\alpha<1$ with

$$\tau\leq 1/MC, \qquad e^{(L_0+M/r_1)\tau}\leq 1+\alpha\frac{L_0+M/r_1}{NM}.$$

Next, we define the set

$$\mathcal{E}_T := \left\{\rho\in C^0\left([0,T+\tau],\mathbb{L}^1(\Omega)\right) \mid \begin{array}{l}\rho \text{ unique solution to (39) on } [0,T],\\ 0\leq\rho\leq 1,\ \sup\limits_{t\in[T,T+\tau]}\|\rho(t)\|_1\leq\|\rho(T,\cdot)\|_1+1\end{array}\right\}.$$

Let us introduce the mapping $\Phi:\mathcal{E}_T\to C^0\left([0,T+\tau],\mathbb{L}^1(\Omega)\right)$ defined in the following way: for $\rho^0\in\mathcal{E}_T$, $\Phi(\rho^0)$ is the unique viscosity solution to

$$\left\{\begin{array}{lll}\frac{\partial\rho}{\partial t}=\left(\bar{c}_1+I(\rho^0)\right)|\nabla\rho| & \text{in} & (0,T+\tau)\times\Omega,\\ \rho(0,x)=\rho_0(x)\equiv\mathbf{1}_{K_0}(x) & \text{on} & \Omega,\\ \frac{\partial\rho}{\partial\nu}=0 & \text{on} & (0,T+\tau)\times\partial\Omega.\end{array}\right. \tag{48}$$

The well-posedness of Φ is established in Lemma 4.2. We claim that Φ maps $\mathcal{E}_T$ into itself. Combining the representation formula (44) of Lemma 4.2 with Corollary 2.11, we have, for all $t\in[T,T+\tau]$,

$$\|\rho(t,\cdot)\|_1\leq\|\rho(t,\cdot)-\rho(T,\cdot)\|_1+\|\rho(T,\cdot)\|_1\leq\|\rho(T,\cdot)\|_1+1$$

(in the last inequality, relation $\tau\leq 1/MC$ has been used).

Now we claim that Φ is a contraction. Fix $\rho_1^0,\rho_2^0\in\mathcal{E}_T$ and set

$$c_i := \bar{c}_1+I(\rho_i^0) \qquad \text{and} \qquad \rho_i:=\Phi(\rho_i^0) \qquad (i=1,\,2).$$

We note that since $\rho_1^0=\rho_2^0$ on $[0,T]$, we have $c_1=c_2$ on $[0,T]$. Moreover,

$$\|c_1-c_2\|_\infty = \sup_{t\in[0,T+\tau]}\left\|I(\rho_1^0-\rho_2^0)(t,\cdot)\right\|_\infty \leq M\sup_{t\in[0,T+\tau]}\left\|\rho_1^0(t,\cdot)-\rho_2^0(t,\cdot)\right\|_1.$$

Therefore, using Proposition 3.1 and the fact that $c_1 = c_2$ on $[0,T]$, we have, for $t \in [T, T+\tau]$,

$$\begin{aligned}\|\rho_1(t,\cdot) - \rho_2(t,\cdot)\|_1 &\le N \|c_1 - c_2\|_\infty \int_T^t e^{(L_0+M/r_1)(t-s)} ds \\ &\le NM \frac{e^{(L_0+M/r_1)\tau} - 1}{L_0 + M/r_1} \sup_{t\in[T,T+\tau)} \left\|\rho_1^0(t,\cdot) - \rho_2^0(t,\cdot)\right\|_1 \\ &\le \alpha \sup_{t\in[T,T+\tau)} \left\|\rho_1^0(t,\cdot) - \rho_2^0(t,\cdot)\right\|_1\end{aligned}$$

(in the last inequality, relation $e^{(L_0+M/r_1)\tau} \le 1 + \alpha \frac{L_0+M/r_1}{NM}$ has been used). In particular, we have proved that Φ is a contraction, and by the Banach fixed point theorem, there exists a unique solution in the strip $[0, T+\tau]$. □

Remark 4.4. This result can be immediately extended to the case of functions $\bar{c}_2$ satisfying only (41)(i)–(iv). Actually, in the proof, one just needs the constant

$$\tilde{M} := \max\{\bar{c}_2(t,x,y) |\, t \in [0, T+1],\, x,\, y \in (K_0 + B_{M(T+1)})\},$$

which gives an upper estimate of $\bar{c}_2$ in the reachable set in time $T+1$. In this case, τ needs to be chosen in the following way:

$$\tau \le 1/MC, \qquad e^{(L_0+M/r_1)\tau} \le 1 + \alpha \frac{L_0 + M/r_1}{N\tilde{M}}.$$

5. Appendix: Proof of Lemma 2.3. Without loss of generality, we assume that the function c is defined and fulfills conditions (8) on the whole space $\mathbb{R}^+ \times \mathbb{R}^2$. For $\epsilon > 0$, we introduce the unconstrained controlled problem:

$$x' = c_\epsilon(t,x)b, \qquad |b| \le 1, \qquad \text{where } c_\epsilon(t,x) = \left(c(t,x) - \frac{d_\Omega(x)}{\epsilon}\right)_+,$$

where $(s)_+ = \max(0,s)$. We denote by $\mathcal{R}_\epsilon(t)$ the reachable set at time t starting from K_0 for this control system. Let us point out that it is a subset of $\mathbb{R}^2$ (and not only of $\overline{\Omega}$, as is $\mathcal{R}(t)$). From standard arguments, we have the inclusion $\mathcal{R}(t) \subset \mathcal{R}_\epsilon(t)$ for any $t \ge 0$. We are going to prove that for $\epsilon > 0$ small enough, extremal trajectories for the perturbed problem ending in $\overline{\Omega}$ are actually extremal trajectories for the initial problem.

Let $\epsilon > 0$ small to be chosen later. Also let $x_\epsilon(\cdot)$ be an extremal trajectory of the perturbed problem ending at some $x \in \overline{\Omega}$ at time T. From the definition of c_ϵ, the Pontryagin maximum principle for Lipschitz continuous dynamics (see, for instance, [10]) states that there are $p_\epsilon : [0,T] \to (\mathbb{R}^N \backslash \{0\})$ Lipschitz continuous and $\lambda_\epsilon : [0,T] \to [0,1]$ measurable such that

$$\text{(49)} \quad \begin{cases} x_\epsilon'(t) = c_\epsilon(t, x_\epsilon(t)) \dfrac{p_\epsilon(t)}{|p_\epsilon(t)|} & \forall t \in (0,T), \\ p_\epsilon'(t) = -\left[D_x c(t, x(t)) - \dfrac{\lambda_\epsilon(t)}{\epsilon} Dd(x(t))\right] |p(t)| & \text{for a.e. } t \in (0,T), \\ \lambda_\epsilon(t) = 0 \text{ if } x_\epsilon(t) \in \Omega, & \lambda_\epsilon(t) = 1 \text{ if } x_\epsilon(t) \notin \overline{\Omega}. \end{cases}$$

We claim that $x_\epsilon(t) \in \overline{\Omega}$ for $t \in [0,T]$. Indeed, suppose on the contrary that $x_\epsilon(t) \notin \overline{\Omega}$ for some $t \in [0,T]$. Let (a,b) be the largest interval containing t such that

$x_\epsilon(s) \notin \overline{\Omega}$ for $s \in (a,b)$. We note that $x_\epsilon(a)$ and $x_\epsilon(b)$ belong to $\partial\Omega$ because $x(T) \in \overline{\Omega}$. Moreover, at $s = a$ we have

$$\langle Dd(x_\epsilon(a)), x_\epsilon'(a)\rangle = \left\langle Dd(x_\epsilon(a)), c_\epsilon(a, x_\epsilon(a))\frac{p_\epsilon(a)}{|p_\epsilon(a)|}\right\rangle \ \geq\ 0.$$

In particular, $\langle Dd(x_\epsilon(a)), p_\epsilon(a)\rangle \geq 0$. In the same way, since $\langle Dd(x_\epsilon(b)), x_\epsilon'(b)\rangle \leq 0$, we have $\langle Dd(x_\epsilon(b)), p_\epsilon(b)\rangle \leq 0$. Hence

$$\begin{aligned} 0 &\geq \int_a^b \frac{d}{dt}\langle Dd(x_\epsilon(s)), p_\epsilon(s)\rangle\, ds \\ &\geq \int_a^b \left[c_\epsilon(s, x_\epsilon(s))\left\langle D^2 d(x_\epsilon(s))\frac{p_\epsilon(s)}{|p_\epsilon(s)|}, \frac{p_\epsilon(s)}{|p_\epsilon(s)|}\right\rangle |p_\epsilon(s)| \right. \\ &\qquad \left. - \langle D_x c(s, x_\epsilon(s)), Dd(x_\epsilon(s))\rangle |p_\epsilon(s)| + \tfrac{1}{\epsilon}|p_\epsilon(s)|\right] ds, \end{aligned}$$

where we have used the fact that for $\epsilon > 0$ sufficiently small and $s \in (a,b)$, $x_\epsilon(s)$ is close to $\partial\Omega$, and therefore d is of class $\mathcal{C}^2$ at $x_\epsilon(s)$ with $|Dd(x_\epsilon(s))| = 1$. From our assumptions, c_ϵ, Dc, and $D^2 d$ are bounded by some constant M, and therefore

$$0 \geq \int_a^b \left[-M^2 - M + \frac{1}{\epsilon}\right] |p_\epsilon(s)| ds.$$

This leads to a contradiction for $\epsilon < 1/[M + M^2]$ because $p_\epsilon(s) \neq 0$ for $s \in [0,T]$. So we have proved that for $\epsilon > 0$ sufficiently small and $t \in [0,T]$, $x_\epsilon(t) \in \overline{\Omega}$. In particular, x_ϵ is actually an extremal trajectory of the unperturbed problem on $[0,T]$.

We now fix ϵ as above. We now claim that, conversely, any extremal trajectory $x(\cdot)$ of the unperturbed problem on the time interval $[0,T]$ is an extremal trajectory of the perturbed problem. Indeed, let T_ϵ be the minimal time for which there is a solution of the perturbed problem reaching $x(T)$. Obviously, $T_\epsilon \leq T$. Moreover, there is an extremal trajectory $x_\epsilon(\cdot)$ for the perturbed problem reaching $x(T)$ at time T_ϵ. But we have already proved that this trajectory is actually a trajectory of the unperturbed problem. So $T_\epsilon \geq T$, whence we get an equality: $T_\epsilon = T$. This proves that $x(T) \in \partial_{\mathbb{R}^2}\mathcal{R}_\epsilon(T)$, and thus $x(\cdot)$ is actually an extremal trajectory for the perturbed problem.

Now let x be an extremal trajectory of the unperturbed problem on the time interval $[0,T]$. Since it is also extremal for the perturbed problem on $[0,T]$, we can use again the Pontryagin maximum principle: there are $p : [0,T] \to (\mathbb{R}^N \backslash \{0\})$ Lipschitz continuous and $\lambda_\epsilon : [0,T] \to [0,1]$ measurable such that

$$\begin{cases} x'(t) = c_\epsilon(t, x(t))\dfrac{p(t)}{|p(t)|} & \forall t \in (0,T), \\ p'(t) = -\left[D_x c(t, x(t)) - \dfrac{\lambda_\epsilon(t)}{\epsilon} Dd(x(t))\right] |p(t)| & \text{for a.e. } t \in (0,T), \\ \lambda_\epsilon(t) = 0 \text{ if } x(t) \in \Omega, & \lambda_\epsilon(t) = 1 \text{ if } x(t) \notin \overline{\Omega}. \end{cases}$$

Note that $c_\epsilon(t, x(t)) = c(t, x(t))$ because $x(t) \in \overline{\Omega}$ for any t. Let us now identify λ_ϵ. For this we set

$$E = \{t \in [0,T] \mid x(t) \in \partial\Omega\}.$$

Then, for almost every $t \in E$, we have $\langle Dd(x(t)), x'(t)\rangle = 0$, and thus $\langle Dd(x(t)), p(t)\rangle = 0$ thanks to (49). Therefore, for almost every $t \in E$, we also have

$$\left\langle D^2 d(x(t))c(t,x(t))\frac{p(t)}{|p(t)|}, p(t)\right\rangle + \left\langle Dd(x(t)), -\left(D_x c(t,x(t)) - \frac{\lambda_\epsilon(t)}{\epsilon} Dd(x(t))\right)|p(t)|\right\rangle = 0.$$

Since $|Dd| = 1$ and $\lambda_\epsilon \geq 0$, we get

$$\frac{\lambda_\epsilon(t)}{\epsilon} = \Big[-\left\langle D^2 d(x(t))c(t,x(t))\frac{p(t)}{|p(t)|}, \frac{p(t)}{|p(t)|}\right\rangle + \langle Dd(x(t)), D_x c(t,x(t))\rangle\Big]_+ \mathbf{1}_{\partial\Omega}(x(t)).$$

Setting $\lambda(t) = \lambda_\epsilon(t)/\epsilon$, we get the desired result.

Acknowledgments. The authors are grateful to the anonymous referees for their helpful comments and suggestions. This paper was started while the second author was visiting the Laboratoire de Mathématiques UMR 6205 at the Université de la Bretagne Occidentale (Brest, France); he is grateful to this institution for the warm hospitality and support.

REFERENCES

[1] O. Alvarez, P. Cardaliaguet, and R. Monneau, *Existence and uniqueness for dislocation dynamics with nonnegative velocity*, Interfaces Free Bound., 7 (2005), pp. 415–434.

[2] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Résolution en temps court d'une équation de Hamilton-Jacobi non locale décrivant la dynamique d'une dislocation*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 679–684.

[3] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Dislocation Dynamics: Short Time Existence and Uniqueness of the Solution*, Arch. Ration. Mech. Anal., to appear.

[4] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.

[5] G. Barles, *Solutions de Viscosité des Equations de Hamilton-Jacobi*, Springer-Verlag, Berlin, 1994.

[6] G. Barles, H. M. Soner, and P. E. Souganidis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[7] P. Cannarsa and H. Frankowska, *Interior sphere property of attainable sets and time optimal control*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 350–370.

[8] P. Cannarsa and C. Sinestrari, *Semiconcave Functions, Hamilton-Jacobi Equations and Optimal Control*, Birkhäuser, Boston, 2004.

[9] P. Cardaliaguet, *On front propagation problems with nonlocal terms*, Adv. Differential Equations, 5 (2000), pp. 213–268.

[10] F. H. Clarke, *The maximum principle under minimal hypotheses*, SIAM J. Control Optim., 14 (1976), pp. 1078–1091.

[11] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[12] H. Frankowska, *Regularity of minimizers and of adjoint states in optimal control under state constraints*, J. Convex Anal., to appear.

[13] D. Rodney, Y. Le Bouar, and A. Finel, *Phase field methods and dislocations*, Acta Materialia, 51 (2003), pp. 17–30.

[14] O.-S. Serea, *On reflecting boundary problem for optimal control*, SIAM J. Control Optim., 42 (2003), pp. 559–575.

[15] H. J. Sussmann, ed., *Nonlinear Controllability and Optimal Control*, Marcel Dekker, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 1039–1073

CLASSIFICATION OF ESTIMATION ALGEBRAS WITH STATE DIMENSION 2*

XI WU† AND STEPHEN S.-T. YAU†

Abstract. This paper considers general finite-dimensional estimation algebras associated with nonlinear filtering systems. General considerations and approaches toward the classification of finite-dimensional estimation algebras are proposed. Some structural results are obtained. The properties of Euler operator and the solutions to an underdetermined partial differential equation, which inevitably arise in an estimation algebra, are studied. These tools and techniques are applied to the study of finite-dimensional estimation algebras with state dimension 2 to obtain a complete classification result. It is shown that a finite-dimensional estimation algebra with state dimension 2 can only have dimension less than or equal to 6. Moreover, the Mitter conjecture and the Levine conjecture hold for finite-dimensional estimation algebras with state dimension 2.

Key words. finite-dimensional filter, estimation algebra, nonlinear drift

AMS subject classifications. 17B30, 35J15, 60G35, 93E11

DOI. 10.1137/S036301290443991X

1. Introduction.

The Lie algebraic method, pioneered by Brockett [3], Brockett and Clark [5], and Mitter [15], provides an important research direction for nonlinear filtering theory. By interpreting the Duncan–Mortensen–Zakai equation or its robust form as a partial differential equation with time varying parameters, one derives an approach to filtering based on Lie algebra as well as the theory of linear differential operators. The search and construction of the finite-dimensional filter are turned into the study of the structure of the estimation algebra. In return, the theory of estimation algebra provides a systematic tool to deal with questions concerning finite-dimensional filters. It has led to a number of new results concerning finite-dimensional filters and to a deeper understanding of the structure of nonlinear filtering in general. It explains convincingly in [6] and [11] why it is easy to find exact recursive filters for linear dynamical systems, while it is very hard to handle the cubic sensor problem. Some new filters have been discovered using the estimation algebra [1], [16], [20], [21], [24]. More importantly, the finite-dimensionality of the estimation algebra guarantees the explicit construction of the finite-dimensional recursive filter, and the filter so constructed is universal in the sense of [7]. However, as an inherited difficulty from the nonlinear filtering, many questions are not satisfyingly answered and still open. Two excellent survey papers [10], [14] provide the detailed material on the Wei–Norman approach, the connections between estimation algebra and nonlinear filtering, the application of the Lie algebra method to a variety of specific nonlinear filtering problems, and the construction of the approximate filters, as well as sufficient reviews of the requisite background and a sufficient number of examples. For more contemporary results, see [23].

*Received by the editors January 15, 2004; accepted for publication (in revised form) January 16, 2006; published electronically August 16, 2006.

http://www.siam.org/journals/sicon/45-3/43991.html

†Department of MSCS (M/C 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (xiwu@pobox.com, yau@uic.edu). The research of the first author was partially supported by U.S. Army grant DAAD 19-02-10292 and NSF grant 0503868. The research of the second author was partially supported by U.S. Army grant DAAD 19-02-10292, W911NF-04-1-0085, and NSF grant 0503868.

At the International Congress of Mathematics in 1983, Brockett [4] proposed classifying all the finite-dimensional estimation algebras. In [26], the concept of estimation algebra of maximal rank was introduced by Yau and Chiou. The complete classification of such a subclass finite-dimensional estimation algebra was done by Yau and his coworkers in a series of papers, and the final results were reported in [25]. In [27], the conditions by which a general estimation algebra has dimension less than or equal to 4 were discussed. In [17], a new class of finite-dimensional estimation algebras were constructed when the state dimension $n \geq 3$. In this paper, the finite-dimensional estimation algebras that can arise when the underlying stochastic systems have state dimension 2 are completely classified. It is shown that an estimation algebra with state dimension 2 can have only 1, 2, 4, 5, or 6 dimensions. The finite-dimensional estimation algebras with state dimension 1 were considered in [5] and [16].

This paper is organized as follow: Section 2 introduces some basic concepts of estimation algebra for a nonlinear filtering system and gives the classification result on the finite-dimensional estimation algebra of maximal rank. In section 3, some tools and techniques are discussed for the classification of general estimation algebras. The complete classification of estimation algebras with state dimension 2 is given in section 4.

2. Basic concepts. In this paper, the filtering system to be considered is based on the following continuous signal-observation model:

$$\begin{cases} dx(t) = f\big(x(t)\big)dt + g\big(x(t)\big)dv(t), & x(0) = x_0, \\ dy(t) = h\big(x(t)\big)dt + dw(t), & y(0) = 0. \end{cases} \tag{2.1}$$

Here x, v, y, and w are, respectively, R^n, R^n, R^m, and R^m valued processes, and v and w have components which are independent, standard Brownian processes. Moreover, $f = (f_1, \ldots, f_n)$ and $h = (h_1, \ldots, h_m)$ are assumed to be C^∞ smooth and g is an orthogonal matrix. $x(t)$ is referred to as the state of the system at time t and $y(t)$ as the observation at time t.

Let $\rho(t, x)$ denote the conditional probability density of the state $x(t)$ given the observation $\{y(s) : 0 \leq s \leq t\}$. $\rho(t, x)$ is governed by the well-known Duncan–Mortensen–Zakai (DMZ) equation, which is a stochastic partial differential equation in terms of an unnormalized version of $\rho(t, x)$, denoted by $\sigma(t, x)$ (see [13], for example). Under the Stratonovich calculus, the DMZ equation for the filtering system (2.1) can be written as

$$\begin{cases} d\sigma(t, x) = L_0\sigma(t, x)dt + \displaystyle\sum_{i=1}^{m} h_i\sigma(t, x)dy_i(t), \\ \sigma(0, x) = \sigma_0, \end{cases} \tag{2.2}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} - \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{m} h_i^2. \tag{2.3}$$

The term σ_0 is the probability density of the initial state x_0.

The Wei–Norman result [19] on constructing explicit solutions to complex partial differential equations suggests a formal approach to constructing a finite-dimensional filter by defining the estimation algebra as a Lie algebra generated by the operators occurring in the DMZ equation.

DEFINITION 2.1. *If X and Y are differential operators, the Lie bracket of X and Y, $[X,Y]$, is defined by $[X,Y]\phi = X(Y\phi) - Y(X\phi)$ for any C^∞-function ϕ.*

DEFINITION 2.2. *The estimation algebra E of a filtering system* (2.1) *is defined as the Lie algebra generated by $\{L_0, h_1, \dots, h_m\}$, i.e., $E = \langle L_0, h_1, \dots, h_m\rangle_{L.A.}$, where L_0 is defined in* (2.3).

The following theorem due to Ocone [16] is the first result to show what kind of functions can appear in a finite-dimensional estimation algebra.

THEOREM 2.3 (see [16]). *Let E be a finite-dimensional estimation algebra. If ϕ is a function in E, then ϕ is a polynomial of degree at most* 2.

In fact, Mitter and Levine have conjectured the following.

MITTER CONJECTURE. *Let E be a finite-dimensional estimation algebra. If ϕ is a function in E, then ϕ is a polynomial of degree at most* 1.

LEVINE CONJECTURE (see [12]). *Let E be a finite-dimensional estimation algebra. The differential operators in E have orders at most* 2.

A fundamental step in the Lie algebraic approach as introduced in [21] is to represent the elliptic operator L_0 in (2.3) in a more compact form by defining

$$D_i = \frac{\partial}{\partial x_i} - f_i, \tag{2.4}$$

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2. \tag{2.5}$$

Then

$$L_0 = \frac{1}{2}\Big(\sum_{i=1}^{n} D_i^2 - \eta\Big). \tag{2.6}$$

DEFINITION 2.4. *The Ω-matrix of a filtering system* (2.1) *is an $n \times n$ matrix $\Omega = (\omega_{ij})$ defined by*

$$\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} \qquad \forall\ 1 \le i, j \le n. \tag{2.7}$$

When $\Omega = O$, i.e., f is a gradient vector field in view of the Poincaré lemma, the filtering system and the corresponding estimation algebra are called exact. In [9] and [18], the finite-dimensional exact estimation algebras are completely classified. In [24], Yau classified the finite-dimensional estimation algebras when the Ω-matrix has constant entries.

THEOREM 2.5 (see [24]). *Let E be an estimation algebra of* (2.1) *whose Ω-matrix has constant entries.*

(1) *If η (in* (2.5)*) is a polynomial of degree at most* 2 *and $h_1, \dots, h_m$ are affine in x, then E is finite-dimensional and has a basis consisting of $E_0 = L_0$, $E_1, \dots, E_p, E_{p+1}, \dots, E_q, 1$ (for some $p < q$). The differential operators $E_1, \dots, E_p$ have the form*

$$\sum_{j=1}^{n} \alpha_{ij} D_j + \beta_j, \qquad 1 \le i \le p,$$

where α_{ij}'s are constants and β_i's are affine in x, and the differential operators $E_{p+1}, \dots, E_q$ are affine in x. Moreover, the quadratic part of $\eta - \sum_{i=1}^{m} h_i^2$ is positive semidefinite.

(2) *Conversely, if E is finite-dimensional, then $h_1, \dots, h_m$ are affine in x, i.e., the observation matrix $H = [\nabla h_1, \dots, \nabla h_m]$ is a constant matrix. Furthermore, if the observation matrix has rank n, then η is a polynomial of degree at most 2 and E is of dimension $2n+2$ with a basis given by $1, x_1, \dots, x_n, D_1, \dots, D_n$, and L_0.*

DEFINITION 2.6. *An estimation algebra E is said to be of maximal rank if, for any $1 \le i \le n$, there exists a constant c_i, such that $x_i + c_i \in E$.*

By direct calculation, if E is of maximal rank, $[L_0, x_i + c_i] = D_i \in E$, and $[D_i, x_i + c_i] = 1 \in E$. Thus, $\langle 1, x_1, \dots, x_n \rangle$ form a vector space of E. Any degree 1 polynomial is an element of E.

THEOREM 2.7 (see [25]). *Let E be an estimation algebra of (2.1). If E is of maximal rank and finite-dimensional, then*

(1) $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ $\forall 1 \le i, j \le n$, *where c_{ij}'s are constants;*

(2) $h_1, \dots, h_m$ *are affine in x;*

(3) η *is a polynomial of degree 2 and the quadratic part of $\eta - \sum_{i=1}^m h_i^2$ is positive semidefinite.*

Moreover, E is of dimension $2n+2$ with a basis given by $1, x_1, \dots, x_n, D_1, \dots, D_n$, and L_0.

For the convenience of the reader, the following elementary lemma on the calculations of the Lie bracket of the estimation algebra is listed without proof.

LEMMA 2.8 (see [8], [24]). *Let E be an estimation algebra for the filtering system (2.1). $\Omega = (\omega_{ij})$ is defined as in (2.7). $X, Y, Z \in E$ and $g, h \in C^\infty(R^n)$. Then*

(1) $[XY, Z] = X[Y, Z] + [X, Z]Y$;

(2) $[gD_i, h] = g\frac{\partial h}{\partial x_i}$;

(3) $[gD_i, hD_j] = -gh\omega_{ij} + g\frac{\partial h}{\partial x_i}D_j - h\frac{\partial g}{\partial x_j}D_i$;

(4) $[gD_i^2, h] = 2g\frac{\partial h}{\partial x_i}D_i + g\frac{\partial^2 h}{\partial x_i^2}$;

(5) $[D_i^2, hD_j] = 2\frac{\partial h}{\partial x_i}D_iD_j - 2h\omega_{ij}D_i + \frac{\partial^2 h}{\partial x_i^2}D_j - h\frac{\partial \omega_{ij}}{\partial x_i}$;

(6) $[D_i^2, D_j^2] = 4\omega_{ji}D_jD_i + 2\frac{\partial \omega_{ji}}{\partial x_j}D_i + 2\frac{\partial \omega_{ji}}{\partial x_i}D_j + \frac{\partial^2 \omega_{ji}}{\partial x_i \partial x_j} + 2\omega_{ji}^2$;

(7) $[D_k^2, hD_iD_j] = 2\frac{\partial h}{\partial x_k}D_kD_iD_j + 2h\omega_{jk}D_iD_k + 2h\omega_{ik}D_kD_j + \frac{\partial^2 h}{\partial x_k^2}D_iD_j + 2h\frac{\partial \omega_{jk}}{\partial x_i}D_k + h\frac{\partial \omega_{jk}}{\partial x_k}D_i + h\frac{\partial \omega_{ik}}{\partial x_k}D_j + h\frac{\partial^2 \omega_{jk}}{\partial x_i \partial x_k}$;

(8) $[gD_iD_j, hD_k] = g\frac{\partial h}{\partial x_j}D_iD_k + g\frac{\partial h}{\partial x_i}D_jD_k + gh\omega_{kj}D_i + gh\omega_{ki}D_j + g\frac{\partial^2 h}{\partial x_i \partial x_j}D_k + gh\frac{\partial \omega_{kj}}{\partial x_i} - h\frac{\partial g}{\partial x_k}D_iD_j$.

3. Preliminary results.

3.1. Differential operator. Let U be the set of differential operators in the form

$$A = \sum_{(i_1, \dots, i_n) \in I_A} a_{i_1, \dots, i_n} D_1^{i_1} \cdots D_n^{i_n}, \tag{3.1}$$

where nonzero functions $a_{i_1, \dots, i_n} \in C^\infty(R^n)$ and I_A is the finite index set of A. Each element of the index set is an n-tuple $(i_1, \dots, i_n)$ of nonnegative integers. The norm of an index $i = (i_1, \dots, i_n)$ is defined to be $|i| = \sum_{l=1}^n i_l$. The order of A is denoted by $ord\ A = \max_{i=(i_1, \dots, i_n) \in I_A} |i|$. If $A = 0$, $ord\ A$ is defined to be $-\infty$. It is clear that for $A, B \in U$

$$ord\ (AB) = ord\ (BA) = ord\ A + ord\ B, \tag{3.2}$$

$$ord\ (A \pm B) \le \max(ord\ A, ord\ B). \tag{3.3}$$

U is a Lie algebra under the Lie bracket $[\cdot,\cdot]$ defined earlier. Two differential operators A and B in U are equal if they have the identical index sets $I_A = I_B$ and $a_{i_1,\ldots,i_n} = b_{i_1,\ldots,i_n} \forall (i_1,\ldots,i_n) \in I_A$. Let U_k denote the subspace of U consisting of elements with order less than or equal to k. In particular, $U_0 = C^\infty(R^n)$. As usual, *mod* is used to denote the equivalence class, i.e., if V is a subspace of U,

$$A = B, \qquad \text{mod } V \Longleftrightarrow A - B \in V. \tag{3.4}$$

If $A, B \in U$, define

$$Ad_A B = [A, B], \qquad Ad_A^l B = [A, Ad_A^{l-1} B], \quad l \geq 1, \tag{3.5}$$

where Ad_A^0 is the identity operator by standard convention.

Recall that

$$D_i = \frac{\partial}{\partial x_i} - f_i, \qquad 1 \leq i \leq n, \tag{3.6}$$

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^n D_i^2 - \eta\right), \qquad \eta = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2. \tag{3.7}$$

The estimation algebra $E = \langle L_0, h_1, \ldots, h_m \rangle_{L.A.}$. Since $L_0 \in U_2$ and $h_i \in U_0, 1 \leq i \leq m$, $E \subset U$. Every element of E will have representation in the form of (3.1).

The finite-dimensionality of E can be measured in terms of the finite order of the elements in E. If E is finite-dimensional, then the orders of its element will have an upper bound. In particular, if there exists a sequence of elements $A_j \in E$ such that the orders of A_j's are strictly increasing, E is not finite-dimensional.

The following is a very useful lemma in the computation of the orders of the differential operators generated under the Lie bracket.

LEMMA 3.1. *Let $g, h \in C^\infty(R^n)$ and let $i_1, \ldots, i_n, j_1, \ldots, j_n$ be nonnegative integers with $\sum_{l=1}^n i_l = r$, $\sum_{l=1}^n j_l = s$, and $r + s \geq 2$. Let δ_{ij} be the Kronecker symbol, i.e., $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. Then*

$$\begin{aligned}[gD_1^{i_1} \cdots D_n^{i_n}, hD_1^{j_1} \cdots D_n^{j_n}] \\ = \sum_{l=1}^n \left(i_l g \frac{\partial h}{\partial x_l} - j_l h \frac{\partial g}{\partial x_l}\right) D_1^{i_1+j_1-\delta_{1l}} \cdots D_n^{i_n+j_n-\delta_{nl}}, \quad \text{mod } U_{r+s-2}.\end{aligned} \tag{3.8}$$

Proof. If $r + s = 2$, there are two cases:

$$[gD_iD_j, h] = -[h, gD_iD_j] = g\frac{\partial h}{\partial x_j}D_i + g\frac{\partial h}{\partial x_i}D_j + g\frac{\partial^2 h}{\partial x_i \partial x_j}, \tag{3.9}$$

$$[gD_i, hD_j] = -[hD_j, gD_i] = gh\omega_{ij} + g\frac{\partial h}{\partial x_i}D_j - h\frac{\partial g}{\partial x_j}D_i. \tag{3.10}$$

Clearly, (3.8) holds for these two cases. Induction on $r + s$ is used to prove (3.8).

If $r + s > 2$, at least one of r and s is greater than 0. Without loss of generality, assume $r > 0$. (In the case where $r = 0$ and $s > 0$, by using the property $[A, B] = -[B, A]$, the same procedure applies below.) Since $\sum_{l=1}^n i_l = r > 0$, at least one i_l is no less than 1. Assume $i_k \geq 1$. By Lemma 2.8, $gD_1^{i_1} \cdots D_k^{i_k} \cdots D_n^{i_n} = D_k g D_1^{i_1} \cdots D_k^{i_k - 1} \cdots D_n^{i_n}$, mod U_{r-1}. Hence,

$$gD_1^{i_1} \cdots D_k^{i_k} \cdots D_n^{i_n} = D_k g D_1^{i_1} \cdots D_k^{i_k - 1} \cdots D_n^{i_n} + A, \tag{3.11}$$

where $A \in U_{r-1}$. By the assumption of the induction, $[A, hD_1^{j_1} \cdots D_n^{j_n}] \in U_{r+s-2}$, since each term of A has order less than or equal to $r-1$. Therefore,

$$
\begin{aligned}
&[gD_1^{i_1} \cdots D_n^{i_n}, hD_1^{j_1} \cdots D_n^{j_n}] \\
&= [D_k g D_1^{i_1} \cdots D_k^{i_k - 1} \cdots D_n^{i_n}, hD_1^{j_1} \cdots D_n^{j_n}] \\
&= D_k [gD_1^{i_1} \cdots D_k^{i_k - 1} \cdots D_n^{i_n}, hD_1^{j_1} \cdots D_n^{j_n}] \\
&\qquad + [D_k, hD_1^{j_1} \cdots D_n^{j_n}] g D_1^{i_1} \cdots D_k^{i_k - 1} \cdots D_n^{i_n} \\
&= D_k \sum_{l=1}^{n} \left((i_l - \delta_{lk}) g \frac{\partial h}{\partial x_l} - j_l h \frac{\partial g}{\partial x_l} \right) D_1^{i_1 + j_1 - \delta_{1l}} \cdots D_k^{i_k + j_k - 1 - \delta_{kl}} \cdots D_n^{i_n + j_n - \delta_{nl}} \\
&\qquad + \frac{\partial h}{\partial x_k} D_1^{j_1} \cdots D_k^{j_k} \cdots D_n^{j_n} g D_1^{i_1} \cdots D_k^{i_k - 1} \cdots D_n^{i_n} \\
&= \sum_{l=1}^{n} \left((i_l - \delta_{lk}) g \frac{\partial h}{\partial x_l} - j_l h \frac{\partial g}{\partial x_l} \right) D_1^{i_1 + j_1 - \delta_{1l}} \cdots D_k^{i_k + j_k - \delta_{kl}} \cdots D_n^{i_n + j_n - \delta_{nl}} \\
&\qquad + g \frac{\partial h}{\partial x_k} D_1^{i_1 + j_1} \cdots D_k^{i_k + j_k - 1} \cdots D_n^{i_n + j_n} \\
&= \sum_{l=1}^{n} \left(i_l g \frac{\partial h}{\partial x_l} - j_l h \frac{\partial g}{\partial x_l} \right) D_1^{i_1 + j_1 - \delta_{1l}} \cdots D_n^{i_n + j_n - \delta_{nl}}, \quad \mod U_{r+s-2}. \qquad \square
\end{aligned}
$$

In fact, (3.8) is still valid for $r + s = 0$ or 1 if the definition of U_k is extended to when $k < 0$, $U_k = \{0\}$, i.e., U_k has just one element 0 when $k < 0$.

A direct result from Lemma 3.1 is that for $A, B \in U$,

$$
ord\ [A, B] \leq ord\ A + ord\ B - 1. \tag{3.12}
$$

Moreover, if $ord\ [A, B] = ord\ A + ord\ B - 1$, the highest order terms of $[A, B]$ are uniquely determined in (3.8) by those of A and B.

The following is a very important theorem on the polynomial structure of coefficient functions of the differential operators in E which generalizes Wong's [22] result.

THEOREM 3.2. *Let E be a finite-dimensional estimation algebra and let the D_i's be defined as in* (3.6). *If $l \geq 0$ and*

$$
A = \sum_{|(i_1, \ldots, i_n)| = l+1} a_{i_1, \ldots, i_n} D_1^{i_1} \cdots D_n^{i_n}, \quad \mod U_l,
$$

is in E, then $a_{i_1, \ldots, i_n}$'s are polynomials.

Proof. If not all the $a_{i_1, \ldots, i_n}$'s are polynomials, at least one, say, $a_{j_1, \ldots, j_n}$, is transcendental. Then there exists a variable x_k in which $a_{j_1, \ldots, j_n}$ is transcendental, i.e., $\frac{\partial^s a_{j_1, \ldots, j_n}}{\partial x_k^s} \neq 0\ \forall s \geq 0$.

Now, among the indices $(j_1, \ldots, j_n)$ of all $a_{j_1, \ldots, j_n}$ that are transcendental in x_k, there exists one $(j_1, \ldots, j_n)$ having the largest kth index j_k. Without loss of generality, assume $k = 1$ and $a_{j_1, \ldots, j_n}$ is transcendental in x_1 and j_1 is the largest among the first indices whose coefficient functions are transcendental in x_1, i.e., if $a_{i_1, \ldots, i_n}$ is

transcendental in x_1 as well, then $j_1 \geq i_1$. Let

$$A_1 = [L_0, A] = \left[L_0, \sum_{|(i_1,\ldots,i_n)|=l+1} a_{i_1,\ldots,i_n} D_1^{i_1} \cdots D_n^{i_n} \right]$$

$$= \sum_{|(i_1,\ldots,i_n)|=l+1} \sum_{r=1}^{n} \frac{\partial a_{i_1,\ldots,i_n}}{\partial x_r} D_1^{i_1} \cdots D_{r-1}^{i_{r-1}} D_r^{i_r+1} D_{r+1}^{i_{r+1}} \cdots D_n^{i_n}, \mod U_{l+1}.$$

The coefficient function of $D_1^{j_1+1} D_2^{j_2} \cdots D_n^{j_n}$ in A_1 is

$$\frac{\partial a_{j_1,j_2,\ldots,j_n}}{\partial x_1} + \frac{\partial a_{j_1+1,j_2-1,\ldots,j_n}}{\partial x_2} + \cdots + \frac{\partial a_{j_1+1,j_2,\ldots,j_n-1}}{\partial x_n}. \tag{3.13}$$

Since j_1 is the largest first index such that a is transcendental in x_1, except that the first term in (3.13) is transcendental in x_1, the other $(n-1)$ terms are all polynomials in x_1. Hence, the coefficient function (3.13) of $D_1^{j_1+1} D_2^{j_2} \cdots D_n^{j_n}$ is transcendental in x_1.

Similarly, let $|(i_1, \ldots, i_n)| = l+2$, $i_1 \geq j_1+2$, and consider the coefficient functions of $D_1^{i_1} D_2^{i_2} \cdots D_n^{i_n}$ in A_1. Their coefficient functions are the sum of $\frac{\partial a_{p_1,p_2,\ldots,p_n}}{\partial x_q}$, where $|(p_1, \ldots, p_n)| = l+1$ and $p_1 \geq j_1 + 1$. By assumption on the a's, these functions are all polynomials in x_1.

Thus, from A to A_1, the differential orders increase by 1, while the transcendental structure on x_1 remains unchanged. The coefficient function of $D_1^{j_1+1} D_2^{j_2} \cdots D_n^{j_n}$ in A_1 is transcendental in x_1 and $j_1 + 1$ is the largest first indices whose coefficient functions are transcendental in x_1. (Note that there may exist other a's having this property, but they must have different numbers in another index, say the nth index. These terms are linearly independent.)

By keeping this process, $A_2 = [L_0, A_1], A_3 = [L_0, A_2], \ldots$. A_k has order of $l + k + 1 \to \infty$. Contradiction! □

3.2. Linear rank and quadratic rank. Unlike estimation algebras of maximal rank (cf. Definition 2.6), it is not clear whether x_i's are elements of a general estimation algebra E. Moreover, 1 may not be in E either. To simplify the situation, the concept of linear rank is introduced in this section together with quadratic rank to explore the structure of the function elements in E. In what follows, dim E is denoted the dimension of the estimation algebra E. For a polynomial ϕ, $\phi^{(k)}$ denotes its homogeneous degree k part.

THEOREM 3.3. *Let $E = \langle L_0, h_1, \ldots, h_m \rangle_{L.A.}$ and $\bar{E} = \langle 1, L_0, h_1, \ldots, h_m \rangle_{L.A.}$. Then E is finite-dimensional if and only if $\bar{E}$ is finite-dimensional.*

Proof. Since $E \subset \bar{E}$, it suffices to show dim $\bar{E}$ − dim $E \leq 1$ if dim $E < \infty$.

Start from the generators of $\bar{E}$ to construct increasing subsets A_k of $\bar{E}$, as follows:

- $A_0 = \{1, L_0, h_1, \ldots, h_m\}$;
- for $k \geq 1$, $A_k = \{a_1 B_1 + a_2 B_2 + a_3 [B_3, B_4] : a_i \in R, B_j \in A_{k-1}, i = 1, 2, 3, j = 1, 2, 3, 4\}$.

Clearly $A_k \nearrow \bar{E}$ as $k \to \infty$. Now, assume that dim $E < \infty$. Let vector space $E^* = E + 1 := \{aX + b : X \in E, a, b \in R\}$. It is easy to show by induction that

$A_k \subset E^*$ for $k \geq 0$. Thus,

$$\bar{E} = \cup_{k\geq 0} A_k \subset E^*,$$

which means

$$\dim \bar{E} \leq \dim E^* \leq \dim E + 1. \qquad \square$$

Thus, the estimation algebra E of a filtering problem being finite-dimensional is equivalent to the finite-dimensionality of $\langle E, 1\rangle_{L.A.}$. In the discussion of the finite-dimensionality of E, 1 can always be assumed an element of E.

Under the assumption that $1 \in E$, any degree 1 polynomial in $E \Longrightarrow$ its homogeneous degree 1 part is in E.

DEFINITION 3.4. *Let $L(E) \subset E$ be the vector space (Lie subalgebra of E) consisting of all the homogeneous degree* 1 *polynomials in E. Then $\nu(E) := \dim L(E)$ is called the linear rank of the estimation algebra E.*

From this definition, an estimation algebra of maximal rank is in fact an estimation algebra with linear rank n.

Note that an estimation algebra is associated with a filtering system and is coordinate-dependent. The recognition of the structurally equivalent estimation algebras is very important in the classification problem. Since an estimation algebra is essentially a Lie algebra, the definitions of homomorphism and isomorphism of estimation algebras follow from those of Lie algebras. It is shown in [2], [8], and [14] that orthogonal variable transformation and affine transformations extend to estimation algebra isomorphisms.

Let $\nu(E) = r$. Clearly $r \leq n$ and there exist r independent linear functions (the basis of L) $l_1(x), \ldots, l_r(x)$ such that

$$(l_1(x), \ldots, l_r(x))^T = Ax,$$

where A is an $r \times n$ matrix with rank r. From the singular value decomposition theorem, there exist orthogonal matrices $U(r \times r)$, $V(n \times n)$ such that

$$A = U[D \quad 0]V^T,$$

where $D = \operatorname{diag}(d_1, \ldots, d_r)$, with $d_1, \ldots, d_r \neq 0$ the singular values of A. Thus,

$$U^T Ax = U^T AV(V^T x) = [D \quad 0]V^T x.$$

After an orthogonal change of the variable $y = V^T x$, it is easy to see that for $1 \leq i \leq r$, y_i is a linear combination of $l_j(x)$ $(1 \leq j \leq r)$. In view of $\nu(E) = r$, $\{y_i, 1 \leq i \leq r\}$ is the basis of $L(E)$.

Hence, by an orthogonal variable transformation, if necessary, a linear function l is in E if and only if

$$l(x) \in L(E) := span\{x_1, \ldots, x_r\}.$$

DEFINITION 3.5. *For a given function $h \in E$, its quadratic part $h^{(2)} = x^T Ax$ for a symmetric matrix A. The quadratic rank $\lambda(h)$ of h is defined as the rank of A. The quadratic rank $\lambda(E)$ of E is the greatest quadratic rank of $h \in E$, i.e., $\lambda(E) = \max_{h\in E} \lambda(h)$.*

LEMMA 3.6. *Let E be an estimation algebra with linear rank r and $\phi \in E$. Then the quadratic part of ϕ is*

$$\phi^{(2)} = x^T \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} x,$$

where A_1 and A_2 are symmetric matrices with dimensions $r \times r$ and $(n-r)\times(n-r)$.

Proof. $\phi^{(2)}$ is a degree 2 homogeneous polynomial; therefore it can be written as $\phi^{(2)} = x^T A x$ for a symmetric matrix $A = (a_{ij})$ of dimension $n \times n$.

Since $\nu(E) = r$, $x_i \in E$ if and only if $1 \le i \le r$. For $1 \le i \le r$, $[L_0, x_i] = D_i \in E$, and

$$[D_i, \phi] = \frac{\partial \phi}{\partial x_i} = 2\sum_{j=1}^{n} a_{ij} x_j \in E, \qquad \text{mod } R.$$

Hence, $a_{ij} = 0$ for $r+1 \le j \le n$, $1 \le i \le r$. By the symmetry of A, the lemma follows. □

Now, consider an estimation algebra E with linear rank r and quadratic rank k. From the definition of the quadratic rank of E, there exists $p_0 \in E$ such that the quadratic rank of p_0 is k. By Lemma 3.6, $p_0^{(2)}(x) = x^T \left(\begin{smallmatrix} A_1 & 0 \\ 0 & A_2 \end{smallmatrix}\right) x$. Let $k_1 = \operatorname{rank}(A_1)$, $k_2 = \operatorname{rank}(A_2)$. Then $k = k_1 + k_2$, $0 \le k_1 \le r$, and $0 \le k_2 \le n-r$. Since A_1 and A_2 are real symmetric, there are orthogonal matrices U_1 and U_2 such that $A_1 = U_1 \left(\begin{smallmatrix} D_1 & 0 \\ 0 & 0 \end{smallmatrix}\right) U_1^T$ and $A_2 = U_2 \left(\begin{smallmatrix} 0 & 0 \\ 0 & D_2 \end{smallmatrix}\right) U_2^T$, where D_1 and D_2 are nonsingular diagonal matrices with dimensions $k_1 \times k_1$ and $k_2 \times k_2$. By taking the orthogonal variable transformation $T = \left(\begin{smallmatrix} U_1 & 0 \\ 0 & U_2 \end{smallmatrix}\right)$, $p_0^{(2)} = \sum_{i=1}^{k_1} d_i x_i^2 + \sum_{i=n-k_2+1}^{n} d_i x_i^2$, where $d_i \ne 0$. Moreover, by an affine variable transformation, if necessary,

$$p_0 = \sum_{i=1}^{k_1} d_i x_i^2 + \sum_{i=n-k_2+1}^{n} d_i x_i^2 + \sum_{i=k_1+1}^{n-k_2} c_i x_i + c_0 \in E \tag{3.14}$$

$$\Longrightarrow [[L_0, p_0], p_0] = \sum_{i=1}^{k_1} 4 d_i^2 x_i^2 + \sum_{i=n-k_2+1}^{n} 4 d_i^2 x_i^2 + \sum_{i=k_1+1}^{n-k_2} c_i^2 \in E$$

$$\Longrightarrow q_0 := \sum_{i=1}^{k_1} d_i^2 x_i^2 + \sum_{i=n-k_2+1}^{n} d_i^2 x_i^2 \in E, \tag{3.15}$$

$$\Longrightarrow q_j := [[L_0, q_{j-1}], q_{j-1}] = \sum_{i=1}^{k_1} 4^j d_i^{2j+2} x_i^2 + \sum_{i=n-k_2+1}^{n} 4^j d_i^{2j+2} x_i^2 \in E, \quad j \ge 1. \tag{3.16}$$

If no d_i^2's are equal, then the coefficient matrix of (3.16) for x_i^2 forms a Vandermonde matrix. By the invertibility of the Vandermonde matrix, x_i^2 can be represented as a linear combination of q_j's, and therefore $x_i^2 \in E$. Hence,

$$\sum_{i=1}^{k_1} x_i^2 + \sum_{i=n-k_2+1}^{n} x_i^2 \in E. \tag{3.17}$$

If some d_i^2's are equal, for example, d_i^2's equal for $l_1 \le i \le l_2$ in (3.14), they can be grouped to be solved as one variable. Instead of individual $x_i^2 \in E$ for $l_1 \le i \le l_2$, $x_{l_1}^2 + \cdots + x_{l_2}^2 \in E$ is obtained as a group under the above Vandermonde argument.

In any case, (3.17) can be constructed as long as the quadratic rank of E is k. An important observation is that both orthogonal variable transformation so used and affine variable transformation do not change the basis of $L(E)$. In summary, we have the following.

THEOREM 3.7. *Let E be a finite-dimensional estimation algebra with linear rank r and quadratic rank k.*

There exists $p_0 = \sum_{i=1}^{k_1} x_i^2 + \sum_{i=n-k_2+1}^{n} x_i^2 \in E$, where $k_1 + k_2 = k$, $k_1 \leq r$, and $k_2 \leq n - r$.

If $\phi \in E$ is a degree 1 polynomial, then ϕ is independent of $x_{r+1}, \ldots, x_n$.

If $\phi \in E$ is a degree 2 polynomial, then $\phi^{(2)}$ is independent of $x_{k_1+1}, \ldots, x_{n-k_2}$.

3.3. Euler operator. As discussed in section 3.1, the finite-dimensionality of an estimation algebra relies on the order of the differential operators as well as the coefficient functions. The conditions on the constructible differential operators with bounded order are shifted to those on the coefficient functions to ensure the finite-dimensionality. Thus, it is important to extract the conditions on some primary characteristics such as ω_{ij}'s from some transformations. The Euler operator (see Definition 3.8) is one of such transformations that occur frequently, especially when a quadratic polynomial is assumed to be an element of the estimation algebra. This section studies the properties of the inverse image of the Euler operators. The results obtained in this section generalize the results discussed in [18] and [27].

DEFINITION 3.8. *Let l be a positive integer such that $l \leq n$. The Euler operator $E_l(\cdot)$ is defined to be a differential operator such that*

$$E_l(\phi) = \sum_{i=1}^{l} x_i \frac{\partial \phi}{\partial x_i}$$

for any $\phi \in C^\infty(R^n)$.

In the case $l = n$, $E(\cdot)$ is used instead of $E_n(\cdot)$.

LEMMA 3.9. *Let $m > 0$ be a positive integer and $\zeta \in C^\infty(R^n)$. If $E_l(\zeta) + m\zeta = 0$, then $\zeta = 0$.*

Proof. Let $\phi = x_1^m \zeta$. Then

$$E_l(\phi) = \sum_{i=1}^{l} x_i \frac{\partial}{\partial x_i}(x_1^m \zeta) = m x_1^m \zeta + x_1^m \sum_{i=1}^{l} x_i \frac{\partial \zeta}{\partial x_i} = x_1^m (m\zeta + E_l(\zeta)) = 0, \tag{3.18}$$

$$\begin{aligned} &\phi(x_1, \ldots, x_l, \ldots, x_n) - \phi(\epsilon x_1, \ldots, \epsilon x_l, x_{l+1}, \ldots, x_n) \\ &= \int_\epsilon^1 \frac{d\phi}{dt}(tx_1, \ldots, tx_l, x_{l+1}, \ldots, x_n) dt \\ &= \int_\epsilon^1 \sum_{i=1}^{l} x_i \frac{\partial \phi}{\partial x_i}(tx_1, \ldots, tx_l, x_{l+1}, \ldots, x_n) dt \end{aligned}$$

$$= \int_\epsilon^1 \frac{1}{t}(E_l(\phi))(tx_1, \ldots, tx_l, x_{l+1}, \ldots, x_n) dt = \int_\epsilon^1 \frac{0}{t} dt = 0. \tag{3.19}$$

Hence, $\phi(x_1, \ldots, x_l, \ldots, x_n) = \epsilon^m x_1^m \zeta(\epsilon x_1, \ldots, \epsilon x_l, x_{l+1}, \ldots, x_n)$. Since $\zeta \in C^\infty(R^n)$, the limit $\lim_{\epsilon \to 0} \zeta(\epsilon x_1, \ldots, \epsilon x_l, x_{l+1}, \ldots, x_n)$ exists. By letting $\epsilon \to 0$, one has that $\phi = 0 \Longrightarrow \zeta = 0$. □

LEMMA 3.10. *Let $\zeta \in C^\infty(R^n)$. If $E_l(\zeta) = 0$, then ζ is a C^∞-function in $x_{l+1}, \ldots, x_n$ variables.*

Proof. For $1 \le j \le l$,

$$\frac{\partial}{\partial x_j} E_l(\zeta) = \frac{\partial \zeta}{\partial x_j} + E_l\left(\frac{\partial \zeta}{\partial x_j}\right) = 0. \tag{3.20}$$

By Lemma 3.9, $\frac{\partial \zeta}{\partial x_j} = 0$. Therefore, ζ is independent of $x_1, \ldots, x_l$. □

LEMMA 3.11. *Let $m > 0$ be a positive integer and $\zeta \in C^\infty(R^n)$. If $E_l(\zeta) - m\zeta = 0$, then ζ is either 0 or a degree m homogeneous polynomial in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables.*

Proof. Let $D = (\frac{\partial}{\partial x_1})^{\alpha_1} \cdots (\frac{\partial}{\partial x_l})^{\alpha_l}$ with $\sum_{i=1}^{l} \alpha_i = m$.

By induction,

$$D(E_l(\zeta)) = E_l(D\zeta) + mD\zeta. \tag{3.21}$$

Hence, $E_l(D\zeta) = D(E_l(\zeta) - m\zeta) = 0$. By Lemma 3.10 $D\zeta$ is a C^∞-function in $x_{l+1}, \ldots, x_n$ variables. This means that ζ is a polynomial of degree $s \le m$ in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables.

Let $\zeta = \sum_{0 \le |i| \le s} a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l}$. Then

$$\begin{aligned}
E_l(\zeta) - m\zeta &= E_l\left(\sum_{0 \le |i| \le s} a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l}\right) \\
&\quad - m \sum_{0 \le |i| \le s} a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l} \\
&= \sum_{0 \le |i| \le s} |i| a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l} \\
&\quad - m \sum_{0 \le |i| \le s} a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l} \\
&= \sum_{0 \le |i| \le s} (|i| - m) a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l} \\
&= 0. \qquad (3.22)
\end{aligned}$$

Thus, every term $(|i| - m) a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n)$ must be 0. If $\zeta \neq 0$, s must be equal to m while $a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) = 0$ for $|i| < m$. □

LEMMA 3.12. *Let m be a constant integer and $\zeta \in C^\infty(R^n)$ such that $E_l(\zeta) + m\zeta$ is a polynomial of degree k (≥ 0) in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables. If ζ is a polynomial of degree s in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$, then $s \ge k$.*

Proof. Let $\zeta = \sum_{0 \le |i| \le s} a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l}$. By repeating the same process in (3.22),

$$E_l(\zeta) + m\zeta = \sum_{0 \le |i| \le s} (|i| + m) a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l}. \tag{3.23}$$

For $E_l(\zeta) + m\zeta$ to be a polynomial of degree k in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables, s must be no less than k. □

THEOREM 3.13. *Let m be a constant integer and $\zeta \in C^\infty(R^n)$ such that $E_l(\zeta) + m\zeta$ is a polynomial of degree k (≥ 0) in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables.*

If $m + k + 1 > 0$, ζ is a polynomial of degree k in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$.

If $m + k + 1 \leq 0$, ζ is a polynomial of degree k or degree $-m$ ($\geq k + 1 > 0$) in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$.

Proof. Let $D = (\frac{\partial}{\partial x_1})^{\alpha_1} \cdots (\frac{\partial}{\partial x_l})^{\alpha_l}$ with $\sum_{i=1}^{l} \alpha_i = k + 1 \geq 1$. Since $E_l(\zeta) + m\zeta$ is a polynomial of degree k in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables,

$$D(E_l(\zeta) + m\zeta) = D(E_l(\zeta)) + mD\zeta = 0. \tag{3.24}$$

On the other hand, by (3.21),

$$D(E_l(\zeta)) = E_l(D\zeta) + (k + 1)D\zeta. \tag{3.25}$$

By (3.24) and (3.25),

$$E_l(D\zeta) + (m + k + 1)D\zeta = 0. \tag{3.26}$$

Now, there are three cases:

(i) $m + k + 1 > 0$. By Lemma 3.9, $D\zeta = 0$. This means that ζ is a polynomial of degree $\leq k$ in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables. By Lemma 3.12, ζ is a polynomial of degree k in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables.

(ii) $m + k + 1 = 0$. By Lemma 3.10, $D\zeta$ is 0 or a nonzero C^∞-function in $x_{l+1}, \ldots, x_n$. Thus, ζ is a polynomial of degree $\leq k + 1$ in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$. By Lemma 3.12, ζ is a polynomial of degree k or $k + 1$ ($= -m$) in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables.

(iii) $m + k + 1 < 0$ ($m < -k - 1 < 0$). Let $m' = -(m + k + 1) > 0$. Then (3.26) becomes $E_l(D\zeta) - m'D\zeta = 0$. By Lemma 3.11, $D\zeta$ is either 0 or a degree m' homogeneous polynomial in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables. If it is the latter case, ζ is a polynomial of degree $(k+1)+m' = -m > 0$ in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$. If $D\zeta = 0$, ζ is a polynomial of degree at most k in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables. By combining these two situations with Lemma 3.12, ζ is a polynomial of degree k or degree $-m$ in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables since $-m > k$. ∎

THEOREM 3.14. *Let m be a constant integer and $\zeta \in C^\infty(R^n)$ such that $E_l(\zeta) + m\zeta = \varphi$, where φ is a C^∞-function in $x_{l+1}, \ldots, x_n$ variables.*

If $m > 0$, then $\zeta = \frac{1}{m}\varphi$.

If $m = 0$, then $\varphi = 0 \Longrightarrow \zeta$ is a C^∞-function in $x_{l+1}, \ldots, x_n$ variables.

If $m \leq -1$, then $\zeta = \xi + \frac{1}{m}\varphi$, where ξ is either 0 or a degree $-m$ homogeneous polynomial in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables.

Proof. (i) If $m + 1 > 0$, by Theorem 3.13, ζ is a C^∞-function in $x_{l+1}, \ldots, x_n$ variables $\Longrightarrow E_l(\zeta) = 0 \Longrightarrow m\zeta = \varphi \Longrightarrow \zeta = \frac{1}{m}\varphi$ if $m \neq 0$; $\varphi = 0$ if $m = 0$.

(ii) $m \leq -1$. Let $\xi = \zeta - \frac{1}{m}\varphi$. Then $E_l(\xi) + m\xi = 0$. By Lemma 3.11, ξ is either 0 or a degree $-m$ homogeneous polynomial in $x_1, \ldots, x_l$ variables with coefficients in C^∞-functions of $x_{l+1}, \ldots, x_n$ variables. ∎

THEOREM 3.15. *Let m be a constant integer and $\zeta \in C^\infty(R^n)$ such that $E_l(\zeta) + m\zeta \in P_k(x_1, \ldots, x_n)$, a polynomial of degree $k \geq 0$ in $x_1, \ldots, x_n$ variables.*

If $m > 0$, $\zeta \in P_k(x_1, \ldots, x_n)$.

If $m = 0$, $\zeta \in P_k(x_1, \ldots, x_n) + a(x_{l+1}, \ldots, x_n)$, where $a(x_{l+1}, \ldots, x_n)$ is a C^∞-function in $x_{l+1}, \ldots, x_n$ variables.

Proof. By Theorem 3.13, $\zeta = \sum_{0 \leq |i| \leq k} a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l}$ for $m \geq 0$. By (3.23)

$$E_l(\zeta) + m\zeta = \sum_{0 \leq |i| \leq k} (|i| + m) a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n) x_1^{i_1} \cdots x_l^{i_l}. \tag{3.27}$$

If $m > 0$, $|i| + m > 0 \Longrightarrow a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n)$ must be a polynomial of degree less than or equal to $k - \sum_{j=1}^{l} i_j$ in $x_{l+1}, \ldots, x_n$ variables since $x_1^{i_1} \cdots x_l^{i_l}$'s are linearly independent.

If $m = 0$, only when $|i| = 0$, i.e., when $i_1 = \cdots = i_l = 0$, $a_{i_1 \cdots i_l}(x_{l+1}, \ldots, x_n)$ may not be a polynomial. □

3.4. Underdetermined partial differential equation. From Theorems 2.5 and 2.7, for the finite-dimensional estimation algebras of maximal rank, $F = \eta - \sum_{i=1}^{m} h_i^2$ must be a polynomial of degree at most 2. Notice that the filtering system (2.1) is completely parameterized by the pair, (f, h). It follows by (2.5) that the underdetermined partial differential equation

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = F \tag{3.28}$$

provides a complete characterization of the realization set of such systems. Therefore, it is of primary interest to investigate the solution and solution properties for this class of equations.

In (3.28), $f_1, \ldots, f_n$ and F are C^∞-functions on R^n. F is given and $f_1, \ldots, f_n$ are treated as unknown. Although there is only one equation with n unknowns, (3.28) may not have solutions. The following important result can be found in [24].

THEOREM 3.16 (see [24]). *Let $F(x_1, \ldots, x_n)$ be a C^∞-function on R^n. Suppose that there exists a path $c : R \to R^n$ and $\delta > 0$ such that $\lim_{t \to \infty} \|c(t)\| = \infty$ and $\lim_{t \to \infty} \sup_{B_\delta(c(t))} F = -\infty$, where $B_\delta(c(t)) = \{x \in R^n : \|x - c(t)\| < \delta\}$. Then there are no C^∞-functions $f_1, f_2, \ldots, f_n$ on R^n satisfying* (3.28).

The following results are applications of Theorem 3.16 that will be used in the classification problems.

THEOREM 3.17. *Let $F(x_1, \ldots, x_n)$ be a polynomial on R^n, let $c(t) = (c_1(t), \ldots, c_n(t))$ be a polynomial path (i.e., $c_i(t)$'s are polynomials in t), and let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Then $F(c(t) + \epsilon)$ is a polynomial in t, i.e., $F(c(t) + \epsilon) = \sum_{j=0}^{d} a_j(\epsilon) t^j$, where $a_j(\epsilon)$'s are polynomials of $\epsilon_1, \ldots, \epsilon_n$. If there exists a path $c(t)$ such that the coefficient $a_d(\epsilon)$ of the leading term of $F(c(t) + \epsilon)$ is a negative constant and $d \geq 1$, there are no C^∞-functions $f_1, f_2, \ldots, f_n$ on R^n satisfying the equation*

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = F. \tag{3.29}$$

Proof. Suppose $a_d(\epsilon) = a_d$ is a negative constant with $d \geq 1$. It is clear that $\lim_{t\to\infty} \|c(t)\| = \infty$.

For $0 \leq j \leq d-1$, $a_j(\epsilon)$ is a polynomial in $\epsilon_1, \ldots, \epsilon_n$. By continuity, there exists a $\delta > 0$ and a ball $B_\delta(0) = \{\epsilon \in R^n \ : \ \|\epsilon\| < \delta\}$, such that for any $\epsilon \in B_\delta(0)$, the following bounds hold:

$$|a_j(\epsilon) - a_j(0)| \leq 1 \qquad \forall 0 \leq j \leq d-1. \tag{3.30}$$

It follows that for $t > 0$,

$$\sup_{B_\delta(c(t))} F(x_1, \ldots, x_n) = \sup_{\epsilon\in B_\delta(0)} F(c(t)+\epsilon) = \sup_{\epsilon\in B_\delta(0)} a_d t^d + \sum_{j=0}^{d-1} a_j(\epsilon) t^j$$

$$\leq a_d t^d + \sum_{j=0}^{d-1} (1 + a_j(0)) t^j \longrightarrow -\infty \quad \text{as } t \to \infty,$$

since $a_d < 0$ and $d \geq 1$. The assertion follows immediately by Theorem 3.16. □

THEOREM 3.18. *Let $F(x_1, \ldots, x_n)$ be a degree $d \geq 1$ polynomial on R^n. The homogeneous degree d part of F is denoted by $F_d = \sum_{|i|=d} a_i x_1^{i_1} \cdots x_n^{i_n}$, where $i = (i_1, \ldots, i_n)$. If there exist n numbers $b_1, \ldots, b_n$ such that $F_d(b_1, \ldots, b_n) < 0$, there are no C^∞-functions $f_1, f_2, \ldots, f_n$ on R^n satisfying* (3.29).

Proof. By taking the polynomial path $c(t) = (b_1 t, \ldots, b_n t)$ one has that $F(c(t)+\epsilon)$ is a degree d polynomial in t and the coefficient of t^d in $F(c(t)+\epsilon)$ is a negative constant $F_d(b_1, \ldots, b_n) < 0$. The proof is concluded by using Theorem 3.17. □

COROLLARY 3.19. *Let $F(x_1, \ldots, x_n)$ be a polynomial on R^n. If the degree of F is odd, there are no C^∞-functions $f_1, f_2, \ldots, f_n$ on R^n satisfying* (3.29).

COROLLARY 3.20. *Let $F(x_1, \ldots, x_n)$ be a degree $d \geq 1$ polynomial on R^n. If there exist C^∞-functions $f_1, f_2, \ldots, f_n$ on R^n satisfying* (3.29)*, the homogeneous degree d part of F is a nonnegative function.*

In fact, the above theorems can be further generalized to the case when F is a polynomial in $x_1, \ldots, x_r$ variables with C^∞ coefficient functions in $x_{r+1}, \ldots, x_n$. For example, we have the following.

THEOREM 3.21. *Let d and $r \leq n$ be two positive integers and*

$$F(x_1, \ldots, x_n) = \sum_{|i|\leq d} a_i(x_{r+1}, \ldots, x_n) x_1^{i_1} \cdots x_r^{i_r},$$

where $i = (i_1, \ldots, i_r)$, and where a_i's are C^∞-functions in $x_{r+1}, \ldots, x_n$ variables. The homogeneous degree d part in $x_1, \ldots, x_r$ variables of F is denoted by $F_d = \sum_{|i|=d} a_i(x_{r+1}, \ldots, x_n) x_1^{i_1} \cdots x_r^{i_r}$. If there exist n numbers $b_1, \ldots, b_n$ such that $F_d(b_1, \ldots, b_n) < 0$, there are no C^∞-functions $f_1, f_2, \ldots, f_n$ on R^n satisfying (3.29).

Proof. Take the path $c(t) = (b_1 t, \ldots, b_r t, b_{r+1}, \ldots, b_n)$. Clearly $\lim_{t\to\infty} \|c(t)\| \to \infty$ since $F_d(b_1, \ldots, b_n) < 0$.

Let $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Then $F(c(t) + \epsilon)$ is a polynomial in t,

$$F(c(t) + \epsilon) = F_d(b_1, \ldots, b_r, b_{r+1} + \epsilon, \ldots, b_n + \epsilon) t^d + \sum_{j=0}^{d-1} c_j(\epsilon) t^j.$$

By continuity, there exist a $\delta > 0$ and a ball $B_\delta(0) = \{\epsilon \in R^n \ : \ \|\epsilon\| < \delta\}$, such that for any $\epsilon \in B_\delta(0)$, the following bounds hold:

$$F_d(b_1,\dots,b_r,b_{r+1}+\epsilon,\dots,b_n+\epsilon) \le \frac{1}{2}F_d(b_1,\dots,b_n) < 0,$$
$$|c_j(\epsilon)-c_j(0)| \le 1 \qquad \forall 0 \le j \le d-1.$$

It follows that for $t > 0$,

$$\begin{aligned}\sup_{B_\delta(c(t))} F(x_1,\dots,x_n) &= \sup_{\epsilon\in B_\delta(0)} F(c(t)+\epsilon)\\ &= \sup_{\epsilon\in B_\delta(0)} F_d(b_1,\dots,b_r,b_{r+1}+\epsilon,\dots,b_n+\epsilon)t^d + \sum_{j=0}^{d-1} c_j(\epsilon)t^j\\ &\le \frac{1}{2}F_d(b_1,\dots,b_n)t^d + \sum_{j=0}^{d-1}(1+c_j(0))t^j \longrightarrow -\infty, \quad \text{as } t\to\infty,\end{aligned}$$

since $F_d(b_1,\dots,b_n) < 0$ and $d \ge 1$. The assertion follows immediately by Theorem 3.16. □

4. Classification of estimation algebras $n = 2$. In this section, the state dimension is assumed to be $n = 2$, i.e., there are two state variables x_1 and x_2. The estimation algebra $E = \langle L_0, h_1, \dots, h_m\rangle_{L.A.}$, where

$$L_0 = \frac{1}{2}(D_1^2 + D_2^2 - \eta), \qquad D_i = \frac{\partial}{\partial x_i} - f_i, \quad i = 1,2, \tag{4.1}$$

$$\eta = \sum_{i=1}^{2}\frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{2} f_i^2 + \sum_{i=1}^{m} h_i^2. \tag{4.2}$$

m is assumed to be a positive integer; otherwise $E = \langle L_0\rangle_{L.A.}$ is a one-dimensional estimation algebra. The Ω-matrix is a 2×2 antisymmetric matrix. Therefore, only $\omega_{12} = -\omega_{21}$ is unknown.

By Theorem 3.3, 1 is assumed to be an element of E in the discussion.

4.1. Linear structure of the Ω-matrix.

THEOREM 4.1. *Suppose* $\dim E < \infty$ *and* $Y = p(x)D_2, \mod U_0 \in E$. *Then* p *is a polynomial in* x_1, x_2 *of degree at most* 1.

Proof. By Theorem 3.2, p is a polynomial in x_1, x_2. Let $l = deg\ p$ and $p^{(l)}$ be the homogeneous degree l part of p. Then

$$p^{(l)} = \sum_{i=0}^{l} a_i x_1^{l-i}x_2^i = \sum_{i=s}^{t} a_i x_1^{l-i}x_2^i,$$

where a_i's are constants, $0 \le s \le t \le l$, $a_s \ne 0$, $a_t \ne 0$, and $a_i = 0$ for $i < s$ or $i > t$. Let $Y_k = Ad_{L_0}^k Y$ for $k \ge 0$. By induction,

$$Y_k = Ad_{L_0}^k Y = \sum_{j=0}^{k}\binom{k}{j}\frac{\partial^k p}{\partial x_1^{k-j}\partial x_2^j}D_1^{k-j}D_2^{j+1}, \quad \mod U_k,$$

where $\binom{k}{j}$'s are binomial numbers. In particular,

$$Y_l = \sum_{j=0}^{l}\binom{l}{j}(l-j)!j!a_j D_1^{l-j}D_2^{j+1} = l!\sum_{j=s}^{t} a_j D_1^{l-j}D_2^{j+1}, \quad \mod U_l,$$

and

$$Y_{l-1} = \sum_{j=0}^{l-1} \binom{l-1}{j} \frac{\partial^{l-1} p}{\partial x_1^{l-j-1} \partial x_2^j} D_1^{l-j-1} D_2^{j+1}$$

$$= \sum_{j=0}^{l-1} (l-1)! \Big((l-j) a_j x_1 + (j+1) a_{j+1} x_2 + c_j \Big) D_1^{l-j-1} D_2^{j+1}, \quad \mod U_{l-1},$$

where c_j's are constants from the $(l-1)$th partial derivatives of the homogeneous degree $l-1$ part of p.

Now, depending on whether or not $s = 0$, and in the case when $s = 0$ whether a_1, a_2 are 0, there are four cases for which similar constructions of sequences with different calculations will show that l must be less than 2 if E is finite-dimensional.

(i) *Case* 1. $s \neq 0$.

$$\begin{aligned}
A_0 := Y_{l-1} &= (d_0 x_2 + (l-1)! c_{s-1}) D_1^{l-s} D_2^s \\
&\quad + \text{terms with lower order in } D_1, \quad \mod U_{l-1}, \\
A_1 := Y_l &= d_1 D_1^{l-s} D_2^{s+1} \ + \text{terms with lower order in } D_1, \quad \mod U_l, \\
A_2 := [A_1, A_0] &= (s+1) d_1 d_0 D_1^{2(l-s)} D_2^{2s} \\
&\quad + \text{terms with lower order in } D_1, \quad \mod U_{2l-1}, \\
&\vdots \\
A_{r+1} := [A_r, A_0] &= d_{r+1} D_1^{(r+1)(l-s)} D_2^{(r+1)s-r+1} \\
&\quad + \text{terms with lower order in } D_1, \quad \mod U_{(r+1)l-r},
\end{aligned}$$

where $d_0 = (l-1)! s a_s \neq 0$, $d_1 = l! a_s \neq 0$, and $d_{r+1} = (rs - r + 2) d_r d_0 \neq 0$ for $r \geq 1$. The orders of A_{r+1}'s are $(r+1)l - r + 1 \to \infty$ unless $l < 2$.

(ii) *Case* 2. $s = 0$ and $a_1 \neq 0$.

$$\begin{aligned}
A_0 := Y_{l-1} &= (l! a_0 x_1 + d_0 x_2 + (l-1)! c_0) D_1^{l-1} D_2 \\
&\quad + \text{terms with lower order in } D_1, \quad \mod U_{l-1}, \\
A_1 := Y_l &= d_1 D_1^l D_2 \ + \text{terms with lower order in } D_1, \quad \mod U_l, \\
A_2 := [A_1, A_0] &= d_1 d_0 D_1^{2l-1} D_2 \\
&\quad + \text{terms with lower order in } D_1, \quad \mod U_{2l-1}, \\
&\vdots \\
A_{r+1} := [A_r, A_0] &= d_{r+1} D_1^{(r+1)l-r} D_2 \\
&\quad + \text{terms with lower order in } D_1, \quad \mod U_{(r+1)l-r},
\end{aligned}$$

where $d_0 = (l-1)! a_1 \neq 0$, $d_1 = l! a_0 \neq 0$, and $d_{r+1} = d_r d_0 \neq 0$ for $r \geq 1$. The orders of A_{r+1}'s are $(r+1)l - r + 1 \to \infty$ unless $l < 2$.

(iii) *Case* 3. $s = 0$ and $a_1 = a_2 = 0$.

$$\begin{aligned}
A_0 &:= Y_{l-1} = (d_0x_1 + (l-1)!c_0)D_1^{l-1}D_2 + (l-1)!c_2D_1^{l-2}D_2^2 \\
&\qquad + \text{degree 1 coeff. terms with lower order in } D_1, \quad \mod U_{l-1}, \\
A_1 &:= Y_l = d_1D_1^lD_2 \\
&\qquad + \text{constant coeff. terms with lower order in } D_1, \quad \mod U_l, \\
A_2 &:= [A_1, A_0] = ld_1d_0D_1^{2l-2}D_2^2 \\
&\qquad + \text{constant coeff. terms with lower order in } D_1, \quad \mod U_{2l-1}, \\
&\vdots \\
A_{r+1} &:= [A_r, A_0] = d_{r+1}D_1^{(r+1)l-2r}D_2^{r+1} \\
&\qquad + \text{constant coeff. terms with lower order in } D_1, \quad \mod U_{(r+1)l-r},
\end{aligned}$$

where $d_0 = l!a_0 \neq 0$, $d_1 = l!a_0 \neq 0$, and $d_{r+1} = (rl - 2r + 2)d_rd_0 \neq 0$ for $r \geq 1$ and $l \geq 2$. The orders of A_{r+1}'s are $(r+1)l - r + 1 \to \infty$ unless $l < 2$.

(iv) *Case* 4. $s = 0$, $a_1 = 0$, but $a_2 \neq 0$.

$$Y_l = d_1D_1^lD_2 + d_3D_1^{l-2}D_2^3 \ + \text{terms with lower order in } D_1, \quad \mod U_l,$$

where $d_1 = l!a_0 \neq 0$ and $d_3 = l!a_2 \neq 0$. If $l \geq 2$, consider $Z = [Y_l, pD_2]$ and $A_0 = Ad_{L_0}^{l-2}Z$:

$$\begin{aligned}
Z = [Y_l, pD_2] &= d_1\frac{\partial p}{\partial x_2}D_1^lD_2 + ld_1\frac{\partial p}{\partial x_1}D_1^{l-1}D_2^2 \\
&\qquad + \text{terms with lower order in } D_1, \quad \mod U_l, \\
A_0 = Ad_{L_0}^{l-2}Z &= d_1\frac{\partial^{l-1}p}{\partial x_1^{l-2}\partial x_2}D_1^{2l-2}D_2 \ + \text{terms with lower order in } D_1, \quad \mod U_{2l-2}.
\end{aligned}$$

Since $p^{(l)} = a_0x_1^l + a_2x_1^{l-2}x_2^2 \ +$ terms with lower degree in x_1, $\frac{\partial^{l-1}p}{\partial x_1^{l-2}\partial x_2} = 2(l-2)!a_2x_2 + c_2$, where c_2 is $(l-2)!$ multiplied by the coefficient of $x_1^{l-2}x_2$ in p. Hence,

$$\begin{aligned}
A_0 &= (e_0x_2 + d_1c_2)D_1^{2l-2}D_2 \ + \text{ terms with lower order in } D_1, \quad \mod U_{2l-2}, \\
A_1 &:= Y_l = d_1D_1^lD_2 + d_3D_1^{l-2}D_2^3 \\
&\qquad + \text{terms with lower order in } D_1, \quad \mod U_l, \\
A_2 &:= [A_1, A_0] = d_1e_0D_1^{3l-2}D_2 \\
&\qquad + \text{terms with lower order in } D_1, \quad \mod U_{3l-2}, \\
&\vdots \\
A_{r+1} &:= [A_r, A_0] = e_{r+1}D_1^{(2r+1)l-2r}D_2 \\
&\qquad + \text{terms with lower order in } D_1, \quad \mod U_{(2r+1)l-2r},
\end{aligned}$$

where $e_0 = 2(l-2)!a_2d_1 \neq 0$, $e_1 = d_1 \neq 0$, and $e_{r+1} = e_re_0 \neq 0$. The orders of A_{r+1}'s are $(2r+1)l - 2r + 1 \to \infty$ unless $l < 2$. □

THEOREM 4.2. *Suppose* $\dim E < \infty$. $x_1 + c \in E \Rightarrow \omega_{12}$ *is a polynomial of degree at most* 1.

Proof. Clearly, $L_0, x_1, D_1 \in E$.

$$H_0 := [L_0, D_1] = \omega_{12} D_2 + \frac{1}{2}\frac{\partial \omega_{12}}{\partial x_2} + \frac{1}{2}\frac{\partial \eta}{\partial x_1} \in E. \tag{4.3}$$

By Theorem 4.1, ω_{12} is a polynomial of degree at most 1. □

THEOREM 4.3. *Suppose* $\dim E < \infty$. $x_1^2 + c \in E \Rightarrow \omega_{21}$ *is a constant.*

Proof.

$$K_0 = \left[L_0, \frac{1}{2}(x_1^2 + c)\right] - \frac{1}{2} = x_1 D_1 \in E,$$

$$K_1 = [L_0, K_0] = D_1^2 - \alpha_2 D_2 + \frac{1}{2}E_1(\eta) - \frac{1}{2}\frac{\partial \alpha_2}{\partial x_2} \in E,$$

$$K_2 = [K_1, K_0] = 2D_1^2 + E_1(\alpha_2)D_2 + \alpha_2^2 - \frac{1}{2}E_1^2(\eta) + \frac{1}{2}E_1\left(\frac{\partial \alpha_2}{\partial x_2}\right) \in E,$$

$$Z_0 = K_2 - 2K_1 = (E_1(\alpha_2) + 2\alpha_2)D_2 + \gamma(x) \in E, \tag{4.4}$$

where $\alpha_2 = x_1\omega_{21}$, $E_1(\cdot) = x_1\frac{\partial}{\partial x_1}$, and $\gamma(x) = \alpha_2^2 - E_1(\eta) - \frac{1}{2}E_1^2(\eta) + \frac{\partial \alpha_2}{\partial x_2} + \frac{1}{2}E_1(\frac{\partial \alpha_2}{\partial x_2})$.

By Theorem 4.1, $E_1(\alpha_2) + 2\alpha_2$ is a polynomial of at most 1, and so is $\alpha_2 = x_1\omega_{21}$ by Theorem 3.15. Since $\omega_{21} \in C^\infty(R^2)$, ω_{21} must be a constant. □

LEMMA 4.4. *Suppose* $K_0 = x_1D_1 + x_2D_2 \in E$ *and* $Y = \sum_{i=0}^k b_i(x)D_1^{k-i}D_2^i$, mod $U_{k-1} \in E$. *Let* $b_i^{(r)}$ *denote the homogeneous degree* r *part of* b_i *for* $0 \le i \le k$. *Then* $\sum_{i=0}^k b_i^{(r)}D_1^{k-i}D_2^i$, mod $U_{k-1} \in E$, *for* $r \ge 0$.

Proof. Let $E(\cdot) = \sum_{i=1}^2 x_i\frac{\partial}{\partial x_i}$.

$$\left[K_0, \sum_{i=0}^k b_i D_1^{k-i}D_2^i\right] = \sum_{i=0}^k (E(b_i) - kb_i)D_1^{k-i}D_2^i, \qquad \text{mod } U_{k-1}. \tag{4.5}$$

Let $l = \max_{0\le i\le k} deg\ b_i$. Use (4.5) to construct a sequence of elements

$$Z_r = \sum_{j=0}^{l-r} c_{rj} \sum_{i=0}^k b_i^{(j)} D_1^{k-i}D_2^i \in E, \qquad \text{mod } U_{k-1}, \tag{4.6}$$

in E as follows:

(1) $Z_0 = Y$, i.e., $c_{0j} = 1, 0 \le j \le l$;

(2) $Z_{r+1} = (l - r - k)Z_r - [K_0, Z_r] \Longrightarrow c_{(r+1)j} = (l - r - j)c_{rj}, 0 \le j \le l - r - 1$.

Note that in (4.6) $c_{rj} \ne 0$ for $r \le l$. Starting from $Z_l = c_{l0}\sum_{i=0}^k b_i^{(0)}D_1^{k-i}D_2^i$, mod U_{k-1}, $Z_{l-1} = c_{(l-1)0}\sum_{i=0}^k b_i^{(0)}D_1^{k-i}D_2^i + c_{(l-1)1}\sum_{i=0}^k b_i^{(1)}D_1^{k-i}D_2^i$, mod $U_{k-1}, \ldots,$ $Z_r, \ldots,$ one can solve $\sum_{i=0}^k b_i^{(r)}D_1^{k-i}D_2^i$ successively. This means

$$\sum_{i=0}^k b_i^{(r)}D_1^{k-i}D_2^i, \quad \text{mod } U_{k-1} \in E. \qquad \square$$

LEMMA 4.5. *Suppose* $\dim E < \infty$. *Let* $Y = \sum_{i=i_0}^{i_k} b_i(x)D_1^{k-i}D_2^i$ mod $U_{k-1} \in E$ *be a differential operator with the highest order* k *in* E *(obviously* $k \ge 2$*), where* $0 \le i_0 \le i_k \le k$. *Then*

$$\frac{\partial b_{i_0}}{\partial x_1} = \frac{\partial b_{i_k}}{\partial x_2} = 0, \qquad \frac{\partial b_i}{\partial x_1} + \frac{\partial b_{i-1}}{\partial x_2} = 0, \quad i_0 + 1 \le i \le i_k. \tag{4.7}$$

Proof.

$$[L_0, Y] = \sum_{i=i_0}^{i_k} \frac{\partial b_i}{\partial x_1} D_1^{k+1-i} D_2^i + \sum_{i=i_0}^{i_k} \frac{\partial b_i}{\partial x_2} D_1^{k-i} D_2^{i+1} \tag{4.8}$$
$$= \frac{\partial b_{i_0}}{\partial x_1} D_1^{k+1} + \sum_{i=i_0+1}^{i_k} \left(\frac{\partial b_i}{\partial x_1} + \frac{\partial b_{i-1}}{\partial x_2} \right) D_1^{k+1-i} D_2^i + \frac{\partial b_{i_k}}{\partial x_2} D_2^{k+1}, \text{ mod } U_k.$$

Since E's elements have the highest differential order k, all the coefficient functions of the order $k+1$ terms in the above equation must be 0. □

LEMMA 4.6. *Suppose* dim $E < \infty$ *and* $Y = p_1 D_1 + p_2 D_2$, mod $U_0 \in E$, *where* $p_1 = \sum_{j=1}^{l} a_j x_1^{l-j} x_2^j$ *and* $p_2 = -\sum_{j=1}^{l} a_j x_1^{l+1-j} x_2^{j-1}$, a_j*'s are constants and* $a_j \neq 0$ *for some* j. *Then* $l \leq 2$.

Proof. Since dim $E < \infty$, there exists an element $P_0 = \sum_{i=i_0}^{i_k} b_i(x) D_1^{k-i} D_2^i$, mod $U_{k-1} \in E$, that has the highest differential order k, where $0 \leq i_0 \leq i_k \leq k$, $b_{i_0} \neq 0$, and $b_{i_k} \neq 0$. By Lemma 4.4, b_i's can be assumed to be homogeneous polynomials. More precisely, P_0 can be assumed to have the following properties:

(C1) $P_0 = \sum_{i=i_0}^{i_k} b_i(x) D_1^{k-i} D_2^i$, mod $U_{k-1} \in E$, $0 \leq i_0 \leq i_k \leq k$, $b_{i_0} \neq 0$, and $b_{i_k} \neq 0$.

(C2) P_0 has the highest differential order in E, i.e., if $A \in E$, then *ord* $A \leq k$.

(C3) $\frac{\partial b_{i_0}}{\partial x_1} = \frac{\partial b_{i_k}}{\partial x_2} = 0$, $\frac{\partial b_i}{\partial x_1} + \frac{\partial b_{i-1}}{\partial x_2} = 0$, $i_0 + 1 \leq i \leq i_k$.

(C4) b_i's are homogeneous degree r polynomials.

Moreover, among the elements in E satisfying (C1) and (C2), P_0 can be chosen such that b_i's have the highest degree:

(C5) Among all elements in E with order k, b_i's have the highest degree.

Assume $l \geq 3$.

Assume either $i_0 > 0$ or $i_k < k$. Without loss of generality, assume $i_0 > 0$.

$$[P_0, Y] = \left[P_0, \sum_{i=1}^{2} p_i D_i \right] = \left[b_{i_0} D_1^{k-i_0} D_2^{i_0} + \cdots \sum_{i=1}^{2} p_i D_i \right]$$
$$= i_0 b_{i_0} \frac{\partial p_1}{\partial x_2} D_1^{k+1-i_0} D_2^{i_0-1}$$
$$+ \text{ terms with lower order in } D_1, \qquad \text{mod } U_{k-1}.$$

Since $\frac{\partial p_1}{\partial x_2} \neq 0$ is a degree $l-1 \geq 2$ polynomial, while b_{i_0} is a polynomial independent of x_1 by (C3), $i_0 b_{i_0} \frac{\partial p_1}{\partial x_2}$ is a polynomial of degree higher than b_{i_0}. This contradicts with (C5). Thus,

(C6) $i_0 = 0$ and $i_k = k$, i.e., $P_0 = \sum_{i=0}^{k} b_i(x) D_1^{k-i} D_2^i$, mod U_{k-1}, where $b_0 \neq 0$ and $b_k \neq 0$.

Again, consider $[P_0, Y]$,

$$[P_0, Y] = \left[P_0, \sum_{i=1}^{2} p_i D_i \right] = \left[b_0 D_1^k + b_1 D_1^{k-1} D_2 + \cdots \sum_{i=1}^{2} p_i D_i \right]$$
$$= \left(k b_0 \frac{\partial p_1}{\partial x_1} - p_2 \frac{\partial b_0}{\partial x_2} + b_1 \frac{\partial p_1}{\partial x_2} \right) D_1^k$$
$$+ \text{ terms with lower order in } D_1, \qquad \text{mod } U_{k-1}. \tag{4.9}$$

By (C3), (C4), and (C6), one can assume that

$$b_0 = x_2^r, \quad \text{where } r \geq 0, \tag{4.10}$$

$$b_1 = -rx_1x_2^{r-1} + cx_2^r, \quad \text{where } c \text{ is an unknown constant.} \tag{4.11}$$

Using (4.10) and (4.11), we obtain the coefficient function of the D_1^k term in $[P_0, \sum_{i=1}^2 p_i D_i]$ given in (4.9):

$$\begin{aligned} &kb_0\frac{\partial p_1}{\partial x_1} - p_2\frac{\partial b_0}{\partial x_2} + b_1\frac{\partial p_1}{\partial x_2} \\ &= (ca_1 - ra_2)x_1^{l-1}x_2^r + \sum_{j=2}^{l-1}(-jra_{j+1} + jca_j + k(l-j+1)a_{j-1})x_1^{l-j}x_2^{r+j-1} \\ &\qquad + (lca_l + ka_{l-1})x_2^{l+r-1}. \end{aligned} \tag{4.12}$$

Equation (4.12) is a degree $l+r-1 > r$ homogeneous polynomial. To meet condition (C5), every term in (4.12) must be 0. That is, $a_1, \ldots, a_l$ must satisfy

$$\begin{aligned} &ca_1 - ra_2 = 0, \qquad lca_l + ka_{l-1} = 0, \\ &-jra_{j+1} + jca_j + k(l-j+1)a_{j-1} = 0 \quad \text{for } 2 \leq j \leq l-1. \end{aligned} \tag{4.13}$$

The above relations can be summarized as a matrix equation:

$$\begin{pmatrix} c & -r & & & & & \\ k(l-1) & 2c & -2r & & & & \\ & k(l-2) & 3c & -3r & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 2k & (l-1)c & -(l-1)r & \\ & & & & k & lc & \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{l-1} \\ a_l \end{pmatrix} = 0. \tag{4.14}$$

Let A denote the $l \times l$ tridiagonal matrix in the above equation. Since all the $A(i, i+1)$ entries are nonpositive while all $A(i, i-1)$ entries are positive, it is easy by induction that $c = 0$ is a necessary condition for the determinant of A to be 0. More specifically, let B_i denote an $i \times i$ matrix having the entries $B_i(s,t) = A(s,t)$ for $1 \leq s, t \leq i$, i.e., B_i is the upper left block of A with size $i \times i$. The determinants of B_i's have the following recursion:

$$|B_1| = c, \qquad |B_2| = 2c^2 + d_1, \tag{4.15}$$

$$|B_{i+1}| = (i+1)c|B_i| + d_i|B_{i-1}|, \tag{4.16}$$

where $d_i = (l-i)ikr > 0$. If l is odd, $|A| = |B_l| = c \cdot q(c^2)$, where q is a nondegenerate polynomial of degree $(l-1)/2$ with nonnegative coefficients. If l is even, $|A| = |B_l| = q(c^2)$, where q is a nondegenerate polynomial of degree $l/2$ with nonnegative coefficients. Thus, to ensure the existence of nontrivial solutions of $a_1, \ldots, a_l$, c must be 0.

Similarly, by assuming $b_k = x_1^r$ and repeating the same computation for the coefficient function of D_2^k term in $[P, \sum_{i=1}^2 p_i D_i]$, one will get

$$\begin{pmatrix} c_1 & -r & & & & & \\ k(l-1) & 2c_1 & -2r & & & & \\ & k(l-2) & 3c_1 & -3r & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 2k & (l-1)c_1 & -(l-1)r & \\ & & & & k & lc_1 & \end{pmatrix} \begin{pmatrix} a_l \\ a_{l-1} \\ a_{l-2} \\ \vdots \\ a_2 \\ a_1 \end{pmatrix} = 0, \tag{4.17}$$

and c_1 must be 0 as well.

This first observation from (4.14) and (4.17) is that $r \neq 0$. Otherwise, one will have $a_1 = \cdots = a_{l-1} = 0$ from (4.14) and $a_l = \cdots = a_2 = 0$ from (4.17) since $c = c_1 = 0$.

Now, $r \neq 0$ and $c = c_1 = 0$, again by solving (4.14) and (4.17),

$$a_j = \begin{cases} 0, & j = 2, l-1, \\ \frac{k(l-j+2)}{r(j-1)} a_{j-2}, & 3 \leq j \leq l, \end{cases} \qquad a_j = \begin{cases} 0, & j = 2, l-1, \\ \frac{r(l-j+2)}{k(j-1)} a_{j-2}, & 3 \leq j \leq l. \end{cases} \tag{4.18}$$

Hence, l must be an odd number and $k = r$ so that (4.18) holds. Without loss of generality, a_1 can be assumed to be 1. That is,

$$\begin{gathered} l \text{ is an odd number } \geq 3, \quad a_1 = a_l = 1, \quad a_2 = 0, \\ a_j = \frac{l-j+2}{j-1} a_{j-2} \text{ for } 3 \leq j \leq l. \end{gathered} \tag{4.19}$$

(C7) $k = r$. This means the degree of the coefficient function b_i in P_0 must be the same as the order of P_0. That is,

$$P_0 = x_2^k D_1^k - k x_1 x_2^{k-1} D_1^{k-1} D_2 \ + \text{terms with lower order in } D_1, \text{ mod } U_{k-1}. \tag{4.20}$$

Consider any $N = \sum_{j=0}^{t} n_j(x) D_1^{t-j} D_2^j$, mod $U_{t-1} \in E$,

$$\begin{aligned} [P_0, N] &= \left[\sum_{i=0}^{k} b_i(x) D_1^{k-i} D_2^i, \sum_{j=0}^{t} n_j(x) D_1^{t-j} D_2^j \right] \\ &= \left(k b_0 \frac{\partial n_0}{\partial x_1} - t n_0 \frac{\partial b_0}{\partial x_1} + b_1 \frac{\partial n_0}{\partial x_2} - n_1 \frac{\partial b_0}{\partial x_2} \right) D_1^{k+t-1} \\ &\qquad + \text{terms with lower order in } D_1 \\ &= k x_2^{k-1} \left(x_2 \frac{\partial n_0}{\partial x_1} - x_1 \frac{\partial n_0}{\partial x_2} - n_1 \right) D_1^{k+t-1} \\ &\qquad + \text{terms with lower order in } D_1, \text{ mod } U_{k+t-2}. \end{aligned} \tag{4.21}$$

If $t \geq 2$, then

$$n_1 = x_2 \frac{\partial n_0}{\partial x_1} - x_1 \frac{\partial n_0}{\partial x_2}. \tag{4.22}$$

Otherwise, $[P_0, N]$ will have order $k + t - 1 > k$, contradicting (C2).

In particular, by (4.22),

$$\text{if } n_0 = x_2^s \Longrightarrow n_1 = -s x_1 x_2^{s-1} \quad \text{for } s \geq 0. \tag{4.23}$$

Now, suppose that $N = x_2^s D_1^t - s x_1 x_2^{s-1} D_1^{t-1} D_2 +$ terms with lower order in D_1, mod $U_{t-1} \in E$. By using (4.18)

$$\begin{aligned} [N, Y] &= [\ x_2^s D_1^t - s x_1 x_2^{s-1} D_1^{t-1} D_2 + \cdots p_1 D_1 + p_2 D_2\ ] \\ &= \left(t x_2^s \frac{\partial p_1}{\partial x_1} - s x_2^{s-1} p_2 - s x_1 x_2^{s-1} \frac{\partial p_1}{\partial x_2} \right) D_1^t \\ &\qquad + \text{terms with lower order in } D_1 \\ &= (t-s) \sum_{j=1}^{l-2} (l-j) a_j x_1^{l-j-1} x_2^{j+s} D_1^t \\ &\qquad + \text{terms with lower order in } D_1, \text{ mod } U_{t-1}, \end{aligned} \tag{4.24}$$

$$\begin{aligned} Ad_{L_0}^{l-2}[N, Y] &= (t-s)(l-1)! x_2^{s+1} D_1^{t+l-2} \\ &\qquad + \text{terms with lower order in } D_1, \text{ mod } U_{t+l-1}. \end{aligned} \tag{4.25}$$

Equations (4.23), (4.24), and (4.25) show that if $s < t$,

$$x_2^s D_1^t \ + \text{terms with lower order in } D_1, \ \mod U_{t-1} \in E$$

$$\implies x_2^{s+1} D_1^{t+l-2} \ + \text{terms with lower order in } D_1, \ \mod U_{t+l-3} \in E. \tag{4.26}$$

Since $l \geq 3$, $s+1 < t+l-2$, and repeating (4.26) will make E not finite-dimensional. Now, choose $N = 2L_0$. Then the above procedure (4.23)–(4.26) applies. Contradiction! □

THEOREM 4.7. *Suppose* $\dim E < \infty$. $x_1^2 + x_2^2 + c \in E \Rightarrow \omega_{12}$ *is a polynomial of degree at most* 1.

Proof.

$$K_0 = \left[L_0, \frac{1}{2}(x_1^2 + x_2^2 + c)\right] - 1 = x_1 D_1 + x_2 D_2 \in E,$$

$$K_1 = [L_0, K_0] = \sum_{i=1}^{2} D_i^2 - \sum_{i=1}^{2} \alpha_i D_i + \frac{1}{2} E(\eta) - \frac{1}{2} \sum_{i=1}^{2} \frac{\partial \alpha_i}{\partial x_i} \in E,$$

$$K_2 = 2L_0 - K_1 = \sum_{i=1}^{2} \alpha_i D_i + \frac{1}{2} \sum_{i=1}^{2} \frac{\partial \alpha_i}{\partial x_i} - \eta - \frac{1}{2} E(\eta) \in E,$$

where $\alpha_1 = x_2 \omega_{12}$, $\alpha_2 = x_1 \omega_{21}$, $E(\cdot) = \sum_{i=1}^{2} x_i \frac{\partial}{\partial x_i}$. By Theorem 3.2, α_i is a polynomial. Since ω_{12} is smooth, ω_{12} must be a polynomial as well.

Let $\alpha_i^{(r)}$ be the homogeneous degree r part of α_i. By Lemma 4.4,

$$\sum_{i=1}^{2} \alpha_i^{(r)} D_i \in E, \quad \mod U_0.$$

On the other hand,

$$\sum_{i=1}^{2} x_i \alpha_i = \sum_{i,j=1}^{2} x_i x_j \omega_{ij} = 0 \Longrightarrow x_1 \alpha_1^{(r)} + x_2 \alpha_2^{(r)} = 0.$$

By Lemma 4.6, $\alpha_i^{(r)} = 0$ for $r > 2$. Therefore, ω_{12} is a polynomial of degree at most 1. □

THEOREM 4.8. *Suppose* $\dim E < \infty$. h_i*'s are polynomials at most of degree* 1.

Proof. Without loss of generality, h_1 is assumed to be a polynomial of degree 2. By orthogonal and affine transformations, h_1 is either $ax_1^2 + cx_2 + d$ or $ax_1^2 + bx_2^2 + d$, where $a, b \neq 0$.

(i) If $h_1 = ax_1^2 + cx_2 + d$, then $[[L_0, h_1], h_1] = 4a^2 x_1^2 + c^2 \in E$. By Theorem 4.3, ω_{21} is a constant. Hence, E is not finite-dimensional according to Theorem 2.5.

(ii) Assume $h_1 = ax_1^2 + bx_2^2 + d$. $[[L_0, h_1], h_1] = 4a^2 x_1^2 + 4b^2 x_2^2 \in E$. If $a \neq b$, both $x_1^2 + c$ and $x_2^2 + c$ are in E. From the discussion (i), E is not finite-dimensional. Hence, $h_1 = ax_1^2 + ax_2^2 + c$. By Theorem 4.7, ω_{12} is a polynomial of degree at most 1. Moreover, ω_{12} must be a nondegenerate degree 1 polynomial by Theorem 2.5, i.e., $\omega_{12} = c_1 x_1 + c_2 x_2 + c_0$, where $c_1 \neq 0$ or $c_2 \neq 0$.

From $K_0, K_2 \in E$ in Theorem 4.7 and by the construction of (4.6), let $Z_0 = K_2$ and $Z_{r+1} = (1-r)Z_r - [K_0, Z_r]$ for $r = 0, 1$,

$$Z_0 = \sum_{i=1}^{2} \alpha_i D_i + q_0 \in E,$$

$$Z_1 = c_0 x_2 D_1 - c_0 x_1 D_2 + q_0 - E(q_0) - \sum_{i=1}^{2} \alpha_i^2 \in E,$$

$$Z_2 = -c_0 x_2 \alpha_1 + c_0 x_1 \alpha_2 + E\left(\sum_{i=1}^{2} \alpha_i^2\right) - E(q_0 - E(q_0)) \in E,$$

where $q_0 = \frac{1}{2}\sum_{i=1}^{2} \frac{\partial \alpha_i}{\partial x_i} - \eta - \frac{1}{2}E(\eta)$. Since Z_2 must be a polynomial of degree less than or equal to 2, regardless of whether c_0 is 0 or not, $q_0 - E(q_0)$ or $E(q_0 - E(q_0))$ is a degree 4 polynomial by Theorem 3.15. So η is a degree 4 polynomial (by Theorems 3.13 and 3.15). Moreover, the homogeneous degree 4 part of η is

$$\eta^{(4)} = c\left(\sum_{i=1}^{2} \alpha_i^2\right)^{(4)} = c(x_1^2 + x_2^2)(c_1 x_1 + c_2 x_2)^2.$$

Hence, the homogeneous degree 4 part of $\eta - \sum h_i^2$ is

$$c(x_1^2 + x_2^2)(c_1 x_1 + c_2 x_2)^2 - a^2(x_1^2 + x_2^2)^2 - \sum_{i>1}^{m} (h_i^{(2)})^2.$$

By taking $x_1 = -c_2$ and $x_2 = c_1$, the above expression results in a negative number. By Theorem 3.18, there are no smooth solutions in f_i. □

THEOREM 4.9. *Suppose* $\dim E < \infty$. *If* $\phi \in E$, ϕ *is a polynomial of degree at most* 1.

Proof. (i) If E has linear rank 2, E is of maximal rank, and therefore no degree 2 polynomials are in E by Theorem 2.7.

(ii) If E has linear rank 0, then h_i's must be constants. E is finite-dimensional and $E \subset \langle L_0, 1\rangle_{L.A.}$. Any function element in E must be a constant.

(iii) Let E have linear rank 1, i.e., there exists a function $d_i x_1 + e_i \in E \Longrightarrow D_1 \in E$.

From the proof of Theorem 4.8, by orthogonal and affine transformations, a degree 2 polynomial $\phi \in E$ only has the form $ax_1^2 + ax_2^2 + d$, where $a \neq 0$. Again by Theorem 4.8, $\omega_{12} = c_1 x_1 + c_2 x_2 + c_0$, where $c_1 \neq 0$ or $c_2 \neq 0$ (otherwise ω_{12} is a constant).

By (4.3), $[L_0, D_1] = \omega_{12} D_2 + \varphi \in E$. Thus

$$[\omega_{12} D_2 + \varphi, ax_1^2 + ax_2^2 + d] = 2ax_2\omega_{12} = 2ac_1 x_1 x_2 + 2ac_2 x_2^2 + 2ac_0 x_2 \in E.$$

If $c_1 \neq 0$, $[D_1, 2ac_1 x_1 x_2 + 2ac_2 x_2^2 + 2ac_0 x_2] = 2ac_1 x_2$ is dependent on x_2. This contradicts the linear rank of E being 1. If $c_1 = 0$ and $c_2 \neq 0$, from $2ac_2 x_2^2 + 2ac_0 x_2$ and $ax_1^2 + ax_2^2 + d$, are in E, one has that $x_1^2 + c_4 x_2 + c_5 \in E$. From the discussion (i) in the proof of Theorem 4.8, E is not finite-dimensional. □

4.2. Constant structure of the Ω-matrix. In this section, E is assumed to be finite-dimensional and have linear rank 1. As discussed in the proof of Theorem 4.9, the structure of E is very clear when E's linear rank is 0 or 2.

Assume $x_1 \in E$. Then if a function $p \in E$, p is a degree 1 polynomial in x_1 as a result of Theorem 4.9 as well as the linear rank assumption. By Theorem 4.2, $\omega_{12} = c_1x_1 + c_2x_2 + c_0$. It will be shown that ω_{12} must be a constant in this section.

By assumptions, $L_0, x_1, D_1, 1 \in E$. Moreover, the following elements are in E:

$$H_0 = [L_0, D_1] = \omega_{12}D_2 + \frac{1}{2}\frac{\partial \eta}{\partial x_1} + \frac{1}{2}c_2, \tag{4.27}$$

$$H_1 = [D_1, H_0] = c_1D_2 - \omega_{12}^2 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2}, \tag{4.28}$$

$$H_2 = [D_1, H_1] = -3c_1\omega_{12} + \frac{1}{2}\frac{\partial^3 \eta}{\partial x_1^3}, \tag{4.29}$$

$$\begin{aligned} H_3 &= [H_1, H_0] - c_2H_1 \\ &= 3c_2\omega_{12}^2 + \frac{1}{2}c_1\frac{\partial^2 \eta}{\partial x_1 \partial x_2} - \frac{1}{2}c_2\frac{\partial^2 \eta}{\partial x_1^2} - \frac{1}{2}\omega_{12}\frac{\partial^3 \eta}{\partial x_1^2 \partial x_2}, \end{aligned} \tag{4.30}$$

$$\begin{aligned} X_0 &= [L_0, H_0] \\ &= c_1D_1D_2 + c_2D_2^2 + \left(-\omega_{12}^2 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2}\right)D_1 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_2}D_2 \\ &\quad + \left(-\frac{1}{2}c_1\omega_{12} + \frac{1}{2}\omega_{12}\frac{\partial \eta}{\partial x_2} + \frac{1}{4}\frac{\partial^3 \eta}{\partial x_1^3} + \frac{1}{4}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2}\right). \end{aligned} \tag{4.31}$$

LEMMA 4.10. *η is a polynomial in x_1 of degree at most 4 with coefficients being functions of x_2. Let $\eta = \sum_{i=0}^4 a_i x_1^i$, where a_i's are functions of x_2. Then*

(i) *a_4 is a constant;*

(ii) *$a_3 = c_1c_2x_2 +$ constant;*

(iii) *$(c_1^2 - 4a_4)c_2 = 0$;*

(iv) *$3c_2^3x_2^2 + 6c_0c_2^2x_2 - c_2a_2 + \frac{1}{2}c_1a_1' - c_0a_2' - c_2a_2'x_2 =$ constant.*

Proof. Since H_2 is a function $\in E$, H_2 is independent of x_2 and a degree 1 polynomial in x_1. Therefore, η is a polynomial in x_1 of degree at most 4 with coefficient functions in x_2. Let $\eta = \sum_{i=0}^4 a_i x_1^i$, where a_i's are functions of x_2.

$H_2 \in E \Longrightarrow a_4$ is a constant, and $a_3 = c_1c_2x_2 +$ constant.

(iii) and (iv) follow from $H_3 \in E$ by substituting $\eta = \sum_{i=0}^4 a_i x_1^i$ into (4.30). ∎

Now, from (4.31), $X_0 \in E$.

$$\begin{aligned} X &= [L_0, X_0] \\ &= \left[L_0, c_1D_1D_2 + c_2D_2^2 + \left(-\omega_{12}^2 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2}\right)D_1 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_2}D_2\right] \\ &= \left(-3c_1\omega_{12} + \frac{1}{2}\frac{\partial^3 \eta}{\partial x_1^3}\right)D_1^2 + \left(-4c_2\omega_{12} + \frac{\partial^3 \eta}{\partial x_1^2 \partial x_2}\right)D_1D_2 \\ &\quad + \left(c_1\omega_{12} + \frac{1}{2}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2}\right)D_2^2 \\ &= q_0D_1^2 + q_1D_1D_2 + q_2D_2^2 \in E, \quad \text{mod } U_1, \end{aligned} \tag{4.32}$$

where $q_0 = -3c_1\omega_{12} + \frac{1}{2}\frac{\partial^3 \eta}{\partial x_1^3}$, $q_1 = -4c_2\omega_{12} + \frac{\partial^3 \eta}{\partial x_1^2 \partial x_2}$, $q_2 = c_1\omega_{12} + \frac{1}{2}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2}$. By

Lemma 4.10 (i) and (ii),

$$\frac{\partial q_0}{\partial x_1} = 12a_4 - 3c_1^2, \quad \frac{\partial q_0}{\partial x_2} = 0, \tag{4.33}$$

$$\frac{\partial q_1}{\partial x_1} = 2c_1c_2, \quad \frac{\partial q_1}{\partial x_2} = -4c_2^2 + 2a_2'', \tag{4.34}$$

$$\frac{\partial q_2}{\partial x_1} = c_1^2 + a_2'', \quad \frac{\partial q_2}{\partial x_2} = c_1c_2 + \frac{1}{2}a_1''' + a_2'''x_1. \tag{4.35}$$

LEMMA 4.11. $a_4 = \frac{1}{4}c_1^2$.

Proof. If $a_4 \neq \frac{1}{4}c_1^2$, i.e., $r := \frac{\partial q_0}{\partial x_1} = 12a_4 - 3c_1^2 \neq 0$, then from X in (4.32) and L_0 one can construct a sequence of elements in E whose orders strictly increase as follows:

$$\begin{aligned}
Y_1 &= [L_0, X] = rD_1^3 + \sum_{i=1}^{3} p_{1i}D_1^{3-i}D_2^i, \quad \text{mod } U_2, \\
Y_2 &= [Y_1, X] = [rD_1^3 + p_{11}D_1^2D_2 + \cdots q_0D_1^2 + q_1D_1D_2 + q_2D_2^2] \\
&= \left(3r\frac{\partial q_0}{\partial x_1} + p_{11}\frac{\partial q_0}{\partial x_2}\right)D_1^4 + \sum_{i=1}^{4} p_{2i}D_1^{4-i}D_2^i \\
&= 3r^2D_1^4 + \sum_{i=1}^{4} p_{2i}D_1^{4-i}D_2^i, \quad \text{mod } U_3, \\
&\vdots
\end{aligned}$$

Assume that $Y_k = r_kD_1^{k+2} + \sum_{i=1}^{k+2} p_{ki}D_1^{k+2-i}D_2^i$, mod U_{k+1}, with $r_k \neq 0$:

$$\begin{aligned}
Y_{k+1} &= [Y_k, X] = [r_kD_1^{k+2} + p_{k1}D_1^{k+1}D_2 + \cdots q_0D_1^2 + q_1D_1D_2 + q_2D_2^2] \\
&= \left((k+2)r_k\frac{\partial q_0}{\partial x_1} + p_{k1}\frac{\partial q_0}{\partial x_2}\right)D_1^{k+3} + \sum_{i=1}^{k+3} p_{(k+1)i}D_1^{k+3-i}D_2^i \\
&= r_{k+1}D_1^{k+3} + \sum_{i=1}^{k+3} p_{(k+1)i}D_1^{k+3-i}D_2^i, \quad \text{mod } U_{k+2},
\end{aligned}$$

where $r_{k+1} = (k+2)rr_k \neq 0$. □

THEOREM 4.12. *If $c_1c_2 \neq 0$, E is not finite-dimensional.*

Proof. Assume that $c_1c_2 \neq 0$.

(i) From Lemma 4.11 and (4.32), (4.33), q_0 is a constant.

(ii) Let $r_1 := \frac{\partial q_1}{\partial x_1} = 2c_1c_2$ in (4.34):

$$\begin{aligned}
Z &= [L_0, X] = [L_0, q_0D_1^2 + q_1D_1D_2 + q_2D_2^2] \\
&= r_1D_1^2D_2 \ + \text{terms with lower order in } D_1, \quad \text{mod } U_2.
\end{aligned}$$

Suppose q_1 in (4.32) is a polynomial of degree k in x_2, i.e., $\frac{\partial^k q_1}{\partial x_2^k}$ is a nonzero constant.

If $k \geq 1$,

$$\begin{aligned}
A_1 &= [Z, X] \\
&= [r_1 D_1^2 D_2 \ + \text{terms with lower order in } D_1, q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2] \\
&= r_1 \frac{\partial q_1}{\partial x_2} D_1^3 D_2 + \text{terms with lower order in } D_1, \quad \mod U_3, \\
A_2 &= [Z, A_1] = \Big[r_1 D_1^2 D_2 \ + \text{terms with lower order in } D_1, \\
&\qquad r_1 \frac{\partial q_1}{\partial x_2} D_1^3 D_2 \ + \text{terms with lower order in } D_1 \Big] \\
&= r_1^2 \frac{\partial^2 q_1}{\partial x_2^2} D_1^5 D_2 \ + \text{terms with lower order in } D_1, \quad \mod U_5, \\
&\ \vdots \\
A_k &= [Z, A_{k-1}] = r_1^k \frac{\partial^k q_1}{\partial x_2^k} D_1^{2k+1} D_2 \ + \text{terms with lower order in } D_1, \quad \mod U_{2k+1}.
\end{aligned}$$

Now one can repeat the above process by letting $Z = A_k$ to construct a sequence of elements in E whose orders strictly increase.

Hence, k must be zero, which means $\frac{\partial q_1}{\partial x_2} = 0 \Longrightarrow a_2'' = 2c_2^2$.

(iii) By taking twice the derivative with respect to x_2 in Lemma 4.10 (iv) and substituting $a_2'' = 2c_2^2$, one has $c_1 a_1''' = 0 \Longrightarrow a_1''' = 0$.

Thus, $X = q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2$, mod U_1, is in E, with

$$\frac{\partial q_0}{\partial x_1} = \frac{\partial q_0}{\partial x_2} = 0, \tag{4.36}$$

$$\frac{\partial q_1}{\partial x_1} = 2c_1 c_2 = r_1 \neq 0, \qquad \frac{\partial q_1}{\partial x_2} = 0, \tag{4.37}$$

$$\frac{\partial q_2}{\partial x_1} = c_1^2 + 2c_2^2 = r_2 > 0, \qquad \frac{\partial q_2}{\partial x_2} = c_1 c_2 = \frac{1}{2} r_1, \tag{4.38}$$

$$\begin{aligned}
Y_0 &= [L_0, X] = [L_0, q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2] \\
&= p_{01} D_1^2 D_2 + p_{02} D_1 D_2^2 + p_{03} D_2^3, \quad \mod U_2,
\end{aligned} \tag{4.39}$$

where $p_{01} = r_1, p_{02} = r_2, p_{03} = \frac{1}{2} r_1$. For $l \geq 0$, suppose $Y_l = p_{l1} D_1^2 D_2^{l+1} + p_{l2} D_1 D_2^{l+2} + p_{l3} D_2^{l+3} \in E$, mod U_{l+2}, where p_{l1}, p_{l2}, p_{l3} are constants.

$$\begin{aligned}
Y_{l+1} &= [Y_l, X] = [p_{l1} D_1^2 D_2^{l+1} + p_{l2} D_1 D_2^{l+2} + p_{l3} D_2^{l+3}, q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2] \\
&= 2p_{l1} \frac{\partial q_1}{\partial x_1} D_1^2 D_2^{l+2} + p_{l2} \frac{\partial q_1}{\partial x_1} D_1 D_2^{l+3} + (l+1) p_{l1} \frac{\partial q_2}{\partial x_2} D_1^2 D_2^{l+2} \\
&\quad + 2p_{l1} \frac{\partial q_2}{\partial x_1} D_1 D_2^{l+3} + (l+2) p_{l2} \frac{\partial q_2}{\partial x_2} D_1 D_2^{l+3} + p_{l2} \frac{\partial q_2}{\partial x_1} D_2^{l+4} \\
&\quad + (l+3) p_{l3} \frac{\partial q_2}{\partial x_2} D_2^{l+4}
\end{aligned}$$

$$= \frac{l+5}{2} r_1 p_{l1} D_1^2 D_2^{l+2} + (2r_2 p_{l1} + \frac{l+4}{2} r_1 p_{l2}) D_1 D_2^{l+3}$$
$$+ \left(r_2 p_{l2} + \frac{l+3}{2} r_1 p_{l3} \right) D_2^{l+4}$$
$$= p_{(l+1)1} D_1^2 D_2^{l+2} + p_{(l+1)2} D_1 D_2^{l+3} + p_{(l+1)3} D_2^{l+4}, \quad \text{mod } U_{l+3},$$

where $p_{(l+1)1} = \frac{l+5}{2} r_1 p_{l1}$, $p_{(l+1)2}, p_{(l+1)3}$ are constants. Since $p_{01} = r_1 \neq 0$, $p_{l1} \neq 0$ for any l. Thus, Y_l is a differential operator of degree $l+3$. Contradiction! □

LEMMA 4.13. *If $c_1 = 0, c_2 \neq 0$, we have* (i) $a_4 = 0$*;* (ii) $a_3 = 0$; (iii) $a_2 = \omega_{12}^2 +$ *constant;* (iv) $a_1''' = 0$.

Proof. (i) follows from Lemma 4.11.

By Lemma 4.10 (ii), a_3 is a constant. Since a_3 is the coefficient of an odd order term in η with the highest order in x_1, $a_3 = 0$ by Theorem 3.21.

(iii) follows from (4.28). H_1 is a function in E when $c_1 = 0$.

Now, q_0, q_1 are actually constants in (4.32), while $\frac{\partial q_2}{\partial x_1} = a_2'' = 2c_2^2$ and $\frac{\partial q_2}{\partial x_2} = \frac{1}{2} a_1'''$.

Let a_1'' be a degree k polynomial in x_2, i.e., $\frac{d^{k+2} a_1}{dx_2^{k+2}}$ is a nonzero constant. If $k > 1$,

$$Z_0 = Ad_{L_0}^k X = \frac{1}{2} \frac{d^{k+2} a_1}{dx_2^{k+2}} D_2^{k+2}, \quad \text{mod } U_{k+1},$$
$$Y_0 = [Z_0, X] = (k+2) \left(\frac{1}{2} \right)^2 \frac{d^{k+2} a_1}{dx_2^{k+2}} \frac{d^3 a_1}{dx_2^3} D_2^{k+3}, \quad \text{mod } U_{k+2}.$$

By letting $Z_l = Ad_{L_0}^{k-1} Y_{l-1}$ and $Y_l = [Z_l, X]$ for $l \geq 1$, one can construct a sequence of elements in E with strictly increasing orders.

If $k = 1$, consider $Z_0 = [L_0, X] = \frac{\partial q_2}{\partial x_1} D_1 D_2^2 + \frac{\partial q_2}{\partial x_2} D_2^3$, mod U_2, and $Z_{l+1} = [Z_l, X]$ for $l \geq 0$. Assume

$$Z_l = p_{l0} D_1 D_2^{l+2} + p_{l1} D_2^{l+3}, \qquad \text{mod } U_{l+2},$$

where $p_{00} = \frac{\partial q_2}{\partial x_1} = 2c_2^2$ and $p_{01} = \frac{\partial q_2}{\partial x_2} = \frac{1}{2} \frac{d^3 a_1}{dx_2^3}$. Then

$$Z_{l+1} = [Z_l, X] = [p_{l0} D_1 D_2^{l+2} + p_{l1} D_2^{l+3}, q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2]$$
$$= (l+2) p_{l0} \frac{\partial q_2}{\partial x_2} D_1 D_2^{l+3} + \left(p_{l0} \frac{\partial q_2}{\partial x_1} + (l+3) p_{l1} \frac{\partial q_2}{\partial x_2} \right) D_2^{l+4}, \quad \text{mod } U_{l+3}.$$

Therefore,

$$p_{(l+1)0} = (l+2) p_{l0} \frac{\partial q_2}{\partial x_2} \Longrightarrow p_{l0} = (l+1)! \frac{\partial q_2}{\partial x_1} \left(\frac{\partial q_2}{\partial x_2} \right)^l \neq 0.$$

Z_l's have strictly increasing orders. Contradiction!

Hence k must be zero $\Longrightarrow a_1''' = 0$. □

THEOREM 4.14. *If $c_1 = 0, c_2 \neq 0$, E is not finite-dimensional.*

Proof. In the proof of this theorem, $r_1, r_2, \ldots$ are used to denote constants. The exact values of these r_i's are not important. Some r_i's may be used repeatedly to denote different constants.

By substituting Lemma 4.13 (i)–(iv) into (4.31), we obtain

$$
\begin{aligned}
Z_0 &= X_0 - r_1 D_1 - r_2 x_1 - r_3 \\
&= c_2 D_2^2 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_2} D_2 + \frac{1}{2}\omega_{12}\frac{\partial \eta}{\partial x_2}, \qquad (4.40)\\
Z_1 &= [L_0, Z_0] - r_5 D_1 \\
&= \frac{1}{2}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2} D_2^2 + \left(\frac{3}{2} c_2 \frac{\partial \eta}{\partial x_2} + \frac{1}{2}\omega_{12}\frac{\partial^2 \eta}{\partial x_2^2}\right) D_2, \quad \mod U_0, \\
Z_2 &= [L_0, Z_1] = a_2'' D_1 D_2^2 + \left(\frac{3}{2} c_2 \frac{\partial^2 \eta}{\partial x_1 \partial x_2} - \frac{1}{2}\omega_{12}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2}\right) D_1 D_2 \\
&\quad + \left(2 c_2 \frac{\partial^2 \eta}{\partial x_2^2} + \frac{1}{2}\omega_{12}\frac{\partial^3 \eta}{\partial x_2^3}\right) D_2^2, \quad \mod U_1, \\
Z_3 &= [L_0, Z_2] = 3 c_2 \frac{\partial^3 \eta}{\partial x_1 \partial x_2^2} D_1 D_2^2 + \left(a_2'' \omega_{12} + \frac{5}{2} c_2 \frac{\partial^3 \eta}{\partial x_2^3} + \frac{1}{2}\omega_{12}\frac{\partial^4 \eta}{\partial x_2^4}\right) D_2^3 \\
&= p_1 D_1 D_2^2 + p_2 D_2^3, \quad \mod U_2, \qquad (4.41)
\end{aligned}
$$

where $p_1 = 3c_2 \frac{\partial^3 \eta}{\partial x_1 \partial x_2^2} = 3c_2(a_1'' + 2a_2'' x_1)$ is degree 1 in x_1 and independent of x_2 with $\frac{\partial p_1}{\partial x_1} = 6c_2 a_2'' = 12 c_2^3 \neq 0$, and $p_2 = a_2'' \omega_{12} + \frac{5}{2} c_2 \frac{\partial^3 \eta}{\partial x_2^3} + \frac{1}{2}\omega_{12}\frac{\partial^4 \eta}{\partial x_2^4}$ is a function of x_2 and independent of x_1. By Theorem 3.2, p_2 must be a polynomial in x_2; otherwise E is not finite-dimensional.

By (4.40), $Y_0 = [Z_0, Z_3] = 2c_2 \frac{\partial p_2}{\partial x_2} D_2^4$, $\mod U_3 \in E$, $Y_1 = [Z_0, Y_0] = (2c_2)^2 \frac{\partial^2 p_2}{\partial x_2^2} D_2^5$, $\mod U_4 \in E$, $\ldots$, $Y_l = Ad_{Z_0}^{l+1} Z_3 = (2c_2)^{l+1} \frac{\partial^{l+1} p_2}{\partial x_2^{l+1}} D_2^{l+4}$, $\mod U_{l+3} \in E$. If $deg\ p_2 = k \geq 1$, one can construct an infinite sequence of elements in E whose orders strictly increase as follows:

$$
\begin{aligned}
W_0 &= Ad_{Z_0}^k Z_3 = (2c_2)^k \frac{\partial^k p_2}{\partial x_2^k} D_2^{k+3} \\
&= d_0 D_2^{e_0}, \quad \mod U_{e_0 - 1}, \quad \text{where } d_0 = (2c_2)^k \frac{\partial^k p_2}{\partial x_2^k} \neq 0, e_0 = k + 3, \\
W_{l+1} &= Ad_{Z_0}^{k-1}[W_l, Z_3] = Ad_{Z_0}^{k-1}[d_l D_2^{e_l}, p_1 D_1 D_2^2 + p_2 D_2^3] \\
&= Ad_{Z_0}^{k-1} e_l d_l \frac{\partial p_2}{\partial x_2} D_2^{e_l + 2} \\
&= e_l d_l (2c_2)^{k-1} \frac{\partial^k p_2}{\partial x_2^k} D_2^{e_l + k + 1} \\
&= d_{l+1} D_2^{e_{l+1}}, \quad \mod U_{e_{l+1} - 1},
\end{aligned}
$$

where $d_{l+1} = e_l d_l (2c_2)^{k-1} \frac{\partial^k p_2}{\partial x_2^k} \neq 0$, and $e_{l+1} = e_l + k + 1 > e_l$. Therefore, p_2 is a constant.

Now, p_1 is a degree 1 polynomial in x_1, and p_2 is a constant in (4.41).

$$V_0 = [L_0, Z_3] = \frac{\partial p_1}{\partial x_1} D_1^2 D_2^2, \quad \text{mod } U_3,$$

$$V_1 = [V_0, Z_3] = 2\left(\frac{\partial p_1}{\partial x_1}\right)^2 D_1^2 D_2^4, \quad \text{mod } U_5,$$

$$V_{l+1} = [V_l, Z_3] = \left[2^l \left(\frac{\partial p_1}{\partial x_1}\right)^{l+1} D_1^2 D_2^{2(l+1)}, p_1 D_1 D_2^2 + p_2 D_2^3\right]$$

$$= 2^{l+1}\left(\frac{\partial p_1}{\partial x_1}\right)^{l+2} D_1^2 D_2^{2(l+2)}, \quad \text{mod } U_{2l+5}.$$

Again, V_l's have increasing orders. Hence, E is not finite-dimensional. □

LEMMA 4.15. *If $c_1 \neq 0, c_2 = 0$, we have* (i) $a_4 = \frac{1}{4}c_1^2$; (ii) a_3 *is a constant;* (iii) $a_2'' = 0$; (iv) $a_1'' = 0$.

Proof. (i) follows from Lemma 4.11.

(ii) follows from Lemma 4.10 (ii).

Substituting $c_2 = 0$ into Lemma 4.10 (iv), $c_1 a_1' - 2c_0 a_2' =$ constant. Hence,

$$a_1'' = 2\frac{c_0}{c_1} a_2''. \tag{4.42}$$

From (4.32)–(4.35), q_0 is a constant, $\frac{\partial q_1}{\partial x_1} = 0$, $\frac{\partial q_1}{\partial x_2} = 2a_2''$. Consider

$$Y_0 = [L_0, [X, x_1]] = [L_0, 2q_0 D_1 + q_1 D_2] = \frac{\partial q_1}{\partial x_2} D_2^2 = 2a_2'' D_2^2, \quad \text{mod } U_1. \tag{4.43}$$

Let a_2'' be a degree k polynomial in x_2, i.e., $\frac{d^{k+2}a_2}{dx_2^{k+2}}$ is a nonzero constant. Then the orders of the elements $Z_0 = Ad_{L_0}^k Y_0$, $Y_l = [Z_{l-1}, Y_0]$, $Z_{l+1} = Ad_{L_0}^{k-1} Y_l$ for $l \geq 1$ will be strictly increasing unless $k = 0$. Hence, a_2'' must be a constant.

By (4.35) and (4.42), $\frac{\partial q_2}{\partial x_2} = 0$. Thus, among q_0, q_1, and q_2 in X in (4.32), q_0 is a constant, $\frac{\partial q_1}{\partial x_1} = \frac{\partial q_2}{\partial x_2} = 0$, $\frac{\partial q_1}{\partial x_2} = 2a_2''$, and $\frac{\partial q_2}{\partial x_1} = c_1^2 + a_2''$ are constants. Consider $Y_l = Ad_{-X}^l Y_0$, where Y_0 is in (4.43).

$$Y_1 = [Y_0, X] = \left[\frac{\partial q_1}{\partial x_2} D_2^2, q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2\right] = 2\left(\frac{\partial q_1}{\partial x_2}\right)^2 D_1 D_2^2, \quad \text{mod } U_2,$$

$$Y_2 = [Y_1, X] = \left[2\left(\frac{\partial q_1}{\partial x_2}\right)^2 D_1 D_2^2, q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2\right]$$

$$= 4\left(\frac{\partial q_1}{\partial x_2}\right)^3 D_1^2 D_2^2 \; + \; \text{terms with lower order in } D_1$$

$$\text{with constant coefficients}, \quad \text{mod } U_3,$$

$$\vdots$$

$$
\begin{aligned}
Y_{l+1} &= [Y_l, X] \\
&= \left[2^l \left(\frac{\partial q_1}{\partial x_2} \right)^{l+1} D_1^l D_2^2 \ + \ \text{terms with lower order in } D_1 \text{ with constant coefficients,} \right. \\
&\qquad\qquad \left. q_0 D_1^2 + q_1 D_1 D_2 + q_2 D_2^2 \right] \\
&= 22^l \left(\frac{\partial q_1}{\partial x_2} \right)^{l+1} \frac{\partial q_1}{\partial x_2} D_1^{l+1} D_2^2 \ + \ \text{terms with lower order in } D_1 \\
&\qquad\qquad \text{with constant coefficients} \\
&= 2^{l+1} \left(\frac{\partial q_1}{\partial x_2} \right)^{l+2} D_1^{l+1} D_2^2 \ + \ \text{terms with lower order in } D_1 \\
&\qquad\qquad \text{with constant coefficients } \mathrm{mod} U_{l+2}.
\end{aligned}
$$

Thus, the order of Y_l goes to infinity unless $\frac{\partial q_1}{\partial x_2} = 0 \Longrightarrow a_2'' = 0 \Longrightarrow a_1'' = 0$ by (4.42). □

LEMMA 4.16. *If $c_1 \neq 0, c_2 = 0$, then η is a degree 4 polynomial in x_1, x_2 with its principal part $\eta^{(4)} = \frac{1}{4} c_1^2 x_1^4$.*

Proof. Again, $r_1, r_2, \ldots$ are used to denote constants. The exact values of these r_i's are not important. Some r_i's may be used repeatedly to denote different constants.

By substituting Lemma 4.15 (i)–(iv) into (4.31),

$$
\begin{aligned}
Z_0 &= X_0 - r_1 x_1 - r_2 \\
&= c_1 D_1 D_2 + \left(-\omega_{12}^2 + \frac{1}{2} \frac{\partial^2 \eta}{\partial x_1^2} \right) D_1 + \frac{1}{2} \frac{\partial^2 \eta}{\partial x_1 \partial x_2} D_2 + \frac{1}{2} \omega_{12} \frac{\partial \eta}{\partial x_2}, \\
Z_1 &= [L_0, Z_0] - r_3 D_1 \\
&= r_4 D_1^2 + c_1 \omega_{12} D_2^2 + 2a_2' D_1 D_2 + c_1 \frac{\partial \eta}{\partial x_2} D_1 \\
&\quad + \left(-\omega_{12}^3 + \frac{1}{2} \omega_{12} \frac{\partial^2 \eta}{\partial x_1^2} + \frac{1}{2} \omega_{12} \frac{\partial^2 \eta}{\partial x_2^2} + \frac{1}{2} c_1 \frac{\partial \eta}{\partial x_1} \right) D_2, \quad \mathrm{mod}\ U_0, \\
Z_2 &= [L_0, Z_1] = c_1^2 D_1 D_2^2 + r_5 D_1^2 + \left(3a_2' \omega_{12} + \frac{1}{2} \omega_{12} \frac{\partial^3 \eta}{\partial x_2^3} + \frac{1}{2} c_1 \frac{\partial^2 \eta}{\partial x_1 \partial x_2} \right) D_2^2 \\
&\quad + \left(2r_4 \omega_{12} - 5c_1 \omega_{12}^2 + \frac{3}{2} c_1 \frac{\partial^2 \eta}{\partial x_2^2} + c_1 \frac{\partial^2 \eta}{\partial x_1^2} + \frac{1}{2} \omega_{12} \frac{\partial^3 \eta}{\partial x_1^3} \right) D_1 D_2, \quad \mathrm{mod}\ U_1, \\
Z_3 &= [L_0, Z_2] = r_6 D_1^2 D_2 + \left(r_7 + 2c_1 \frac{\partial^3 \eta}{\partial x_2^3} \right) D_1 D_2^2 + \left(c_1^2 \omega_{12} + \frac{1}{2} \omega_{12} \frac{\partial^4 \eta}{\partial x_2^4} \right) D_2^3 \\
&= r_6 D_1^2 D_2 + (r_7 + 2c_1 a_0''') D_1 D_2^2 + \left(c_1^2 + \frac{1}{2} a_0'''' \right) \omega_{12} D_2^3, \quad \mathrm{mod}\ U_2, \\
Y_0 &= [Z_3, x_1] = 2r_6 D_1 D_2 + (r_7 + 2c_1 a_0''') D_2^2, \quad \mathrm{mod}\ U_1.
\end{aligned}
$$

From the proof of Lemma 4.13 (iii), $\frac{\partial}{\partial x_2}(r_7 + 2c_1 a_0''') = 0 \rightarrow a_0'''' = 0$. By Lemma 4.15, η is a degree 4 polynomial in x_1, x_2 with principal part $\eta^{(4)} = \frac{1}{4} c_1^2 x_1^4$. □

THEOREM 4.17. *If $c_1 \neq 0, c_2 = 0$, E is not finite-dimensional.*

Proof. From Lemma 4.15, (4.28), and (4.27),

$$H_1 = [D_1, H_0] = c_1 D_2 - \omega_{12}^2 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2}$$

$$= c_1 D_2 + \frac{1}{2}c_1^2 x_1^2 + (3a_3 - 2c_1 c_0)x_1 + a_2 - c_0^2, \tag{4.44}$$

$$H_0 = [L_0, D_1] = \omega_{12} D_2 + \frac{1}{2}\frac{\partial \eta}{\partial x_1} + \frac{1}{2}c_2$$

$$= (c_1 x_1 + c_0)D_2 + \frac{1}{2}(a_1 + 2a_2 x_1 + 3a_3 x_1^2 + 4a_4 x_1^3)$$

$$= x_1 H_1 + d_0 H_1 + \left(\frac{1}{2}d x_1^2 + d_0 d x_1 + d_2 x_2 + d_3\right), \tag{4.45}$$

where

$$d = 3c_0 c_1 - 3a_3,\ d_0 = \frac{c_0}{c_1},\ d_2 = \frac{1}{2}a_1' - d_0 a_2', \tag{4.46}$$

and d_3 is a constant with unknown value. (Note that all the d's here are constants.)

From (4.44), D_2 can be represented in terms of H_1, x_1, x_2, i.e.,

$$D_2 = \frac{1}{c_1}\left(H_1 - \left(\frac{1}{2}c_1^2 x_1^2 + (3a_3 - 2c_1 c_0)x_1 + a_2 - c_0^2\right)\right). \tag{4.47}$$

Thus, any element in E can be represented in terms of D_1, H_1, x_1, x_2. The following relations hold:

$$[D_1, x_1] = 1, \qquad [D_1, x_2] = 0, \tag{4.48}$$

$$[H_1, x_1] = 0, \qquad [H_1, x_2] = c_1, \tag{4.49}$$

$$[D_1, H_1] = 3a_3 - 3c_1 c_0 = -d. \tag{4.50}$$

Definition. Let A be a differential operator of order k in E. Then A has a unique representation:

$$A = \sum_{i=0}^{k}\sum_{j=0}^{i} h_{ij} D_1^{i-j} H_1^j, \tag{4.51}$$

where h_{ij}'s are polynomials in x_1, x_2. (It is worth mentioning that in the scope of this theorem, η is a polynomial by Lemma 4.16. In general, it is not clear whether h_{ij} is a polynomial or not.) The *degree* of A is defined to be $deg\ A = \max_i \max_j \{deg\ h_{ij} + 2i\}$. In other words, the degree of the differential operator A is the degree of the polynomial (of x_1, x_2, x_3, x_4 variables) obtained by replacing D_1, H_1 by x_3^2, x_4^2 in the form (4.51) of A.

For example, in (4.45), H_0 is a sum of three terms. The first, $x_1 H_1$, is of degree 3; the second, $d_0 H_1$, is of degree 2; the order 0 term $\frac{1}{2}d x_1^2 + d_0 d x_1 + d_2 x_2 + d_3$ is of degree 2. Hence, $deg\ H_0 = 3$. Similarly, $deg\ D_2 = 2$ by (4.47).

Since $D_1 H_1 = H_1 D_1 - d$, $H_1 h = h H_1 + c_1 \frac{\partial h}{\partial x_2}$, and $D_1 h = h D_1 + \frac{\partial h}{\partial x_1}$, it is easy to show by induction that *deg* is commutative. Moreover, for any $X, Y \in E$,

$$deg\ X + Y \leq \max\{deg\ X, deg\ Y\}, \quad deg\ XY = deg\ YX, \quad deg\ XY \leq deg\ X + deg\ Y. \tag{4.52}$$

Let $Y_l = Ad^l_{L_0} D_1$. The following will show that $deg\ Y_l = l+2$ and

$$Y_{2l} = D_1 H_1^l + Z_{2l}, \qquad Y_{2l+1} = x_1 H_1^{l+1} + Z_{2l+1}, \tag{4.53}$$

where Z_k is a differential operator with $deg\ Z_k \le k+1$:

- $l = 0$, $Y_0 = D_1$, $deg\ Y_0 = 2$, $Z_0 = 0$.
- $l = 1$, $Y_1 = [L_0, D_1] = H_0 = x_1 H_1 + Z_1$, where $Z_1 = d_0 H_1 + (\frac{1}{2} dx_1^2 + d_0 dx_1 + d_2 x_2 + d_3)$ by (4.45). $deg\ Y_1 = 3$, $deg\ Z_1 \le 2$.

Suppose $Y_{2l-1} = x_1 H_1^l + Z_{2l-1}$ with $deg\ Z_{2l-1} \le 2l$.

$$\begin{aligned} Y_{2l} &= [L_0, Y_{2l-1}] = [L_0, x_1 H_1^l + Z_{2l-1}] \\ &= [L_0, x_1] H_1^l + x_1 [L_0, H_1^l] + [L_0, Z_{2l-1}] = D_1 H_1^l + Z_{2l}, \end{aligned}$$

where $Z_{2l} = x_1 [L_0, H_1^l] + [L_0, Z_{2l-1}]$. By Lemma 4.18,

$$deg\ Z_{2l} \le \max\{1 + deg\ [L_0, H_1^l], deg\ [L_0, Z_{2l-1}]\} \le 2l+1.$$

$$\begin{aligned} Y_{2l+1} &= [L_0, Y_{2l}] = [L_0, D_1 H_1^l + Z_{2l}] = [L_0, D_1] H_1^l + D_1 [L_0, H_1^l] + [L_0, Z_{2l}] \\ &= (x_1 H_1 + Z_1) H_1^l + D_1 [L_0, H_1^l] + [L_0, Z_{2l}] = x_1 H_1^{l+1} + Z_{2l+1}, \end{aligned}$$

where $Z_{2l+1} = Z_1 H_1^l + D_1 [L_0, H_1^l] + [L_0, Z_{2l}]$ with

$$deg\ Z_{2l+1} \le \max\{deg\ Z_1 H_1^l, 2 + deg\ [L_0, H_1^l], deg\ [L_0, Z_{2l}]\} = 2l+2.$$

Hence, (4.53) holds by induction. $deg\ Y_l = l+2 \to \infty$. E is not finite-dimensional. □

LEMMA 4.18.

(i) *Let* $A_l = [L_0, H_1^l]$, $B_l = [L_0, D_1^l]$. *Then* $deg\ A_l \le 2l$ *and* $deg\ B_l \le 2l+1$.

(ii) *Let* $C = [L_0, h D_1^{i_1} H_1^{i_2}]$. *Then* $deg\ C \le deg\ h + 2(i_1 + i_2) + 1$.

(iii) $deg\ [L_0, A] \le deg\ A + 1$.

Proof. (i) $l = 1$,

$$\begin{aligned} A_1 = [L_0, H_1] &= \left[L_0, c_1 D_2 + \frac{1}{2} c_1^2 x_1^2 + (3a_3 - 2c_1 c_0) x_1 + a_2 - c_0^2 \right] \\ &= -d D_1 + d_1 H_1 + (d_1 d + c_1 d_2) x_1 + d_4, \end{aligned}$$

where d, d_2 are in (4.46), and $d_1 = \frac{a_2'}{c_1}$ and $d_4 = \frac{1}{2} c_1 a_0' - \frac{1}{c_1} a_2 a_2' + d_1 c_0^2$ are constants. (Note: Actually one extra condition has been obtained on a_0 that $a_0'' = 2(\frac{a_2'}{c_1})^2 = 2d_1^2$ since d_4 is in E and independent of x_2.) Thus, $deg\ A_1 = 2$. By (4.45), $deg\ B_1 = deg\ H_0 = 3$.

$$\begin{aligned} A_{l+1} = [L_0, H_1^{l+1}] &= [L_0, H_1^l] H_1 + H_1^l [L_0, H_1] \\ &= A_l H_1 + H_1^l A_1, \\ B_{l+1} = [L_0, D_1^{l+1}] &= [L_0, D_1^l] D_1 + D_1^l [L_0, D_1] \\ &= B_l D_1 + D_1^l B_1. \end{aligned}$$

By induction and (4.52), (i) holds since $deg\ A_{l+1} \le \max\{deg\ A_l + deg\ H_1, deg\ H_1^l + deg\ A_1\} \le 2(l+1)$, $deg\ B_{l+1} \le \max\{deg\ B_l + deg\ D_1, deg\ D_1^l + deg\ B_1\} \le 2(l+1)+1$.

(ii)

$$\begin{aligned} C &= [L_0, h D_1^{i_1} H_1^{i_2}] \\ &= [L_0, h] D_1^{i_1} H_1^{i_2} + h [L_0, D_1^{i_1}] H_1^{i_2} + h D_1^{i_1} [L_0, H_1^{i_2}]. \\ &= \sum_{j=1}^{2} \frac{\partial h}{\partial x_j} D_j D_1^{i_1} H_1^{i_2} + \frac{1}{2} \sum_{j=1}^{2} \frac{\partial^2 h}{\partial x_j^2} D_1^{i_1} H_1^{i_2} + h B_{i_1} H_1^{i_2} + h D_1^{i_1} A_{i_2}. \end{aligned}$$

Since $deg\ D_2 = 2$ by (4.47),

$$\begin{aligned}
&deg\ C\\
&\le \max\left\{deg\ \frac{\partial h}{\partial x_1}D_1^{i_1+1}H_1^{i_2}, deg\ \frac{\partial h}{\partial x_2}D_2D_1^{i_1}H_1^{i_2}, deg\ \frac{\partial^2 h}{\partial x_1^2}D_1^{i_1}H_1^{i_2},\right.\\
&\qquad \left. deg\ \frac{\partial^2 h}{\partial x_2^2}D_1^{i_1}H_1^{i_2}, deg\ hB_{i_1}H_1^{i_2}, deg\ hD_1^{i_1}A_{i_2}\right\}\\
&\le \max\{deg\ h - 1 + 2(i_1+i_2+1), deg\ h - 1 + 2(i_1+i_2+1), deg\ h - 2 + 2i_1 + 2i_2,\\
&\qquad deg\ h - 2 + 2i_1 + 2i_2, deg\ h + 2i_1 + 1 + 2i_2, deg\ h + 2i_1 + 2i_2\}\\
&= deg\ h + 2(i_1+i_2) + 1.
\end{aligned}$$

(iii) Let $A = \sum_{i=0}^{k}\sum_{j=0}^{i} h_{ij}D_1^{i-j}H_1^j$. Then by (ii)

$$\begin{aligned}
deg\ [L_0, A] &= deg\ \left[L_0, \sum_{i=0}^{k}\sum_{j=0}^{i} h_{ij}D_1^{i-j}H_1^j\right]\\
&\le \max_{i,j} deg\ [L_0, h_{ij}D_1^{i-j}H_1^j] \le \max_{i,j} deg\ h_{ij} + 2i + 1\\
&= deg\ A + 1. \qquad \square
\end{aligned}$$

By putting Theorems 4.12, 4.14, and 4.17 together, we obtain the following.

THEOREM 4.19. *If E has linear rank 1 and is finite-dimensional, then the Ω-matrix has constant entries.*

4.3. Classification results. First assume E has linear rank 1. By Theorem 4.9, h_i's must be degree at most 1 polynomials in x_1. By Theorem 4.19, ω_{12} is a constant.

(i) If $\omega_{12} = 0$, $[L_0, D_1] = \frac{1}{2}\frac{\partial \eta}{\partial x_1} \in E$. Thus, η must be a degree 2 polynomial in x_1 plus a C^∞-function in x_2. $E = \{L_0, x_1, D_1, 1\}$. For example, $f_1 = x_1$, $f_2 = \sin x_2$, $h_1 = x_1$, $\eta = 2x_1^2 + 1 + \cos x_2 + \sin^2 x_2$.

(ii) If $\omega_{12} \neq 0$,

$$A_1 := [L_0, D_1] = \omega_{12}D_2 + \frac{1}{2}\frac{\partial \eta}{\partial x_1} \in E, \tag{4.54}$$

$$A_2 := [D_1, A_1] = -\omega_{12}^2 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2} \in E, \tag{4.55}$$

$$\begin{aligned}
A_3 := [L_0, A_1] = &\left(\frac{1}{2}\frac{\partial^2 \eta}{\partial x_1^2} - \omega_{12}^2\right)D_1 + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_2}D_2\\
&+\frac{1}{2}\omega_{12}\frac{\partial \eta}{\partial x_2} + \frac{1}{4}\frac{\partial^3 \eta}{\partial x_1^3} + \frac{1}{4}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2} \in E.
\end{aligned} \tag{4.56}$$

By (4.55), $\eta = d_0x_1^3 + d_1x_1^2 + e_2(x_2)x_1 + e_3(x_2)$. By Theorem 3.21, $d_0 = 0$. By (4.56) and $D_1 \in E$,

$$\frac{1}{2}\frac{\partial^2 \eta}{\partial x_1 \partial x_2}D_2 + \frac{1}{2}\omega_{12}\frac{\partial \eta}{\partial x_2} + \frac{1}{4}\frac{\partial^3 \eta}{\partial x_1 \partial x_2^2} \in E. \tag{4.57}$$

By Theorem 4.1, $\frac{\partial^2 \eta}{\partial x_1 \partial x_2} = e_2'(x_2)$ is a polynomial of degree at most 1. Thus, e_2 is a polynomial of degree at most 2 in x_2. By substituting e_2 and e_3 into η and removing

D_1, x_1, and 1 from (4.54) and (4.57), one has

$$\bar{A}_1 = \omega_{12} D_2 + \frac{1}{2} e_2 \in E, \tag{4.58}$$

$$\bar{A}_3 = \frac{1}{2} e_2' D_2 + \frac{1}{2} \omega_{12} (e_2' x_1 + e_3') \in E. \tag{4.59}$$

If e_2 is a degree 2 polynomial in x_2, then $[\bar{A}_1, \bar{A}_3] - \frac{1}{2} e_2'' \bar{A}_1 = \frac{1}{2}\omega_{12}^2 e_3'' - \frac{1}{4}(e_2 e_2'' + e_2' e_2') + \frac{1}{2}\omega_{12}^2 e_2'' x_1 \in E$. Since the degree 2 term of $e_2 e_2'' + e_2' e_2'$ will never be zero, $e_3(x_2)$ must be a degree 4 polynomial. Now, consider

$$B_1 = [L_0, \bar{A}_3] = \frac{1}{2} e_2'' D_2^2 + \frac{1}{2}\omega_{12}(e_2'' x_1 + e_3'') D_2 + \frac{1}{4} e_2'(e_2' x_1 + e_3') + \frac{1}{4}\omega_{12} e_3'' \in E,$$

$$B_2 = [L_0, B_1] = \frac{1}{2} e_2''[L_0, D_2^2] + \left[L_0, \frac{1}{2}\omega_{12}(e_2'' x_1 + e_3'') D_2\right]$$

$$= \frac{1}{2}\omega_{12} e_3''' D_2^2 - \frac{1}{2}\omega_{12} e_2'' D_2 D_1 \in E, \qquad \mod U_1.$$

From the discussion of the proof of Lemma 4.13 (iii), e_3''' must be a constant. Contradiction! Hence, e_2 must be a degree 1 polynomial.

Consider $\omega_{12}\bar{A}_3 - \frac{1}{2} e_2' \bar{A}_1 = \frac{1}{2}\omega_{12}^2 e_2' x_1 + \frac{1}{2}\omega_{12}^2 e_3' - \frac{1}{4} e_2' e_2 \in E \Longrightarrow \frac{1}{2}\omega_{12}^2 e_3' - \frac{1}{4} e_2' e_2$ is independent of x_2 and e_3 must be a degree 2 polynomial. Hence, η is a degree 2 polynomial in x_1 and x_2. E is of dimension 5 and $E = \{L_0, x_1, D_1, D_2 + c x_2, 1\}$.

For example, $f_1 = 5x_1 - 3x_2$, $f_2 = 4x_2$, $h_1 = x_1$. Then $\omega_{12} = 3$ and $\eta = 26x_1^2 - 30 x_1 x_2 + 25 x_2^2 + 9$. It is easy to show that $E = \{L_0, x_1, D_1, D_2 - 5x_2, 1\}$.

If E has linear rank 0, h_i's must be constants, and $E = \{L_0\}$ or $E = \{L_0, 1\}$.

If E has linear rank 2, E is of maximal rank. The Ω-matrix must have constant entries and $E = \{L_0, x_1, x_2, D_1, D_2, 1\}$ by Theorem 2.7. In summary, we have the following.

THEOREM 4.20. *Let $n = 2$. If E is finite-dimensional, then*

(1) *if h_i's are constants, $E = \{L_0\}$ or $E = \{L_0, 1\}$;*

(2) *otherwise, Ω-matrix has constant entries. h_i's must be affine in x_1 and x_2. E has dimension of either 4, 5, or 6.*

Moreover, from the above discussion, it is easy to see that if E is finite-dimensional, it has only elements with order less than or equal to 2. Thus, the Levine conjecture holds for the finite-dimensional estimation algebras with state dimension 2.

REFERENCES

[1] V. E. BENES, *Exact finite dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.

[2] R. W. BROCKETT, *Classification and equivalence in estimation theory*, in Proceedings of the 18th Annual IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1979, pp. 172–175.

[3] R. W. BROCKETT, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, 1981, pp. 479–504.

[4] R. W. BROCKETT, *Nonlinear control theory and differential geometry*, in Proceedings of the International Congress of Mathematics, Warsaw, 1983, pp. 1357–1368.

[5] R. W. BROCKETT AND J. M. C. CLARK, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs et. al., eds., Academic Press, New York, 1980, pp. 299–309.

[6] R. S. BUCY AND J. PAGES, *A prior error bounds for the cubic sensor problem*, IEEE Trans. Automat. Control, 24 (1979), pp. 948–953.

[7] M. Chaleyat-Maurel and D. Michel, *Des resultats de non existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.

[8] W.-L. Chiou and S. S.-T. Yau, *Finite-dimensional filters with nonlinear drift* II: *Brockett's problem on classification of finite-dimensional estimation algebras*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[9] R.-T. Dong, L.-F. Tam, W. S. Wong, and S. S.-T. Yau, *Structure and classification theorems of finite-dimensional exact estimation algebras*, SIAM J. Control Optim., 29 (1991), pp. 866–877.

[10] M. Hazewinkel, *Lectures on linear and nonlinear filtering*, in Analysis and Estimation of Stochastic Mechanical Systems, W. Schiehlen and W. Wedig, eds., Springer-Verlag, New York, 1988, pp. 103–135.

[11] M. Hazewinkel, S. I. Marcus, and H. J. Sussmann, *Nonexistence of finite dimensional filters for conditional statistics of the cubic sensor problem*, Systems Control Lett., 3 (1983), pp. 331–340.

[12] J. Levine, *Finite dimensional realizations of stochastic P.D.E.'s and application to filtering*, Stochastics, 37 (1991), pp. 75–103.

[13] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes* I, Springer-Verlag, New York, 1977.

[14] S. I. Marcus, *Algebraic and geometric methods in nonlinear filtering*, SIAM J. Control Optim., 22 (1984), pp. 817–844.

[15] S. K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, Ricerche Automat., 10 (1979), pp. 163–216.

[16] D. Ocone, *Finite dimensional estimation algebras in nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. S. Willems, eds., Reidel, Dordrecht, 1981, pp. 629–636.

[17] A. Rasoulian and S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* IX: *Construction of finite dimensional estimation algebras of nonmaximal rank*, Systems Control Lett., 30 (1997), pp. 109–118.

[18] L.-F. Tam, W. S. Wong, and S. S.-T. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[19] J. Wei and E. Norman, *On the global representation of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Sci., 15 (1964), pp. 327–334.

[20] W. S. Wong, *New classes of finite dimensional nonlinear filters*, Systems Control Lett., 3 (1983), pp. 155–164.

[21] W. S. Wong, *On a new class of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 79–83.

[22] W. S. Wong, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[23] W. S. Wong and S. S.-T. Yau, *The estimation algebra of nonlinear filtering systems*, in Mathematical Control Theory, Special Volume Dedicated to the 60th Birthday of Brockett, J. Baillieul and J. C. Willems eds., Springer-Verlag, New York, 1998, pp. 33–65.

[24] S. S.-T. Yau, *Finite-dimensional filters with nonlinear drift* I: *A class of filters including both Kalman-Bucy filters and Benes filters*, J. Math. Systems Estim. Control, 4 (1994), pp. 181–203.

[25] S. S.-T. Yau, *Brockett's problem on nonlinear filtering theory*, in Lectures on Systems, Control, and Information, AMS/IP Stud. Adv. Math. 17, AMS, Providence, RI, 2000, pp. 177–212.

[26] S. S.-T. Yau and W. L. Chiou, *Recent results on classification of finite dimensional estimation algebras: Dimension of state space* 2, in Proceedings of the 30th Annual Conference on Decision and Control, Brighton, England, 1991, pp. 2758–2760.

[27] S. S.-T. Yau and A. Rasoulian, *Classification of four-dimensional estimation algebras*, IEEE Trans. Automat. Control, 44 (1999), pp. 2312–2318.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 1074–1083

EXACT BOUNDARY CONTROLLABILITY FOR 1-D QUASILINEAR WAVE EQUATIONS*

LI TATSIEN† AND YU LIXIN‡

Abstract. In this paper we present a unified constructive method to establish the local exact boundary controllability for 1-D quasilinear wave equations with boundary conditions of various types.

Key words. exact boundary controllability, 1-D quasilinear wave equation

AMS subject classifications. 35L50, 35L20, 49J20, 93B05, 93C20

DOI. 10.1137/S0363012903427300

1. Introduction and main results. There are many publications concerning the exact controllability for linear hyperbolic systems (see Lions [11], Russell [12], and the references therein). In the semilinear case, some results on the exact boundary controllability for semilinear wave equations are obtained by Zuazua [15], [16], [17], Lasiecka and Triggiani [3], Emanuilov [2], etc. However, even in the 1-D case, only a few results are known for quasilinear wave equations.

Consider the following quasilinear wave equation:

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial}{\partial x}\left(K\left(u, \frac{\partial u}{\partial x}\right)\right) = F\left(u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial t}\right), \tag{1.1}$$

where $K = K(u, v)$ is a given C^2 function of u and v such that

$$K_v(u, v) > 0, \tag{1.2}$$

and $F = F(u, v, w)$ is a given C^1 function of u, v and w satisfying

$$F(0, 0, 0) = 0. \tag{1.3}$$

Moreover, without loss of generality, we may assume that

$$K(0, 0) = 0. \tag{1.4}$$

On one end $x = 0$ we prescribe any one of the following boundary conditions:

$$u = h(t) \qquad \text{(of Dirichlet type)}, \tag{1.5.1}$$

$$u_x = h(t) \qquad \text{(of Neumann type)}, \tag{1.5.2}$$

$$u_x - \alpha u = h(t) \qquad \text{(of the third type)}, \tag{1.5.3}$$

*Received by the editors May 9, 2003; accepted for publication (in revised form) January 20, 2006; published electronically August 16, 2006. Research supported by the Special Funds for Major State Basic Research Projects of China.
http://www.siam.org/journals/sicon/45-3/42730.html
†School of Mathematical Sciences, Fudan University, Shanghai 200433, People's Republic of China (dqli@fudan.edu.cn).
‡Institute of Mathematics, Fudan University, Shanghai 200433, People's Republic of China. (fdylx01@sina.com)

or

$$u_x - \bar{\alpha} u_t = h(t) \qquad \text{(of the dissipative type)}, \tag{1.5.4}$$

where α and $\bar{\alpha}$ are given positive constants and $h(t)$ is a C^2 function (in case (1.5.1)) or a C^1 function (in cases (1.5.2)–(1.5.4)).

Similarly, on another end, $x = 1$, the boundary condition is

$$u = \bar{h}(t) \qquad \text{(of Dirichlet type)}, \tag{1.6.1}$$

$$u_x = \bar{h}(t) \qquad \text{(of Neumann type)}, \tag{1.6.2}$$

$$u_x + \beta u = \bar{h}(t) \qquad \text{(of the third type)}, \tag{1.6.3}$$

or

$$u_x + \bar{\beta} u_t = \bar{h}(t) \qquad \text{(of the dissipative type)}, \tag{1.6.4}$$

where β and $\bar{\beta}$ are given positive constants and $\bar{h}(t)$ is a C^2 function (in case (1.6.1)) or a C^1 function (in cases (1.6.2)–(1.6.4)).

In the case that K and F in (1.1) are independent of u, if on one end, say, on $x = 0$, the boundary condition is given by (1.5.1), by means of the local exact boundary controllability for quasilinear hyperbolic systems without zero eigenvalues, Li and Rao [6] established the corresponding local exact boundary controllability with the boundary control $\bar{h}(t)$ on another end, $x = 1$; while, if on $x = 0$, the boundary condition is given by (1.5.3), by means of the theory on the local exact boundary controllability for quasilinear hyperbolic systems without zero eigenvalues together with a kind of nonlocal boundary conditions, Li and Xu [8] established the corresponding local exact boundary controllability with the boundary control $\bar{h}(t)$ on another end, $x = 1$. However, the method used in [6] and [8] cannot be applied to the following cases: (1) K and F in (1.1) depend explicitly on u; for instance, (1.1) is a linear or nonlinear Klein–Gordon equation. (2) The boundary condition on $x = 0$ is of Neumann type or of the dissipative type. (3) Boundary controls are simultaneously given on both ends $x = 0$ and $x = 1$.

In this paper, we will present a unified method to get the local exact boundary controllability with boundary controls on one end or on two ends for (1.1) with boundary conditions (1.5) and (1.6) of different types. The main results are the following two theorems.

THEOREM 1.1 (exact boundary controllability with boundary controls acting on two ends). *Let*

$$T > \frac{1}{\sqrt{K_v(0,0)}}. \tag{1.7}$$

For any given initial data (φ, ψ) *and final data* (Φ, Ψ) *with small norms*

$$\|(\varphi,\psi)\|_{C^2[0,1]\times C^1[0,1]} \text{ and } \|(\Phi,\Psi)\|_{C^2[0,1]\times C^1[0,1]},$$

there exist boundary controls $h(t)$ *and* $\bar{h}(t)$ *with small norms* $\|h\|_{C^2[0,T]}$ *and* $\|\bar{h}\|_{C^2[0,T]}$ *(for* (1.5.1) *and* (1.6.1)*) or with small norms* $\|h\|_{C^1[0,T]}$ *and* $\|\bar{h}\|_{C^1[0,T]}$ *(for* (1.5.2)–(1.5.4) *and* (1.6.2)–(1.6.4)*) such that the mixed initial-boundary value problem for* (1.1) *with the initial condition*

$$t = 0: \quad u = \varphi(x), \quad u_t = \psi(x), \quad 0 \le x \le 1, \tag{1.8}$$

one of the boundary conditions (1.5) *on* $x = 0$*, and one of the boundary conditions* (1.6) *on* $x = 1$ *admits a unique* C^2 *solution* $u = u(t, x)$ *with small* C^2 *norm on the domain*

$$R(T) = \{(t, x) |\ \ 0 \le t \le T, \ \ 0 \le x \le 1\}, \tag{1.9}$$

which exactly satisfies the final condition

$$t = T : \ \ u = \Phi(x), \ \ u_t = \Psi(x), \ \ 0 \le x \le 1. \tag{1.10}$$

THEOREM 1.2 (exact boundary controllability with boundary control acting on one end). *Let*

$$T > \frac{2}{\sqrt{K_v(0,0)}}. \tag{1.11}$$

Suppose that

$$\bar{\alpha} \ne \frac{1}{\sqrt{K_v(0,0)}}, \tag{1.12}$$

where $\bar{\alpha}$ *is given in* (1.5.4)*. For any given initial data* (φ, ψ) *and final data* (Φ, Ψ) *with small norms* $\|(\varphi, \psi)\|_{C^2[0,1]\times C^1[0,1]}$ *and* $\|(\Phi, \Psi)\|_{C^2[0,1]\times C^1[0,1]}$ *and for any given function* $h(t)$ *with small norm* $\|h\|_{C^2[0,T]}$ *(in case* (1.5.1)*) or with small norm* $\|h\|_{C^1[0,T]}$ *(in cases* (1.5.2)–(1.5.4)*) such that the conditions of* C^2 *compatibility (for details, see Remark* 1.1 *below) are satisfied at the points* $(0, 0)$ *and* $(T, 0)$*, respectively, there exists a boundary control* $\bar{h}(t)$ *with small norm* $\|\bar{h}\|_{C^2[0,T]}$ *(in case* (1.6.1)*) or with small norm* $\|\bar{h}\|_{C^1[0,T]}$ *(in cases* (1.6.2)–(1.6.4)*) such that the mixed initial-boundary value problem for* (1.1) *with the initial condition* (1.8)*, one of the boundary conditions* (1.5) *on* $x = 0$*, and one of the boundary conditions* (1.6) *on* $x = 1$ *admits a unique* C^2 *solution* $u = u(t, x)$ *with small* C^2 *norm on the domain* $R(T)$ *(see* (1.9)*), which exactly satisfies the final condition* (1.10)*.*

Remark 1.1. Precisely speaking, the conditions of C^2 compatibility at the point $(0, 0)$ are given, respectively, for boundary conditions (1.5.1)–(1.5.4) as follows:

(1.13.1)
$$\begin{cases} h(0) = \varphi(0), \\ h'(0) = \psi(0), \\ h''(0) - K_u(\varphi(0), \varphi'(0))\varphi'(0) - K_v(\varphi(0), \varphi'(0))\varphi''(0) = F(\varphi(0), \varphi'(0), \psi(0)); \end{cases}$$

(1.13.2)
$$\begin{cases} h(0) = \varphi'(0), \\ h'(0) = \psi'(0); \end{cases}$$

(1.13.3)
$$\begin{cases} \varphi'(0) - \alpha\varphi(0) = h(0), \\ \psi'(0) - \alpha\psi(0) = h'(0); \end{cases}$$

or

(1.13.4)
$$\begin{cases} \varphi'(0) - \bar{\alpha}\psi(0) = h(0), \\ \psi'(0) - \bar{\alpha}(K_u(\varphi(0), \varphi'(0))\varphi'(0) + K_v(\varphi(0), \varphi'(0))\varphi''(0) + F(\varphi(0), \varphi'(0), \psi(0))) = h'(0). \end{cases}$$

The conditions of C^2 compatibility at the point $(T, 0)$ are similar.

2. C^2 solution on a preassigned domain for 1-D quasilinear wave equations. Since the hyperbolic wave has a finite speed of propagation, the exact boundary controllability for the quasilinear wave equation requires that the controllability time T must be suitably large so that two maximum determinate domains associated with the initial data and the final data, respectively, are separated. Hence, in order to get the exact boundary controllability, we should first prove the existence and uniqueness of the classical solution on the time interval $0 \le t \le T_0$, where $T_0 > 0$ is a preassigned and possibly quite large number.

In order to get the existence and uniqueness of the C^2 solution on a preassigned domain for 1-D quasilinear wave equations with boundary conditions of various types, the best way is to reduce the equation to a first order quasilinear hyperbolic system, then apply the corresponding theory on the existence and uniqueness of the C^1 solution to this system with general nonlinear boundary conditions on a preassigned domain (see [4], [7], [14]), and finally obtain the desired result in a unified way; alternatively, we may apply a careful extension method based on some well-known results on the local existence and uniqueness of the C^2 solution for quasilinear wave equations with boundary conditions of some prescribed types (for instance, see [13]) to get the desired result in an exhaustive way for all possible cases, respectively.

As a result, we can get the following lemma.

LEMMA 2.1. *Under the assumptions given at the beginning of section* 1*, suppose furthermore that the usual conditions of C^2 compatibility (cf.* (1.13)*) are satisfied at the points* $(0,0)$ *and* $(0,1)$*, respectively. Then, for a given $T_0 > 0$, the forward mixed initial-boundary value problem* (1.1)*,* (1.5)–(1.6)*, and* (1.8) *admit a unique C^2 solution $u = u(t,x)$ with small C^2 norm on the domain*

$$R(T_0) = \{(t,x) |\ \ 0 \le t \le T_0, \quad 0 \le x \le 1\}, \tag{2.1}$$

provided that the norms $\|(\varphi,\psi)\|_{C^2[0,1]\times C^1[0,1]}$ and $\|h\|_{C^2[0,T_0]}, \|\bar{h}\|_{C^2[0,T_0]}$ (for (1.5.1) *and* (1.6.1)*) or $\|h\|_{C^1[0,T_0]}, \|\bar{h}\|_{C^1[0,T_0]}$ (for* (1.5.2)–(1.5.4) *and* (1.6.2)–(1.6.4)*) are small enough (depending on T_0).*

Similar results can be obtained for the backward mixed initial-boundary value problem. We then have the following lemma.

LEMMA 2.2. *Under the assumptions given at the beginning of section* 1*, suppose that* (1.12) *holds and*

$$\bar{\beta} \ne \frac{1}{\sqrt{K_v(0,0)}}. \tag{2.2}$$

For a given $T_0 > 0$, suppose furthermore that the conditions of C^2 compatibility are satisfied at the points $(T_0,0)$ *and* $(T_0,1)$*, respectively. Then the backward mixed initial-boundary value problem* (1.1)*,* (1.5)–(1.6)*, and the final condition*

$$t = T_0: \qquad u = \Phi(x), \quad u_t = \Psi(x), \qquad 0 \le x \le 1, \tag{2.3}$$

admit a unique C^2 solution $u = u(t,x)$ with small C^2 norm on the domain (2.1)*, provided that the norms $\|(\Phi,\Psi)\|_{C^2[0,1]\times C^1[0,1]}$ and $\|h\|_{C^2[0,T_0]}, \|\bar{h}\|_{C^2[0,T_0]}$ (for* (1.5.1) *and* (1.6.1)*) or $\|h\|_{C^1[0,T_0]}, \|\bar{h}\|_{C^1[0,T_0]}$ (for* (1.5.2)–(1.5.4) *and* (1.6.2)–(1.6.4)*) are small enough (depending on T_0).*

3. Proof of Theorem 1.1. As in [5] (also see [7] and [9]), in order to prove Theorem 1.1, it suffices to prove the following lemma.

LEMMA 3.1. *Let $T>0$ be defined by* (1.7). *For any given initial data (φ,ψ) and final data (Φ,Ψ) with small norms $\|(\varphi,\psi)\|_{C^2[0,1]\times C^1[0,1]}$ and $\|(\Phi,\Psi)\|_{C^2[0,1]\times C^1[0,1]}$, the quasilinear wave equation* (1.1) *admits a C^2 solution $u=u(t,x)$ with small C^2 norm on the domain $R(T)=\{(t,x)|\ 0\le t\le T,\ 0\le x\le 1\}$, which satisfies simultaneously the initial condition* (1.8) *and the final condition* (1.10).

Proof. By (1.7), there exists an $\varepsilon_0>0$ so small that

$$T>\max_{|u|+|v|\le\varepsilon_0}\frac{1}{\sqrt{K_v(u,v)}}. \tag{3.1}$$

Let

$$T_1=\frac{1}{2}\max_{|u|+|v|\le\varepsilon_0}\frac{1}{\sqrt{K_v(u,v)}}. \tag{3.2}$$

We divide the proof into several steps.

(i) We first consider the following forward mixed initial-boundary value problem of (1.1) with the initial condition (1.8) and the boundary conditions

$$x=0:\qquad u=f_0(t), \tag{3.3}$$

$$x=1:\qquad u=f_1(t), \tag{3.4}$$

where f_0 and f_1 are any given C^2 functions of t with small $C^2[0,T_1]$ norm such that the conditions of C^2 compatibility are satisfied at the points $(0,0)$ and $(0,1)$, respectively. By Lemma 2.1, on the domain

$$\{(t,x)|\ \ 0\le t\le T_1,\ \ 0\le x\le 1\} \tag{3.5}$$

there exists a unique C^2 solution $u=u^{(1)}(t,x)$ with small C^2 norm and

$$|u^{(1)}(t,x)|+\left|\frac{\partial u^{(1)}}{\partial x}(t,x)\right|\le\varepsilon_0. \tag{3.6}$$

Thus, we can determine the corresponding value of $(u^{(1)}(t,x),\frac{\partial u^{(1)}}{\partial x}(t,x))$ on $x=\frac{1}{2}$ as

$$x=\frac{1}{2}:\qquad (u,u_x)=(a(t),\bar a(t)),\quad 0\le t\le T_1, \tag{3.7}$$

the $C^2[0,T_1]$ norm of $a(t)$ and the $C^1[0,T_1]$ norm of $\bar a(t)$ are small.

(ii) Similarly, we consider the following backward mixed initial-boundary value problem of (1.1) with the final condition (1.10) and the boundary conditions

$$x=0:\qquad u=g_0(t), \tag{3.8}$$

$$x=1:\qquad u=g_1(t), \tag{3.9}$$

where g_0 and g_1 are any given C^2 functions of t with small $C^2[T-T_1,T]$ norm such that the conditions of C^2 compatibility are satisfied at the points $(T,0)$ and $(T,1)$, respectively. By Lemma 2.2, on the domain

$$\{(t,x)|\ T-T_1\le t\le T,\ \ 0\le x\le 1\} \tag{3.10}$$

there exists a unique C^2 solution $u = u^{(2)}(t,x)$ with small C^2 norm and

$$|u^{(2)}(t,x)| + \left|\frac{\partial u^{(2)}}{\partial x}(t,x)\right| \le \varepsilon_0. \tag{3.11}$$

Thus, we can determine the corresponding value of $(u^{(2)}(t,x), \frac{\partial u^{(2)}}{\partial x}(t,x))$ on $x = \frac{1}{2}$ as

$$x = \frac{1}{2}: \qquad (u, u_x) = (b(t), \bar{b}(t)), \quad T - T_1 \le t \le T, \tag{3.12}$$

where the $C^2[T-T_1, T]$ norm of $b(t)$ and the $C^1[T-T_1, T]$ norm of $\bar{b}(t)$ are small.

(iii) We can find $c(t) \in C^2[0,T]$ with small C^2 norm and $\bar{c}(t) \in C^1[0,T]$ with small C^1 norm such that

$$(c(t), \bar{c}(t)) = \begin{cases} (a(t), \bar{a}(t)), & 0 \le t \le T_1, \\ (b(t), \bar{b}(t)), & T - T_1 \le t \le T. \end{cases} \tag{3.13}$$

Noting (1.2), we now change the status of t and x, and on the domain

$$R_l(T) = \left\{ (t,x) |\ 0 \le t \le T, \quad 0 \le x \le \frac{1}{2} \right\} \tag{3.14}$$

we consider the following leftward mixed initial-boundary value problem of (1.1) with the initial condition

$$x = \frac{1}{2}: \qquad u = c(t), \quad u_x = \bar{c}(t), \quad 0 \le t \le T, \tag{3.15}$$

and the boundary conditions

$$t = 0: \qquad u = \varphi(x), \qquad 0 \le x \le \frac{1}{2}, \tag{3.16}$$

$$t = T: \qquad u = \Phi(x), \qquad 0 \le x \le \frac{1}{2}, \tag{3.17}$$

where $\varphi(x)$ and $\Phi(x)$ are given by (1.8) and (1.10), respectively.

Obviously, the corresponding conditions of C^2 compatibility are satisfied at the points $(0, \frac{1}{2})$ and $(T, \frac{1}{2})$, respectively. Hence, by Lemma 2.2, there exists a unique C^2 solution $u = u_l(t,x)$ with small C^2 norm on the domain $R_l(T)$ and

$$|u_l(t,x)| + \left|\frac{\partial u_l}{\partial x}(t,x)\right| \le \varepsilon_0 \qquad \text{on } R_l(T), \tag{3.18}$$

provided that the C^2 norms of $\varphi(x)$, $\Phi(x)$, $f_0(t)$, $f_1(t)$, $g_0(t)$, and $g_1(t)$ are small enough.

(iv) Similarly, the rightward mixed initial-boundary value problem of (1.1) with the initial condition (3.15) and the boundary conditions

$$t = 0: \qquad u = \varphi(x), \qquad \frac{1}{2} \le x \le 1, \tag{3.19}$$

$$t = T: \qquad u = \Phi(x), \qquad \frac{1}{2} \le x \le 1, \tag{3.20}$$

admits a unique C^2 solution $u = u_r(t,x)$ with small C^2 norm on the domain

$$R_r(T) = \left\{(t,x)|\ \ 0 \le t \le T,\ \ \frac{1}{2} \le x \le 1\right\} \tag{3.21}$$

and

$$|u_r(t,x)| + \left|\frac{\partial u_r}{\partial x}(t,x)\right| \le \varepsilon_0 \qquad \text{on } R_r(T). \tag{3.22}$$

(v) Let

$$u(t,x) = \begin{cases} u_l(t,x), & (t,x) \in R_l(T), \\ u_r(t,x), & (t,x) \in R_r(T). \end{cases} \tag{3.23}$$

We prove

$$t = 0: \qquad u_t = \psi(x), \qquad 0 \le x \le 1, \tag{3.24}$$

$$t = T: \qquad u_t = \Psi(x), \qquad 0 \le x \le 1. \tag{3.25}$$

In fact, the C^2 solutions $u = u_l(t,x)$ (resp., $u = u_r(t,x)$) and $u = u^{(1)}(t,x)$ satisfy simultaneously (1.1), the initial condition

$$x = \frac{1}{2}: \qquad u = a(t),\ \ u_x = \bar{a}(t), \qquad 0 \le t \le T_1, \tag{3.26}$$

and the boundary condition (3.16) (resp., (3.19)). By uniqueness (cf. [10]) and the choice of T_1 given in (3.2), the mixed initial-boundary value problem (1.1), (3.26), and (3.16) (resp., (3.19)) has a unique C^2 solution on the domain

$$\left\{(t,x)|\ 0 \le t \le 2T_1 x,\ \ 0 \le x \le \frac{1}{2}\right\} \tag{3.27}$$

$$\left(\text{resp., } \left\{(t,x)|\ 0 \le t \le 2T_1(1-x),\ \ \frac{1}{2} \le x \le 1\right\}\right).$$

(In fact, this domain is included in the corresponding maximum determinate domain.) Then

$$u(t,x) \equiv u^{(1)}(t,x) \tag{3.28}$$

on these domains and, in particular, on the interval $0 \le x \le 1$ on the x-axis. Hence, noting (1.8), we immediately get (3.24). In a similar manner we obtain (3.25).

Thus, $u = u(t,x)$ satisfies all the requirements of Lemma 3.1.

4. Proof of Theorem 1.2. As in [6] (also see [9]), in order to get Theorem 1.2 it suffices to prove the following lemma.

LEMMA 4.1. *Let $T > 0$ be defined by* (1.11). *Under the assumptions of Theorem* 1.2, *the quasilinear wave equation* (1.1) *with the boundary condition* (1.5) *on $x = 0$ admits a C^2 solution $u = u(t,x)$ with small C^2 norm on the domain $R(T) = \{(t,x)|\ 0 \le t \le T,\ 0 \le x \le 1\}$, which satisfies the initial condition* (1.8) *and the final condition* (1.10).

Proof. Noting (1.11), there exists an $\varepsilon_0 > 0$ so small that

$$T > 2 \max_{|v|+|v|\le\varepsilon_0} \frac{1}{\sqrt{K_v(u,v)}}. \tag{4.1}$$

Let

$$T_1 = \max_{|u|+|v|\le\varepsilon_0} \frac{1}{\sqrt{K_v(u,v)}}. \tag{4.2}$$

(i) We first consider the following forward mixed initial-boundary value problem of (1.1) with the initial condition (1.8) and the boundary conditions (1.5) and (3.4) (in which $f_1 \in C^2[0,T_1]$ with small C^2 norm and the conditions of C^2 compatibility at the point $(0,1)$ are assumed to be satisfied). By the assumptions of Theorem 1.2, the conditions of C^2 compatibility are also satisfied at the points $(0,0)$. Then, by Lemma 2.1, on the domain (3.5) there exists a unique C^2 solution $u = u^{(1)}(t,x)$ with small C^2 norm, and (3.6) holds. Thus, we can determine the corresponding value of $(u^{(1)}(t,x), \frac{\partial u^{(1)}}{\partial x}(t,x))$ on $x = 0$ as

$$x = 0: \qquad (u, u_x) = (a(t), \bar{a}(t)), \quad 0 \le t \le T_1, \tag{4.3}$$

and the $C^2[0,T_1]$ norm of $a(t)$ and the $C^1[0,T_1]$ norm of $\bar{a}(t)$ are small.

(ii) Similarly, we consider the following backward mixed initial-boundary value problem of (1.1) with the final condition (1.10) and the boundary conditions (1.5) and (3.9) (in which $g_1 \in C^2[T-T_1,T]$ with small C^2 norm and the conditions of C^2 compatibility at the point $(T,1)$ are assumed to be satisfied). By the assumptions of Theorem 1.2, the conditions of C^2 compatibility are also satisfied at the point $(T,0)$. Then, by Lemma 2.2, on the domain (3.10) there exists a unique C^2 solution $u = u^{(2)}(t,x)$ with small C^2 norm, and (3.11) holds. Thus, we can determine the corresponding value of $(u^{(2)}(t,x), \frac{\partial u^{(2)}}{\partial x}(t,x))$ on $x = 0$ as

$$x = 0: \qquad (u, u_x) = (b(t), \bar{b}(t)), \quad T - T_1 \le t \le T, \tag{4.4}$$

and the $C^2[T-T_1,T]$ norm of $b(t)$ and the $C^1[T-T_1,T]$ norm of $\bar{b}(t)$ are small.

(iii) Noting that both $(u, u_t, u_x) = (a(t), a'(t), \bar{a}(t))$ and $(b(t), b'(t), \bar{b}(t))$ satisfy the boundary condition (1.5), we can find $c(t) \in C^2[0,T]$ with small C^2 norm and $\bar{c}(t) \in C^1[0,T]$ with small C^1 norm such that (3.13) holds and $(u, u_t, u_x) = (c(t), c'(t), \bar{c}(t))$ satisfies the boundary condition (1.5) on the whole interval $[0,T]$.

We now change the status of t and x and consider the following rightward mixed initial-boundary value problem of (1.1) with the initial condition

$$x = 0: \qquad u = c(t), \quad u_x = \bar{c}(t), \quad 0 \le t \le T, \tag{4.5}$$

and the boundary conditions

$$t = 0: \qquad u = \varphi(x), \quad 0 \le x \le 1, \tag{4.6}$$

$$t = T: \qquad u = \Phi(x), \quad 0 \le x \le 1, \tag{4.7}$$

where $\varphi(x)$ and $\Phi(x)$ are given in (1.8) and (1.10), respectively.

Obviously, the conditions of C^2 compatibility at the points $(0,0)$ and $(T,0)$ are satisfied. Then, by Lemma 2.1, there exists a unique C^2 solution $u = u(t,x)$ with small C^2 norm on the domain $R(T) = \{(t,x)|\ 0 \le t \le T,\ 0 \le x \le 1\}$ and

$$|u(t,x)| + \left|\frac{\partial u}{\partial x}(t,x)\right| \le \varepsilon_0 \qquad \text{on } R(T). \tag{4.8}$$

The C^2 solutions $u = u(t,x)$ and $u = u^{(1)}(t,x)$ satisfy simultaneously (1.1), the initial condition

$$x = 0: \qquad u = a(t), \quad u_x = \bar{a}(t), \quad 0 \le t \le T_1, \tag{4.9}$$

and the boundary condition (4.6). Noting the choice of T_1 given by (4.2), by the uniqueness (cf. [10]), the mixed initial-boundary value problem (1.1), (4.9), and (4.6) has a unique C^2 solution on the domain

$$\{(t,x)|\ 0 \le t \le T_1(1-x), \quad 0 \le x \le 1\}. \tag{4.10}$$

Then

$$u(t,x) \equiv u^{(1)}(t,x) \tag{4.11}$$

on this domain and, in particular, on the interval $0 \le x \le 1$ on the x-axis. Hence, noting (1.8), we get (3.24). Similarly, we obtain (3.25).

Thus, $u = u(t,x)$ satisfies all the requirements of Lemma 4.1.

5. Remarks.

Remark 5.1. The exact controllability time given in Theorem 1.1 or Theorem 1.2 is optimal.

Remark 5.2. If the boundary conditions (1.5.2)–(1.5.4) on $x = 0$ are replaced, respectively, by

$$x = 0: \qquad K(u, u_x) = h(t), \tag{5.1}$$

$$x = 0: \qquad K(u, u_x) - \alpha u = h(t), \tag{5.2}$$

and

$$x = 0: \qquad K(u, u_x) - \bar{\alpha} u_t = h(t), \tag{5.3}$$

and the boundary conditions (1.6.2)–(1.6.4) on $x = 1$ are replaced, respectively, by

$$x = 1: \qquad K(u, u_x) = \bar{h}(t), \tag{5.4}$$

$$x = 1: \qquad K(u, u_x) + \beta u = \bar{h}(t), \tag{5.5}$$

and

$$x = 1: \qquad K(u, u_x) + \bar{\beta} u_t = \bar{h}(t), \tag{5.6}$$

the conclusions of Theorems 1.1 and 1.2 are still valid provided that the conditions of C^2 compatibility (1.13.2)–(1.13.4), etc., at the points $(0,0)$ and $(T,0)$ are changed correspondingly and (1.12) is replaced by

$$\bar{\alpha} \ne \sqrt{K_v(0,0)}. \tag{5.7}$$

Remark 5.3. For more general boundary conditions such as

$$x = 0: \qquad R(t, u, u_x, u_t) = 0, \tag{5.8}$$

similar results can be obtained under suitable conditions.

Remark 5.4. The method presented in this paper works also for some kinds of second order quasilinear hyperbolic systems and higher order quasilinear hyperbolic equations.

Acknowledgment. The authors would like to express their sincere thanks to the referees for their valuable comments and suggestions.

REFERENCES

[1] M. Cirinà, *Nonlinear hyperbolic problems with solutions on preassigned sets*, Michigan Math. J., 17 (1970), pp. 193–209.

[2] O. Yu. Emanuilov, *Boundary control by semilinear evolution equations*, Russian Math. Surveys, 44 (3) (1989), pp. 183–184.

[3] I. Lasiecka and R. Triggiani, *Exact controllability of semilinear abstract systems with applications to waves and plates boundary control problems*, Appl. Math. Optim., 23 (1991), pp. 109–154.

[4] T. Li and Y. Jin, *Semi-global C^1 solution to the mixed initial-boundary value problem for quasilinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 22 (2001), pp. 325–336.

[5] T. Li and B. Rao, *Exact boundary controllability for quasi-linear hyperbolic systems*, SIAM J. Control Optim., 41 (2003), pp. 1748–1755.

[6] T. Li and B. Rao, *Local exact boundary controllability for a class of quasilinear hyperbolic systems*, Chinese Ann. Math. Ser. B, 23 (2002), pp. 209–218.

[7] T. Li, B. Rao, and Y. Jin, *Solution C^1 semi-globale et contrôlabilité exacte frontière de systèmes hyperboliques quasi linéaires*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 219–224.

[8] T. Li and Y. Xu, *Local exact boundary controllability for nonlinear vibrating string equations*, Internat. J. Modern Phys. B, 17 (2003), pp. 4062–4071.

[9] T. Li and L. Yu, *Exact controllability for first order quasilinear hyperbolic systems with zero eigenvalues*, Chinese Ann. Math. Ser. B, 24 (2003), pp. 415–422.

[10] T. Li and W. Yu, *Boundary Value Problems for Quasilinear Hyperbolic Systems*, Duke Univ. Math. Ser. 5, Duke University, Durham, NC, 1985.

[11] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Vol. I, Masson, Paris, 1988.

[12] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[13] Y. Shibata and G. Nakamura, *On a local existence theorem of Neumann problem for some quasilinear hyperbolic systems of 2nd order*, Math. Z., 202 (1989), pp. 1–64.

[14] L. Yu, *Semi-global C^1 solutions to the mixed initial-boundary value problem for a general class of quasilinear hyperbolics systems*, Chinese Ann. Math. Ser. A, 25 (2004), pp. 549–560 (in Chinese).

[15] E. Zuazua, *Exact controllability for the semilinear wave equation*, J. Math. Pures Appl. (9), 69 (1990), pp. 1–31.

[16] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 10 (1993), pp. 109–129.

[17] E. Zuazua, *Controllability of partial differential equations and its semi-discrete approximations*, Discrete Contin. Dyn. Syst., 8 (2002), pp. 469–513.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 1084–1115

LINEAR SYSTEMS SUBJECT TO INPUT SATURATION AND TIME DELAY: FINITE-GAIN L^p-STABILIZATION*

KARIM YAKOUBI† AND YACINE CHITOUR†

Abstract. This paper deals with finite-gain stabilization of linear systems by static feedbacks which are bounded and time-delayed. For neutrally stable continuous-time linear systems, we provide a saturated linear feedback law ensuring finite-gain stability with respect to every L^p-norm, $p \in [1, \infty]$, with an arbitrary small bound on the control and large (constant) delay. Moreover, we provide upper bounds for the corresponding L^p-gains which are delay-independent.

Key words. input saturation, time-delay systems, Lyapunov functions, finite-gain stability

AMS subject classifications. 93D25, 34K50, 93D05, 93D15, 34H05

DOI. 10.1137/050626582

1. Introduction.

In this paper, we address the finite-gain L^p-stabilization issue for continuous-time delay linear systems subject to input saturation, i.e., of the type

$$(S) : \dot{x}(t) = Ax(t) - rB\sigma\left(u(t-h)\right), \tag{1}$$

with state $x \in R^n$, input value $u \in R^m$, and where $r \in (0, 1]$ depends only on (S), the constant delay $h \geq 0$ appears in the input, and $\sigma : R^m \to R^m$ is of "saturation" type (definitions are given in section (2)). The present paper is a continuation of [12], where the global asymptotic stabilization issue for systems (S) was successfully addressed. It also has to be noted that the present paper is technically independent of [12].

We use $(S)_h^r$, $r \in (0, 1]$, $h \geq 0$, to denote the control system (S) with input bound r and input time delay h. We omit the index r if it is equal to one; similarly for the index h if it is equal to zero.

In the zero-delay case, the above problem has been widely investigated in the last ten years; see, for example, [4], [6], [1], and references therein, for both continuous-time and discrete-time linear systems. It was shown in [4] ([1], [3], respectively, for the discrete case) that, for neutrally stable open-loop systems, linear feedback laws can be used to achieve finite-gain stability, with respect to every L^p-norm (l_p-norm, respectively). For the continuous case, it is proved in [6] that finite-gain L^p-stability is possible if and only if (S) satisfies

$$(\mathbf{C}) : \begin{cases} \text{(i) all eigenvalues of } A \text{ are in the closed left-half plane,} \\ \\ \text{(ii) the pair } (A, B) \text{ is stabilizable.} \end{cases}$$

An $n \times n$ matrix A is said to be neutrally stable if all solutions of $\dot{x} = Ax$ in the continuous case ($x(t+1) = Ax(t)$ in the discrete case, respectively) are bounded; equivalently, A has no eigenvalues with positive real part and each Jordan block corresponding to a purely imaginary eigenvalue has size one (all the eigenvalues of A

*Received by the editors March 11, 2005; accepted for publication (in revised form) January 28, 2006; published electronically August 25, 2006.
http://www.siam.org/journals/sicon/45-3/62658.html

†Laboratoire des signaux et systèmes, Université Paris-Sud, CNRS, Supélec, 91192 Gif-sur-Yvette cedex, France (karim.yakoubi@lss.supelec.fr, yacine.chitour@lss.supelec.fr).

are inside or on the unit circle, with those on the unit circle having all Jordan blocks of size one, respectively).

In this paper, we focus on the continuous case in the presence of time delay in the input. It is trivial to see that condition (C) is also necessary in the nonzero-delay case and it seems natural to expect condition (C) to also be sufficient.

Our objective here consists of showing that the results of [4] carry over to continuous linear time-delay systems. More specifically, we show that, for neutrally stable continuous linear time-delay systems subject to input saturation, finite-gain L^p-stabilization can be achieved by the use of linear feedbacks for every $p \in [1, \infty]$. The issue of L^p-stabilization for delay systems was already addressed in the literature by a systematic use of linear matrix inequalities (LMIs)(cf. [2], [10]). While many of the arguments of the present paper are conceptually similar to those of [4], there are technical aspects that are different and not obvious. Indeed, as in [4], the proof to get finite-gain L^p-stability relies on passivity techniques. We determine a suitable "storage" function V_p and establish for it a "dissipation inequality" of the form $\frac{dV_p(x_u(t))}{dt} \leq -\|x_u(t)\|^p + \lambda_p \|u(t)\|^p$ for some constant $\lambda_p > 0$ possibly depending on the input bound r and the delay h. For more discussion on passivity, see [11], for instance. Recall that the "storage function" in [4], V_p^0, is nonsmooth. In the present situation, the "storage function" V_p will be the sum of a term similar to V_p^0 and a Lyapunov–Krasovskii functional, in order to take care of the delay. However, unlike in [4], the saturation in (1) needs to be multiplied by a small factor r dependent on the delay h in order to ensure finite-gain L^p-stability.

Moreover, by carefully choosing the factor r and the linear feedback inside the saturation, we are able to provide upper bounds for the L^p-gains of $(S)_h^r$ which are *independent* of $h > 0$. We refer to that property as the *unrestricted finite-gain L^p-stability*. The argument corresponding to that uniformity result is specific to the nonzero-delay case and constitutes the most technical part of the paper. Moreover, that argument requires new technical ingredients with respect to those of [4]. We first start with the single-input case, where it amounts to estimating the behavior of the solution P_r of a parameterized Lyapunov equation (L_r), $r \in (0, 1]$, as the parameter r tends to zero. The multi-input case requires additional work. We first rewrite the original system as an appropriate cascade of single-input subsystems, all of them except one being perturbed by an external disturbance, appearing outside the saturation (see Theorem 2). We then proceed inductively on the number of distinct algebraic multiplicities of the eigenvalues of A.

Generally speaking, our treatment of the aforementioned issues on time-delay systems follows a common pattern. We always try to reformulate them as problems for perturbed *delay-free* systems and handle the perturbation with Lyapunov techniques. That strategy works well because the input saturation makes the perturbation uniformly bounded with respect to the delay.

Organization of the paper. In section 2, we provide the main notation used in the paper and state the theorem relative to the unrestricted finite-gain L^p-stabilization. Section 3 is devoted to the proofs of the main theorems. We start with that of the single-input case. Then, before starting the proof of the multi-input case, we illustrate the cascade argument through the detailed analysis of an eight-dimensional example and then proceed with the general case. Finally, we gather in the appendices the proofs of the two main technical results. The first one describes the asymptotic behavior of the solution of a Lyapunov equation (single-input case) and the second one is devoted to the cascade argument necessary to address the multi-input case.

2. Notation and statements of the main results. For $x \in R^n, \|x\|$ and x^T denote, respectively, the Euclidean norm of x and the transpose of x. Similarly, for any $n \times m$ matrix K, K^T and $\|K\|$ denote, respectively, the transpose of K and the induced 2-norm of K. Moreover, $\lambda_{\min}(K)$ and $\lambda_{\max}(K)$ denote the minimal and the maximal singular eigenvalues of the matrix K. For a square matrix K, we use $\mathrm{Tr}(K)$ to denote the trace of K. If $f(.)$ and $g(.)$ are two real-valued functions, we mean by $f(r) \asymp_0 g(r)$ that there are positive constants ξ_1 and ξ_2 such that for r small enough,

$$\xi_1 g(r) \leq f(r) \leq \xi_2 g(r).$$

For $h > 0$, let $x_t(\theta) := x(t+\theta)$, $-h \leq \theta \leq 0$. Initial conditions for time-delay systems are elements of $C_h := C([-h, 0], R^n)$, the Banach space of continuous R^n-valued functions defined on $[-h, 0]$, equipped with the supremum norm, $\|x_t\|_h = \sup_{-h\leq\theta\leq 0} \|x(t+\theta)\|$.

We next introduce the class of saturation functions to be considered and state the main results on finite-gain stability.

2.1. Saturation functions. We next formally define what we mean by a saturation.

DEFINITION 1. *We call $\sigma : R \longrightarrow R$ a saturation function (S-function) if there exist two real numbers $0 < a \leq K_\sigma$ such that for all $t, t' \in R$,*

(i) $|\sigma(t) - \sigma(t')| \leq K_\sigma \inf(1, |t - t'|)$,

(ii) $|\sigma(t) - at| \leq K_\sigma t\sigma(t)$.

It is assumed here that the function is normalized at the origin, i.e., $a = \sigma'(0) = 1$. A constant K_σ defined as above is called an S-bound for σ. Note that (i) is equivalent to the fact that σ is bounded and globally Lipschitz, while (ii) is equivalent to

$$t\sigma(t) > 0 \text{ for } t \neq 0, \quad \liminf_{|t|\to\infty} |\sigma(t)| > 0, \quad \limsup_{t\to 0} \frac{\sigma(t) - t}{t^2} < \infty.$$

If m is a positive integer, we say that σ is an R^m-valued S-function if $\sigma = (\sigma_1, \ldots, \sigma_m)^T$, where each component σ_i is an S-function and, for $s = (s_1, \ldots, s_m)^T \in R^m$,

$$\sigma(s) := \Big(\sigma_1(s_1), \sigma_2(s_2), \ldots, \sigma_m(s_m)\Big)^T.$$

2.2. L^p-stability. For $p \in [1, \infty]$ and $0 \leq h$ (and $T \geq h$, respectively), we use L^p to denote $L^p(-h, \infty)$ and we let $\|y\|_{L^p}$ ($\|y\|_{L^p}([-h, T])$, respectively) denote the L^p-norm

$$\|y\|_{L^p} = \left(\int_{-h}^{\infty} \|y(t)\|^p dt\right)^{\frac{1}{p}} \quad \left(\|y\|_{L^p([-h,T])} = \left(\int_{-h}^{T} \|y(t)\|^p dt\right)^{\frac{1}{p}}, \quad \text{respectively}\right)$$

if $p < \infty$ and

$$\|y\|_{L^\infty} = ess \sup_{-h\leq t<\infty} \|y(t)\|, \quad \left(\|y\|_{L^\infty([-h,T])} = ess \sup_{-h\leq t\leq T} \|y(t)\|, \quad \text{respectively}\right).$$

Consider the control system with delay in the input given by

$$(\Sigma)_h : \dot{x} = f\left(x(t), u(t-h)\right), \quad t \geq 0, \tag{2}$$

where the state x and the control u take, respectively, values in R^n and R^m and $f : R^n \times R^m \to R^n$ is locally Lipschitz in (x, u), with $f(0, 0) = 0$. Trajectories of $(\Sigma)_h$ starting at $x_0 \in C_h$ and corresponding to an input $u \in L^p$ are defined for a time interval I of R^+ (depending on x_0 and u) and verify (2) for a.e. $t \in I$. Let $\bar{0}$ be the zero function in C_h.

DEFINITION 2 (L^p-stability). *Given $p \in [1, \infty]$, the continuous-time delay system $(\Sigma)_h$ is said to be L^p-stable if, for every $u \in L^p$, we have $x_u \in L^p$, where x_u denotes the solution of $(\Sigma)_h$ corresponding to u with initial condition $x_0 = \bar{0}$.*

DEFINITION 3 (finite-gain L^p-stability). *Given $p \in [1, \infty]$, the continuous-time delay system $(\Sigma)_h$ is said to be finite-gain L^p-stable if it is L^p-stable, and there exists a positive constant M_p such that, for every $u \in L^p$,*

$$\|x_u\|_{L^p} \leq M_p \|u\|_{L^p}.$$

Furthermore, the infimum over such numbers M_p will be called the L^p-gain of the system.

We now state our main results.

THEOREM 1. *Let A, B be $n \times n, n \times m$ matrices, respectively. Let σ be an R^m-valued S-function. Assume that A is neutrally stable and (A, B) is controllable. Then, for every $h \geq 0$, there exists an $n \times m$ matrix F_h such that the system*

$$(S)_h^r : \dot{x} = Ax - rB\sigma\left(F_h^T x(t-h) + u(t-h)\right) \text{ for } t \geq 0$$

has the unrestricted finite-gain L^p-stability property for every $p \in [1, \infty]$; i.e., for every $h > 0$, there exists $r^*(h) \in (0, 1]$ such that for every $p \in [1, \infty]$, $(S)_h^r$, $r \in (0, r^*(h)]$, is finite-gain L^p-stable.

Remark 1. In the absence of u, the equilibrium point $\bar{0}$ is globally asymptotically stable for the delayed system $\dot{x} = Ax - rB\sigma\left(F_h^T x(t-h)\right)$.

Theorem 1 is a particular case of a stronger result given next.

THEOREM 2. *With the same hypothesis on A, B, and σ, consider the delayed system (still denoted $(S)_h^r$)*

$$(S)_h^r : \dot{x} = Ax - rB\sigma\left(F_h^T x(t-h) + u_1(t-h)\right) + ru_2(t-h) \text{ for } t \geq 0,$$

where F_h is defined as in Theorem 1 *and the input u_2 takes values in R^n. Then, there exist a constant $C_0 > 0$ and, for every $1 \leq p \leq \infty$, a constant $M_p > 0$ such that for every $h > 0$ there is an $r^*(h) \in (0, 1]$ for which the trajectories x_{u_1,u_2} of $(S)_h^r$, $r \in (0, r^*(h)]$, starting at $\bar{0}$ and corresponding to $u_1, u_2 \in L^p$ with $\|u_2\|_{L^\infty} \leq C_0$, verify*

$$\|x_{u_1,u_2}\|_{L^p} \leq M_p(\|u_1\|_{L^p} + \|u_2\|_{L^p}).$$

The proof of Theorem 2 is deferred to section 3.

Remark 2. It will be clear from our argument that we can in fact obtain the stronger input-to-state stable (ISS)-like property (see [7] and references therein)

$$\|x^{\psi}_{u_1,u_2}\|_{L^p} \leq \theta_p(\|\psi\|_h) + M_p(\|u_1\|_{L^p} + \|u_2\|_{L^p}),$$

where $\psi \in C_h$ is the initial condition for the trajectory $x^{\psi}_{u_1,u_2}$ corresponding to u_1, u_2 and θ_p is a $\mathcal{K}$-function (i.e., $\theta_p : R^+ \to R^+$ is continuous and strictly increasing and satisfies $\theta_p(0) = 0$).

3. Proof of Theorem 2. From elementary linear algebra, a neutrally stable matrix A is similar to a matrix

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}, \tag{3}$$

where A_1 is a $q \times q$ Hurwitz matrix and A_2 is an $(n-q) \times (n-q)$ skew-symmetric matrix. So, up to a change of coordinates, we may assume that A is already in the form (3). In these coordinates, we write $B = (B_1^T \quad B_2^T)^T$, where B_2 is an $(n-q) \times m$ matrix, and we write vectors as $x = (x_1^T, x_2^T)^T$ and $u_2 = (u_{21}^T, u_{22}^T)^T$.

For $r \in (0,1]$ and $h > 0$, consider the feedback law $(0, F_h^T)$. Then system $(S)_h^r$, with this choice of F_h^T, can be written as

$$\dot{x}_1(t) = A_1 x_1(t) - rB_1\sigma[F_h^T x_2(t-h) + u_1(t-h)] + ru_{21}(t-h),$$

$$\dot{x}_2(t) = A_2 x_2(t) - rB_2\sigma\left(F_h^T x_2(t-h) + u_1(t-h)\right) + ru_{22}(t-h).$$

Since A_1 is Hurwitz, it will be sufficient to show that there exists an $r^*(h) \in (0,1]$, such that the x_2-subsystem is finite-gain L^p-stable for all $r \in (0, r^*(h)]$.

The controllability assumption on (A,B) implies that the pair (A_2, B_2) is also controllable. Therefore, the theorem is a consequence of the following proposition.

PROPOSITION 1. *Let σ, u_1, u_2 be as in Theorem 2. Let (A,B) be a controllable pair with A skew-symmetric. Then, for every $h \geq 0$, there exist an $n \times m$ matrix F_h and $r^*(h) \in (0,1]$ such that, for every $r \in (0, r^*(h)]$, the system*

$$(S)_h^r : \dot{x}(t) = Ax(t) - rB\sigma\left(F_h^T x(t-h) + u_1(t-h)\right) + ru_2(t-h) \text{ for } t \geq 0$$

verifies the conclusion of Theorem 2.

Remark 3. In the single-input case ($m = 1$), F_h can be chosen as $e^{-Ah}B$, which corresponds, up to the delay h, to the linear feedback law suggested by the passivity approach and used in [4]. A simple adaptation of the proof to the multi-input case shows that such a feedback can also be used to get finite-gain L^p-stability but the corresponding L^p-gain is, in general, delay-dependent.

Proof. We start the proof by zooming in on the single-input case. The general proof first starts with algebraic transformations and proceeds by induction on the number of distinct algebraic multiplicities of the eigenvalues of A.

3.1. The single-input case. Let $h > 0$ and consider the feedback law $F_h = e^{-Ah}B = e^{-Ah}b$. Then, $\|F_h\| = \|b\|$. Let y be the solution of

$$\begin{aligned} &\dot{y}(t) = (A - rbb^T)y(t) + ru_2(t-h) \text{ for } t \geq 0, \\ &y_0 = \bar{0} \quad \text{ on } [-h, 0], \end{aligned} \tag{4}$$

with $u_2 \in L^p$ and $\|u_2\|_{L^\infty} \leq C_0$, where C_0 is a positive constant to be determined later. (By an obvious abuse of notation, L^p denotes here $L^p([-h,\infty), R^n)$.) Since A is skew-symmetric, the matrix $A_r := A - rbb^T$ is Hurwitz for every $r > 0$. Then (4) is L^p-stable for any $1 \leq p \leq \infty$ (see [4, Remark 4]). Let γ_p be its L^p-gain, so $\|y\|_{L^p} \leq \gamma_p \|u_2\|_{L^p}$. We need the next lemma, whose proof is given in Appendix A. □

LEMMA 1. *Let A and b be as in Proposition 1. Then, the following properties hold.*

(i) *There exist an* $r^* \in (0,1]$ *and positive constants* C_1, C_1', C_2, *and* C_2' *such that, for all* $t \geq 0$ *and* $r \in (0, r^*]$,

$$C_1' e^{-C_2' r t} \leq ||e^{(A - r b b^T)t}|| \leq C_1 e^{-C_2 r t}. \tag{5}$$

(ii) *For* $r \geq 0$, *let* P_r *be the unique positive-definite solution to the Lyapunov equation*

$$(L_r): \quad P_r(A - rbb^T) + (A - rbb^T)^T P_r = -I_n. \tag{6}$$

Then,

$$\lambda_{\max}(P_r) \asymp_0 \lambda_{\min}(P_r) \asymp_0 \frac{1}{r}. \tag{7}$$

We also need the following lemma.

LEMMA 2. *Let* γ_p *be the* L^p*-gain of* (4). *So,* $\gamma_p \leq \frac{C_1}{C_2}$.

Proof. For $t \geq 0$, set $A_r := A - rbb^T$. Then, $y(t) = r \int_0^t e^{A_r(t-s)} u_2(s-h) ds$. From (5), we have

$$\begin{aligned} \|y(t)\| &\leq r \int_0^t \|e^{A_r(t-s)}\| \|u_2(s-h)\| ds \\ &\leq r C_1 \int_0^t e^{-rC_2(t-s)} \|u_2(s-h)\| ds. \end{aligned} \tag{8}$$

Let G_p be the L^p-gain of the control system with state $\chi \in R$ given by

$$(S_\chi): \dot{\chi}(t) = -rC_2 \chi(t) + rC_1 v(t-h),$$

where $v \in L^p([-h, \infty], R)$. It is clear from (8) that $\gamma_p \leq G_p$. So, for each $1 < p < \infty$, we have

$$|\chi(t)|^{p-2} \chi(t) \dot{\chi}(t) = -rC_2 |\chi(t)|^p + rC_1 |\chi(t)|^{p-2} \chi(t) v(t-h).$$

For every $t \geq 0$, integrating the last equation from zero to ∞ and using Hölder's inequality (with $p' = \frac{p}{p-1}$ and $q' = p$), we have

$$rC_2 \|\chi\|_{L^p}^p \leq rC_1 \|\chi\|_{L^p}^{p-1} \|v\|_{L^p},$$

because $\chi(t) \to 0$ if $t \to \infty$ (see [4, Lemma 4]). Then $\|\chi\|_{L^p} \leq \frac{C_1}{C_2} \|v\|_{L^p}$. For $p = 1$, if we integrate (S_χ) from zero to ∞, we have the result. It remains to study the case $p = \infty$. We have

$$\|y(t)\| \leq rC_1 \|v\|_{L^\infty} \int_0^\infty e^{-rC_2(t-s)} ds \leq \frac{C_1}{C_2} \|v\|_{L^\infty}. \qquad \square$$

Let x be the solution of $(S)_h^r$ (given in Proposition 1) starting at $\bar{0} \in C_h$ and corresponding to u_1, u_2. Set $z := x - y$. Then, z satisfies

$$\dot{z}(t) = Az(t) - rb\Big[\sigma\Big(F_h^T z(t-h) + F_h^T y(t-h) + u_1(t-h)\Big) - b^T y(t)\Big] \text{ for } t \geq 0, \tag{9}$$

$$z_0 = \bar{0}, \quad \text{on } [-h, 0].$$

Let $\tilde{u}(t-h) = F_h^T y(t-h) + u_1(t-h)$ and $\tilde{v}(t) = b^T y(t)$. Choosing the constant $C_0 > 0$ such that $\gamma_\infty \|b\| C_0 \leq \Gamma$, with $\Gamma = \lim_{|\xi|\to\infty} \inf \sigma(\xi) > 0$, we get

$$\|\tilde{v}\|_{L^\infty} \leq \|b\| \|y\|_{L^\infty} \leq \gamma_\infty \|b\| \|u_2\|_{L^\infty} \leq \gamma_\infty \|b\| C_0 \leq \Gamma. \tag{10}$$

Equation (9) can be written as

$$\begin{aligned} &\dot{z}(t) = Az(t) - rb\Big[\sigma\Big(F_h^T z(t-h) + \tilde{u}(t-h)\Big) - \tilde{v}(t)\Big] \text{ for } t \geq 0, \\ &z_0 = \bar{0}, \quad \text{on } [-h, 0]. \end{aligned} \tag{11}$$

From (11), we have

$$z(t) = e^{Ah} z(t-h) - r \int_{t-h}^{t} e^{A(t-\xi)} b\Big[\sigma\Big(F_h^T z(\xi - h) + \tilde{u}(\xi - h)\Big) - \tilde{v}(\xi)\Big] d\xi.$$

Then, $F_h^T z(t-h) + \tilde{u}(t-h) = b^T z(t) + \tilde{d}(t)$, where

$$\tilde{d}(t) = \tilde{u}(t-h) + r \textstyle\int_{t-h}^{t} g(t-\xi) \left[\sigma\left(F_h^T z(\xi - h) + \tilde{u}(\xi - h)\right) - \tilde{v}(\xi)\right] d\xi,$$

$$g(v) = b^T e^{Av} b.$$

According to (11), we obtain

$$\begin{aligned} &\dot{z}(t) = Az(t) - rb\Big[\sigma\Big(b^T z(t) + \tilde{d}(t)\Big) - \tilde{v}(t)\Big] \text{ for } t \geq 0. \\ &z_0 = \bar{0}, \quad \text{on } [-h, 0]. \end{aligned} \tag{12}$$

Let $\tilde{z}(t) = b^T z(t) + \tilde{d}(t)$. As in [4], we consider, for each $1 \leq p < \infty$, the function $V_{0,p} : R^n \longrightarrow R$ given by

$$V_{0,p}(z(t)) = \frac{\|z(t)\|^{p+1}}{p+1}. \tag{13}$$

We next majorize the time derivative of $V_{0,p}$ along the trajectories of (12). We follow the computations leading to (22) of [4] and obtain

$$\begin{aligned} \dot{V}_{0,p}(z(t)) \quad &\leq \quad -r\|z(t)\|^{p-1} \tilde{z}^T(t) \sigma(\tilde{z}(t)) \\ &\quad + r\|z(t)\|^{p-1} \Big((K + \Gamma)\|\tilde{d}(t)\| + M_1 \|\tilde{v}(t)\|\Big), \end{aligned} \tag{14}$$

where, by definition of Γ, there is some $M_1 \geq 1$ so that $\inf_{|\xi| \geq M_1} |\sigma(\xi)| \geq \frac{1}{2}\Gamma$.

Since A_r is Hurwitz, we consider for each $1 < p < \infty$ the differentiable function $V_{1,p} : R^n \to R$ defined by

$$V_{1,p}(z(t)) = (z^T(t) P_r z(t))^{\frac{p}{2}}, \tag{15}$$

where P_r is the solution of (6). We rewrite (12) in the form

$$\begin{aligned} &\dot{z}(t) = A_r z(t) - rb\left[\sigma(\tilde{z}(t)) - \tilde{z}(t) + \tilde{d}(t) - \tilde{v}(t)\right] \text{ for } t \geq 0, \\ &z_0 = \bar{0}, \quad \text{on } [-h, 0]. \end{aligned} \tag{16}$$

The derivative of $V_{1,p}$ along the trajectories of (16) verifies

$$\begin{aligned}
\dot{V}_{1,p} &= p\Big(z^T(t)P_rz(t)\Big)^{\frac{p}{2}-1}z^T(t)P_r\dot{z}(t)\\
&= p\Big(z^T(t)P_rz(t)\Big)^{\frac{p}{2}-1}\Big\{z^T(t)\Big(-I-A_r^TP_r\Big)z(t)-rz^T(t)P_rb\Big[\sigma(\tilde{z}(t))-\tilde{z}(t)\\
&\qquad\qquad +\tilde{d}(t)-\tilde{v}(t)\Big]\Big\}\\
&\leq -p\Big(z^T(t)P_rz(t)\Big)^{\frac{p}{2}-1}\|z(t)\|^2\\
&\quad +p\Big(z^T(t)P_rz(t)\Big)^{\frac{p}{2}-1}rK\|P_rb\|\|z(t)\|\Big\{\tilde{z}^T(t)\sigma(\tilde{z}(t))+\tfrac{1}{K}\Big(\|\tilde{d}(t)\|+\|\tilde{v}(t)\|\Big)\Big\}\\
&\leq \begin{cases}
-\dfrac{p}{\Big(\lambda_{\max}(P_r)\Big)^{1-\frac{p}{2}}}\|z(t)\|^p+\dfrac{rpK\|P_rb\|}{\Big(\lambda_{\min}(P_r)\Big)^{1-\frac{p}{2}}}\|z(t)\|^{p-1}\tilde{z}^T(t)\sigma(\tilde{z}(t))\\
+\dfrac{rp\|P_rb\|}{\Big(\lambda_{\min}(P_r)\Big)^{1-\frac{p}{2}}}\|z(t)\|^{p-1}\Big(\|\tilde{d}(t)\|+\|\tilde{v}(t)\|\Big) \qquad \text{if } 1<p\leq 2,\\
-p\Big(\lambda_{\min}(P_r)\Big)^{\frac{p}{2}-1}\|z(t)\|^p+rpK\|P_rb\|\Big(\lambda_{\max}(P_r)\Big)^{\frac{p}{2}-1}\|z(t)\|^{p-1}\tilde{z}^T(t)\sigma(\tilde{z}(t))\\
+rp\|P_rb\|\Big(\lambda_{\max}(P_r)\Big)^{\frac{p}{2}-1}\|z(t)\|^{p-1}\Big(\|\tilde{d}(t)\|+\|\tilde{v}(t)\|\Big) \quad \text{if } 2\leq p<\infty.
\end{cases}
\end{aligned}$$

For $1\leq p<\infty$, set

$$\lambda_{p,r}:=\begin{cases}
\dfrac{pK\|P_rb\|}{\Big(\lambda_{\min}(P_r)\Big)^{1-\frac{p}{2}}} & \text{if } 1\leq p\leq 2,\\
pK\|P_rb\|\Big(\lambda_{\max}(P_r)\Big)^{\frac{p}{2}-1} & \text{if } 2\leq p<\infty.
\end{cases}$$

Finally, consider, for each $1\leq p<\infty$, the following function:

$$V_{2,p}(z(t)):=\Big(\lambda_{p,r}V_{0,p}+V_{1,p}\Big)(z(t)). \tag{17}$$

Along trajectories of (12), we obtain, for $1<p<\infty$,

$$\dot{V}_{2,p}\leq -\zeta_{p,r}\|z(t)\|^p+\kappa_{p,r}\|z(t)\|^{p-1}\Big(\|\tilde{d}(t)\|+\|\tilde{v}(t)\|\Big), \tag{18}$$

where

$$\begin{cases}
\zeta_{p,r}=\begin{cases}
\dfrac{p}{(\lambda_{\max}(P_r))^{1-\frac{p}{2}}} & \text{if } 1<p\leq 2,\\
p(\lambda_{\min}(P_r))^{\frac{p}{2}-1} & \text{if } 2\leq p<\infty,
\end{cases}\\
\kappa_{p,r}=r\lambda_{p,r}\max\{1+K+\Gamma,1+M_1\}.
\end{cases}$$

There exists $T > h$ such that the differentiable function $V_{1,p}$ verifies $\|DV_{1,p}(z)\| \leq c\lambda_{p,r}\|z\|^{p-1}$ for all $z \in R^n$ and $t \geq T$, where c is some positive constant independent of p. Moreover, $V_{1,p}$ can be chosen so that $\limsup_{p\to 1^+} \lambda_{p,r} = \lambda_{1,r} < \infty$, and the limit $V_{1,1}(z) = \lim_{p\to 1^+} V_{1,p}(z)$ exists for all $z \in R^n$.

By the Lebesgue dominated convergence theorem (applied to a sequence $\{p^j\}_{j=1}^{\infty}$ decreasing to 1) we know that the above inequality is also true for $p = 1$.

Consider the Lyapunov–Krasovskii functional, given by

$$H_\mu(t) = \mu \int_{t-2h}^{t} \left(\int_s^t \|z(l)\|^p dl \right) ds = \mu \int_0^{2h} \left(\int_{t-s}^t \|z(l)\|^p dl \right) ds \tag{19}$$

for $t \geq 2h$, where $\mu > 0$ will be chosen later. Its derivative is equal to

$$\dot{H}_\mu(t) = 2h\mu\|z(t)\|^p - \mu \int_{t-2h}^{t} \|z(s)\|^p ds. \tag{20}$$

Finally, consider the following function:

$$V_p(z(t)) := V_{2,p}(z(t)) + H_\mu(t). \tag{21}$$

Along the trajectories of (12), from (18) and (20), we have

$$\begin{array}{rcl} \dot{V}_p(z(t)) & \leq & -(\zeta_{p,r} - 2h\mu)\|z(t)\|^p + \kappa_{p,r}\|z(t)\|^{p-1}(\|\tilde{d}(t)\| \\ & + & \|\tilde{v}(t)\|) - \mu \int_{t-2h}^{t} \|z(s)\|^p ds. \end{array} \tag{22}$$

From the definition of $\tilde{d}(t)$ and Hölder's inequality, it follows that

$$\begin{array}{rcl} \|\tilde{d}(t)\| + \|\tilde{v}(t)\| & \leq & \|\tilde{u}(t-h)\| + \|\tilde{v}(t)\| + r\max\{1,K\}\|b\|^2 \int_{t-h}^{t} \Big(\|\tilde{u}(\xi - h)\| \\ & + & \|\tilde{v}(\xi)\| \Big) d\xi + rK\|b\|^3 (2h)^{\frac{p-1}{p}} \left(\int_{t-2h}^{t} \|z(s)\|^p ds \right)^{\frac{1}{p}}. \end{array}$$

From (22) and the Hölder inequality (with $p' = \frac{p}{p-1}$ and $q' = p$) we have

$$\begin{array}{rcl} \dot{V}_p(z(t)) & \leq & -(\zeta_{p,r} - 2h\mu)\|z(t)\|^p + \kappa_{p,r}\|z(t)\|^{p-1} \left(\|\tilde{u}(t-h)\| + \|\tilde{v}(t)\|\right) \\ & + & r\max\{1,K\}\|b\|^2\kappa_{p,r}\|z(t)\|^{p-1} \int_{t-h}^{t} \left(\|\tilde{u}(\xi - h)\| + \|\tilde{v}(\xi)\|\right) d\xi \\ & + & rK\|b\|^3\kappa_{p,r}(2h)^{\frac{p-1}{p}} \|z(t)\|^{p-1} \left(\int_{t-2h}^{t} \|z(s)\|^p ds \right)^{\frac{1}{p}} - \mu \int_{t-2h}^{t} \|z(s)\|^p ds \\ & \leq & -\left(\zeta_{p,r} - 2^{\frac{2p-1}{p}} h\mu\right) \|z(t)\|^p + \kappa_{p,r}\|z(t)\|^{p-1} \left(\|\tilde{u}(t-h)\| + \|\tilde{v}(t)\|\right) \\ & + & r\max\{1,K\}\|b\|^2\kappa_{p,r}\|z(t)\|^{p-1} \int_{t-h}^{t} \left(\|\tilde{u}(\xi - h)\| + \|\tilde{v}(\xi)\|\right) d\xi \\ & - & \left(\mu - \frac{1}{p}\mu^{1-p} r^p \kappa_{p,r}^p K^p \|b\|^{3p}\right) \int_{t-2h}^{t} \|z(s)\|^p ds. \end{array}$$

We choose $\mu^*(h)$ and $r^*(h) \in (0,1]$ such that

$$\begin{aligned} \mu^*(h) &= \min\left\{\tfrac{1}{4h}\tfrac{p}{2p-1}\zeta_{p,r}, \tfrac{K\kappa_{p,r}\|b\|^3}{p^{\frac{1}{p}}}\right\}, \\ r^*(h) &= \min\left\{1, \tfrac{p^{\frac{1}{p}+1}}{K(2p-1)\|b\|^3}\tfrac{1}{4h}\tfrac{\zeta_{p,r}}{\kappa_{p,r}}\right\}, \end{aligned} \tag{23}$$

since $\frac{\zeta_{p,r}}{\kappa_{p,r}} \asymp_0 O(1)$. So, for $\mu \le \mu^*(h)$ and $r \le r^*(h)$, we get

$$\begin{aligned} \dot V_p(z(t)) &\le -\tfrac{1}{2}\zeta_{p,r}\|z(t)\|^p + \kappa_{p,r}\|z(t)\|^{p-1}\Big(\|\tilde u(t-h)\| + \|\tilde v(t)\|\Big) \\ &+ r\max\{1,K\}\|b\|^2\kappa_{p,r}\|z(t)\|^{p-1}\int_{t-h}^{t}\Big(\|\tilde u(\xi-h)\| + \|\tilde v(\xi)\|\Big)d\xi \\ &:= -\tfrac{1}{2}\zeta_{p,r}\|z(t)\|^p + \kappa_{p,r}\|z(t)\|^{p-1} \\ &\times\Big[\|\tilde u(t-h)\| + \|\tilde v(t)\| + r\max\{1,K\}\|b\|^2\left(\tilde m(t) + \tilde n(t)\right)\Big], \end{aligned} \tag{24}$$

where $\tilde m(t) = \int_{t-h}^{t}\|\tilde u(\xi-h)\|d\xi$ and $\tilde n(t) = \int_{t-h}^{t}\|\tilde v(\xi)\|d\xi$. A simple computation shows that

$$\|\tilde m\|_{L^p} \le h\|\tilde u\|_{L^p} \text{ and } \|\tilde n\|_{L^p} \le h\|\tilde v\|_{L^p}. \tag{25}$$

For every $t \ge 0$, integrating (24) from 0 to t, we have

$$\begin{aligned} V_p(z(t)) + \tfrac{1}{2}\zeta_{p,r}\int_0^t\|z(s)\|^p ds &\le \kappa_{p,r}\int_0^t\|z(s)\|^{p-1}[\|\tilde u(s-h)\| + \|\tilde v(s)\| \\ &+ r\max\{1,K\}\|b\|^2(\tilde m(s) + \tilde n(s))]ds. \end{aligned}$$

Applying Hölder's inequality to

$$\int_0^t\|z(s)\|^{p-1}\left[\|\tilde u(s-h)\| + \|\tilde v(s)\| + r\max\{1,K\}\|b\|^2\left(\tilde m(s) + \tilde n(s)\right)\right]ds,$$

we conclude that, for all $1 < p < \infty$ and $t \ge 0$,

$$\begin{aligned} V_p(z(t)) + \tfrac{1}{2}\zeta_{p,r}\|z\|^p_{L^p[0,t]} &\le \kappa_{p,r}\|z\|^{p-1}_{L^p[0,t]}[\|\tilde u\|_{L^p} + \|\tilde v\|_{L^p} \\ &+ r\max\{1,K\}\|b\|^2(\|\tilde m\|_{L^p} + \|\tilde n\|_{L^p})]. \end{aligned} \tag{26}$$

Since $V_p \ge 0$, we get that $z \in L^p([0,\infty), R^n)$ and, by using (25),

$$\|z\|_{L^p} \le \frac{2\kappa_{p,r}}{\zeta_{p,r}}\left(1 + rh\max\{1,K\}\|b\|^2\right)\left(\|\tilde u\|_{L^p} + \|\tilde v\|_{L^p}\right). \tag{27}$$

Since $z = x - y, \tilde u = F_h^T y + u_1, \tilde v = b^T y$, we have

$$\|\tilde v\|_{L^p} \le \|b\|\|y\|_{L^p} \le \gamma_p\|b\|\|u_2\|_{L^p},$$

$$\|\tilde u\|_{L^p} \le \|u_1\|_{L^p} + \gamma_p\|b\|\|u_2\|_{L^p},$$

$$\|z\|_{L^p} \ge \|x\|_{L^p} - \|y\|_{L^p} \ge \|x\|_{L^p} - \gamma_p\|u_2\|_{L^p}.$$

Combining this with (27) we have

$$\begin{aligned} \|x\|_{L^p} &\leq 2\tfrac{\kappa_{p,r}}{\zeta_{p,r}}(1+rh\max\{1,K\}\|b\|^2)\|u_1\|_{L^p} \\ &+ \gamma_p\left[1+4\|b\|\tfrac{\kappa_{p,r}}{\zeta_{p,r}}(1+rh\max\{1,K\}\|b\|^2)\right]\|u_2\|_{L^p} \\ &\leq M_p(\|u_1\|_{L^p}+\|u_2\|_{L^p}), \end{aligned} \tag{28}$$

where

$$\begin{aligned} M_p &= \max\Big\{2\tfrac{\kappa_{p,r}}{\zeta_{p,r}}\Big(1+rh\max\{1,K\}\|b\|^2\Big), \\ &\qquad \gamma_p\Big[1+4\|b\|\tfrac{\kappa_{p,r}}{\zeta_{p,r}}(1+rh\max\{1,K\}\|b\|^2)\Big]\Big\}. \end{aligned} \tag{29}$$

It remains to show the independence of the upper bound for M_p with respect to the parameter h. Multiplying both sides of (6) by P_r on the right and taking the trace, one gets

$$\mathrm{Tr}(P_r) = 2r||P_rb||^2.$$

Since $\mathrm{Tr}(P_r) \asymp_0 \lambda_{\max}(P_r)$, by Lemma 1, $||P_rb|| \asymp_0 \frac{1}{r}$. In addition,

$$M_p \asymp_0 \frac{\kappa_{p,r}}{\zeta_{p,r}} = Kr||P_rb|| \left(\frac{\lambda_{\max}(P_r)}{\lambda_{\min}(P_r)}\right)^{|\frac{p}{2}-1|} \asymp_0 O(1). \tag{30}$$

Then, the upper bound for the finite-gain is independent of h for all $p \in [1,\infty)$.

In the case $p=\infty$, finite-gain stabilization was already proved in Lemma 4 of [12] for $n=2$ (just take $f := (\alpha,\beta)^T = 0$ in that lemma). That argument trivially extends to every $n \geq 2$.

Remark 4. In the next subsection, we will proceed with the argument of the multi-input case. Before doing so, we can remark that the feedback $F_h = e^{-hA}B$ L^p-stabilizes the system $(S)_h^r$. However, the L^p-gain depends on the delay. To see that, we simply follow the previous argument. The difference comes up with (7). In this case, there are n eigenvalues of $A_r = A - rbb^T$, $\lambda_1(r),\dots,\lambda_n(r)$, defining continuous functions, which are not analytic in general. These functions, though, can be written as Puiseux series (cf. [5]),

$$\lambda_i(r) = \lambda_i(0) + \sum_{j=1}^{\infty} \alpha_j^{(i)} r^{\frac{j}{p_i}},$$

where $\lambda_i(0)$ is a root of multiplicity ξ of A and p_i is a positive integer eventually larger than one. Therefore, by (30), it is clear that M_p cannot be delay-independent.

3.2. The general case. First, we present an example to clarify the proof of Proposition 1 in the general case, i.e., the multi-input case. The procedure for computing stabilizing feedbacks brings the matrices A and B into a special form.

Example. Consider the following system where eight oscillators are coupled:

$$(S)_h^r : \dot{x}(t) = Ax(t) - rB\sigma(u(t-h)) + rv(t-h), x \in R^{16}, u \in R^3, v \in R^{16}, \tag{31}$$

where $A = \text{ diag } \left(\omega_1 A^{(1)}, \omega_2 A^{(2)}, \omega_3 A^{(2)}, \omega_4 A^{(3)}\right)$, $A^{(i)} = \text{ diag } (A_0)$ repeated i times, $i \in \{1,2,3\}$, $A_0 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ $\omega_i \neq \omega_j$ for $i \neq j$, $\omega_i > 0$ for all $i = 1, \ldots, 4$, and

$$B = \begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} b^1_{1,3} & b^2_{1,3} & b^3_{1,3} \end{pmatrix} \\ \begin{pmatrix} b^1_{1,2} & b^2_{1,2} & b^3_{1,2} \\ b^1_{2,2} & b^2_{2,2} & b^3_{2,2} \end{pmatrix} \\ \begin{pmatrix} \tilde{b}^1_{1,2} & \tilde{b}^2_{1,2} & \tilde{b}^3_{1,2} \\ \tilde{b}^1_{2,2} & \tilde{b}^2_{2,2} & \tilde{b}^3_{2,2} \end{pmatrix} \\ \begin{pmatrix} b^1_{1,1} & b^2_{1,1} & b^3_{1,1} \\ b^1_{2,1} & b^2_{2,1} & b^3_{2,1} \\ b^1_{3,1} & b^2_{3,1} & b^3_{3,1} \end{pmatrix} \end{pmatrix}.$$

In order to prove the result of finite-gain L^p-stabilization for all $p \in [1, \infty]$ of the system $(S)^r_h$, we rewrite the original system as an appropriate cascade of single-input subsystems. Then, there exists a linear transformation, given below for the general case, that puts (31) in the special form, with $b_0 = (0 \quad 1)^T$:

(32)

$$\begin{cases} (1) \begin{cases} \dot{Y}_{1,1} &= \omega_1 A_0 Y_{1,1} - r b_0 \sigma_1(u_1(t-h)) - r\hat{b}^2_{1,3}\sigma_2(u_2(t-h)) \\ &- r\hat{b}^3_{1,3}\sigma_3(u_3(t-h)) + r v_{1,1}(t-h), \end{cases} \\ (2) \begin{cases} \dot{Y}_{1,2} = \omega_2 A_0 Y_{1,2} - r b_0 \sigma_1(u_1(t-h)) - r\hat{b}^3_{1,2}\sigma_3(u_3(t-h)) + r v_{1,2}(t-h), \\ \dot{Y}_{2,2} = \omega_2 A_0 Y_{2,2} - r b_0 \sigma_2(u_2(t-h)) - r\hat{b}^3_{2,2}\sigma_3(u_3(t-h)) + r v_{2,2}(t-h), \end{cases} \\ (3) \begin{cases} \dot{\tilde{Y}}_{1,2} = \omega_3 A_0 \tilde{Y}_{1,2} - r b_0 \sigma_1(u_1(t-h)) - r\hat{\tilde{b}}^3_{1,2}\sigma_3(u_3(t-h)) + r\tilde{v}_{1,2}(t-h), \\ \dot{\tilde{Y}}_{2,2} = \omega_3 A_0 \tilde{Y}_{2,2} - r b_0 \sigma_2(u_2(t-h)) - r\hat{\tilde{b}}^3_{2,2}\sigma_3(u_3(t-h)) + r\tilde{v}_{2,2}(t-h), \end{cases} \\ (4) \begin{cases} \dot{Y}_{1,3} = \omega_4 A_0 Y_{1,3} - r b_0 \sigma_1(u_1(t-h)) + r v_{1,3}(t-h), \\ \dot{Y}_{2,3} = \omega_4 A_0 Y_{2,3} - r b_0 \sigma_2(u_2(t-h)) + r v_{2,3}(t-h), \\ \dot{Y}_{3,3} = \omega_4 A_0 Y_{3,3} - r b_0 \sigma_3(u_3(t-h)) + r v_{3,3}(t-h). \end{cases} \end{cases}$$

Now, we transform the system $(S)^r_h$ in some appropriate cascade of single-input subsystems $(S_i)^r_h$, $i \geq 1$, where the number of subsystems is equal to the number of distinct algebraic multiplicities corresponding to the eigenvalues of A, and all of them

except one are perturbed by an external disturbance appearing outside the saturation and of sufficiently small amplitudes. These perturbations are characterized by external exogenous signals which are proper perturbations and by other ones coming from the subsystems $(S_j)_h^r$, $j > i$. So, we first rearrange the order of equations defined by $Z_1 = Y_{1,1}, Z_2 = Y_{1,2}, Z_3 = \tilde{Y}_{1,2}, Z_4 = Y_{1,3}, Z_5 = Y_{2,2}, Z_6 = \tilde{Y}_{2,2}, Z_7 = Y_{2,3}$, and $Z_8 = Y_{3,3}$, for which the system (32) transforms into

$$
\left\{
\begin{array}{l}
(S_1)_h^{r_1}\left\{
\begin{array}{rcl}
\dot{Z}_1(t) &=& \omega_1 A_0 Z_1(t) - r_1 b_0 \sigma_1(u_1(t-h)) + r_1 v_1(t-h) \\
&-& r_2 \hat{b}^2_{1,3} \sigma_2(u_2(t-h)) - r_3 \hat{b}^3_{1,3} \sigma_3(u_3(t-h)), \\
\dot{Z}_2(t) &=& \omega_2 A_0 Z_2(t) - r_1 b_0 \sigma_1(u_1(t-h)) + r_1 v_2(t-h) \\
&-& r_3 \hat{b}^3_{1,2} \sigma_3(u_3(t-h)), \\
\dot{Z}_3(t) &=& \omega_3 A_0 Z_3(t) - r_1 b_0 \sigma_1(u_1(t-h)) + r_1 v_3(t-h) \\
&-& r_3 \hat{\tilde{b}}^3_{1,2} \sigma_3(u_3(t-h)), \\
\dot{Z}_4(t) &=& \omega_4 A_0 Z_4(t) - r_1 b_0 \sigma_1(u_1(t-h)) + r_1 v_4(t-h),
\end{array}
\right. \\
(S_2)_h^{r_2}\left\{
\begin{array}{rcl}
\dot{Z}_5(t) &=& \omega_2 A_0 Z_5(t) - r_2 b_0 \sigma_2(u_2(t-h)) + r_2 v_5(t-h) \\
&-& r_3 \hat{b}^3_{2,2} \sigma_3(u_3(t-h)), \\
\dot{Z}_6(t) &=& \omega_3 A_0 Z_6(t) - r_2 b_0 \sigma_2(u_2(t-h)) + r_2 v_6(t-h) \\
&-& r_3 \hat{\tilde{b}}^3_{2,2} \sigma_3(u_3(t-h)), \\
\dot{Z}_7(t) &=& \omega_4 A_0 Z_7(t) - r_2 b_0 \sigma_2(u_2(t-h)) + r_2 v_7(t-h),
\end{array}
\right. \\
(S_3)_h^{r_3}: \quad \dot{Z}_8(t) = \omega_4 A_0 Z_8(t) - r_3 b_0 \sigma_3(u_3(t-h)) + r_3 v_8(t-h).
\end{array}
\right. \tag{33}
$$

To stabilize this system, we start with the *bidimensional single-input controllable* subsystem

$$
(S_3)_h^{r_3}: \quad \dot{Z}_8(t) = \omega_4 A_0 Z_8(t) - r_3 b_0 \sigma_3(u_3(t-h)) + r_3 v_8(t-h). \tag{34}
$$

According to Proposition 1 in the single-input case, there exist $C_{3,0} > 0$ and, for every $1 \le p \le \infty$, a constant $M_{3,p} > 0$ such that, for every $h > 0$, there is an $\tilde{r}_3(h) \in (0,1]$, for which the system $(S_3)_h^{r_3}$, $r_3 \in (0, \tilde{r}_3(h)]$, with the feedback

$$
u_3(t-h) = b_0^T e^{\omega_4 A_0 h} Z_8(t-h) + v_1^3(t-h), \tag{35}
$$

is finite-gain L^p-stable, i.e., the trajectories $(Z_8)_{\Pi_3}$, $\Pi_3 := (v_1^3, v_8)$, and $r_3 \in (0, \tilde{r}_3(h)]$, starting at $\bar{0}$ and corresponding to $v_1^3, v_8 \in L^p$ with $\|v_8\|_{L^\infty} \le C_{3,0}$, verify

$$
\Big\| (Z_8)_{\Pi_3} \Big\|_{L^p} \le M_{3,p} \Big(\|v_1^3\|_{L^p} + \|v_8\|_{L^p} \Big). \tag{36}
$$

Recall that the upper bound for the L^p-gain $M_{3,p}$ is independent of h because the system $(S_i)_h^{r_i}$, $1 \le i \le 3$, is in cascade and the amplitude $r_3 \le \tilde{r}_3(h)$ has to be chosen smaller than $r_3^*(h)$, the final input bound for the subsystem $(S_3)_h^{r_3}$, i.e., $r_3 \le r_3^*(h)$, which will be determined later. In a second step, we stabilize the subsystem $(S_2)_h^{r_2}$ defined by

$$\begin{aligned} \dot{Z}_{5,6,7}(t) &= A_0^{(3)} Z_{5,6,7}(t) - r_2 b_0^{(3)} \sigma_2(u_2(t-h)) + r_2 v_{5,6,7}(t-h) \\ &- r_3 \tilde{b}_3 \sigma_3(u_3(t-h)), \end{aligned} \tag{37}$$

where $\theta_{5,6,7} := [\theta_5, \theta_6, \theta_7]^T$, $\theta = Z, v$; $b_0^{(3)} := [b_0, b_0, b_0]^T$, $A_0^{(3)} := \operatorname{diag}\,[\omega_2 A_0, \omega_3 A_0, \omega_4 A_0]$, and $\tilde{b}_3 := (\hat{b}_{2,2}^3, \hat{\hat{b}}_{2,2}^3, 0)^T$.

It is clear that $(S_2)_h^{r_2}$ is a *finite dimensional single-input controllable* subsystem with *three distinct eigenvalues*, but perturbed by an external disturbance of sufficiently small amplitude, appearing outside the saturation. According to Proposition 1 (the single-input case), there exist $C_{2,0} > 0$ and, for every $1 \le p \le \infty$, a constant $M_{2,p} > 0$ such that, for every $h > 0$, there is an $\tilde{r}_2(h) \in (0, 1]$ for which the system $(S_2)_h^{r_2}$, $r_2 \in (0, \tilde{r}_2(h)]$, with the perturbed feedback

$$u_2(t-h) = (b_0^{(3)})^T e^{A_0^{(3)} h} Z_{5,6,7}(t-h) + v_1^2(t-h), \tag{38}$$

is finite-gain L^p-stable; i.e., the trajectories $(Z_{5,6,7})_{\Pi_2}$, $\Pi_2 := (v_1^2, v_{5,6,7}, v_1^3, v_8)$, and $r_2 \in (0, \tilde{r}_2(h)]$, starting at $\bar{0}$ and corresponding to $v_1^2, v_{5,6,7}, v_1^3, v_8 \in L^p$ with

$$\|v_{5,6,7}\|_{L^\infty} + \frac{r_3}{r_2} K \|\tilde{b}_3\| \le C_{2,0}, \tag{39}$$

since $\|\sigma_3(.)\|_{L^\infty} \le K$, and where $\|v_{5,6,7}\|_{L^\infty}$ is to be chosen later, verify

$$\left\| (Z_{5,6,7})_{\Pi_2} \right\|_{L^p} \le M_{2,p} \left(\|v_1^2\|_{L^p} + \|v_{5,6,7}\|_{L^p} + \|v_1^3\|_{L^p} + \|v_3\|_{L^p} \right). \tag{40}$$

In addition, the upper bound for the L^p-gain $M_{2,p}$ is independent of the delay $h > 0$. Here, (39) is a necessary condition on r_2, r_3, and $v_{5,6,7}$ to get the final result. The amplitude $r_2 \le r_2^*(h)$, the final input bound of the subsystem $(S_2)_h^{r_2}$, is to be chosen below.

Finally, we need to stabilize the *finite dimensional single-input controllable* system with *four distinct eigenvalues*, $(S_1)_h^{r_1}$. Let $\theta_{1,2,3,4} := [\theta_1, \theta_2, \theta_3, \theta_4]^T$, $\theta = Z, v$; $b_0^{(4)} := [b_0, b_0, b_0, b_0]^T$, $A_0^{(4)} := \operatorname{diag}[\omega_1 A_0, \omega_2 A_0, \omega_3 A_0, \omega_4 A_0]$, $b_2 := [\hat{b}_{1,3}^2, 0, 0, 0]^T$, and $b_3 := [\hat{b}_{1,3}^3, \hat{b}_{1,2}^3, \hat{\hat{b}}_{1,2}^3, 0]^T$. With the previous notation, $(S_1)_h^{r_1}$ becomes

$$\begin{aligned} \dot{Z}_{1,2,3,4}(t) &= A_0^{(4)} Z_{1,2,3,4}(t) - r_1 b_0^{(4)} \sigma_1(u_1(t-h)) + r_1 v_{1,2,3,4}(t-h) \\ &- r_2 b_2 \sigma_2(u_2(t-h)) - r_3 b_3 \sigma_3(u_3(t-h)). \end{aligned} \tag{41}$$

According to Proposition 1 (single-input case), there exist $C_{1,0} > 0$ and, for every $1 \le p \le \infty$, $M_{1,p} > 0$ such that, for every $h > 0$, one has an $\tilde{r}_1(h) \in (0, 1]$ for which the system $(S_1)_h^{r_1}$, $r_1 \in (0, \tilde{r}_1(h)]$, with the feedback

$$u_1(t-h) = (b_0^{(4)})^T e^{A_0^{(4)} h} Z_{1,2,3,4}(t-h) + v_1^1(t-h), \tag{42}$$

is finite-gain L^p-stable; i.e., the trajectories $(Z_{1,2,3,4})_{\Pi_1}$, $\Pi_1 := (v_1^1, v_{1,2,3,4}, v_1^2, v_{5,6,7}, v_1^3, v_8)$, and $r_1 \in (0, r_1^*(h)]$, starting at $\bar{0}$ and corresponding to $v_1^1, v_{1,2,3,4}, v_1^2, v_{5,6,7}, v_1^3, v_8 \in L^p$ with

$$\|v_{1,2,3,4}\|_{L^\infty} + \frac{r_2}{r_1} K\|b_2\| + \frac{r_3}{r_1} K\|b_3\| \leq C_{1,0}, \tag{43}$$

where $\|v_{1,2,3,4}\|_{L^\infty}$ is to be chosen later, verify

$$\begin{aligned} \left\|(Z_{1,2,3,4})_{\Pi_1}\right\|_{L^p} \quad &\leq \quad M_{1,p}\Big(\|v_1^1\|_{L^p} + \|v_{1,2,3,4}\|_{L^p} + \|v_1^2\|_{L^p} \\ &\qquad + \|v_{5,6,7}\|_{L^p} + \|v_1^3\|_{L^p} + \|v_8\|_{L^p}\Big). \end{aligned} \tag{44}$$

Note that the upper bound for the L^p-gain, $M_{1,p}$, is independent of h. In order to satisfy (39) and (43), we first give the following bounds for $\|v_{1,2,3,4}\|_{L^\infty}$ and $\|v_{5,6,7}\|_{L^\infty}$:

$$\|v_{1,2,3,4}\|_{L^\infty} \leq \frac{C_{1,0}}{2} \quad \text{and} \quad \|v_{5,6,7}\|_{L^\infty} \leq \frac{C_{2,0}}{2}. \tag{45}$$

Then, the input bounds r_1, r_2, and r_3 must verify

$$\begin{aligned} &\forall\ 1 \leq i \leq 3, \quad r_i \leq \tilde{r}_i(h), \quad \tfrac{r_3}{r_2} K\|\tilde{b}_3\| \leq \tfrac{C_{2,0}}{2}, \quad \text{and} \\ &\tfrac{r_2}{r_1} K\|b_2\| + \tfrac{r_3}{r_1} K\|b_3\| \leq \tfrac{C_{1,0}}{2}. \end{aligned} \tag{46}$$

It is clear that (46) is verified for $r_i \leq r_i^*(h)$, $1 \leq i \leq 3$, where

$$\begin{aligned} &r_1^*(h) := \tilde{r}_1(h), \quad r_2^*(h) = \min\left\{ \tfrac{C_{1,0}\|\tilde{b}_3\|}{C_{2,0}\|b_3\| + 2K\|b_2\|\|\tilde{b}_3\|} r_1^*(h), \tilde{r}_2(h) \right\}, \\ &r_3^*(h) = \min\left\{ \tfrac{C_{1,0}C_{2,0}}{2K(C_{2,0}\|b_3\| + 2K\|b_2\|\|\tilde{b}_3\|)} r_1^*(h), \tfrac{C_{2,0}}{2K\|\tilde{b}_3\|} \tilde{r}_2(h), \tilde{r}_3(h) \right\}. \end{aligned} \tag{47}$$

Set $r^*(h) := \min_{1 \leq i \leq 3} r_i^*(h)$ and $M_p := M_{1,p} + M_{2,p} + M_{3,p}$. We have shown that there exists a constant $C_0 > 0$ and, for every $1 \leq p \leq \infty$, a constant $M_p > 0$ such that for every $h > 0$ there is an $r^*(h) \in (0, 1]$ for which the system (31), $r \in (0, r^*(h)]$, with the bounded feedback

$$u(t-h) = \begin{pmatrix} (b_0^{(4)})^T e^{A_0^{(4)} h} (Z_{1,2,3,4})_{\Pi_1}(t-h) + v_1^1(t-h) \\ (b_0^{(3)})^T e^{A_0^{(3)} h} (Z_{5,6,7})_{\Pi_2}(t-h) + v_1^2(t-h) \\ b_0^T e^{v_4 A_0 h} (Z_8)_{\Pi_3}(t-h) + v_1^3(t-h) \end{pmatrix}, \tag{48}$$

is finite-gain L^p-stable; i.e., the trajectories $[(Z_{1,2,3,4})_{\Pi_1}, (Z_{5,6,7})_{\Pi_2}, (Z_8)_{\Pi_3}]$, $r \in (0, r^*(h)]$, starting at $\bar{0}$ and corresponding to $[(v_1^1; v_{1,2,3,4}; v_1^2; v_{5,6,7}; v_1^3; v_8), (v_1^2; v_{5,6,7}; v_1^3; v_8), (v_1^3; v_8)] \in L^p$, with

$$\|v_{1,2,3,4}\|_{L^\infty} + \|v_{5,6,7}\|_{L^\infty} + \|v_8\|_{L^\infty} + K\left(\|b_2\| + \|b_3\| + \|\tilde{b}_3\|\right) \leq C_0, \tag{49}$$

verify

$$\begin{aligned} \left\|\left[\left(Z_{1,2,3,4}\right)_{\Pi_1}, \left(Z_{5,6,7}\right)_{\Pi_2}, (Z_8)_{\Pi_3}\right]\right\|_{L^p} \quad &\leq \quad M_p\Big[\|v_1^1\|_{L^p} + \|v_{1,2,3,4}\|_{L^p} \\ &+ \quad \|v_1^2\|_{L^p} + \|v_{5,6,7}\|_{L^p} + \|v_1^3\|_{L^p} + \|v_8\|_{L^p}\Big], \end{aligned} \tag{50}$$

where the upper bound for the corresponding global L^p-gain M_p is delay-independent.

Proof of Proposition 1 *in the general case.* We deal with the general m-input case $m \geq 2$ and prove Proposition 1 by induction on the number of distinct algebraic multiplicities of the eigenvalues of A.

Let l be the number of conjugate pairs of nonzero purely imaginary eigenvalues of A (counting multiplicity) and $\mu^{(0)}$ be the multiplicity of 0 as an eigenvalue of A. Suppose that the l conjugate pairs of nonzero purely imaginary eigenvalues of A are in the form $\{(-i\omega^{(j)}, i\omega^{(j)})_{1\leq j\leq l}, i^2 = -1\}, \omega^{(j)} > 0$, and of algebraic multiplicity $\mu^{(j)}$ for all $1 \leq j \leq l$. Then, $2\sum_{j=1}^{l} \mu^{(j)} + \mu^{(0)} = n$. More precisely, for each $0 \leq j \leq l$, we consider a relabeling of the $\mu^{(j)}$'s, defined by μ_j, corresponding to the eigenvalue ω_j such that $\mu_l \geq \mu_{l-1} \geq \cdots \geq \mu_\kappa \geq \cdots \geq \mu_1 \geq \mu_0 \geq 1$, where μ_κ is the multiplicity corresponding to 0, i.e., there exists $0 \leq \kappa \leq l$ such that $\mu_\kappa = \mu^{(0)}$.

First, we find nonsingular transformation matrices O and T such that the pair (A, B) is transformed into the block diagonal control canonical form

$$OAO^T = \begin{pmatrix} A_{0,\mu_0} & & & & \\ & \ddots & & & \\ & & A_{0,\mu_\kappa} & & (0) \\ & & & \ddots & \\ & (0) & & & A_{0,\mu_l} \end{pmatrix}, \tag{51}$$

where, for every $0 \leq j \leq l$, $j \neq \kappa$, A_{0,μ_j} is a $2\mu_j \times 2\mu_j$ matrix and is in the form

$$A_{0,\mu_j} = \begin{pmatrix} \omega_j A_0 & 0 & \dots & 0 \\ 0 & \omega_j A_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_j A_0 \end{pmatrix}, \quad \text{with } A_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and A_{0,μ_κ} is a $\mu_\kappa \times \mu_\kappa$ zero matrix; and

$$OBT = \Big(B_0, \dots, B_\kappa, \dots, B_{l-1}, B_l\Big)^T, \tag{52}$$

where

$$B_j = \begin{bmatrix} \begin{pmatrix} \begin{pmatrix} b_0 & & \\ (0) & \ddots & (0) \\ & & b_0 \end{pmatrix} & (*)_j \end{pmatrix} & (**)_j \end{bmatrix}, \quad \text{with } b_0 = (0 \quad 1)^T,$$

$0 \leq j \leq l, j \neq \kappa$, and

$$B_\kappa = \begin{bmatrix} \begin{pmatrix} \begin{pmatrix} 1 & & \\ (0) & \ddots & (0) \\ & & 1 \end{pmatrix} & (*)_\kappa \end{pmatrix} & (**)_\kappa \end{bmatrix}.$$

Here $(*)'_j s$, $(**)'_j s, 0 \leq j \leq l$, $j \neq \kappa$, represent, respectively, submatrices of size $2\mu_j \times (\mu_l - \mu_j)$ and $2\mu_j \times (m - \mu_l)$; $(*)_\kappa$ and $(**)_\kappa$ are submatrices of dimensions $\mu_\kappa \times (\mu_l - \mu_\kappa)$ and $\mu_\kappa \times (m - \mu_l)$.

We note that the existence of the above canonical form is shown in Appendix B.

There exists a linear transformation of the system $(S)_h^r$, given in Lemma 4, which simplifies the analysis of the problem. Our method consists of rewriting the original system $(S)_h^r$ as an appropriate l cascade of single-input subsystems of dimension $2\mu_j, 0 \leq j \leq l$, $j \neq \kappa$, and one subsystem of dimension μ_κ, since A has l conjugate pairs of nonzero purely imaginary eigenvalues of the forms $(\pm i\omega_j)_{1\leq j\leq l}$ and 0 is an eigenvalue of A with multiplicity μ_κ. The system $(S)_h^r$ becomes

(53)
$$\begin{aligned}
\dot{Y}_{\theta,\mu_s}(t) &= \omega_s A_0 Y_{\theta,\mu_s}(t) - rb_0\sigma_\theta(u_\theta(t-h)) + rv_{\theta,\mu_s}(t-h) \\
&\qquad -r\textstyle\sum_{j=\mu_s+1}^{m} \hat{b}^j_{\theta,s}\sigma_j(u_j(t-h)), 1 \leq \theta \leq \mu_s, 0 \leq s \leq l, s \neq \kappa, \\
\dot{y}_{\theta,\mu_\kappa}(t) &= -r\sigma_\theta(u_\theta(t-h)) + rv_{\theta,\mu_\kappa}(t-h) \\
&\qquad -r\textstyle\sum_{j=\mu_\kappa+1}^{m} \hat{a}^j_\theta\sigma_j(u_j), 1 \leq \theta \leq \mu_\kappa,
\end{aligned}$$

where $\hat{b}^j_{\theta,s} \in R^2$ and $\hat{a}^j_\theta \in R$ are arbitrarily vectors and real numbers. Without loss of generality, in our proof of the general case, we will suppose that A and B are already in these forms. Consider then

$$\begin{aligned}
(S)_h^r : \dot{Y}(t) &= AY(t) - rB\sigma\left(u(t-h)\right) + rv(t-h) \\
&:= AY(t) - rB\sigma\left(F_h^T Y(t-h) + v^1(t-h)\right) + rv^2(t-h),
\end{aligned}$$

and assume without loss of generality that $(S)_h^r$ is of the form (53). Next, we deal with the general m-input case and prove Proposition 1 by induction on $\bar{\delta} := 1 + \#\{\delta_k > 0, 1 \leq k \leq l\}$, where $\delta_k = \mu_k - \mu_{k-1}$ for all $1 \leq k \leq l$ and $\#$ is the cardinal of distinct algebraic multiplicities of the eigenvalues of the skew-symmetric matrix A.

Let μ^j, $0 \leq j \leq \psi \leq l$, denote the distinct algebraic multiplicities corresponding to the eigenvalues ω^j such that $\mu^0 < \cdots < \mu^\psi$, and let χ^j be the number of equal algebraic multiplicities in each inequality. Then, $\sum_{j=0}^{k} \chi^j = l+1$ and ω^j is composite by χ^j eigenvalues $(\omega_1^j, \ldots, \omega_{\chi^j}^j)$ for all $0 \leq j \leq \psi$. We can show that, for every $0 \leq j \leq \psi$, μ^j corresponds to the jth strict inequality.

Induction hypothesis. For every $1 \leq p \leq \infty$, there exist $M_p, C_0 > 0$ such that, for every $h > 0$, one has an $n \times m$ matrix F_h and $r^*(h) \in (0, 1]$, for which the trajectories Y_{v^1,v^2} of $(S)_h^r$, $r \in (0, r^*(h)]$, starting at $\bar{0}$ and corresponding to $v^1, v^2 \in L^p$ with $\|v^2\|_{L^\infty} \leq C_0$, verify $\|Y_{v^1,v^2}\|_{L^p} \leq M_p\left(\|v^1\|_{L^p} + \|v^2\|_{L^p}\right)$.

Notation. First, we introduce some notation. Let $\mu^j := \mu_1^j = \cdots = \mu_{\chi^j}^j$; for any vector $b_0, b_{0,1}^{(j+1)} := [b_0, \ldots, b_0, 1, b_0, \ldots, b_0]^T$, $b_0^{(j)} := [b_0, \ldots, b_0]^T$, b_0 repeated j times; $A_{0,0}^{(\chi^j+\cdots+\chi^\psi)}$ denotes diag $(\omega_1^j A_0, \ldots, \omega_{\chi^j}^j A_0; \ldots; \omega_1^\psi A_0, \ldots, \omega_{\chi^\psi}^\psi A_0)$ if $\omega_\kappa := 0 \in \{\omega_1^j, \ldots, \omega_{\chi^j}^j; \ldots; \omega_1^\psi, \ldots, \omega_{\chi^\psi}^\psi\}$; $A_0^{(\chi^j+\cdots+\chi^\psi)}$ denotes diag $(\omega_1^j A_0, \ldots, \omega_{\chi^j}^j A_0; \ldots; \omega_1^\psi A_0, \ldots, \omega_{\chi^\psi}^\psi A_0)$ if not, and for every $\theta \leq \mu^j$, $Y^\theta_{\mu^j,\ldots,\mu^\psi} := [Y_{\theta,\mu^j}, \ldots, Y_{\theta,\mu^\psi}]^T$, where $Y^\theta_{\mu^j,\ldots,\mu^j} = Y^\theta_{\mu^j} = (Y_{\theta,\mu_1^j}, \ldots, Y_{\theta,\mu_{\chi^j}^j})^T$ for all $0 \leq j \leq \psi$.

Step 1. Verify that the induction hypothesis is satisfied for $\bar{\delta} = 1$, i.e., $\psi = 0$. Then, $\mu_0 = \mu_1 = \cdots = \mu_l := \rho \geq 1$ and $(*)_j = 0$ for all $0 \leq j \leq l$. In this step, we transform the system $(S)_h^r$ into ρ single-input subsystems made by distinct eigenvalues and by equations which have a common number θ, $1 \leq \theta \leq \rho$, in each subsystem (j),

$0 \le j \le l$; i.e., $(S)_h^r$ is the reunion of ρ-subsystems: $(S)_h^r = \bigcup_{\theta=1}^{\rho}(S_\theta)_h^{r_\theta}$ with

$$(54)\qquad \begin{aligned}(S_\theta)_h^{r_\theta}:\ \dot{Y}_\rho^\theta(t) &= A_{0,0}^{(l+1)}Y_\rho^\theta(t) - r_\theta b_{0,1}^{(l+1)}\sigma_\theta(u_\theta(t-h)) + r_\theta v_\rho^\theta(t-h)\\ &\quad - r_\theta \textstyle\sum_{j=\rho+1}^{m}\hat{b}^j_{\theta,0,\dots,\kappa,\dots,l}\sigma_j(u_j(t-h)),\end{aligned}$$

where $\hat{b}_{\theta,0,\dots,\kappa,\dots,l} := (\hat{b}_{\theta,0},\dots,\hat{b}_{\theta,\kappa},\dots,\hat{b}_{\theta,l})^T$ and $\hat{b}_{\theta,\kappa} := \hat{a}_\theta$. So, $(S_\theta)_h^{r_\theta}$ is *a finite dimensional single-input controllable system* with *distinct eigenvalues* for all $1 \le \theta \le \rho$.

Put $u_j = 0$, $j \ge \rho+1$, and take into account the result previously proved for the single-input case. Then, for every $1 \le \theta \le \rho$, there exist a constant $C_{\theta,0}$ and, for every $1 \le p \le \infty$, a constant $M_{\theta,p} > 0$ such that, for every $h > 0$, there is an $r_\theta^*(h) \in (0,1]$ for which the trajectories $(Y_\rho^\theta)_{v_1^\theta,v_\rho^\theta}$ of $(S_\theta)_h^{r_\theta}$, $r_\theta \in (0, r_\theta^*(h)$, starting at $\bar{0}$ and corresponding to $v_1^\theta, v_\rho^\theta \in L^p$ with $\|v_\rho^\theta\|_{L^\infty} \le C_{\theta,0}$, are finite-gain L^p-stable, with the feedback

$$(55)\qquad u_\theta(t-h) = (b_{0,1}^{(l+1)})^T e^{A_{0,0}^{(l+1)}h}Y_\rho^\theta(t-h) + v_1^\theta(t-h).$$

Moreover,

$$(56)\qquad \left\|(Y_\rho^\theta)_{v_1^\theta,v_\rho^\theta}\right\|_{L^p} \le M_{\theta,p}\left(\|v_1^\theta\|_{L^p} + \|v_\rho^\theta\|_{L^p}\right),\quad \theta = 1,\dots,\rho.$$

Recall that for every $1 \le \theta \le \rho$, the upper bound for the L^p-gain $M_{\theta,p}$ is delay-independent.

Step 2. Assume now that the induction hypothesis is satisfied at the step $\bar{\delta} = k$, i.e., $\psi = k-1$ for all $1 \le k \le l$. Take the $k+1$ distinct algebraic multiplicity μ^k, corresponding to the eigenvalues $\omega^k := (\omega_1^k,\dots,\omega_{\chi^k}^k)$ such that $\mu^0 < \cdots < \mu^k$ and χ^k is the number of equal algebraic multiplicities in the last inequality. Then, ω^k is made of χ^k eigenvalues. Note that the original system $(S)_h^r$ is a cascade of single-input subsystems, denoted $(0),\dots,(k)$, each of them corresponding to the algebraic multiplicity μ^j, $0 \le j \le k$. We can show that each subsystem (j), $j = 0$ to k is made of χ^j subsystems $(\nu)_j$, $\nu = 1$ to χ^j. So, our method consists of transforming the system $(S)_h^r$ in a cascade of single-input $(k+1)$-subsystems; i.e., $(S)_h^r = \cup_{\alpha=0}^{k}(S_{k-\alpha+1})_h^{r^{k-\alpha+1}}$, where each $(S_{k-\alpha+1})_h^{r^{k-\alpha+1}}$ has $\chi^{k-\alpha}+\cdots+\chi^k$ distinct imaginary eigenvalues and is of dimension $(2(\chi^{k-\alpha}+\cdots+\chi^k)-1)(\mu^{k-\alpha}-\mu^{k-\alpha-1})$ if 0 is an eigenvalue or $2(\chi^{k-\alpha}+\cdots+\chi^k)(\mu^{k-\alpha}-\mu^{k-\alpha-1})$ if not for all $1 \le \alpha \le k+1$ with $\mu^{-1} := 0$.

Remark 5. We can see that each subsystem $(S_{k-\alpha+1})_h^{r^{k-\alpha+1}}$, $0 \le \alpha \le k$, is made of equations having the same order number θ, $\mu^{k-\alpha-1}+1 \le \theta \le \mu^{k-\alpha}$, in the subsystems $(k),(k-1),\dots,(k-\alpha)$, where it has distinct imaginary eigenvalues, but perturbed by an external disturbance appearing outside the saturation, except the subsystem $(S_{k+1})_h^{r^{k+1}}$.

For $0 \le \alpha \le k$, each subsystem $(S_{k-\alpha+1})_h^{r^{k-\alpha+1}}$ takes the form

$$(S_{k-\alpha+1})_h^{r^{k-\alpha+1}} \begin{cases} \dot{Y}_{\theta,\mu^{k-\alpha}}(t) = A_{0,0}^{(\chi^{k-\alpha})} Y_{\theta,\mu^{k-\alpha}}(t) - r_\theta^{k-\alpha+1} b_{0,1}^{(\chi^{k-\alpha})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{k-\alpha}} - \sum_{j=\mu^{k-\alpha}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{k-\alpha}}^{j} \sigma_j(\bar{u}_j), \\ \vdots \\ \dot{Y}_{\theta,\mu^{k-1}}(t) = A_0^{(\chi^{k-1})} Y_{\theta,\mu^{k-1}}(t) - r_\theta^{k-\alpha+1} b_0^{(\chi^{k-1})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{k-1}} - \sum_{j=\mu^{k-1}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{k-1}}^{j} \sigma_j(\bar{u}_j), \\ \dot{Y}_{\theta,\mu^{k}}(t) = A_0^{(\chi^{k})} Y_{\theta,\mu^{k}}(t) - r_\theta^{k-\alpha+1} b_0^{(\chi^{k})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{k}} - \sum_{j=\mu^{k}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{k}}^{j} \sigma_j(\bar{u}_j) \end{cases}$$

if $\mu_\kappa \neq \{\mu^{k-\alpha}, \dots, \mu^k\}$ and takes the form

$$(S_{k-\alpha+1})_h^{r^{k-\alpha+1}} \begin{cases} \dot{Y}_{\theta,\mu^{k-\alpha}}(t) = A_{0,0}^{(\chi^{k-\alpha})} Y_{\theta,\mu^{k-\alpha}}(t) - r_\theta^{k-\alpha+1} b_{0,1}^{(\chi^{k-\alpha})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{k-\alpha},0} - \sum_{j=\mu^{k-\alpha}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{k-\alpha},0}^{j} \sigma_j(\bar{u}_j), \\ \vdots \\ \dot{Y}_{\theta,\mu^{j_0}}(t) = A_{0,0}^{(\chi^{j_0})} Y_{\theta,\mu^{j_0}}(t) - r_\theta^{k-\alpha+1} b_{0,1}^{(\chi^{j_0})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{j_0}} - \sum_{j=\mu^{j_0}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{j_0},0}^{j} \sigma_j(\bar{u}_j), \\ \dot{Y}_{\theta,\mu^{j_0-1}}(t) = A_0^{(\chi^{j_0-1})} Y_{\theta,\mu^{j_0-1}}(t) - r_\theta^{k-\alpha+1} b_0^{(\chi^{j_0-1})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{j_0-1}} - \sum_{j=\mu^{j_0-1}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{j_0-1}}^{j} \sigma_j(\bar{u}_j), \\ \vdots \\ \dot{Y}_{\theta,\mu^{k-1}}(t) = A_0^{(\chi^{k-1})} Y_{\theta,\mu^{k-1}}(t) - r_\theta^{k-\alpha+1} b_0^{(\chi^{k-1})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{k-1}} - \sum_{j=\mu^{k-1}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{k-1}}^{j} \sigma_j(\bar{u}_j), \\ \dot{Y}_{\theta,\mu^{k}}(t) = A_0^{(\chi^{k})} Y_{\theta,\mu^{k}}(t) - r_\theta^{k-\alpha+1} b_0^{(\chi^{k})} \sigma_\theta(\bar{u}_\theta) \\ \qquad + r_\theta^{k-\alpha+1} \bar{v}_{\theta,\mu^{k}} - \sum_{j=\mu^{k}+1}^{m} r_j^{k-\alpha+1} \hat{b}_{\theta,\xi^{k}}^{j} \sigma_j(\bar{u}_j) \end{cases}$$

if not, for all $\mu^{k-\alpha-1} + 1 \leq \theta \leq \mu^{k-\alpha}$, where $\bar{\pi} = \pi(t-h)$, $\hat{b}_{\theta,\xi^s}^{j} = (\hat{b}_{\theta,1}^{j}, \dots, \hat{b}_{\theta,\chi^s}^{j})^T$, $k-\alpha \leq s \leq k$, and $\hat{b}_{\theta,\xi^s,0}^{j} = (\hat{b}_{\theta,1}^{j}, \dots, \hat{a}_\theta^j, \dots, \hat{b}_{\theta,\chi^s}^{j})^T$, $k-\alpha \leq s \leq j_0$, and $\mu_\kappa = \mu^{j_0}$.

With the notation, each subsystem $(S_{k-\alpha+1})_h^{r^{k-\alpha+1}}$, $0 \le \alpha \le k$, is given by

(57)
$$\dot{Y}^{\theta}_{\mu^{k-\alpha},\dots,\mu^k}(t)=\begin{cases} A_{0,0}^{(\chi^{k-\alpha}+\cdots+\chi^k)}Y^{\theta}_{\mu^{k-\alpha},\dots,\mu^k}(t)-r_\theta^{k-\alpha+1}b_{0,1}^{(\chi^{k-\alpha}+\cdots+\chi^k)}\sigma_\theta\Big(u_\theta(t-h)\Big) \\ \qquad + r_\theta^{k-\alpha+1}v^{\theta}_{\mu^{k-\alpha},\dots,\mu^k,0}(t-h)+V_{k-\alpha+2}(t-h) \\ \qquad\qquad \text{if } \mu_\kappa \in \{\mu^{k-\alpha},\dots,\mu^k\}, \\ A_{0}^{(\chi^{k-\alpha}+\cdots+\chi^k)}Y^{\theta}_{\mu^{k-\alpha},\dots,\mu^k}(t)-r_\theta^{k-\alpha+1}b_{0}^{(\chi^{k-\alpha}+\cdots+\chi^k)}\sigma_\theta\Big(u_\theta(t-h)\Big) \\ \qquad + r_\theta^{k-\alpha+1}v^{\theta}_{\mu^{k-\alpha},\dots,\mu^k}(t-h)+V_{k-\alpha+2}(t-h) \end{cases}$$

for every $\mu^{k-\alpha-1}+1 \le \theta \le \mu^{k-\alpha}$, where

$$\begin{aligned} V_{k-\alpha+2}(.) &:= -\Big(\textstyle\sum_{j=\mu^{k-\alpha}+1}^{m} r_j^{k-\alpha+2}\hat{b}^j_{\theta,\xi^{k-\alpha}}\sigma_j\Big(u_j(.)\Big) \quad \dots \\ &\qquad\qquad \textstyle\sum_{j=\mu^k+1}^{m} r_j^{k+1}\hat{b}^j_{\theta,\xi^k}\sigma_j\Big(u_j(.)\Big)\Big)^T \\ &:= -r^{k-\alpha+2}_{\mu^{k-\alpha}+1}b_{\mu^{k-\alpha}+1}\sigma_{\mu^{k-\alpha}+1}\Big(u_{\mu^{k-\alpha}+1}(.)\Big)-\cdots-r^{k-\alpha+2}_{\mu^{k-\alpha+1}}b_{\mu^{k-\alpha+1}} \\ &\times \quad \sigma_{\mu^{k-\alpha+1}}\Big(u_{\mu^{k-\alpha+1}}(.)\Big)-\cdots-r^{k+1}_{\mu^{k-1}+1}b_{\mu^{k-1}+1}\sigma_{\mu^{k-1}+1}\Big(u_{\mu^{k-1}+1}(.)\Big) \\ &- \quad r^{k+1}_{\mu^k}b_{\mu^k}\sigma_{\mu^k}\Big(u_{\mu^k}(.)\Big)-\cdots-r^m_m b_m\sigma_m\Big(u_m(.)\Big), \end{aligned}$$

with $\mu^m = m$. First, for $j \ge \mu^k+1$, put $u_j = 0$, then $V_{k+2} = 0$. $(S_{k-\alpha+1})_h^{r^{k-\alpha+1}}$ is a *finite dimensional single-input controllable* system with *distinct eigenvalues* but perturbed by the vector $V_{k-\alpha+2}$ for all $0 \le \alpha \le k$. The problem is that the perturbation may cause instability. In this case, our method of finite-gain stabilizability proceeds step by step, meaning that, first, we begin to finite-gain L^p-stabilize the subsystem $(S_{k+1})_h^{r^{k-\alpha+1}}$ corresponding to $\alpha = 0$, $\mu^{k-1}+1 \le \theta \le \mu^k$, $V_{k+2} = 0$, and made of χ^k distinct eigenvalues $(\omega_1^k,\dots,\omega_{\chi^k}^k)$ of dimension $(2\chi^k-1)(\mu^k-\mu^{k-1})$ if $0 \in \{\omega_1^k,\dots,\omega_{\chi^k}^k\}$ or $2\chi^k(\mu^k-\mu^{k-1})$ if not. As in the first step, we transform the subsystem $(S_{k+1})_h^{r^{k+1}}$ into the union of $\mu^k-\mu^{k-1}$ single-input subsystems made of distinct eigenvalues and equations which have the same order number θ, $\mu^{k-1}+1 \le \theta \le \mu^k$ in each subsystem (ν), $1 \le \nu \le \chi^k$, where the subsystem (ν) corresponds to the eigenvalue ω_ν^k; i.e., $(S_{k+1})_h^{r^{k+1}} = \cup_{\theta=\mu^{k-1}+1}^{\mu^k}(S^\theta_{k+1})_h^{r^{k+1}_\theta}$ with

$$(S^\theta_{k+1})_h^{r^{k+1}_\theta} : \dot{Y}^\theta_{\mu^k}(t) = \begin{cases} A_{0,0}^{(\chi^k)}Y^\theta_{\mu^k}(t) - r_\theta^{k+1}b_{0,1}^{(\chi^k)}\sigma_\theta(u_\theta(t-h)) + r_\theta^{k+1}v^\theta_{\mu^k,0}(t-h) \\ \qquad\qquad \text{if } \mu^k = \mu_\kappa, \\ A_{0}^{(\chi^k)}Y^\theta_{\mu^k}(t) - r_\theta^{k+1}b_{0}^{(\chi^k)}\sigma_\theta(u_\theta(t-h)) + r_\theta^{k+1}v^\theta_{\mu^k}(t-h) \\ \qquad\qquad \text{if not.} \end{cases}$$

$(S^\theta_{k+1})^{r^{k+1}_\theta}_h$ is a *finite dimensional single-input controllable* system with *distinct eigenvalues* for every $\mu^{k-1}+1 \leq \theta \leq \mu^k$. According to the result previously proved for the single-input case ((Step 1), respectively), this subsystem can be finite-gain L^p-stabilized, for all $1 \leq p \leq \infty$ and $\mu^{k-1}+1 \leq \theta \leq \mu^k$, with the feedback

$$u_\theta(t-h) = \begin{cases} (b_0^{(\chi^k)})^T e^{A_0^{(\chi^k)}h} Y^\theta_{\mu^k}(t-h) + v^\theta_1(t-h) \text{ if } \mu^k \neq \mu_\kappa, \\ \\ (b_{0,1}^{(\chi^k)})^T e^{A_{0,0}^{(\chi^k)}h} Y^\theta_{\mu^k}(t-h) + v^\theta_{1,0}(t-h) \text{ if not;} \end{cases} \tag{58}$$

i.e., for every $\mu^{k-1}+1 \leq \theta \leq \mu^k$, there exists a constant $C^{k+1}_{\theta,0} > 0$ and, for every $1 \leq p \leq \infty$, a constant $M^{k+1}_{\theta,p} > 0$ such that, for every $h > 0$, there is an $(\tilde{r}_\theta)^{k+1}(h) \in (0,1]$ for which the trajectories $(Y^\theta_{\mu^k})_{v^\theta_1, v^\theta_{\mu^k}}$ ($(Y^\theta_{\mu^k})_{v^\theta_{1,0}, v^\theta_{\mu^k,0}}$, respectively) of $(S_{k+1})^{r^{k+1}_\theta}_h$, $r^{k+1}_\theta \in (0, (\tilde{r}_\theta)^{k+1}(h)]$, starting at $\bar{0}$ and corresponding to $v^\theta_1, v^\theta_{\mu^k}$ ($v^\theta_{1,0}, v^\theta_{\mu^k,0}$, respectively) $\in L^p$ with $\|v^\theta_{\mu^k}\|_{L^\infty}$ ($\|v^\theta_{\mu^k,0}\|_{L^\infty}$, respectively) $\leq C^{k+1}_{\theta,0}$, verify

$$\begin{aligned} \left\| (Y^\theta_{\mu^k})_{v^\theta_1, v^\theta_{\mu^k}} \right\|_{L^p} &\leq M^{k+1}_{\theta,p} \left(\|v^\theta_1\|_{L^p} + \|v^\theta_{\mu^k}\|_{L^p} \right) \text{ if } \mu_\kappa \neq \mu^k, \\ \left\| (Y^\theta_{\mu^k})_{v^\theta_{1,0}, v^\theta_{\mu^k,0}} \right\|_{L^p} &\leq M^{k+1}_{\theta,p} \left(\|v^\theta_{1,0}\|_{L^p} + \|v^\theta_{\mu^k,0}\|_{L^p} \right) \text{ if } \mu_\kappa = \mu^k. \end{aligned} \tag{59}$$

Recall that, for every $\mu^{k-1}+1 \leq \theta \leq \mu^k$, the upper bound for the corresponding L^p-gain, $M^{k+1}_{\theta,p}$, of the subsystem $(S^\theta_{k+1})^{r^{k+1}_\theta}_h$, $r^{k+1}_\theta \in (0, (\tilde{r}_\theta)^{k+1}(h)]$, is independent of the delay $h > 0$. Note that the amplitude $r^{k+1}_\theta \leq (r^*_\theta)^{k+1}$, the final input bound of the subsystem $(S^\theta_{k+1})^{r^{k+1}_\theta}_h$, which will be determined later. Now, since $2\chi^k(\mu^k - \mu^{k-1})$ is the dimension of the subsystem $(S_{k+1})^{r^{k+1}}_h$ ($(2\chi^k-1)(\mu^k-\mu^{k-1})$, respectively, if $\mu_\kappa = \mu^k$) and $2\chi^k\mu^k$ is the dimension of the subsystem (k), ($(2\chi^k-1)\mu^k$, respectively, if $\mu_\kappa = \mu^k$), then after stabilizing $(S_{k+1})^{r^{k+1}}_h$, it is obvious that $2\chi^k\mu^{k-1}$ or $\left((2\chi^k-1)\mu^{k-1}\right)$ is the dimension of the rest of the subsystem (k), which is made of the subsystems $(j)_k$, $1 \leq j \leq \chi^k$, corresponding to the algebraic multiplicities $\mu^k := \mu^k_1 = \cdots = \mu^k_{\chi^k}$ and the eigenvalues $\omega^k_1, \ldots, \omega^k_{\chi^k}$. Then, the algebraic multiplicity of the eigenvalues of the rest of the system (k) is μ^{k-1} which is the algebraic multiplicity of the subsystem $(k-1)$ corresponding to the eigenvalues $\omega^{k-1}_1, \ldots, \omega^{k-1}_{\chi^{k-1}}$. Therefore, the remaining part of the system $(S)^r_h$, denoted $(\hat{S})^{\hat{r}}_h$, is characterized by $\bar{\delta} = k$ but it is perturbed by the vector of sufficiently small amplitude

$$\hat{V}(.) := -r^{k+1}_{\mu^{k-1}+1} b_{\mu^{k-1}+1} \sigma_{\mu^{k-1}+1}(u_{\mu^{k-1}+1}(.)) - \cdots - r^{k+1}_{\mu^k} b_{\mu^k} \sigma_{\mu^k}(u_{\mu^k}(.)).$$

Set

$$\begin{aligned} \dot{\hat{Y}}(t) &= \hat{A}\hat{Y}(t) - \hat{r}\hat{B}\hat{\sigma}\left(\hat{u}(t-h)\right) + \hat{r}\hat{v}^2(t-h) + \hat{V}(t-h) \\ &:= \hat{A}\hat{Y}(t) - \hat{r}\hat{B}\hat{\sigma}\left(\hat{F}^T_h \hat{Y}(t-h) + \hat{v}^1(t-h)\right) + \hat{r}\hat{v}^2(t-h) + V_{k+1}(t-h) \end{aligned}$$

as the subsystem $(\hat{S})^{\hat{r}}_h$ which is characterized by $\bar{\delta} = k$. By the inductive hypothesis, this subsystem can be finite-gain L^p-stabilized for all $1 \leq p \leq \infty$. So, there exists a constant $\hat{C}_0 > 0$ and, for every $1 \leq p \leq \infty$, a constant $\hat{M}_p > 0$ such

that, for every $h > 0$, there is an $(\hat{r})^*(h) \in (0,1]$ for which the trajectories $(\hat{Y})_{\hat{\Pi}}$, $\hat{\Pi} := ((\hat{v}^1, \hat{v}^2), (v_1^\theta, v_{\mu^k}^\theta)_{\mu^{k-1}+1 \le \theta \le \mu^k})$ if $\mu_\kappa \neq \mu^k$, or $((\hat{v}^1, \hat{v}^2), (v_{1,0}^\theta, v_{\mu^k,0}^\theta)_{\mu^{k-1}+1 \le \theta \le \mu^k})$ if $\mu_\kappa = \mu^k$, of $(\hat{S})_h^{\hat{r}}$, $\hat{r} \in (0, (\hat{r})^*(h)]$, starting at $\bar{0}$ and corresponding to $\hat{v}^1, \hat{v}^2, v_1^{\mu^{k-1}+1}, v_{\mu^k}^{\mu^{k-1}+1}, \ldots, v_1^{\mu^k}, v_{\mu^k}^{\mu^k}$ $(\hat{v}^1, \hat{v}^2, v_{1,0}^{\mu^{k-1}+1}, v_{\mu^k,0}^{\mu^{k-1}+1}, \ldots, v_{1,0}^{\mu^k}, v_{\mu^k,0}^{\mu^k}$ if $\mu_\kappa = \mu^k) \in L^p$ with

$$\|\hat{v}^2\|_{L^\infty} + K \sup_{\mu^{k-1}+1 \le \theta \le \mu^k} \|b_\theta\| \sum_{\theta=\mu^{k-1}+1}^{\mu^k} \frac{r_\theta^{k+1}}{\hat{r}} \le \hat{C}_0, \tag{60}$$

since $\|\sigma_\theta(.)\|_{L^\infty} \le K$, $\theta = \mu^{k-1}+1$ to μ^k, $\|\hat{v}^2\|_{L^\infty}$ is to be chosen later and

$$\hat{u}(t-h) = \hat{B}^T e^{\hat{A}h} \hat{Y}(t-h) + \hat{v}^1(t-h), \tag{61}$$

verify

(62)
$$\|\hat{Y}_{\hat{\Pi}}\|_{L^p} \le \hat{M}_p \begin{cases} \left[\|\hat{v}^1\|_{L^p} + \|\hat{v}^2\|_{L^p} + \sum_{\theta=\mu^{k-1}}^{\mu^k} \left(\|v_1^\theta\|_{L^p} + \|v_{\mu^k}^\theta\|_{L^p}\right)\right] & \text{if } \mu_\kappa \neq \mu^k, \\ \left[\|\hat{v}^1\|_{L^p} + \|\hat{v}^2\|_{L^p} + \sum_{\theta=\mu^{k-1}}^{\mu^k} \left(\|v_{1,0}^\theta\|_{L^p} + \|v_{\mu^k,0}^\theta\|_{L^p}\right)\right] & \text{if } \mu_\kappa = \mu^k. \end{cases}$$

In addition, the upper bound for the L^p-gain $\hat{M}_p$ of the system $(\hat{S})_h^{\hat{r}}$, $\hat{r} \le (\hat{r})^*(h)$, is independent of $h > 0$.

In order to satisfy (60), we first give the following bound for $\|\hat{v}^2\|_{L^\infty}$:

$$\|\hat{v}^2\|_{L^\infty} \le \frac{\hat{C}_0}{2}. \tag{63}$$

Then, the input bounds $\hat{r}$ and r_θ^{k+1}, $\mu^{k-1}+1 \le \theta \le \mu^k$, must verify

(64)
$$\begin{aligned} &\hat{r} \le (\hat{r})^*(h) \quad \forall\ \mu^{k-1}+1 \le \theta \le \mu^k, r_\theta^{k+1} \le (\tilde{r}_\theta)^{k+1}(h) \text{ and} \\ &K \sup_{\mu^{k-1}+1 \le \theta \le \mu^k} \|b_\theta\| \sum_{\theta=\mu^{k-1}+1}^{\mu^k} \frac{r_\theta^{k+1}}{\hat{r}} \le \frac{\hat{C}_0}{2}. \end{aligned}$$

It is clear that (64) is verified for $\hat{r} \le (\hat{r})^*(h)$ and $r_\theta^{k+1} \le (r_\theta^*)^{k+1}(h)$ for all $\mu^{k-1}+1 \le \theta \le \mu^k$, where

$$(r_\theta^*)^{k+1}(h) = \min\left\{ \frac{\hat{C}_0}{2K\mu^k \sup_{\mu^{k-1}+1 \le \theta \le \mu^k} \|b_\theta\|} (\hat{r})^*(h), (\tilde{r}_\theta)^{k+1} \right\}. \tag{65}$$

Set $r^*(h) := \min\left\{(\hat{r})^*(h), \min_{\mu^{k-1}+1 \le \theta \le \mu^k} (r_\theta^*)^{k+1}(h)\right\}$ and $M_p = \sum_{\theta=\mu^{k-1}+1}^{\mu^k} M_{\theta,p}^{k+1} + \hat{M}_p$. We have shown that there exist a constant $C_0 > 0$, and for every $1 \le p \le \infty$, a constant $M_p > 0$ such that for every $h > 0$ there is an $r^*(h) \in (0,1]$ for which the global system $(S)_h^r$, $r \in (0, r^*(h)]$, with the feedback given by (58) and (61), is finite-gain L^p-stabile; i.e., the trajectories $[((Y_{\mu^k}^\theta)_{v_1^\theta, v_{\mu^k}^\theta})_{\mu^{k-1}+1 \le \theta \le \mu^k}, (\hat{Y})_{\hat{\Pi}}]$ $([((Y_{\mu^k}^\theta)_{v_{1,0}^\theta, v_{\mu^k,0}^\theta})_{\mu^{k-1}+1 \le \theta \le \mu^k}, (\hat{Y})_{\hat{\Pi}}]$, respectively), $r \in (0, r^*(h)]$, starting at $\bar{0}$ and corresponding to $[v_1^{\mu^{k-1}+1}, v_{\mu^k}^{\mu^{k-1}+1}, \ldots, v_1^{\mu^k}, v_{\mu^k}^{\mu^k}, \hat{v}^1, \hat{v}^2]$, $([v_{1,0}^{\mu^{k-1}+1}, v_{\mu^k,0}^{\mu^{k-1}+1}, \ldots, v_{1,0}^{\mu^k}, v_{\mu^k,0}^{\mu^k},$

$\hat{v}^1, \hat{v}^2]$, respectively) $\in L^p$, with

$$\begin{array}{l} \|\hat{v}^2\|_{L^\infty} + \sum_{\theta=\mu^{k-1}+1}^{\mu^k} \left(\|v^\theta_{\mu^k}\|_{L^\infty} + \|b_\theta\| \right) \leq C_0 \text{ if } \mu^k \neq \mu_\kappa, \\ \|\hat{v}^2\|_{L^\infty} + \sum_{\theta=\mu^{k-1}+1}^{\mu^k} \left(\|v^\theta_{\mu^k,0}\|_{L^\infty} + \|b_\theta\| \right) \leq C_0 \text{ if } \mu^k = \mu_\kappa, \end{array} \tag{66}$$

verify

$$\left\{ \begin{array}{l} \left\| \left((Y^\theta_{\mu^k})_{v^\theta_1, v^\theta_{\mu^k}} \right)_{\mu^{k-1}+1 \leq \theta \leq \mu^k}, (\hat{Y})_{\hat{\Pi}} \right\|_{L^p} \leq M_p \Big\{ \sum_{\theta=\mu^{k-1}+1}^{\mu^k} \left(\|v^\theta_1\|_{L^p} + \|v^\theta_{\mu^k}\|_{L^p} \right) \\ \qquad\qquad + \|\hat{v}^1\|_{L^p} + \|\hat{v}^2\|_{L^p} \Big\} \text{ if } \mu_\kappa \neq \mu^k, \\ \left\| \left((Y^\theta_{\mu^k})_{v^\theta_{1,0}, v^\theta_{\mu^k,0}} \right)_{\mu^{k-1}+1 \leq \theta \leq \mu^k}, (\hat{Y})_{\hat{\Pi}} \right\|_{L^p} \leq M_p \Big\{ \sum_{\theta=\mu^{k-1}+1}^{\mu^k} \left(\|v^\theta_{1,0}\|_{L^p} + \|v^\theta_{\mu^k,0}\|_{L^p} \right) \\ \qquad\qquad + \|\hat{v}^1\|_{L^p} + \|\hat{v}^2\|_{L^p} \Big\} \text{ if } \mu_\kappa = \mu^k, \end{array} \right. \tag{67}$$

where the upper bound for the corresponding global L^p-gain, M_p, is delay-independent.

The proof of Proposition 1 is complete. □

3.3. Construction of the feedback. In this paragraph we provide an explicit expression of the L^p-stabilizing feedback for the system $(S)^r_h$, namely, the solution to the problem.

PROPOSITION 2. *The bounded state feedback*

$$u(.) = \left(\begin{array}{l} (b^{(l+1)}_{0,1})^T e^{A^{(l+1)}_{0,0} h} Y^\theta_{\mu^0,\ldots,\mu^k}(.) + v^0_{1,\theta}(.), \quad 1 \leq \theta \leq \mu^0, \\ (b^{(\chi^1+\cdots+\chi^k)}_{0,1})^T e^{A^{(\chi^1+\cdots+\chi^k)}_{0,0} h} Y^\theta_{\mu^1,\ldots,\mu^k}(.) + v^1_{1,\theta}(.), \quad \mu^0+1 \leq \theta \leq \mu^1, \\ \vdots \\ (b^{(\chi^{j_0}+\cdots+\chi^k)}_{0,1})^T e^{A^{(\chi^{j_0}+\cdots+\chi^k)}_{0,0} h} Y^\theta_{\mu^{j_0},\ldots,\mu^k}(.) + v^{j_0}_{1,\theta}(.), \ \mu^{j_0-1}+1 \leq \theta \leq \mu^{j_0}, \\ (b^{(\chi^{j_0-1}+\cdots+\chi^k)}_{0})^T e^{A^{(\chi^{j_0-1}+\cdots+\chi^k)}_{0} h} Y^\theta_{\mu^{j_0-1},\ldots,\mu^k}(.) + v^{j_0-1}_{1,\theta}(.), \\ \qquad\qquad \mu^{j_0-2}+1 \leq \theta \leq \mu^{j_0-1}, \\ \vdots \\ (b^{(\chi^{k-1}+\chi^k)}_{0})^T e^{A^{(\chi^{k-1}+\chi^k)}_{0} h} Y^\theta_{\mu^{k-1},\mu^k}(.) + v^{k-1}_{1,\theta}(.), \quad \mu^{k-2}+1 \leq \theta \leq \mu^{k-1}, \\ (b^{(\chi^k)}_{0})^T e^{A^{(\chi^k)}_{0} h} Y^\theta_{\mu^k}(.) + v^k_{1,\theta}(.), \quad \mu^{k-1}+1 \leq \theta \leq \mu^k, \end{array} \right) \tag{68}$$

finite-gain L^p-stabilizes the system $(S)^r_h : \dot{x}(t) = Ax(t) - rB\sigma(u(t-h)) + rv(t-h)$, where $v = (v^{\theta^0}_{\mu^0,\ldots,\mu^k}, \ldots, v^{\theta^k}_{\mu^k}) \in L^p$, $\mu^{j-1}+1 \leq \theta^j \leq \mu^j$ for all $0 \leq j \leq k$; i.e., the

trajectories $((Y^{\theta^0}_{\mu^0,\dots,\mu^k})_{\Pi_1},\dots,(Y^{\theta^k}_{\mu^k})_{\Pi_{k+1}})$, $\Pi_\nu := (v^{\nu-1}_{1,\theta},\, v^{\theta}_{\mu^{\nu-1},\dots,\mu^k},\Pi^T_{\nu+1})^T$, $1\le\nu\le k+1$, $r\in(0,\,r^*(h)]$, *starting at* $\bar 0$ *and corresponding to* $v^0_{1,\theta},v^{\theta}_{\mu^0,\dots,\mu^k},\dots,v^k_{1,\theta},v^{\theta}_{\mu^k}\in L^p$, *verify, for all* $j=0,\dots,k$,

$$\left\|\left((Y^{\theta^0}_{\mu^0,\dots,\mu^k})_{\Pi_1},\dots,(Y^{\theta^k}_{\mu^k})_{\Pi_{k+1}}\right)\right\|_{L^p}\le M_p\sum_{\theta^j=\mu^{j-1}+1}^{\mu^j}\left(\|v^0_{1,\theta^j}\|_{L^p}+\|v^{\theta^j}_{\mu^0,\dots,\mu^k}\|_{L^p}\right),$$

where the upper bound for the corresponding global L^p*-gain* M_p *is delay-independent.*

Appendix A. Proof of Lemma 1. Since the pair (A,b) is controllable and b is a column vector, it is clear that A has n distinct eigenvalues. So, from [5, Theorem XII.1], there exists an $r_1^*\in(0,1)$ such that $A_r=A-rbb^T$ has n different analytic eigenvalues $(\lambda_j(r))_{1\le j\le n}$ for every $r\in(0,r_1^*]$, with $i\omega_j=\lambda_{2j-1}(0)$, $-i\omega_j=\lambda_{2j}(0)$, $1\le j\le[n]$, and $\lambda_n(0)=0$ if n is odd. Note that one can choose the ω_j's to be all positive, except the last one if n is odd. Moreover, A verifies (51), with two-dimensional diagonal blocks equal to ω_jA_0 if $\omega_j>0$ and equal to zero otherwise. Set $D_0:=OAO^T$.

We first show that there exist n positive constants C_j, $1\le j\le n$, and $0<r_2^*\le r_1^*$, such that, for every $r\in(0,r_2^*]$ and $1\le j\le\left[\frac{n}{2}\right]$, we have

$$\lambda_{2j-1}(r)=-(C_j+\varepsilon^1_j(r))r+i(\omega_j+\varepsilon^2_j(r)) \tag{69}$$

and

$$\lambda_{2j}(r)=-(C_j+\varepsilon^1_j(r))r-i(\omega_j+\varepsilon^2_j(r)) \tag{70}$$

(with the obvious extension for $\lambda_n(r)$ in the case n odd), where the functions $\varepsilon^l_j(r)$, $1\le j\le n$, and $l=1,2$, are real-valued and tend to zero as r tends to zero.

We next finish the proof of Lemma 1 assuming that (69) and (70) hold. First, it is clear that, when n is odd, $\omega_n=0$, and in that case, $\varepsilon^2_n\equiv0$. In addition, we deduce, by a simple continuity argument, that the analytic matrix function $A_r:=A-rbb^T$ has distinct eigenvalues for $r\in[0,r_1^*)$ and is thus diagonalizable over C. Therefore, one has $Q(r)A(r)Q(r)^{-1}=D(r)$, where

$Q(0)=O$ and $r\mapsto Q(r)$ is real-valued and continuous; therefore, $Q(r)=O+q(r)$ with $q(r)$ tending to zero as r tends to zero;

$D(0)=D_0$, $r\mapsto D(r)$ is real-valued, continuous, and block diagonal with two-dimensional diagonal blocks equal to $\mathrm{Re}(\lambda_j(r))I_2+Im(\lambda_j(r))A_0$ if $\lambda_j(0)=i\omega_j\neq0$ and equal to the real number $\lambda_j(r)$ if $\lambda_j(0)=0$.

It follows that $Q(r)e^{A(r)t}Q(r)^{-1}=e^{D(r)t}$, and an easy computation yields (5).

One clearly has that $Q(r)(Q(r)^T)^{-1}=I_n+\tilde q(r)$, with $\tilde q(r)$ tending to zero as r tends to zero. For $t\ge0$ and $r\in[0,r_1^*)$, one has

$$Q(r)e^{A(r)t}e^{A(r)^Tt}(Q(r)^T)^{-1}=e^{D(r)t}Q(r)(Q(r)^T)^{-1}e^{D(r)^Tt}=e^{D(r)t}e^{D(r)^Tt}+R(r,t),$$

where $R(r,t)=e^{D(r)t}\tilde q(r)e^{D(r)^Tt}$. In addition, $D(r)$ commutes with $D(r)^T$ and, then, $e^{D(r)t}e^{D(r)^Tt}$ is equal to $e^{(D(r)+D(r)^T)t}$, which is block diagonal with two-dimensional diagonal blocks equal to $e^{2\mathrm{Re}(\lambda_j(r)t}I_2$ if $\lambda_j(0)=i\omega_j\neq0$ and the real number $e^{2\lambda_j(r)t}$ if $\lambda_j(0)=0$. Taking now into account (69) and integrating over R^+, one gets that

$$Q(r)P_r(Q(r)^T)^{-1}=P^1(r)+P^2(r),$$

so that

$P^1(r) = \int_0^\infty e^{D(r)t} e^{D(r)^T t} dt$ where $P^1(r)$ is a diagonal matrix where every diagonal element is bigger than $\frac{C_1}{r}$ for some $C_1 > 0$;
$P^2(r) = \int_0^\infty R(r,t)dt$ with $\|P^2(r)\| \leq \frac{\bar{q}(r)}{r}$, $\bar{q}(r)$ tending to zero as r tends to zero. It is therefore clear that, for r small enough, (7) follows readily.

We now prove (69). Since A is skew-symmetric, there exists an $r_2^* \in (0,1]$ such that for all $r \in (0, r_2^*]$,

$$e^{(A-rbb^T)t} = \sum_{j=1}^{N} M_j(r,t) e^{\lambda_j(r)t}, \tag{71}$$

with $N = [\frac{n+1}{2}]$. As for all $r \in (0, r_3^* = \min\{r_1^*, r_2^*\}]$, $\mathrm{Re}(\lambda_j(r)) < 0$ for $j = 1, \ldots, N$ and $M_j(r,t), j = 1, \ldots, N$, are uniformly bounded (since $A - rbb^T$ is Hurwitz and A is skew-symmetric). It turns out that we have the desired result for every $r \in (0, r_3^*]$, provided that

$$-C_2' \leq \frac{\mathrm{Re}(\lambda_j(r))}{r} \leq -C_2 < 0, \qquad j = 1, \ldots, N,$$

uniformly for some positive constants C_2', $C_2 > 0$ independent of r. Indeed this is true if

$$\mathrm{Re}\left(\frac{\partial \lambda_j(r)}{\partial r}\right)_{|r=0} = \frac{\partial \mathrm{Re}(\lambda_j(r))}{\partial r}_{|r=0} \neq 0, \qquad j = 1, \ldots, N.$$

We have

$$\mathcal{X}_{A-rbb^T}(s) = \det\,(A - rbb^T - sI) = \mathcal{X}_A(s) + r\mathcal{X}_{\bar{A}}(s), \tag{72}$$

with

$$\mathcal{X}_{\bar{A}}(s) = \mathcal{X}_{A-bb^T}(s) - \mathcal{X}_A(s).$$

Set $\lambda_j(0) := iw_j$, $i^2 = -1$, $w_j \geq 0$, $1 \leq j \leq N$, as simple roots of A; then $\mathcal{X}_A'(iw_j) \neq 0, 1 \leq j \leq N$.

For $1 \leq j \leq N$, the relation $\mathcal{X}_{A-rbb^T}(\lambda_j(r)) = 0$ leads to $\mathcal{X}_A(\lambda_j(r)) + r\mathcal{X}_{\bar{A}}(\lambda_j(r)) = 0$, and then

$$\frac{\mathcal{X}_A(\lambda_j(r)) - \mathcal{X}_A(\lambda_j(0))}{r} = -\mathcal{X}_{\bar{A}}(\lambda_j(r)). \tag{73}$$

On the other hand,

$$\mathcal{X}_A(\lambda_j(r)) = \mathcal{X}_A(\lambda_j(0)) + r\mathcal{X}_A'(\lambda_j(0))\lambda_j'(0) + o(r),$$

where $o(r) \to 0$ as $r \to 0$. From (73), it follows that

$$\mathcal{X}_A'(\lambda_j(0))\lambda_j'(0) + o(1) = -\mathcal{X}_{\bar{A}}(\lambda_j(r)),$$

where $o(1) \to 0$ as $r \to 0$. Now, if r converges to zero, we get

$$\lambda_j'(0) = -\frac{\mathcal{X}_{\bar{A}}(i\omega_j)}{\mathcal{X}_A'(i\omega_j)} = -\frac{\mathcal{X}_{A-bb^T}(i\omega_j)}{\mathcal{X}_A'(i\omega_j)} \quad \text{for every } j = 1, \ldots, N.$$

This completes the proof of (69) since $A - bb^T$ is Hurwitz. ∎

Appendix B. Canonical form. Since A is skew-symmetric, there exists an orthogonal $n \times n$ matrix O such that

$$OAO^T = \begin{pmatrix} A_{0,\mu_0} & & & & \\ & \ddots & & & \\ & & A_{0,\mu_\kappa} & & (0) \\ & & & \ddots & \\ & (0) & & & A_{0,\mu_l} \end{pmatrix}, \tag{74}$$

where, for every $0 \le j \le l$, $j \ne \kappa$, A_{0,μ_j} is a $2\mu_j \times 2\mu_j$ matrix in the form

$$A_{0,\mu_j} = \begin{pmatrix} \omega_j A_0 & 0 & \dots & 0 \\ 0 & \omega_j A_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \omega_j A_0 \end{pmatrix}, \quad \text{with } A_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and A_{0,μ_κ} is a $\mu_\kappa \times \mu_\kappa$ zero matrix.

So, up to an orthonormal change of basis, we can assume that $A = \text{diag}\,(A_{0,\mu_j})$, $1 \le j \le l$. Consider now $OX = (X_0, \dots, X_\kappa, \dots, X_l)^T$ and $OB = (B_0, \dots, B_\kappa, \dots, B_l)^T$, where $X_j \in R^{2\mu_j}, X_\kappa \in R^{\mu_\kappa}$, the size of B_j is $2\mu_j \times m$, and the size of B_κ is $\mu_\kappa \times m$ for all $0 \le j \le l$, $j \ne \kappa$. Then, for all $0 \le j \le l$, the system $(S)_h^r$ becomes

$$\dot{X}_j = A_{0,\mu_j} X_j - rB_j \sigma\Big(F_{j,h}^T X_j(t-h) + u_{1,j}(t-h)\Big) + ru_{2,j}(t-h), \tag{75}$$

where $F_{j,h} \in R^{2\mu_j \times m}, F_{\kappa,h} \in R^{\mu_\kappa \times m}, u_{1,j} \in R^m, u_{1,\kappa} \in R^m, u_{2,j} \in R^{2\mu_j}$, and $u_{2,\kappa} \in R^{\mu_\kappa}$ for all $0 \le j \le l$, $j \ne \kappa$, are to be chosen later. Then, it is easy to see that

$$(A, B) \text{ is controllable iff } (A_{0,\mu_j}, B_j) \text{ is for } 0 \le j \le l, j \ne \kappa, \text{ and rank } B_\kappa = \mu_\kappa.$$

Indeed, for all $0 \le j \le l$, $j \ne \kappa$, let $R(A_{0,\mu_j}, B_j)$ be the controllability matrix of the pair (A_{0,μ_j}, B_j). We have $R(A_{0,\mu_j}, B_j) = [B_j, A_{0,\mu_j} B_j]$ since $A_{0,\mu_j}^2 = -\omega_j^2 I_{2\mu_j}$. Then, (A_{0,μ_j}, B_j) is controllable if and only if $\text{rank} R(A_{0,\mu_j}, B_j) = \text{rank}[B_j, A_{0,\mu_j} B_j] = 2\mu_j$.

Let $B = (B_0, \dots, B_\kappa, \dots, B_l)^T = (b_1, \dots, b_m) \in R^{n \times m}$, with for each $1 \le \theta \le m, b_\theta = (b_0^\theta, \dots, b_\kappa^\theta, \dots, b_l^\theta)^T \in R^n$, and $b_j^\theta = (b_{1,j}^\theta, b_{2,j}^\theta, \dots, b_{\mu_j,j}^\theta)^T \in R^{2\mu_j}, 0 \le j \le l$, $j \ne \kappa$, and $b_\kappa^\theta = (a_1^\theta, \dots, a_{\mu_\kappa}^\theta)^T \in R^{\mu_\kappa}$. In the next lemma, we consider a linear transformation which is crucial for the analysis of $(S)_h^r$.

LEMMA 3. *For each $\mu_{k-1} + 1 \le \theta \le \mu_k$, $0 \le k \le l$ fixed, with $\mu_{-1} = 0$, there exists $\alpha^\theta := (\alpha_\theta^\theta, \alpha_{\theta+1}^\theta, \dots, \alpha_m^\theta) \in R^{m-\theta+1}$ such that*

$$\tilde{b}_j^\theta := \sum_{s=\theta}^{m} \alpha_s^\theta b_j^s \ne 0 \quad \forall k \le j \le l. \tag{76}$$

Remark 6. This change of coordinates transforms the matrix B into $\tilde{B} = (\tilde{B}_0, \dots, \tilde{B}_\kappa, \dots, \tilde{B}_l)^T = (\tilde{b}_1, \tilde{b}_2, \dots, \tilde{b}_{\mu_l}, b_{\mu_l+1}, \dots, b_m)$ with, for any $1 \le \theta \le \mu_l$, $\tilde{b}_\theta = (\tilde{b}_0^\theta, \dots, \tilde{b}_\kappa^\theta, \dots, \tilde{b}_l^\theta)^T$and for each $0 \le j \le l, j \ne \kappa, \tilde{b}_j^\theta = (\tilde{b}_{1,j}^\theta, \tilde{b}_{2,j}^\theta, \dots, \tilde{b}_{\mu_j,j}^\theta)^T \ne 0$ and $\tilde{b}_\kappa^\theta = (\tilde{a}_1^\theta, \dots, \tilde{a}_{\mu_\kappa}^\theta)^T \ne 0$.

Proof. There exists $T \in GL(R^m)$ such that $BT = \tilde{B}$ with

$$T = \begin{pmatrix} \alpha_1^1 & 0 & \dots & 0 \\ \alpha_2^1 & \alpha_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_m^1 & \alpha_m^2 & \dots & \alpha_m^m \end{pmatrix}.$$

The proof is by induction in the number of distinct algebraic multiplicity $\mu_j, 0 \leq j \leq l$, of the eigenvalues of A, i.e., in $\bar{\delta} := 1 + \#\{\delta_k > 0, 1 \leq k \leq l\}$, where $\delta_k := \mu_k - \mu_{k-1}$ for all $1 \leq k \leq l$, since we have the controllability property if and only if (A_{0,μ_j}, B_j), $0 \leq j \leq l$, $j \neq \kappa$, are controllable and $\mathrm{rank}B_\kappa = \mu_\kappa$. First, let $\mu^0 < \mu^1 < \dots < \mu^\psi$, $0 \leq \psi \leq l$, denote the distinct algebraic multiplicities of the eigenvalues of A. Moreover, $\bar{\delta} = 1 + \psi$.

Step 1. Verify that the desired result holds for $\bar{\delta} = 1$; nevertheless $\psi = 0$. Here, when 1 is substituted for $\bar{\delta}$, this gives $\mu_0 = \dots = \mu_\kappa = \dots = \mu_l := \rho$. Specifically, the conclusion of Lemma 3 becomes the following: For each $1 \leq \theta \leq \rho$ fixed, there exists $\alpha^\theta := (\alpha_\theta^\theta, \alpha_{\theta+1}^\theta, \dots, \alpha_m^\theta) \in R^{m-\theta+1}$ such that $\tilde{b}_j^\theta := \sum_{s=\theta}^m \alpha_s^\theta b_j^s \neq 0$ for all $0 \leq j \leq l$. So, $\tilde{B} := (\tilde{b}_1, \dots, \tilde{b}_\rho, b_{\rho+1}, \dots, b_m) = ((\tilde{b}_0^1, \tilde{b}_1^1, \dots, \tilde{b}_l^1)^T, \dots, (\tilde{b}_0^\rho, \tilde{b}_1^\rho, \dots, \tilde{b}_l^\rho)^T, b_{\rho+1}, \dots, b_m)$ and we will prove the lemma by contradiction. The opposite thesis is as follows: For all $\alpha^\theta = (\alpha_\theta^\theta, \dots, \alpha_m^\theta) = R^{m-\theta+1}$, $1 \leq \theta \leq \rho$ fixed, there exists $j(\alpha^\theta) \in [0, l]$ such that $(\tilde{b}_{j(\alpha^\theta)}^\theta)^{(\alpha^\theta)} := \sum_{s=\theta}^m \alpha_s^\theta b_{j(\alpha^\theta)}^s = 0$.

Let, for all $i \leq j \leq m$, $S_{\theta,j} := \{\alpha^\theta \in R^m, 1 \leq \theta \leq \rho / \quad j(\alpha^\theta) = j\}$; then $R^{m-\theta+1} = \bigcup_{j=0}^m S_{\theta,j}$. So, there exists j_0 such that the interior of S_{θ,j_0} is not empty. So, there exist $\bar{r} > 0$ and $\bar{\alpha^\theta}$ such that $B(\bar{\alpha^\theta}, \bar{r})$, the ball of radius $\bar{r}$ centered at $\bar{\alpha^\theta}$, is included in S_{θ,j_0} for all $1 \leq \theta \leq \rho$. Therefore, for all $\alpha^\theta \in B(\bar{\alpha^\theta}, \bar{r})$, $(\tilde{b}_{j_0}^\theta)^{(\alpha^\theta)} = \sum_{s=\theta}^m \alpha_s^\theta b_{j_0}^s = 0, 1 \leq \theta \leq \rho$.

Writing for all $1 \leq \theta \leq \rho$, $\alpha_s^\theta = \bar{\alpha_s^\theta} + (\alpha_s^\theta - \bar{\alpha_s^\theta})$. It follows that $(\tilde{b}_{j_0}^\theta)^{(\alpha^\theta)} = (\tilde{b}_{j_0}^\theta)^{(\bar{\alpha^\theta})} + (\tilde{b}_{j_0}^\theta)^{(\alpha^\theta - \bar{\alpha^\theta})} = (\tilde{b}_{j_0}^\theta)^{(\alpha^\theta - \bar{\alpha^\theta})} = 0$ for all $1 \leq \theta \leq \rho$. Then, for all $\alpha^\theta \in B(0, \bar{r})$, $(\tilde{b}_{j_0}^\theta)^{(\alpha^\theta)} = 0$, $1 \leq \theta \leq \rho$. So, for all $\alpha^\theta \in R^{m-\theta+1}$, $1 \leq \theta \leq \rho$, $(\tilde{b}_{j_0}^\theta)^{(\alpha^\theta)} := \sum_{s=\theta}^m \alpha_s^\theta b_{j_0}^s = 0$. Then, for all $0 \leq s \leq m$, $b_{j_0}^s = 0$. Moreover, $B_{j_0} = (b_{j_0}^1, \dots, b_{j_0}^{\theta-1}, 0, \dots, 0)$. Then, $\mathrm{rank}B_{j_0} \leq \theta - 1 \leq \rho - 1$. It is impossible since $(A_{0,\mu_{j_0}}, B_{j_0})$ is controllable. So, $\mathrm{rank}B_{j_0} = \mu_{j_0} \geq \rho$. This completes step 1.

Step 2. Assume that the desired result holds for $\bar{\delta} = k$; nevertheless $\psi = k - 1$. Then we have k distinct algebraic multiplicities. Let $\mu^0 < \mu^1 < \dots < \mu^{k-1}$ denote the distinct algebraic multiplicities. Therefore, for every $\mu^{\nu-1} + 1 \leq \theta \leq \mu^\nu$, $0 \leq \nu \leq k-1$ with $\mu^{-1} = 0$, there exists $\alpha^\theta := (\alpha_\theta^\theta, \dots, \alpha_m^\theta) \in R^{m-\theta+1}$ such that $\tilde{b}_j^\theta \neq 0$ for all $\nu \leq j \leq l$. Now, we use the assumption from step 2 to show that the results hold for $\bar{\delta} = k + 1$; nevertheless $\sigma = k$. Then $\mu^0 < \mu^1 < \dots < \mu^k$. That is, we want to show that for every $\mu^{k-1} + 1 \leq \theta \leq \mu^k$, there exists $\alpha^\theta := (\alpha_\theta^\theta, \dots, \alpha_m^\theta) \in R^{m-\theta+1}$ such that $\tilde{b}_j^\theta \neq 0$ for all $k \leq j \leq l$.

According to the controllability property of the pair (A_{0,μ^k}, B_ϖ), B_ϖ is the vector corresponding to the algebraic multiplicity μ^k, which implies that $\mathrm{rank}B_\varpi = \mu^k$; it follows that there exists μ^k free columns or lines. Nevertheless, from the assumption of step 2 we have just $\mu^k - \mu^{k-1}$ independent columns. By applying Lemma 3 for all $\mu^{k-1} + 1 \leq \theta \leq \mu^k$ and discarding the vectors $(\tilde{b}_0^{\mu^{j-1}+1}, \dots, \tilde{b}_0^{\mu^j}; \tilde{b}_1^{\mu^{j-1}+1}, \dots, \tilde{b}_1^{\mu^j}; \dots; \tilde{b}_{j-1}^{\mu^{j-1}+1}, \dots, \tilde{b}_{j-1}^{\mu^j})$ for all $1 \leq j \leq k$, one can prove that for every $\mu^{k-1} + 1 \leq \theta \leq \mu^k$, $\tilde{b}_j^\theta \neq 0$ for all $k \leq j \leq l$. ∎

In order to facilitate the analysis of the finite-gain L^p-stability by bounded feedback of continuous-time delay linear skew-symmetric systems, we provide a linear transformation of the system $(S)_h^r$ which simplifies the analysis of the problem. Our method consists of rewriting the original system $(S)_h^r$ as an appropriate l cascade of single-input subsystems of dimension $2\mu_j, 0 \leq j \leq l$, $j \neq \kappa$, and one subsystem of dimension μ_κ, since A has l conjugate pairs of nonzero purely imaginary eigenvalues of the forms $(\pm i\omega_j)_{1\leq j\leq l}$ and 0 is an eigenvalue of A with multiplicity μ_κ. So, we consider the system $(S)_h^r$ with state space R^n, input space R^m, and delay h. More precisely, we consider the system described by the equations

$$
(S)_h^r := \left\{
\begin{array}{l}
(0)\left\{
\begin{array}{rcl}
\dot{X}_{1,\mu_0} &=& \omega_0 A_0 X_{1,\mu_0} - r\tilde{b}_{1,1}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{1,1}^{\mu_0}\sigma_{\mu_0}(u_{\mu_0}) \\
&& + rv_{1,\mu_0} - r\sum_{i=\mu_0+1}^m \tilde{b}_{1,1}^i\sigma_i(u_i), \\
\vdots & & \\
\dot{X}_{\mu_0,\mu_0} &=& \omega_0 A_0 X_{\mu_0,\mu_0} - r\tilde{b}_{\mu_0,1}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{\mu_0,1}^{\mu_0}\sigma_{\mu_0}(u_{\mu_0}) \\
&& + rv_{\mu_0,\mu_0} - \sum_{i=\mu_0+1}^m \tilde{b}_{\mu_0,1}^i\sigma_i(u_i),
\end{array}\right. \\
(1)\left\{
\begin{array}{rcl}
\dot{X}_{1,\mu_1} &=& \omega_1 A_0 X_{1,\mu_1} - r\tilde{b}_{1,1}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{1,1}^{\mu_1}\sigma_{\mu_1}(u_{\mu_1}) \\
&& + rv_{1,\mu_1} - r\sum_{i=\mu_1+1}^m \tilde{b}_{1,1}^i\sigma_i(u_i), \\
\vdots & & \\
\dot{X}_{\mu_1,\mu_1} &=& \omega_1 A_0 X_{\mu_1,\mu_1} - r\tilde{b}_{\mu_1,1}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{\mu_1,1}^{\mu_1}\sigma_{\mu_1}(u_{\mu_1}) \\
&& + rv_{\mu_1,\mu_1} - \sum_{i=\mu_1+1}^m \tilde{b}_{\mu_1,1}^i\sigma_i(u_i),
\end{array}\right. \\
\vdots \\
(\kappa)\left\{
\begin{array}{rcl}
\dot{x}_{1,\mu_\kappa} &=& - r\tilde{a}_1^1\sigma_1(u_1) - r\tilde{a}_1^2\sigma_2(u_2) - \cdots - r\tilde{a}_1^{\mu_\kappa}\sigma_{\mu_\kappa}(u_{\mu_\kappa}) \\
&& + rv_{1,\mu_\kappa} - r\sum_{i=\mu_\kappa+1}^m \tilde{a}_1^i\sigma_i(u_i), \\
\vdots & & \\
\dot{x}_{\mu_\kappa,\mu_\kappa} &=& - r\tilde{a}_{\mu_\kappa}^1\sigma_1(u_1) - r\tilde{a}_{\mu_\kappa}^2\sigma_2(u_2) - \cdots - r\tilde{a}_{\mu_\kappa}^{\mu_\kappa}\sigma_{\mu_\kappa}(u_{\mu_\kappa}) \\
&& + rv_{\mu_\kappa,\mu_\kappa} - r\sum_{i=\mu_\kappa+1}^m \tilde{a}_{\mu_\kappa}^i\sigma_i(u_i),
\end{array}\right. \\
\vdots \\
(l)\left\{
\begin{array}{rcl}
\dot{X}_{1,\mu_l} &=& \omega_l A_0 X_{1,\mu_l} - r\tilde{b}_{1,l}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{1,l}^{\mu_l}\sigma_{\mu_l}(u_{\mu_l}) \\
&& + rv_{1,\mu_l} - r\sum_{i=\mu_l+1}^m \tilde{b}_{1,l}^i\sigma_i(u_i), \\
\vdots & & \\
\dot{X}_{\mu_l,\mu_l} &=& \omega_l A_0 X_{\mu_l,\mu_l} - r\tilde{b}_{\mu_l,l}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{\mu_l,l}^{\mu_l}\sigma_{\mu_l}(u_{\mu_l}) \\
&& + rv_{\mu_l,\mu_l} - r\sum_{i=\mu_l+1}^m \tilde{b}_{\mu_l,l}^i\sigma_i(u_i),
\end{array}\right.
\end{array}
\right.
$$

where $X_{\theta,\mu_s}, b_{\theta,s}^j, \in R^2$, $1 \leq s \leq \mu$, $1 \leq \theta, j \leq \mu_s$, $x_{\theta,\mu_\kappa}, a_\theta^j \in R$, $1 \leq \theta, j \leq \mu_\kappa$, and

there exists $1 \leq s \leq l$, $s \neq \kappa$, such that $\omega_0 = \omega_s$ since $\omega_\kappa = 0$.

After manipulation of lines and columns for the matrix B, we can suppose that $\tilde{b}^\theta_{\theta,j} \neq 0$ for all $1 \leq \theta \leq \mu_j$, $0 \leq j \leq l$, $j \neq \kappa$ and $\tilde{a}^\theta_\theta \neq 0$ for all $1 \leq \theta \leq \mu_\kappa$, since from Lemma 3, $\tilde{b}^\theta_j = (\tilde{b}^\theta_{1,j}, \ldots, \tilde{b}^\theta_{\mu_j,j})^T$ and $\tilde{b}^\theta_\kappa = (\tilde{a}^\theta_1, \ldots, \tilde{a}^\theta_{\mu_\kappa})^T \neq 0$ are for all $1 \leq \theta \leq \mu_j$ and $0 \leq j \leq l$.

LEMMA 4. *Let $\tilde{b}^\theta_{\theta,s} = \tilde{\xi}^\theta_{\theta,s} V^\theta_{\theta,s} b_0$, $1 \leq \theta \leq \mu_s$, $0 \leq s \leq l$, $s \neq \kappa$, with $V^\theta_{\theta,s} \in SO(2)$ the set of orthogonal matrices and $b_0 = (0 \quad 1)^T$. Then, for each $1 \leq \theta, j \leq \mu_s$, $0 \leq s \leq l$, $s \neq \kappa$, the new variables*

$$
\begin{aligned}
&Y_{\theta,\mu_s}(t) = (V^\theta_{\theta,s})^{-1} U^\theta_{j,s} \tfrac{X_{\theta,\mu_s}(t)}{\tilde{\xi}^\theta_{\theta,s}},\\
&Y_{j,\mu_s}(t) = (V^\theta_{\theta,s})^{-1} \tfrac{X_{j,\mu_s}(t)}{\tilde{\xi}^\theta_{j,s}} - (V^\theta_{\theta,s})^{-1} U^\theta_{j,s} \tfrac{X_{\theta,\mu_s}(t)}{\tilde{\xi}^\theta_{\theta,s}}, \theta \neq j,\\
&y_{\theta,\mu_\kappa}(t) = \tfrac{x_{\theta,\mu_\kappa}(t)}{\tilde{a}^\theta_\theta}, 1 \leq \theta \leq \mu_\kappa,\\
&y_{j,\mu_\kappa}(t) = \tfrac{x_{j,\mu_\kappa}(t)}{\tilde{a}^\theta_j} - \tfrac{x_{\theta,\mu_\kappa}(t)}{\tilde{a}^\theta_\theta}, \quad 1 \leq \theta, j \leq \mu_\kappa, \theta \neq j,
\end{aligned}
\tag{77}
$$

where $U^\theta_{j,s} \in SO(2)$, with $U^\theta_{\theta,s} = Id_2$, transform $(S)^r_h$ into the new system

$$
\begin{aligned}
\dot{Y}_{\theta,\mu_s}(t) &= \omega_s A_0 Y_{\theta,\mu_s}(t) - r b_0 \sigma_\theta(u_\theta(t-h)) + r v_{\theta,\mu_s}(t-h)\\
&\quad - r \textstyle\sum_{j=\mu_s+1}^m \hat{b}^j_{\theta,s} \sigma_j(u_j(t-h)), 1 \leq \theta \leq \mu_s, 0 \leq s \leq l, s \neq \kappa,\\
\dot{y}_{\theta,\mu_\kappa}(t) &= -r\sigma_\theta(u_\theta(t-h)) + r v_{\theta,\mu_\kappa}(t-h) - r \textstyle\sum_{j=\mu_\kappa+1}^m \hat{a}^j_\theta \sigma_j(u_j), 1 \leq \theta \leq \mu_\kappa,
\end{aligned}
\tag{78}
$$

where $\hat{b}^j_{\theta,s} \in R^2$ and $\hat{a}^j_\theta \in R$ are arbitrarily vectors and real numbers.

Proof. For each $0 \leq s \leq l$, $(S)^r_h$ becomes

$$
\left\{
\begin{array}{l}
(S_{\neq 0})^r_h \left\{
\begin{array}{rcl}
\dot{X}_{1,\mu_s} &=& \omega_s A_0 X_{1,\mu_s} - r\tilde{b}^1_{1,s}\sigma_1(u_1) - \cdots - r\tilde{b}^{\mu_s}_{1,s}\sigma_{\mu_s}(u_{\mu_s}) + r v_{1,\mu_s}\\
&& - r\sum_{\theta=\mu_s+1}^m \tilde{b}^\theta_{1,s}\sigma_\theta(u_\theta),\\
\dot{X}_{2,\mu_s} &=& \omega_s A_0 X_{2,\mu_s} - r\tilde{b}^1_{2,s}\sigma_1(u_1) - \cdots - r\tilde{b}^{\mu_s}_{2,s}\sigma_{\mu_s}(u_{\mu_s}) + r v_{2,\mu_s}\\
&& - r\sum_{\theta=\mu_s+1}^m \tilde{b}^\theta_{2,s}\sigma_\theta(u_\theta),\\
&\vdots&\\
\dot{X}_{\mu_s,\mu_s} &=& \omega_s A_0 X_{\mu_s,\mu_s} - r\tilde{b}_{\mu_s,s}{}^1\sigma_1(u_1) - \cdots - r\tilde{b}_{\mu_s,s}{}^{\mu_s}\sigma_{\mu_s}(u_{\mu_s}) + r v_{\mu_s,\mu_s}\\
&& - r\sum_{\theta=\mu_s+1}^m \tilde{b}^\theta_{\mu_s,s}\sigma_\theta(u_\theta),
\end{array}
\right.\\
(S_0)^r_h \left\{
\begin{array}{rcl}
\dot{x}_{1,\mu_\kappa} &=& - r\tilde{a}^1_1\sigma_1(u_1) - r\tilde{a}^2_1\sigma_2(u_2) - \cdots - r\tilde{a}^{\mu_\kappa}_1\sigma_{\mu_\kappa}(u_{\mu_\kappa}) + r v_{1,\mu_\kappa}\\
&& - r\sum_{\theta=\mu_\kappa+1}^m \tilde{a}^\theta_1\sigma_\theta(u_\theta),\\
&\vdots&\\
\dot{x}_{\mu_\kappa,\mu_\kappa} &=& - r\tilde{a}^1_{\mu_\kappa}\sigma_1(u_1) - r\tilde{a}^2_{\mu_\kappa}\sigma_2(u_2) - \cdots - r\tilde{a}^{\mu_\kappa}_{\mu_\kappa}\sigma_{\mu_\kappa}(u_{\mu_\kappa}) + r v_{\mu_\kappa,\mu_\kappa}\\
&& - r\sum_{\theta=\mu_\kappa+1}^m \tilde{a}^\theta_{\mu_\kappa}\sigma_\theta(u_\theta).
\end{array}
\right.
\end{array}
\right.
\tag{79}
$$

We have, for each $1 \leq \theta \leq \mu_s, 0 \leq s \leq l$, $s \neq \kappa$, $\tilde{b}^{\theta}_{\theta,s} \neq 0$, and one can suppose that $\|\tilde{b}^{\theta}_{\theta,s}\| = \|\tilde{b}^{\theta}_{j,s}\|, 1 \leq j \leq \mu_s$ (multiplying the equation (θ) of each subsystem (s) by $\tilde{\lambda}^{\theta}_{j,s} := \frac{\|\tilde{b}^{\theta}_{j,s}\|}{\|\tilde{b}^{\theta}_{\theta,s}\|}$). However, there exists $U^{\theta}_{j,s} \in SO(2)$ such that $U^{\theta}_{j,s}\tilde{b}^{\theta}_{\theta,s} = \tilde{b}^{\theta}_{j,s}$.

We can suppose that $\tilde{b}^{\theta}_{j,s} = 0, 1 \leq \theta, j \leq \mu_s, \theta \neq j$. Indeed, simply multiply successively, for each $0 \leq s \leq l$, $s \neq \kappa$, each equation $(\theta), 1 \leq \theta \leq \mu_s$, by $U^{\theta}_{j,s}$, $1 \leq j \leq \mu_s$, and take the linear change of coordinates given by

$$
\begin{aligned}
&Z_{\theta,\mu_s} = U^{\theta}_{j,s} X_{\theta,\mu_s}, \quad 1 \leq \theta, j \leq \mu_s, \quad 0 \leq s \leq l, \quad s \neq \kappa,\\
&Z_{j,\mu_s} = X_{j,\mu_s} - U^{\theta}_{j,s} X_{\theta,\mu_s}, \quad \theta \neq j.
\end{aligned} \tag{80}
$$

But, we have $\tilde{b}^{\theta}_{\theta,s} = \tilde{\xi}^{\theta}_{\theta,s} V^{\theta}_{\theta,s} b_0$ for all $0 \leq s \leq l$, $s \neq \kappa$. Then, for every $1 \leq \theta \leq \mu_s$, $0 \leq s \leq l$, and $s \neq \kappa$, the linear transformation given by $Y_{\theta,\mu_s} = (V^{\theta}_{\theta,s})^{-1} \frac{Z_{\theta,\mu_s}}{\xi^{\theta}_{\theta,s}}$ transforms (81) into the first block of (78). For the subsystem $(S_0)^r_h$ corresponding to $\omega_\kappa = 0$ as an eigenvalue of A with multiplicity μ_κ,

$$
\left\{
\begin{array}{rcl}
\dot{x}_{1,\mu_\kappa} &=& -r\tilde{a}^1_1\sigma_1(u_1) - r\tilde{a}^2_1\sigma_2(u_2) - \cdots - r\tilde{a}^{\mu_\kappa}_1\sigma_{\mu_\kappa}(u_{\mu_\kappa}) + rv_{1,\mu_\kappa}\\
&& - r\sum_{\theta=\mu_\kappa+1}^{m} \tilde{a}^{\theta}_1\sigma_\theta(u_\theta),\\
\vdots &&\\
\dot{x}_{\mu_\kappa,\mu_\kappa} &=& -r\tilde{a}^1_{\mu_\kappa}\sigma_1(u_1) - r\tilde{a}^2_{\mu_\kappa}\sigma_2(u_2) - \cdots - r\tilde{a}^{\mu_\kappa}_{\mu_\kappa}\sigma_{\mu_\kappa}(u_{\mu_\kappa}) + rv_{\mu_\kappa,\mu_\kappa}\\
&& - r\sum_{\theta=\mu_\kappa+1}^{m} \tilde{a}^{\theta}_{\mu_\kappa}\sigma_\theta(u_\theta),
\end{array}
\right.
$$

taking successively the linear transformation $z_{s,\mu_\kappa} = \frac{x_{s,\mu_\kappa}}{\tilde{a}^{\theta}_{s}}$ for all $1 \leq s \leq \mu_\kappa$ and $1 \leq \theta \leq \mu_\kappa$ and using the second part of the transformation given by (77), we have the second equation in (78). So, $(S_0)^r_h$ is transformed in

$$
\begin{aligned}
&\dot{Z}_{1,\mu_s} = \omega_s A_0 Z_{1,\mu_s} - r\tilde{b}^1_{1,s}\sigma_1(u_1) + rv_{1,\mu_s} - r\sum_{\theta=\mu_s+1}^{m} \tilde{b}^{\theta}_{1,s}\sigma_\theta(u_\theta),\\
&\dot{Z}_{2,\mu_s} = \omega_s A_0 Z_{2,\mu_s} - r\tilde{b}^2_{2,s}\sigma_2(u_2) + rv_{2,\mu_s} - r\sum_{\theta=\mu_s+1}^{m} \tilde{b}^{\theta}_{2,s}\sigma_\theta(u_\theta),\\
&\quad\vdots\\
&\dot{Z}_{\mu_s,\mu_s} = \omega_s A_0 Z_{\mu_s,\mu_s} - r\tilde{b}^{\mu_s}_{\mu_s,s}\sigma_{\mu_s}(u_{\mu_s}) + rv_{\mu-s,\mu_s} - r\sum_{\theta=\mu_s+1}^{m} \tilde{b}^{\theta}_{\mu_s,s}\sigma_\theta(u_\theta). \qquad \square
\end{aligned} \tag{81}
$$

Then, with this linear change of coordinates the matrix B can be transformed into

$$B = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_\kappa \\ \vdots \\ B_{l-1} \\ B_l \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} \begin{pmatrix} b_0 & & \\ (0) & \ddots & (0) \\ & & b_0 \end{pmatrix} & (*)_0 \end{pmatrix} & (**)_0 \\ \begin{pmatrix} \begin{pmatrix} b_0 & & & \\ & \ddots & (0) & \\ & (0) & \ddots & \\ & & & b_0 \end{pmatrix} & (*)_1 \end{pmatrix} & (**)_1 \\ \vdots & \vdots \\ \begin{pmatrix} \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & (0) & \\ & (0) & & \ddots & \\ & & & & 1 \end{pmatrix} & (*)_\kappa \end{pmatrix} & (**)_\kappa \\ \vdots & \vdots \\ \begin{pmatrix} \begin{pmatrix} b_0 & & & & \\ & \ddots & & & \\ & & \ddots & (0) & \\ & (0) & & \ddots & \\ & & & & b_0 \end{pmatrix} & (*)_{l-1} \end{pmatrix} & (**)_{l-1} \\ \begin{pmatrix} b_0 & & & & & \\ & \ddots & & & & \\ & & \ddots & (0) & & \\ & (0) & & \ddots & & \\ & & & & \ddots & \\ & & & & & b_0 \end{pmatrix} & (**)_l \end{bmatrix},$$

where $0 \leq \kappa \leq l$. Here $(*)'_j s$, $(**)'_j s, 0 \leq j \leq l$, $j \neq \kappa$, represent, respectively, submatrices of size $2\mu_j \times (\mu_l - \mu_j)$ and $2\mu_j \times (m - \mu_l)$; $(*)_\kappa$ and $(**)_\kappa$ are submatrices of dimensions $\mu_\kappa \times (\mu_l - \mu_\kappa)$ and $\mu_\kappa \times (m - \mu_l)$.

REFERENCES

[1] X. Bao, Z. Lin, and E. Sontag, *Finite-gain stabilization of discrete-time linear systems subject to actuator saturation*, Automatica J. IFAC, 36 (2000), pp. 269–277.

[2] Y.-Y. Cao, Z. Lin, and T. Hu, *Stability analysis of linear time-delay systems subject to input saturation*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 233–240.

[3] Y. Chitour and Z. Lin, *Finite-gain l_p stabilization of discrete-time linear systems subject to actuator saturation: The case $p = 1$*, IEEE Trans. Automat. Control, 49 (2004), pp. 2196–2198.

[4] W. Liu, Y. Chitour, and E. Sontag, *On finite-gain stabilizability of linear systems subject to input saturation*, SIAM J. Control Optim., 34 (1996), pp. 1190–1219.

[5] M. Reed and B. Simon, *Methods of Modern Mathematical Physics,* IV. *Analysis of Operators*, Academic Press, New York, 1978.

[6] A. Saberi, P. Hou, and A. A. Stoorvogel, *On simultaneous global external and global internal stabilization of critically unstable linear systems with saturating actuators*, IEEE Trans. Automat. Control, 45 (2000), pp. 1042–1052.

[7] E. D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 43 (1998), pp. 93–100.

[8] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite-Dimensional Systems*, 2nd ed., Springer-Verlag, New York, 1998.

[9] H. J. Sussmann, E. D. Sontag, and Y. Yang, *A general result on the stabilization of linear systems using bounded controls*, IEEE Trans. Automat. Control, 39 (1994), pp. 2411–2425.

[10] S. Tarbouriech and J. M. Gomes da Silva, *Synthesis of controllers for continuous-time delay systems with saturating controls via LMIs,* IEEE Trans. Automat. Control, 45 (2000), pp. 105–111.

[11] A. J. Van Der Schaft, *L^2-Gain and Passivity Techniques in Nonlinear Control*, Lecture Notes in Control and Inform. Sci. 218, Springer, London, 1996.

[12] K. Yakoubi and Y. Chitour, *Linear systems subject to input saturation and time delay: Global asymptotic stabilization*, IEEE Trans. Automat. Control, submitted.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 3, pp. 1116–1146

NONOCCURRENCE OF THE LAVRENTIEV PHENOMENON FOR MANY OPTIMAL CONTROL PROBLEMS*

ALEXANDER J. ZASLAVSKI†

Abstract. In this paper we study nonoccurrence of the Lavrentiev phenomenon for a large class of nonconvex optimal control problems identified with the corresponding complete metric space of integrands $\mathcal{M}$, which satisfy a growth condition common in the literature and are Lipschitzian on bounded sets. We establish that for most elements of $\mathcal{M}$ (in the sense of Baire category) the infimum on the full admissible class of trajectory-control pairs is equal to the infimum on a subclass of trajectory-control pairs whose controls are bounded by a certain constant.

Key words. Banach space, integrand, Lavrentiev phenomenon, optimal control problem

AMS subject classifications. 49J15, 49J27

DOI. 10.1137/050640370

1. Introduction, assumptions, and main results. In this paper we study nonoccurrence of the Lavrentiev phenomenon for a large class of nonconvex optimal control problems with a state variable and a control variable belonging to Banach spaces. We say that the Lavrentiev phenomenon occurs for an optimal control problem if its infimum on the full admissible class of trajectory-control pairs is less than its infimum on a subclass of trajectory-control pairs with bounded controls.

The Lavrentiev phenomenon in the calculus of variations was discovered in 1926 by M. Lavrentiev in [13]. There it was shown that it is possible for the variational integral of a two-point Lagrange problem, which is sequentially weakly lower semicontinuous on the admissible class of absolutely continuous functions, to possess an infimum on the dense subclass of C^1 admissible functions that is strictly greater than its minimum value on the admissible class. Since this seminal work, the Lavrentiev phenomenon has been of great interest in the calculus of variations [1], [2], [6], [7], [9], [10], [11], [12], [14] [15], [16], [18], [19], [20]. The original example of Lavrentiev was simplified by Mania in [15]. In [6], [7] Ball and Mizel demonstrated that the Lavrentiev phenomenon can occur with fully regular integrands. Sarychev [18] constructed a broad class of integrands that exhibit the Lavrentiev phenomenon. Nonoccurrence of the Lavrentiev phenomenon was studied in [1], [2], [10], [11], [12], [14], [19], [20], [22], [24]. Clarke and Vinter [10] showed that the Lavrentiev phenomenon cannot occur when a variational integrand $f(t, x, u)$ is independent of t. Sychev and Mizel [20] considered a class of integrands $f(t, x, u)$, which are convex with respect to the last variable. For this class of integrands they established that the Lavrentiev phenomenon does not occur. Sarychev and Torres [19] studied a class of optimal control problems with control-affine dynamics and with continuously differentiable integrands $f(t, x, u)$. For this class of problems they established Lipschitzian regularity of minimizers, which implies nonoccurrence of the Lavrentiev phenomenon. Ferriero [12] showed that the Lavrentiev phenomenon cannot occur for a class of higher-order variational problems with integrands which are convex with respect to the last variable.

*Received by the editors September 15, 2005; accepted for publication (in revised form) May 15, 2006; published electronically August 25, 2006.
http://www.siam.org/journals/sicon/45-3/64037.html
†Department of Mathematics, The Technion-Israel Institute of Technology, 32000 Haifa, Israel (ajzasl@tx.technion.ac.il).

In [24] we studied nonoccurrence of the Lavrentiev phenomenon for a large class of nonconvex nonautonomous constrained variational problems with integrands which belong to a complete metric space of functions. We established that for most integrands in the sense of the Baire category the Lavrentiev phenomenon does not occur [24]. In the present paper we consider a large class of optimal control problems identified with the corresponding complete metric space of integrands $\mathcal{M}_B$, which satisfy a growth condition common in the literature and satisfy a Lipschitz condition on bounded sets. We establish that for most elements of $\mathcal{M}_B$ (in the sense of the Baire category) the infimum on the full admissible class of trajectory-control pairs is equal to the infimum on a subclass of trajectory-control pairs whose controls are bounded by a certain constant.

In this paper we use the following definitions and notation. For each $t \in R^1$, set $\exp(t) = e^t$. For each function $h : Z \to R^1 \cup \{\infty\}$, where Z is nonempty, and each nonempty subset $Y \subset Z$, set

$$\inf(h) = \inf\{h(z) : \ z \in Z\}, \ \inf(h; Y) = \inf\{h(z) : \ z \in Y\}.$$

Denote by $\mathrm{mes}(\Omega)$ the Lebesgue measure of a Lebesgue measurable set $\Omega \subset R^1$. If $(Z, ||\cdot||)$ is a normed space, then for each $z \in Z$ and each $r > 0$, set

$$B_Z(z, r) = \{y \in Z : \ ||y - z|| \le r\}, \ B_Z(r) = B_Z(0, r).$$

We say that a property of elements of a complete metric space Z is generic (typical) in Z if the set of all elements of Z which possess this property contains an everywhere dense G_δ subset of Z. In this case we also say that the property holds for a generic (typical) element of Z or that a generic (typical) element of Z possesses the property [5], [17], [23].

Assume that $(E, ||\cdot||)$, $(F, ||\cdot||)$ are Banach spaces.

Let $-\infty < \tau_1 < \tau_2 < \infty$. Denote by $W^{1,1}(\tau_1, \tau_2; E)$ the set of all functions $x : [\tau_1, \tau_2] \to E$ for which there exists a Bochner integrable function $u : [\tau_1, \tau_2] \to E$ such that

$$x(t) = x(\tau_1) + \int_{\tau_1}^{t} u(s)ds, \ t \in (\tau_1, \tau_2]$$

(see, e.g., Brezis [8]). It is known that if $x \in W^{1,1}(\tau_1, \tau_2; E)$, then the equation above defines a unique Bochner integrable function u, which is called the derivative of x and is denoted by x'.

We also use the notation $W^{1,1}(\Omega; E) = W^{1,1}(\tau_1, \tau_2; E)$, where $\Omega = [\tau_1, \tau_2]$.

Let $0 \le T_1 < T_2 < \infty$ and let $U : [T_1, T_2] \to 2^F \setminus \{\emptyset\}$ be a set-valued mapping.

Denote by X the set of all pairs of functions (x, u), where $x \in W^{1,1}(T_1, T_2; E)$ and $u : [T_1, T_2] \to F$ is a strongly measurable function which satisfies

$$u(t) \in U(t), \ t \in [T_1, T_2] \text{ almost everywhere (a.e.).} \tag{1.1}$$

To be more precise, we have to define elements of X as classes of pairs equivalent in the sense that (x_1, u_1) and (x_2, u_2) are equivalent if and only if $x_2(t) = x_1(t)$ for all $t \in [T_1, T_2]$ and $u_1(t) = u_2(t)$ for almost every $t \in [T_1, T_2]$.

Denote by $\mathcal{D}(E)$ the collection of all nonempty bounded subsets of E.

Let $G : [T_1, T_2] \times E \times F \to E$ be a continuous mapping, $\phi : [0, \infty) \to [0, \infty)$ be an increasing function, and c_0 be a positive number such that

$$\lim_{t \to \infty} \phi(t)/t = \infty, \tag{1.2}$$

$$||G(t, y_1, v) - G(t, y_2, v)|| \le c_0 ||y_1 - y_2|| \phi(||v||) \tag{1.3}$$

for each $t \in [T_1, T_2]$, each $y_1, y_2 \in E$, and each $v \in F$ and that

$$||G(t, y, v_1) - G(t, y, v_2)|| \le c_0||v_1 - v_2||\phi(||y||) \tag{1.4}$$

for each $t \in [T_1, T_2]$, each $y \in E$, and each $v_1, v_2 \in F$.

It is clear that the mapping G is bounded in the norm topology on any bounded subset of $[T_1, T_2] \times E \times F$.

We assume that there exist a Bochner integrable function $u_*(t)$, $t \in [T_1, T_2]$, and $N_* > 0$ such that

$$u_*(t) \in U(t),\ t \in [T_1, T_2] \text{ a.e. and } ||u_*(t)|| \le N_*,\ t \in [T_1, T_2] \text{ a.e..} \tag{1.5}$$

For each $D \in \mathcal{D}(E)$ denote by $X(G, D)$ the set of all pairs $(x, u) \in X$ such that

$$x(T_1) \in D \text{ and } x'(t) = G(t, x(t), u(t)),\ t \in [T_1, T_2] \text{ a.e..} \tag{1.6}$$

For each $D \in \mathcal{D}(E)$ denote by $X_L(G, D)$ the set of all $(x, u) \in X(G, D)$ such that $||u(t)|| \le M_u$ for $t \in [T_1, T_2]$ a.e., where M_u is a positive constant depending on u.

In this paper we study the nonoccurrence of the Lavrentiev phenomenon for the optimal control problem

$$\int_{T_1}^{T_2} f(t, x(t), u(t))dt \to \min,\ (x, u) \in X(G, D), \tag{P}$$

where $D \in \mathcal{D}(E)$ and f belongs to one of the two complete metric spaces of integrands $\mathcal{M}_A$ and $\mathcal{M}_B$, which will be described below.

Comparing our work with the results of [12], [19], note that in our paper E and F are general Banach spaces, that we do not impose any convexity assumption and any differentiability assumption on integrands f, that the control u is constrained and the right-hand side of differential equations is determined by mappings $G : [T_1, T_2] \times E \times F \to E$ belonging to a large general class of continuous mappings, while in [12], [19] the following assumptions are imposed:

- The spaces E and F are finite-dimensional;
- in [19] integrands are continuously differentiable and in [12] integrands are convex with respect to the last variable;
- the control u is unconstrained;
- $G(t, x, u) = g_1(t, x) + g_2(t, x)u$ in [19] and $G(t, x, u) = u$ in [12].

Now we define the spaces of integrands $\mathcal{M}_A$ and $\mathcal{M}_B$ considered in our paper.

Let $\lambda_0(t)$, $\lambda_1(t)$, $t \in [T_1, T_2]$ be integrable scalar nonnegative functions and let c_1, c_2 be positive constants.

Denote by $\mathcal{M}_A$ the set of all functions $f : [T_1, T_2] \times E \times F \to R^1$ which satisfy the following assumptions:

(A1) f is measurable with respect to the σ-algebra generated by products of Lebesgue measurable subsets of $[T_2, T_2]$ and Borel subsets of $E \times F$.

(A2) For each $M > 0$ there is $M_0 > 0$ such that for almost every $t \in [T_1, T_2]$ the inequality $|f(t, x, u)| \le M_0$ holds for each $x \in B_E(M)$ and each $u \in B_F(M)$.

(A3) For each $\epsilon, M > 0$ there exists $\delta > 0$ such that for almost every $t \in [T_1, T_2]$ the inequality $|f(t, x_1, u) - f(t, x_2, u)| \le \epsilon$ holds for each $u \in B_F(M)$ and each $x_1, x_2 \in B_E(M)$ satisfying $||x_1 - x_2|| \le \delta$.

(A4)

$$f(t, y, v) \ge c_1||G(t, y, v)|| - \lambda_0(t) \text{ for all } (t, y, v) \in [T_1, T_2] \times E \times F, \tag{1.7}$$

$$f(t, y, v) \ge c_2\phi(||v||) - \lambda_1(t) \text{ for all } (t, y, v) \in [T_1, T_2] \times E \times F. \tag{1.8}$$

We equip the set $\mathcal{M}_A$ with the uniformity determined by the base

$$\mathcal{E}_{As}(N,\epsilon)=\{(f,g)\in\mathcal{M}_A\times\mathcal{M}_A:\ |g(t,x,u)-f(t,x,u)|\le\epsilon \tag{1.9}$$
$$\text{for all } (t,x,u)\in[T_1,T_2]\times B_E(N)\times B_F(N)\}$$
$$\cap\{(f,g)\in\mathcal{M}_A\times\mathcal{M}_A:\ |(f-g)(t,x_1,u_1)-(f-g)(t,x_2,u_2)|\le\epsilon(||x_1-x_2||+||u_1-u_2||)$$
$$\text{for each } t\in[T_1,T_2],\ \text{each } x_1,x_2\in B_E(N),\ \text{and each } u_1,u_2\in B_F(N)\},$$

where $N,\epsilon>0$. Clearly the space $\mathcal{M}_A$ with this uniformity is metrizable and complete. We equip the space $\mathcal{M}_A$ with the topology induced by this uniformity. This topology will be called the strong topology of $\mathcal{M}_A$.

We also equip the set $\mathcal{M}_A$ with the uniformity determined by the base

$$\mathcal{E}_{Aw}(N,\epsilon)=\{(f,g)\in\mathcal{M}_A\times\mathcal{M}_A:\ \text{for almost every } t\in[T_1,T_2] \tag{1.10}$$
$$\text{the inequality } |g(t,x,u)-f(t,x,u)|\le\epsilon \text{ holds}$$
$$\text{for each } x\in B_E(N) \text{ and each } u\in B_F(N)\},$$

where $N,\epsilon>0$. We equip the space $\mathcal{M}_A$ with the topology induced by this uniformity. This topology will be called the weak topology of $\mathcal{M}_A$.

Denote by $\mathcal{M}_A^l$ (respectively, $\mathcal{M}_A^c$) the set of all lower semicontinuous (respectively, continuous) functions f in $\mathcal{M}_A$. Clearly $\mathcal{M}_A^l$, $\mathcal{M}_A^c$ are closed subsets of $\mathcal{M}_A$ with the strong topology.

We equip the topological subspaces $\mathcal{M}_A^l,\mathcal{M}_A^c\subset\mathcal{M}_A$ with the relative weak and strong topologies.

It follows from (A1) that for each $f\in\mathcal{M}_A$, each continuous function $x:[T_1,T_2]\to E$, and each strongly measurable function $u:[T_1,T_2]\to F$, the function $f(t,x(t),u(t))$, $t\in[T_1,T_2]$ is Lebesgue measurable.

For each $f\in\mathcal{M}_A$ and each $(x,u)\in X$ define

$$I^f(x,u)=\int_{T_1}^{T_2}f(t,x(t),u(t))dt. \tag{1.11}$$

For each $f\in\mathcal{M}_A$ and each $D\in\mathcal{D}(E)$ we study the optimal control problem

$$I^f(x,u)\to\min,\ (x,u)\in X(G,D). \tag{1.12}$$

DEFINITION 1.1. *Let $f\in\mathcal{M}_A$ and $D\in\mathcal{D}(E)$. We say that the Lavrentiev phenomenon does not occur for the pair (f,D) if for each $\epsilon>0$ there exist $(x,u)\in X(G,D)$ and $K>0$ such that $||u(t)||\le K$ for almost every $t\in[T_1,T_2]$ and $I^f(x,u)\le\inf(I^f;X(G,D))+\epsilon$.*

DEFINITION 1.2. *Let $f\in\mathcal{M}_A$. We say that the Lavrentiev phenomenon does not occur for the integrand f if it does not occur for (f,D) for all $D\in\mathcal{D}(E)$.*

We will establish (see Theorem 1.7) that the Lavrentiev phenomenon does not occur for a generic integrand $f\in\mathcal{M}_A$.

Let c_3 be a nonnegative constant. Denote by $\mathcal{M}_B$ the set of all functions $f:[T_1,T_2]\times E\times F\to R^1$, which satisfy the following assumptions:

(B1) f is measurable with respect to the σ-algebra generated by products of Lebesgue measurable subsets of $[T_2,T_2]$ and Borel subsets of $E\times F$.

(B2) For each $M>0$ there is $M_0>0$ such that for almost every $t\in[T_1,T_2]$ the inequality $|f(t,x,u)|\le M_0$ holds for each $x\in B_E(M)$ and each $u\in B_F(M)$.

(B3) For each $M > 0$ there exists $L > 0$ such that for almost every $t \in [T_1, T_2]$ the inequality $|f(t, x_1, u) - f(t, x_2, u)| \le L||x_1 - x_2||$ holds for each $x_1, x_2 \in B_E(M)$ and each $u \in B_F(M)$.

(B4) Inequality (1.7) holds for all $(t, y, v) \in [T_1, T_2] \times E \times F$ and

$$f(t, y, v) \ge c_2\phi(||v||) - c_3 \text{ for all } (t, y, v) \in [T_1, T_2] \times E \times F. \tag{1.13}$$

We equip the set $\mathcal{M}_B$ with the uniformity determined by the base

$$\begin{aligned}\mathcal{E}_{Bs}(N, \epsilon) = \{(f, g) \in \mathcal{M}_B \times \mathcal{M}_B :\ & |g(t, x, u) - f(t, x, u)| \le \epsilon \\ & \text{for all } (t, x, u) \in [T_1, T_2] \times B_E(N) \times B_F(N)\} \\ \cap\{(f, g) \in \mathcal{M}_B \times \mathcal{M}_B :\ & |(f-g)(t, x_1, u_1) - (f-g)(t, x_2, u_2)| \le \epsilon(||x_1 - x_2|| + ||u_1 - u_2||) \\ & \text{for each } t \in [T_1, T_2],\ \text{each } x_1, x_2 \in B_E(N),\ \text{and each } u_1, u_2 \in B_F(N)\},\end{aligned} \tag{1.14}$$

where $N, \epsilon > 0$. Clearly the space $\mathcal{M}_B$ with this uniformity is metrizable and complete. We equip the space $\mathcal{M}_B$ with the topology induced by this uniformity. This topology will be called the strong topology of $\mathcal{M}_B$.

We also equip the set $\mathcal{M}_B$ with the uniformity determined by the base

$$\begin{aligned}\mathcal{E}_{Bw}(N, \epsilon) = \{(f, g) \in \mathcal{M}_B \times \mathcal{M}_B :\ & \text{for almost every } t \in [T_1, T_2] \\ & \text{the inequality } |g(t, x, u) - f(t, x, u)| \le \epsilon \text{ holds} \\ & \text{for each } x \in B_E(N) \text{ and each } u \in B_F(N)\} \\ \cap\{(f, g) \in \mathcal{M}_B \times \mathcal{M}_B :\ & \text{for almost every } t \in [T_1, T_2] \text{ the inequality} \\ & |(f-g)(t, x_1, u) - (f-g)(t, x_2, u)| \le \epsilon||x_1 - x_2|| \\ & \text{holds for each } x_1, x_2 \in B_E(N) \text{ and each } u \in B_F(N)\},\end{aligned} \tag{1.15}$$

where $N, \epsilon > 0$.

We equip the space $\mathcal{M}_B$ with the topology induced by this uniformity. This topology will be called the weak topology of $\mathcal{M}_B$. Denote by $\mathcal{M}_B^a$ the set of all functions $f \in \mathcal{M}_B$ such that for almost every $t \in [T_1, T_2]$ the function $f(t, \cdot, \cdot) : E \times F \to R^1$ is lower semicontinuous. Denote by $\mathcal{M}_B^l$ (respectively, $\mathcal{M}_B^c$) the set of all lower semicontinuous (respectively, continuous) functions $f \in \mathcal{M}_B$. Clearly $\mathcal{M}_B^a$, $\mathcal{M}_B^l$, and $\mathcal{M}_B^c$ are closed subsets of $\mathcal{M}_B$ with the strong topology.

We equip the topological subspaces $\mathcal{M}_B^a, \mathcal{M}_B^l, \mathcal{M}_B^c \subset \mathcal{M}_B$ with the relative weak and strong topologies.

Remark 1.1. If $\lambda_1(t) = c_3$, $t \in [T_1, T_2]$, then $\mathcal{M}_B \subset \mathcal{M}_A$.

For each $M > 0$ denote by $\mathcal{D}_M(E)$ the collection of all nonempty subsets $D \subset E$ such that $D \subset B_E(M)$.

DEFINITION 1.3. *Let $f \in \mathcal{M}_B$ and M be a positive number. We say that the integrand f possesses the (M)-strong Lipschitz regularity property, abbreviated as (M)-(SLR) property, if there exists $K > 0$ such that for each $g \in \mathcal{M}_B$ satisfying $(f, g) \in \mathcal{E}_{Bw}(K, M)$, each $D \in \mathcal{D}_M(E)$, and each $(x, u) \in X(G, D)$ satisfying mes$(\{t \in [T_1, T_2] :\ ||u(t)|| > K\}) > 0$ there exists $(y, v) \in X(G, D)$ such that $I^g(y, v) < I^g(x, u)$ and $||v(t)|| \le K$ for almost every $t \in [T_1, T_2]$.*

DEFINITION 1.4. *Let $f \in \mathcal{M}_B$. We say that the integrand f possesses the strong Lipschitz regularity property, abbreviated as (SLR) property, if for any positive number M the integrand f possesses the (M)-(SLR) property.*

It is clear that if an integrand $f \in \mathcal{M}_B$ possesses the (SLR) property, $D \in \mathcal{D}(E)$, and $(x, u) \in X(G, D)$ is a minimizer of problem (P), then there is a positive number

$K > 0$ such that $||u(t)|| \le K$ for almost every $t \in [T_1, T_2]$. It is also clear that if an integrand $f \in \mathcal{M}_B$ possesses the (SLR) property and $D \in \mathcal{D}(E)$, then there exist $K > 0$ and a minimizing sequence of trajectory-control pairs $\{(x_n, u_n)\}_{n=1}^{\infty}$ for problem (P) such that $||u_n(t)|| \le K$ for each natural number n and almost every $t \in [T_1, T_2]$. We will establish (see Theorem 1.6) that the (SLR) property is generic in the space $\mathcal{M}_B$.

Note that assumptions (A1)–(A4) and (B1)–(B4) are not very restrictive. They are common in the literature, and the spaces $\mathcal{M}_A$ and $\mathcal{M}_B$ contain many integrands. Therefore it is natural to ask a question if the Lavrentiev phenomenon does not occur for many integrands in $\mathcal{M}_A$ and if the (SLR) property holds for many integrands in $\mathcal{M}_B$. Our first goal is to find conditions on an integrand $f \in \mathcal{M}_B$ (respectively, $f \in \mathcal{M}_A$) which imply the (SLR) property or the (M)-(SLR) property (respectively, nonoccurrence of the Lavrentiev phenomenon). This goal is achieved by Theorems 1.1–1.4. Our second goal is to show nonoccurrence of the Lavrentiev phenomenon for most integrands in $\mathcal{M}_A$ and that most integrands in $\mathcal{M}_B$ possess the (SLR) property. This is achieved by Theorems 1.5–1.7. In order to meet these goals we introduce the following subspaces of integrands.

Let $M > 0$. Denote by $\mathcal{L}_{A,M}$ the set of all functions $f \in \mathcal{M}_A$ which satisfy the following assumption:

(A5) There exist $M_0, M_1, M_2 > 0$, and an integrable scalar nonnegative function $\psi(t)$, $t \in [T_1, T_2]$ which depend only on f and M such that

(i) for each $D \in \mathcal{D}_M(E)$ and each $(x, u) \in X(G, D)$ which satisfies $||u(t)|| \le M + N_*$ for almost every $t \in [T_1, T_2]$ the inequality $||x(t)|| \le M_0$ holds for all $t \in [T_1, T_2]$;

(ii) for almost every $t \in [T_1, T_2]$ the inequality $f(t, x, u) \le M_1$ holds for each $x \in B_E(M + M_0 + N_*)$ and each $u \in B_F(M + M_0 + N_*)$;

(iii) for each $g \in \mathcal{M}_A$, each $D \in \mathcal{D}_M(E)$, and each $(x, u) \in X(G, D)$ satisfying $I^g(x, u) \le (M + M_1)(T_2 - T_1) + 1$ the inequality $||x(t)|| \le M_2$ holds for all $t \in [T_1, T_2]$;

(iv) for each $\epsilon > 0$ there exist $\Gamma, \delta > 0$ such that for almost every $t \in [T_1, T_2]$ the inequality

$$|f(t, x_1, u) - f(t, x_2, u)| \le \epsilon(f(t, x_1, u) + \psi(t))$$

is valid for each $x_1, x_2 \in B_E(M_2 + 1)$ satisfying $||x_1 - x_2|| \le \delta$ and each $u \in F$ satisfying $||u|| \ge \Gamma$.

Remark 1.2. The existence of M_0 in (A5)(i) follows from Lemma 2.5. The existence of M_1 in (A5)(ii) follows from (A2), and the existence of M_2 in (A5)(iii) follows from Lemma 2.7.

Let $M > 0$. Denote by $\mathcal{L}_{B,M}$ the set of all functions $f \in \mathcal{M}_B$ which satisfy the following assumption:

(B5) There exist positive numbers M_0, M_1, M_2, L, δ and an integrable scalar nonnegative function $\psi(t)$, $t \in [T_1, T_2]$, which depend only on f and M such that

(i) for each $D \in \mathcal{D}_M(E)$ and each $(x, u) \in X(G, D)$ which satisfies $||u(t)|| \le M + N_*$ for almost every $t \in [T_1, T_2]$ the inequality $||x(t)|| \le M_0$ holds for all $t \in [T_1, T_2]$;

(ii) for almost every $t \in [T_1, T_2]$ the inequality $f(t, x, u) \le M_1$ holds for each $x \in B_E(M + M_0 + N_*)$ and each $u \in B_F(M + M_0 + N_*)$;

(iii) for each $g \in \mathcal{M}_B$, each $D \in \mathcal{D}_M(E)$, and each $(x, u) \in X(G, D)$ satisfying $I^g(x, u) \le (M + M_1)(T_2 - T_1) + 1$ the inequality $||x(t)|| \le M_2$ holds for all $t \in [T_1, T_2]$;

(iv) for almost every $t \in [T_1, T_2]$ the inequality

$$|f(t, x_1, u) - f(t, x_2, u)| \le L||x_1 - x_2||(f(t, x_1, u) + \psi(t))$$

holds for each $x_1, x_2 \in B_E(M_2 + 1)$ satisfying $||x_1 - x_2|| \le \delta$ and each $u \in F$.

Remark 1.3. The existence of M_0 in (B5)(i) follows from Lemma 2.5. The existence of M_1 in (B5)(ii) follows from (B2) and the existence of M_2 in (B5)(iii) follows from Lemma 2.7.

The following theorem is our first main result. It is also an important ingredient in the proofs of our other main results.

THEOREM 1.1. *Let M be a positive number. Then each integrand $f \in \mathcal{L}_{B,M}$ possesses the (M)-(SLR) property.*

The next theorem follows from Theorem 1.1 and auxiliary results which are proved in section 2.

THEOREM 1.2. *Let $f \in \mathcal{M}_B$ satisfy the following assumption:*

(B6) *For each $q > 0$ there exist positive numbers L, δ and an integrable scalar nonnegative function $\psi(t)$, $t \in [T_1, T_2]$, such that for almost every $t \in [T_1, T_2]$ the inequality*

$$|f(t, x_1, u) - f(t, x_2, u)| \le L||x_1 - x_2||(f(t, x_1, u) + \psi(t))$$

holds for each $x_1, x_2 \in B_E(q)$ satisfying $||x_1 - x_2|| \le \delta$ and for each $u \in F$.

Then f possesses the strong Lipschitz regularity property.

Theorems 1.1 and 1.2 will be proved in section 3.

The next theorem implies that if $M > 0$ and if an integrand f belongs to $\mathcal{L}_{A,M}$, then the Lavrentiev phenomenon does not occur for the pair (f, D), where $D \in \mathcal{D}_M(E)$.

THEOREM 1.3. *Let M, ϵ be positive numbers and let $f \in \mathcal{L}_{A,M}$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathcal{M}_A$ with the weak topology and $K > 0$ such that for each $g \in \mathcal{U}$ and each $D \in \mathcal{D}_M(E)$ there is $(x, u) \in X(G, D)$ such that $||u(t)|| \le K$ for almost every $t \in [T_1, T_2]$ and $I^g(x, u) \le \inf(I^g; X(G, D)) + \epsilon$.*

The next theorem follows from Theorem 1.3 and auxiliary results which are proved in section 2.

THEOREM 1.4. *Let $f \in \mathcal{M}_A$ satisfy the following assumption:*

(A6) *There exists an integrable scalar nonnegative function $\psi(t)$, $t \in [T_1, T_2]$, such that for each $q, \epsilon > 0$ there are $\Gamma, \delta > 0$ such that for almost every $t \in [T_1, T_2]$ the inequality*

$$|f(t, x_1, u) - f(t, x_2, u)| \le \epsilon(f(t, x_1, u) + \psi(t))$$

is valid for each $x_1, x_2 \in B_E(q)$ satisfying $||x_1 - x_2|| \le \delta$ and each $u \in F$ satisfying $||u|| \ge \Gamma$.

Then for each $M, \epsilon > 0$ there exists a neighborhood $\mathcal{U}$ of f in $\mathcal{M}_A$ with the weak topology and $K > 0$ such that for each $g \in \mathcal{U}$ and each $D \in \mathcal{D}_M(E)$ there is $(x, u) \in X(G, D)$ such that $||u(t)|| \le K$ for almost every $t \in [T_1, T_2]$ and $I^g(x, u) \le \inf(I^g; X(G, D)) + \epsilon$.

Theorem 1.4 implies that if an integrand $f \in \mathcal{M}_A$ satisfies assumption (A6), then the Lavrentiev phenomenon does not occur for f. Theorems 1.3 and 1.4 will be proved in section 4.

The following generic results (Theorems 1.5–1.7) will be obtained by using Theorems 1.1 and 1.3 and an auxiliary density result which will be obtained in section 5. They will be proved in section 6.

THEOREM 1.5. *Let M be a positive number. Then there exists an open (in the weak topology) everywhere dense (in the strong topology) subset $\mathcal{F} \subset \mathcal{M}_B$ such that each $f \in \mathcal{F}$ possesses the (M)-(SLR) property.*

Moreover, $\mathcal{F} \cap \mathcal{M}_B^a$ (respectively, $\mathcal{F} \cap \mathcal{M}_B^l$, $\mathcal{F} \cap \mathcal{M}_B^c$) contains an open (in the weak topology) everywhere dense (in the strong topology) subset of $\mathcal{M}_B^a$ (respectively, $\mathcal{M}_B^l$, $\mathcal{M}_B^c$).

THEOREM 1.6. *There exists a subset $\mathcal{F}_B \subset \mathcal{M}_B$ which is a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) subsets of $\mathcal{M}_B$ such that each $f \in \mathcal{F}_B$ possesses the (SLR) property.*

Moreover, $\mathcal{F}_B \cap \mathcal{M}_B^a$ (respectively, $\mathcal{F}_B \cap \mathcal{M}_B^l$, $\mathcal{F} \cap \mathcal{M}_B^c$) contains a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) subsets of $\mathcal{M}_B^a$ (respectively, $\mathcal{M}_B^l$, $\mathcal{M}_B^c$).

THEOREM 1.7. *There exists a subset $\mathcal{F}_A \subset \mathcal{M}_A$ which is a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) subsets of $\mathcal{M}_A$ such that for each $f \in \mathcal{F}_A$ the Lavrentiev phenomenon does not occur.*

Moreover, $\mathcal{F}_A \cap \mathcal{M}_A^l$ (respectively, $\mathcal{F}_A \cap \mathcal{M}_A^c$) contains a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) subsets of $\mathcal{M}_A^l$ (respectively, $\mathcal{M}_A^c$).

Section 7 contains examples of integrands for which the Lavrentiev phenomenon does not occur and examples of integrands possessing the (SLR) property. These examples include an optimal control problem which is not covered by the results of [12], [19], [24].

Sarychev [18, Theorem 2.1] constructed a large class of integrands which exhibit the Lavrentiev phenomenon. This class includes the famous example of Ball and Mizel [7]. By Theorem 1.4 the Lavrentiev phenomenon does not occur for any integrand $f \in \mathcal{M}_A$ which satisfies assumption (A6). In section 8 we consider the class of integrands studied in [18] and show that these integrands do not satisfy assumption (A6).

Theorem 1.6 establishes that the (SLR) property holds for most integrands in $\mathcal{M}_B$, and Theorem 1.7 establishes nonoccurrence of the Lavrentiev phenomenon for most integrands in $\mathcal{M}_A$ (in the sense of the Baire category) with respect to suitable topologies defined on the spaces of integrands. Such an approach is often used in many situations when a certain property is studied for the whole space rather than for a single element of the space. See, for example, [5], [17], [23] and the references mentioned there. Clearly in this approach a choice of topologies is of great importance. As in [16], [23], [24], in the present paper the topologies are defined on the spaces of integrands, which is quite natural. Note that in the variational analysis and in the optimization theory there are other well-known topologies of variational character, like epiconvergence or bounded Hausdorff convergence [3], [4]. In our case the convergence with respect to these topologies means convergence in the space of functionals $\{I^f : f \in \mathcal{M}\}$ (see (1.11)) on the set $X(G, D)$, where $\mathcal{M}$ is either $\mathcal{M}_A$ or $\mathcal{M}_B$ and D is a nonempty bounded subset of E. Two functionals I^f and I^g with $f, g \in \mathcal{M}$ are close in these topologies if their epigraphs are close in some sense. These topologies are very important and used for many optimization problems [3], [4]. It is not known whether results comparable with those of our paper hold by using such topologies.

In this paper the space of integrands $\mathcal{M}_B$ ($\mathcal{M}_A$, respectively) is equipped with the two topologies where one is weaker than the other. They are called the weak and the strong topologies, respectively. Note that if a set is G_δ and everywhere dense with respect to one of these two topologies, then this fact does imply that the set is G_δ and everywhere dense with respect to the other topology. We show that the (SLR) property (nonoccurrence of the Lavrentiev phenomenon, respectively) holds for all integrands which belong to a set that is a countable intersection of open (in the

weak topology) everywhere dense (in the strong topology) sets. This implies that our properties are generic with respect to any topology which is weaker than the strong topology and is stronger than the weak topology.

The results which will be established in this paper are more complicated than their prototypes in [24], where we study variational problems. Theorem 1.2, one of our main results, establishes that an integrand $f \in \mathcal{M}_B$ which satisfies (B6) possesses the (SLR) property. In order to prove this result, as well as its prototype in [24], we consider a trajectory-control pair $(x, u) \in X(G, D)$ such that $||u(t)||$ is very large on a subset of $[T_1, T_2]$ with a positive Lebesgue measure and construct another admissible trajectory-control pair (y, v), which is an approximation of (x, u). The approximation is constructed as follows: We put $v(t) = u(t)$ for those t where $||u(t)||$ is not too large and use a bounded alternative for $v(t)$ when $||u(t)||$ is large. The trajectory y is obtained as a solution of the differential equation

$$y'(t) = G(t, y(t), v(t)),\ t \in [T_1, T_2], \tag{1.16}$$
$$y(T_1) = x(T_1).$$

Then we need to estimate carefully $||y(t) - x(t)||$ and $f(t, x(t), u(t)) - f(t, y(t), v(t))$ for $t \in [T_1, T_2]$. In [24] and here the control v is constructed explicitly. Since in [24] $G(t, y, v) = v$, (1.16) is very simple and its solution is a function $y(t) = x(T_1) + \int_{T_1}^{t} v(s)ds$, $t \in [T_1, T_2]$. In [24] the explicit form of the trajectory y plays a crucial role in estimations. In the present paper the right-hand side of (1.16) is very general, and this makes the problem more difficult and less understood. The solution of (1.16) cannot be obtained explicitly and even its existence is not obvious. In section 2 we will develop a technique which allows us to obtain the existence of a solution y of (1.16) and to estimate $||x(t) - y(t)||$ for $t \in [T_1, T_2]$. These estimations will be used in order to compare integrals $I^f(x, u)$ and $I^f(y, v)$. In [24], in order to show that the (SLR) is generic in a space of integrands, we consider its subset, which contains all integrands satisfying assumption (B6) (which implies the (SLR) property) and show that this subset is everywhere dense. In this paper we cannot show that the set of all integrands in $\mathcal{M}_B$ satisfying assumption (B6) is an everywhere dense subset of the space $\mathcal{M}_B$ with the strong topology. It happens because the space $\mathcal{M}_B$ is more complicated than its analogue in [24]. In order to overcome this difficulty we introduce the (M)-(SLR) property, which is a weakened version of the (SLR) property, introduce the spaces $\mathcal{L}_{B,M}$, where $M > 0$, and show that any $f \in \mathcal{L}_{B,M}$ possesses the (M)-(SLR) property. Then we manage to show that all the spaces $\mathcal{L}_{B,M}$ are everywhere dense in $\mathcal{M}_B$ with the strong topology. This allows us to obtain our generic result for the space $\mathcal{M}_B$. Analogously we can compare the results for the space $\mathcal{M}_A$ and for its prototype in [24].

2. Auxiliary results.

LEMMA 2.1 (see [21, Lemma 5.1]). *Assume that $t_0 \in [T_1, T_2]$, $y_0 \in E$, and $u : [T_1, T_2] \to F$ is a strongly measurable function satisfying*

$$\int_{T_1}^{T_2} \phi(||u(t)||)dt < \infty,\ \int_{T_1}^{T_2} ||G(t, 0, u(t))||dt < \infty. \tag{2.1}$$

For each $\delta > 0$, set $I_\delta = [t_0 - \delta, t_0 + \delta] \cap [T_1, T_2]$. Then the following assertions hold:

1. *There are $\delta > 0$ and $y \in W^{1,1}(I_\delta; E)$ such that*

$$y(t_0) = y_0,\ y'(t) = G(t, y(t), u(t)),\ t \in I_\delta \text{ a.e.}. \tag{2.2}$$

2. *If* $\delta_1, \delta_2 > 0$ *and if for* $i = 1, 2$, $y_i \in W^{1,1}(I_{\delta_i}; E)$ *satisfies* (2.2) *with* $y = y_i$, $\delta = \delta_i$, *then there is* $\delta_3 \in (0, \min\{\delta_1, \delta_2\})$ *such that* $y_1(t) = y_2(t)$ *for all* $t \in I_{\delta_3}$.

LEMMA 2.2 (see [21, Lemma 5.2]). *Assume that* $u : [T_1, T_2] \to F$ *is a strongly measurable function satisfying* (2.1), $T_1 < S \leq T_2$, $y_1, y_2 \in W^{1,1}(T_1, S; E)$ *satisfy* $y_1(T_1) = y_2(T_1)$, *and that for* $i = 1, 2$,

$$y_i'(t) = G(t, y_i(t), u(t)),\ t \in [T_1, T_2]\ a.e..$$

Then $y_1(t) = y_2(t)$ *for all* $t \in [T_1, S]$.

LEMMA 2.3. *Assume that* $y_0 \in E$, M_0 *is a positive number, and that* $u : [T_1, T_2] \to F$ *is a strongly measurable function such that*

$$||u(t)|| \leq M_0,\ t \in [T_1, T_2]\ a.e.. \tag{2.3}$$

Then there exists $y \in W^{1,1}(T_1, T_2; E)$ *such that*

$$y(T_1) = y_0,\ y'(t) = G(t, y(t), u(t)),\ t \in [T_1, T_2]\ a.e.. \tag{2.4}$$

Proof. Since the function ϕ is increasing and G is bounded on bounded subsets of $[T_1, T_2] \times E \times F$, it follows from (2.3) that (2.1) holds.

Denote by Ω the set of all $S \in (T_1, T_2]$ such that there exists $y \in W^{1,1}(T_1, S; E)$, which satisfies

$$y(T_1) = y_0,\ y'(t) = G(t, y(t), u(t)),\ t \in [T_1, S] \text{ a.e..}$$

In view of (2.1) and Lemma 2.1, $\Omega \neq \emptyset$. Set

$$S_0 = \sup\{\tau :\ \tau \in \Omega\}. \tag{2.5}$$

By (2.1), Lemma 2.2, the definition of Ω, and (2.5) there is a function $y : [T_1, S_0) \to E$ such that for each $\tau \in (T_1, S_0)$ we have $y \in W^{1,1}(T_1, \tau; E)$,

$$y(T_1) = y_0,\ \text{and } y'(t) = G(t, y(t), u(t)),\ t \in [T_1, S_0) \text{ a.e..} \tag{2.6}$$

We show that there exists $\lim_{t \to S_0^-} y(t)$ in the norm topology of E. Let $T_1 \leq \tau < S_0$. It follows from (2.6), (1.3), (2.3), (2.1), and the monotonicity of ϕ that

$$\begin{aligned}
||y(\tau)|| &\leq ||y_0|| + \int_{T_1}^{\tau} ||G(t, y(t), u(t))||dt \leq ||y_0|| + \int_{T_1}^{\tau} ||G(t, 0, u(t))||dt \\
&+ \int_{T_1}^{\tau} ||G(t, y(t), u(t)) - G(t, 0, u(t))||dt \leq ||y_0|| + \int_{T_1}^{\tau} ||G(t, 0, u(t))||dt \\
&+ \int_{T_1}^{\tau} c_0||y(t)||\phi(||u(t)||)dt \leq ||y_0|| + \int_{T_1}^{\tau} ||G(t, 0, u(t))||dt \\
&+ \int_{T_1}^{\tau} c_0||y(t)||\phi(M_0)dt \leq ||y_0|| + \int_{T_1}^{T_2} ||G(t, 0, u(t))||dt + c_0\phi(M_0) \int_{T_1}^{\tau} ||y(t)||dt.
\end{aligned}$$

In view of this inequality and Gronwall's inequality for any $\tau \in [T_1, S_0)$,

$$||y(\tau)|| \leq \left[||y_0|| + \int_{T_1}^{T_2} ||G(t, 0, u(t))||dt\right] \exp(c_0\phi(M_0)(T_2 - T_1)). \tag{2.7}$$

Let $T_1 \le t_1 < t_2 < S_0$. By (2.6), (1.3), (2.3), and the monotonicity of ϕ,

$$||y(t_2) - y(t_1)|| \le \int_{t_1}^{t_2} ||G(t, y(t), u(t))||dt$$
$$\le \int_{t_1}^{t_2} ||G(t, 0, u(t))||dt + \int_{t_1}^{t_2} ||G(t, y(t), u(t)) - G(t, 0, u(t))||dt$$
$$\le \int_{t_1}^{t_2} ||G(t, 0, u(t))||dt + \int_{t_1}^{t_2} c_0\phi(||u(t)||)||y(t)||dt$$
$$\le \int_{t_1}^{t_2} ||G(t, 0, u(t))||dt + c_0\phi(M_0)\int_{t_1}^{t_2} ||y(t)||dt.$$

It follows from this relation, (2.1), and (2.7) that

$$||y(t_2) - y(t_1)|| \to 0 \text{ as } t_1, t_2 \to S_0^-.$$

Therefore there exists $\lim_{t\to S_0^-} y(t)$ in the norm topology which is denoted by $y(S_0)$.

It follows from (1.3), (2.7), the monotonicity of ϕ, (2.1), and (2.7) that

$$\int_{T_1}^{S_0} ||G(t, y(t), u(t))||dt \le \int_{T_1}^{S_0} ||G(t, 0, u(t))||dt + \int_{T_1}^{S_0} c_0||y(t)||\phi(||u(t)||)dt$$
$$\le \int_{T_1}^{S_0} ||G(t, 0, u(t))||dt + c_0\phi(M_0)\int_{T_1}^{S_0} ||y(t)||dt < \infty.$$

Together with (2.6) and the equality $y(S_0) = \lim_{\tau\to S_0^-} y(\tau)$ this inequality implies that

$$y \in W^{1,1}(T_1, S_0; E). \tag{2.8}$$

In order to complete the proof of the lemma it is sufficient to show that $S_0 = T_2$.

Let us assume that $S_0 < T_2$. Then by (2.1) and Lemma 2.1 there exist $\epsilon \in (0, T_2 - S_0)$ and $z \in W^{1,1}(S_0, S_0 + \epsilon; E)$ such that

$$z(S_0) = y(S_0),\ z'(t) = G(t, z(t), u(t)),\ t \in [S_0, S_0 + \epsilon] \text{ a.e..}$$

Combined with (2.8) and (2.6), these relations imply that $S_0 + \epsilon \in \Omega$, a contradiction. The contradiction we have reached proves that $S_0 = T_2$. Lemma 2.3 is proved. □

LEMMA 2.4. *Let $u : [T_1, T_2] \to F$ be a strongly measurable function, $x \in W^{1,1}(T_1, T_2; E)$ satisfy*

$$x'(t) = G(t, x(t), u(t)),\ t \in [T_1, T_2] \text{ a.e.,} \tag{2.9}$$

and let

$$\int_{T_1}^{T_2} ||u(t)||dt < \infty,\ M_0 > N_*,\ \Omega = \{t \in [T_1, T_2] :\ ||u(t)|| > M_0\}. \tag{2.10}$$

Assume that

$$v(t) = u(t),\ t \in [T_1, T_2] \setminus \Omega,\ v(t) = u_*(t),\ t \in \Omega \tag{2.11}$$

and that $y \in W^{1,1}(T_1, T_2; E)$ *satisfies*

$$y(T_1) = x(T_1),\ y'(t) = G(t, y(t), v(t)),\ t \in [T_1, T_2]\ a.e.. \tag{2.12}$$

Then for each $\tau \in [T_1, T_2]$,

$$||y(\tau) - x(\tau)|| \leq 2c_0\phi(\sup\{||x(s)|| :\ s \in [T_1, T_2]\}) \int_\Omega ||u(t)||dt\ exp\left(c_0 \int_{T_1}^{T_2} \phi(||v(t)||)dt\right).$$

Proof. Set

$$M_1 = \sup\{||x(t)|| :\ t \in [T_1, T_2]\}, \tag{2.13}$$

$$d = \int_\Omega ||u(t)||dt. \tag{2.14}$$

Let $\tau \in (T_1, T_2]$. It follows from (2.9), (2.12), (1.3), (2.11), (2.10), (1.4), the monotonicity of ϕ, (1.5), (2.13), and (2.14) that

$$||y(\tau) - x(\tau)|| = || \int_{T_1}^{\tau} G(t, y(t), v(t))dt - \int_{T_1}^{\tau} G(t, x(t), u(t))||$$

$$\leq \int_{T_1}^{\tau} ||G(t, y(t), v(t)) - G(t, x(t), v(t))||dt + \int_{T_1}^{\tau} ||G(t, x(t), v(t)) - G(t, x(t), u(t))||dt$$

$$\leq c_0 \int_{T_1}^{\tau} ||y(t) - x(t)||\phi(||v(t)||)dt + c_0 \int_{T_1}^{\tau} ||u(t) - v(t)||\phi(||x(t)||)dt$$

$$\leq c_0 \int_{T_1}^{\tau} ||y(t) - x(t)||\phi(||v(t)||)dt + c_0\phi(M_1) \int_{[T_1,\tau]\cap\Omega} ||u(t) - v(t)||dt$$

$$\leq c_0 \int_{T_1}^{\tau} ||y(t) - x(t)||\phi(||v(t)||)dt + c_0\phi(M_1) \left[\int_\Omega ||u(t)||dt + N_* \mathrm{mes}(\Omega)\right]$$

$$\leq c_0 \int_{T_1}^{\tau} ||y(t) - x(t)||\phi(||v(t)||)dt + c_0\phi(M_1)2d.$$

By this inequality and Gronwall's inequality, for each $\tau \in [T_1, T_2]$ we get

$$||y(\tau) - x(\tau)|| \leq 2c_0 d\phi(M_1)\exp\left(c_0 \int_{T_1}^{T_2} \phi(||v(t)||)dt\right).$$

This completes the proof of Lemma 2.4. □

LEMMA 2.5. *Let* $M > 0$. *Then there is* $M_0 > 0$ *such that for each* $D \in \mathcal{D}_M(E)$ *and each* $(x, u) \in X(G, D)$ *which satisfies* $||u(t)|| \leq M$ *for almost every* $t \in [T_1, T_2]$ *the inequality* $||x(t)|| \leq M_0$ *holds for all* $t \in [T_1, T_2]$.

Proof. Since G is bounded on any bounded subset of $[T_1, T_2] \times E \times F$ there is $M_1 > 0$ such that

$$||G(t, 0, z)|| \leq M_1 \text{ for each } t \in [T_1, T_2] \text{ and each } z \in B_F(M). \tag{2.15}$$

Set

$$M_0 = [M + M_1(T_2 - T_1)]\exp(c_0\phi(M)(T_2 - T_1)). \tag{2.16}$$

Assume that

$$D \in \mathcal{D}_M(E),\ (x,u) \in X(G,D),\ ||u(t)|| \le M \text{ for } t \in [T_1,T_2] \text{ a.e..} \tag{2.17}$$

By (2.17), (1.6), (2.15), (1.3), and the monotonicity of ϕ for any $t \in [T_1,T_2]$,

$$||x(t)|| \le ||x(T_1)|| + \int_{T_1}^t ||G(s,x(s),u(s))||ds$$

$$\le M + \int_{T_1}^t ||G(s,0,u(s))||ds + \int_{T_1}^t ||G(s,x(s),u(s)) - G(s,0,u(s))||ds$$

$$\le M + M_1(T_2 - T_1) + \int_{T_1}^t c_0||x(s)||\phi(||u(s)||)ds$$

$$\le M + M_1(T_2 - T_1) + c_0\phi(M)\int_{T_1}^t ||x(s)||ds.$$

It follows from this inequality, (2.16), and Gronwall's inequality that for each $t \in [T_1,T_2]$,

$$||x(t)|| \le [M + M_1(T_2 - T_1)]\exp(c_0\phi(M)(T_2 - T_1)) = M_0.$$

Lemma 2.5 is proved. □

Lemmas 2.3 and 2.5, (1.5), and (A2) imply the following auxiliary result.

LEMMA 2.6. *Let $M > 0$ and $f \in \mathcal{M}_A$. Then there is $M_1 > 0$ such that for each $D \in \mathcal{D}_M(E)$,*

$$\inf(I^f; X(D,G)) \le M_1.$$

LEMMA 2.7. *Let $M_0, M_1 > 0$. Then there exists $M_2 > 0$ such that for each $g \in \mathcal{M}_A$, each $D \in \mathcal{D}_{M_0}(E)$, and each $(x,u) \in X(G,D)$ satisfying $I^g(x,u) \le M_1$, the inequality $||x(t)|| \le M_2$ holds for all $t \in [T_1,T_2]$.*

Proof. Set

$$M_2 = M_0 + c_1^{-1}M_1 + c_1^{-1}\int_{T_1}^{T_2} \lambda_0(s)ds. \tag{2.18}$$

Assume that

$$g \in \mathcal{M}_A,\ D \in \mathcal{D}_{M_0}(E),\ (x,u) \in X(G,D),\ I^g(x,u) \le M_1. \tag{2.19}$$

Let $t \in [T_1,T_2]$. It follows from (2.19), (1.6), (1.7), and (2.18) that

$$||x(t)|| \le ||x(T_1)|| + \int_{T_1}^t ||G(s,x(s),u(s))||ds \le M_0 + \int_{T_1}^{T_2} c_1^{-1}[g(s,x(s),u(s)) + \lambda_0(s)]ds$$

$$\le M_0 + c_1^{-1}M_1 + c_1^{-1}\int_{T_1}^{T_2} \lambda_0(s)ds = M_2.$$

Lemma 2.7 is proved. □

3. Proofs of Theorems 1.1 and 1.2. In this section we establish the following result.

THEOREM 3.1. *Let M be a positive number and let $f \in \mathcal{L}_{B,M}$. Then there exists $K > N_*$, $\Delta_1 > 0$ such that for each $g \in \mathcal{M}_B$ satisfying $(f,g) \in \mathcal{E}_{Bw}(K,M)$, each $D \in \mathcal{D}_M(E)$, and each $(x,u) \in X(G,D)$ which satisfies $I^g(x,u) \le \inf(I^g; X(G,D)) + 1$, the following assertion holds: If the set $\Omega := \{t \in [T_1,T_2] : \ ||u(t)|| > K\})$ has a positive Lebesgue measure, then there exists $(y,v) \in X(G,D)$ such that*

$$I^g(y,v) < I^g(x,u) - M\int_\Omega ||u(t)||dt, \ ||x(t)-y(t)|| \le \Delta_1 \int_\Omega ||u(t)||dt, \ t \in [T_1,T_2],$$

$$v(t) = u(t), \ t \in [T_1,T_2] \setminus \Omega, \ v(t) = u_*(t), \ t \in \Omega.$$

It is clear that Theorem 3.1 implies Theorem 1.1.

Proof of Theorem 3.1. Since $f \in \mathcal{L}_{B,M}$, assumption (B5) holds for f. Let $L > 1$, positive numbers M_0, M_1, M_2, δ, and an integrable scalar nonnegative function $\psi(t)$, $t \in [T_1,T_2]$, be as guaranteed by (B5). In view of (B2) there is $M_3 > 0$ such that for almost every $t \in [T_1,T_2]$ the following inequality holds:

(3.1)
$$f(t,h,\xi) \le M_3 \text{ for each } h \in B_E(M_2+N_*+1) \text{ and each } \xi \in B_F(M_2+N_*+1).$$

Set

$$\kappa = 2c_0\phi(M_2)\exp(c_0c_2^{-1}c_3(T_2-T_1) + c_0c_2^{-1}((M+M_1)(T_2-T_1)+1)), \tag{3.2}$$

$$\Delta_1 = (\kappa+1)\left[1 + L + M + L(3M+M_1+c_3)(T_2-T_1) + L\int_{T_1}^{T_2}\psi(t)dt\right] \tag{3.3}$$

and choose $\gamma \in (0,1)$ such that

$$\gamma c_2^{-1}((M+M_1)(T_2-T_1)+1+c_3(T_2-T_1))\kappa < 8^{-1}\min\{1,\delta\}, \tag{3.4}$$

$$(4\gamma)^{-1}c_2 \ge 8\Delta_1.$$

By (1.2) there exists a positive number K such that

$$K > 16(M+M_0+M_1+M_2+M_3+N_*+1+c_3), \tag{3.5}$$

$$\phi(t)t^{-1} \ge \max\{\gamma^{-1}, c_2^{-1}\} \text{ for all } t \ge K. \tag{3.6}$$

Assume that

$$g \in \mathcal{M}_B, \ (f,g) \in \mathcal{E}_{Bw}(K,M), \ D \in \mathcal{D}_M(E), \tag{3.7}$$

$$(x,u) \in X(G,D), \ I^g(x,u) \le \inf(I^g; X(G,D)) + 1, \tag{3.8}$$

$$\text{mes}(\{t \in [T_1,T_2] : \ ||u(t)|| > K\}) > 0. \tag{3.9}$$

Set

$$\Omega = \{t \in [T_1,T_2] : \ ||u(t)|| > K\}, \ d = \int_\Omega ||u(t)||dt. \tag{3.10}$$

Relations (3.9), (3.10), and (3.5) imply that

$$d > 0. \tag{3.11}$$

We will show that $||x(t)|| \le M_2$ for all $t \in [T_1, T_2]$ and $I^g(x,u) \le (M+M_1)(T_2-T_1)+1$. By Lemma 2.3 and (1.5) there exists $x_* \in W^{1,1}(T_1, T_2; E)$ such that

$$(x_*, u_*) \in X(G, D). \tag{3.12}$$

In view of this inclusion, (3.7), (1.5), the choice of M_0, and (B5)(i),

$$||x_*(t)|| \le M_0 \text{ for all } t \in [T_1, T_2]. \tag{3.13}$$

It follows from the choice of M_1, (B2)(ii), (3.13), and (1.5) that for almost every $t \in [T_1, T_2]$,

$$f(t, x_*(t), u_*(t)) \le M_1. \tag{3.14}$$

By (3.14), (3.13), (1.5), (3.5), (3.7), and (1.15) for almost every $t \in [T_1, T_2]$,

$$g(t, x_*(t), u_*(t)) \le f(t, x_*(t), u_*(t)) + M \le M + M_1.$$

Together with (3.8) and (3.12) this inequality implies that

$$I^g(x,u) \le I^g(x_*, u_*) + 1 \le (M + M_1)(T_2 - T_1) + 1. \tag{3.15}$$

It follows from this inequality, the choice of M_2, (B5)(iii), (3.7), and (3.8) that

$$||x(t)|| \le M_2 \text{ for all } t \in [T_1, T_2]. \tag{3.16}$$

By (3.10), (3.6), (1.13), and (3.15),

$$d = \int_\Omega ||u(t)||dt \le \int_\Omega \gamma\phi(||u(t)||)dt \le \gamma \int_{T_1}^{T_2} \phi(||u(t)||)dt$$

$$\le \gamma \int_{T_1}^{T_2} c_2^{-1}[g(t, x(t), u(t)) + c_3]dt = \gamma c_2^{-1}(I^g(x,u) + c_3(T_2 - T_1))$$

$$\le \gamma c_2^{-1}((M + M_1)(T_2 - T_1) + 1 + c_3(T_2 - T_1)). \tag{3.17}$$

Equation (3.10) implies that

$$\text{mes}(\Omega) \le K^{-1} d. \tag{3.18}$$

Set

$$v(t) = u(t),\ t \in [T_1, T_2] \setminus \Omega,\ v(t) = u_*(t),\ t \in \Omega. \tag{3.19}$$

It is clear that $v : [T_1, T_2] \to F$ is a strongly measurable function. Relations (3.19), (1.5), (3.10), and (3.5) imply that

$$||v(t)|| \le K,\ t \in [T_1, T_2] \text{ a.e.}. \tag{3.20}$$

In view of (3.20), Lemma 2.3, and (3.8) there exists $y \in W^{1,1}(T_1, T_2; E)$ such that

$$y(T_1) = x(T_1),\ (y, v) \in X(G, D). \tag{3.21}$$

It follows from Lemma 2.4 (with $M_0 = K$), (3.8), (3.5), (3.17), (3.10), (3.21), (3.19), and (3.16) that for each $t \in [T_1, T_2]$

$$||y(t) - x(t)|| \le \exp\left(c_0 \int_{T_1}^{T_2} \phi(||v(t)||)dt\right) 2c_0 \left(\int_\Omega ||u(t)||dt\right) \phi(M_2). \tag{3.22}$$

By (3.19), (1.5), (3.10), (3.5), the monotonicity of ϕ, (1.13), and (3.15),

$$\int_{T_1}^{T_2} \phi(||v(t)||)dt \le \int_{T_1}^{T_2} \phi(||u(t)||)dt \le c_2^{-1} \int_{T_1}^{T_2} [g(t, x(t), u(t)) + c_3]dt$$
$$\le c_2^{-1} c_3 (T_2 - T_1) + c_2^{-1}[(M + M_1)(T_2 - T_1) + 1].$$

Combined with (3.22), (3.10), and (3.2) this inequality implies that

$$||y(t) - x(t)|| \le d\kappa, \ t \in [T_1, T_2]. \tag{3.23}$$

It follows from (3.23), (3.17), and (3.4) that for all $t \in [T_1, T_2]$,

(3.24)
$$||y(t) - x(t)|| \le d\kappa \le \gamma c_2^{-1}((M + M_1)(T_2 - T_1) + 1 + c_3(T_2 - T_1))\kappa < 8^{-1} \min\{1, \delta\}.$$

Relations (3.24) and (3.16) imply that

$$||y(t)|| \le M_2 + 8^{-1}, \ t \in [T_1, T_2]. \tag{3.25}$$

We will estimate $I^g(x, u) - I^g(y, v)$. It is clear that

$$I^g(x, u) - I^g(y, v) = \int_\Omega g(t, x(t), u(t))dt - \int_\Omega g(t, y(t), v(t))dt \tag{3.26}$$
$$+ \int_{[T_1,T_2]\setminus\Omega} g(t, x(t), u(t))dt - \int_{[T_1,T_2]\setminus\Omega} g(t, y(t), v(t))dt.$$

It follows from (3.25), (3.19), (1.5), and the choice of M_3 (see (3.1)) that for almost every $t \in \Omega$, $f(t, y(t), v(t)) \le M_3$. Together with (3.7), (1.15), (3.25), (3.19), (1.5), and (3.5) this inequality implies that for almost every $t \in \Omega$,

$$g(t, y(t), v(t)) \le M + f(t, y(t), v(t)) \le M + M_3. \tag{3.27}$$

In view of (1.13), the monotonicity of ϕ, (3.10), (3.6), and (3.5) for every $t \in \Omega$,

$$g(t, x(t), u(t)) \ge c_2\phi(||u(t)||) - c_3 \ge c_2\phi(K) - c_3 \ge K - c_3 \ge K/2. \tag{3.28}$$

Relations (3.28), (3.27), and (3.5) imply that for almost every $t \in \Omega$,

$$g(t, x(t), u(t)) - g(t, y(t), v(t)) \ge g(t, x(t), u(t)) - (M + M_3)$$

$$\ge g(t, x(t), u(t)) - K/16 \ge (7/8)g(t, x(t), u(t)).$$

Therefore

$$\int_\Omega g(t, x(t), u(t))dt - \int_\Omega g(t, y(t), v(t))dt \ge (7/8) \int_\Omega g(t, x(t), u(t))dt. \tag{3.29}$$

By (3.19), (3.25), (3.16), (3.24), the choice of L, δ, and (B5)(iv), for almost every $t \in [T_1, T_2] \setminus \Omega$,

$$|f(t, x(t), u(t)) - f(t, y(t), v(t))| = |f(t, x(t), u(t)) - f(t, y(t), u(t))| \tag{3.30}$$

$$\leq L||x(t) - y(t)||(f(t, x(t), u(t)) + \psi(t)).$$

It follows from (3.7), (1.15), (3.5), (3.19), (3.10), (3.30), (3.25), (3.16), and (3.23) that for almost every $t \in [T_1, T_2] \setminus \Omega$,

$$|g(t, x(t), u(t)) - g(t, y(t), v(t))| = |g(t, x(t), u(t)) - g(t, y(t), u(t))|$$

$$\leq |f(t, x(t), u(t)) - f(t, y(t), u(t))| + |(g - f)(t, x(t), u(t)) - (g - f)(t, y(t), u(t))|$$

$$\leq L||x(t) - y(t)||(f(t, x(t), u(t)) + \psi(t)) + M||x(t) - y(t)||$$

$$\leq L||x(t) - y(t)||(g(t, x(t), u(t)) + M + \psi(t)) + M||x(t) - y(t)||$$

$$\leq d\kappa[L(g(t, x(t), u(t)) + M + \psi(t)) + M].$$

By this relation, (1.13), (3.15), and (3.3),

$$\left| \int_{[T_1,T_2]\setminus\Omega} g(t, x(t), u(t))dt - \int_{[T_1,T_2]\setminus\Omega} g(t, y(t), v(t))dt \right| \tag{3.31}$$

$$\leq d\kappa \int_{[T_1,T_2]\setminus\Omega} [L(g(t, x(t), u(t)) + M + \psi(t)) + M]dt$$

$$\leq d\kappa \int_{T_1}^{T_2} [L(g(t, x(t), u(t)) + c_3 + M + \psi(t)) + M]dt$$

$$\leq d\kappa \left[LI^g(x, u) + L(c_3 + M)(T_2 - T_1) + L \int_{T_1}^{T_2} \psi(t)dt + M(T_2 - T_1) \right]$$

$$\leq d\kappa \Bigg[L(M + M_1)(T_2 - T_1) + L(c_3 + M)(T_2 - T_1)$$

$$+ L \int_{T_1}^{T_2} \psi(t)dt + M(T_2 - T_1) + L \Bigg] \leq d\Delta_1.$$

It follows from (1.13), (3.10), (3.6), and (3.5) that for all $t \in \Omega$,

$$g(t, x(t), u(t)) \geq c_2\phi(||u(t)||) - c_3$$

$$\geq c_2\gamma^{-1}||u(t)|| - c_3 \geq c_2\gamma^{-1}||u(t)|| - ||u(t)|| = ||u(t)||(c_2\gamma^{-1} - 1) \geq c_2\gamma^{-1}2^{-1}||u(t)||.$$

Combined with (3.29), (3.10), and (3.4), this inequality implies that

$$\int_\Omega g(t, x(t), u(t))dt - \int_\Omega g(t, y(t), v(t))dt \geq (7/8)c_2\gamma^{-1}2^{-1} \int_\Omega ||u(t)||dt$$

$$\geq (2\gamma)^{-1}(7/8)c_2 d \geq 8\Delta_1 d.$$

Together with (3.31), (3.26), and (3.3) this inequality implies that $I^g(x, u) - I^g(y, v) \geq 7d\Delta_1 > Md$. This completes the proof of Theorem 3.1. □

Proof of Theorem 1.2. Lemma 2.5, (B2), and Lemma 2.7 imply that $f \in \mathcal{L}_{B,M}$ for each $M > 0$. Theorem 1.2 now follows from Theorem 1.1. □

4. Proofs of Theorems 1.3 and 1.4.

Proof of Theorem 1.3. Since $f \in \mathcal{L}_{A,M}$ assumption (A5) holds. Let positive numbers M_0, M_1, M_2 and an integrable scalar nonnegative function $\psi(t)$, $t \in [T_1, T_2]$, be as guaranteed by (A5). Choose a positive number ϵ_0 such that

$$\epsilon_0 < 16^{-1} \min\{\epsilon, 1, M\}(T_2 - T_1 + 1)^{-1}, \tag{4.1}$$

$$\epsilon_0 \left[(M + M_1)(T_2 - T_1) + 3(T_2 - T_1) + \int_{T_1}^{T_2} \lambda_1(t)dt + \int_{T_1}^{T_2} \psi(t)dt \right] < \epsilon/8.$$

Set

$$\kappa = c_2^{-1} \left[(M + M_1)(T_2 - T_1) + 1 + \int_{T_1}^{T_2} \lambda_1(t)dt \right]. \tag{4.2}$$

In view of (A2) there is $M_3 > 0$ such that for almost every $t \in [T_1, T_2]$,

(4.3)
$f(t, h, \xi) \le M_3$ for each $h \in B_E(M_2 + N_* + 1)$ and each $\xi \in B_F(M_2 + N_* + 1)$.

Since the function λ_1 is integrable there exists

$$\epsilon_1 \in (0, \epsilon_0) \tag{4.4}$$

such that

(4.5)
$\int_e \lambda_1(t)dt \le \epsilon/8$ for each Lebesgue measurable set $e \subset [T_1, T_2]$ satisfying $\mathrm{mes}(e) \le \epsilon_1$.

By the choice of M_0, M_1, M_2, ψ, and (A5)(iv) there exist $\Gamma > 0$, $\bar{\delta} \in (0, 1/4)$ such that for almost every $t \in [T_1, T_2]$ the inequality,

$$|f(t, x_1, u) - f(t, x_2, u)| \le \epsilon_0(f(t, x_1, u) + \psi(t)) \tag{4.6}$$

holds for each $x_1, x_2 \in B_E(M_2+1)$ satisfying $||x_1 - x_2|| \le \bar{\delta}$ and each $u \in F$ satisfying $||u|| \ge \Gamma$.

By (A3) there is

$$\delta \in (0, \bar{\delta}) \tag{4.7}$$

such that for almost every $t \in [T_1, T_2]$,

$$|f(t, x_1, u) - f(t, x_2, u)| \le \epsilon_0 \tag{4.8}$$

for each $u \in B_F(\Gamma+M_2+1)$ and each $x_1, x_2 \in B_E(\Gamma+M_2+1)$ satisfying $||x_1-x_2|| \le \delta$.

Choose a number $\gamma \in (0, 1)$ such that

$$\gamma(c_2 + 1)[\exp(c_0\kappa)2c_0\phi(M_2)c_2^{-1}\kappa] < \delta,\ \gamma(c_2 + 1)c_2^{-1}(\kappa + 1) < \epsilon_1. \tag{4.9}$$

By (1.2) we choose a positive number K such that

$$K > 16(M + M_0 + M_1 + M_2 + \Gamma + N_* + 1), \tag{4.10}$$

$$\phi(t)/t \geq \gamma^{-1} \text{ for all } t \geq K \text{ and } \phi(K) > (M_3+1)c_2^{-1}. \tag{4.11}$$

Set

$$\mathcal{U} = \{g \in \mathcal{M}_A : \ (f,g) \in \mathcal{E}_{Aw}(K,\epsilon_0)\}. \tag{4.12}$$

Assume that

$$g \in \mathcal{U}, \ D \in \mathcal{D}_M(E). \tag{4.13}$$

In order to prove the theorem, it is sufficient to show that there exists

$$(y,v) \in X(G,D) \tag{4.14}$$

such that

$$||v(t)|| \leq K, \ t \in [T_1,T_2] \text{ a.e. and } I^g(y,v) \leq \inf(I^g; X(G,D)) + \epsilon. \tag{4.15}$$

There exists a trajectory-control pair (x,u) such that

$$(x,u) \in X(G,D), \ I^g(x,u) \leq \inf(I^g; X(G,D)) + 4^{-1}\min\{\epsilon,1\}. \tag{4.16}$$

We may assume without loss of generality that

$$\text{mes}(\{t \in [T_1,T_2] : \ ||u(t)|| > K\}) > 0. \tag{4.17}$$

Set

$$\Omega_0 = \{t \in [T_1,T_2] : \ ||u(t)|| > K\}, \ \Omega_1 = \{t \in [T_1,T_2] : \ ||u(t)|| \leq \Gamma\},$$
$$\Omega_2 = [T_1,T_2] \setminus (\Omega_0 \cup \Omega_1), \ d = \int_{\Omega_0} ||u(t)||dt. \tag{4.18}$$

By Lemma 2.3 and (1.5) there exists $x_* \in W^{1,1}(T_1,T_2;E)$ such that (3.12) holds. In view of (3.12), (4.13), (1.5), the choice of M_0, and (A5)(i), relation (3.13) holds. It follows from the choice of M_1, (A2)(ii), (3.13), and (1.5) that for almost every $t \in [T_1,T_2]$, relation (3.14) holds.

By (3.14), (3.13), (1.5), (4.10), (4.12), (4.13), (4.1), and (1.10) for almost every $t \in [T_1,T_2]$,

$$g(t,x_*(t),u_*(t)) \leq f(t,x_*(t),u_*(t)) + \epsilon_0 \leq M_1 + M.$$

Together with (4.16) and (3.12) this inequality implies that

$$I^g(x,u) \leq I^g(x_*,u_*) + 1 \leq (M+M_1)(T_2-T_1) + 1. \tag{4.19}$$

It follows from (4.19), (A5)(iii), (4.13), (4.24), and the choice of M_2 that

$$||x(t)|| \leq M_2 \text{ for all } t \in [T_1,T_2]. \tag{4.20}$$

By (4.18), (4.11), (1.8), (4.19), and (4.2),

$$\begin{aligned} d = \int_{\Omega_0} ||u(t)||dt &\leq \int_{\Omega_0} \gamma\phi(||u(t)||)dt \leq \gamma \int_{T_1}^{T_2} \phi(||u(t)||)dt \\ &\leq \gamma \int_{T_1}^{T_2} c_2^{-1}[g(t,x(t),u(t)) + \lambda_1(t)]dt \\ &\leq \gamma c_2^{-1}\left[(M+M_1)(T_2-T_1) + 1 + \int_{T_1}^{T_2} \lambda_1(t)dt\right] \leq \gamma\kappa. \end{aligned} \tag{4.21}$$

Relation (4.18) implies that

$$\text{mes}(\Omega_0) \le K^{-1} \int_{\Omega_0} ||u(t)||dt \le d/K. \tag{4.22}$$

Set

$$v(t) = u(t),\ t \in [T_1, T_2] \setminus \Omega_0,\ v(t) = u_*(t),\ t \in \Omega_0. \tag{4.23}$$

Clearly $v : [T_1, T_2] \to F$ is a strongly measurable function. Relations (4.23), (4.18), (1.5), and (4.10) imply that

$$||v(t)|| \le K,\ t \in [T_1, T_2] \text{ a.e..} \tag{4.24}$$

In view of (4.24), Lemma 2.3, (4.13), and (4.16) there exists $y \in W^{1,1}(T_1, T_2; E)$ such that

$$y(T_1) = x(T_1),\ (y, v) \in X(G, D). \tag{4.25}$$

It follows from Lemma 2.4 (with $M_0 = K$), (4.16), (4.10), (4.18), (4.23), (4.25), (4.21), (4.20), and the monotonicity of ϕ that for each $t \in [T_1, T_2]$,

$$||y(t) - x(t)|| \le \exp\left(c_0 \int_{T_1}^{T_2} \phi(||v(t)||)dt\right) 2c_0 \left(\int_{\Omega_0} ||u(t)||dt\right) \phi(M_2). \tag{4.26}$$

By (4.23), (4.18), (4.10), (1.5), the monotonicity of ϕ, and (1.8),

$$\int_{T_1}^{T_2} \phi(||v(t)||)dt \le \int_{T_1}^{T_2} \phi(||u(t)||)dt \le c_2^{-1} \int_{T_1}^{T_2} [g(t, x(t), u(t)) + \lambda_1(t)]dt$$

$$\le c_2^{-1} \left[(M + M_1)(T_2 - T_1) + 1 + \int_{T_1}^{T_2} \lambda_1(t)dt\right].$$

Combined with (4.2), (4.26), (4.18), (4.21), and (4.9) this inequality implies that for each $t \in [T_1, T_2]$,

$$||y(t) - x(t)|| \le \exp(c_0\kappa)2c_0 d\phi(M_2) \le \exp(c_0\kappa)2c_0\phi(M_2)\kappa\gamma < \delta. \tag{4.27}$$

Together with the inequality $\delta < 1/4$ and (4.20) this inequality implies that

$$||y(t)|| \le M_2 + 1/4 \text{ for all } t \in [T_1.T_2]. \tag{4.28}$$

We will estimate $I^g(x, u) - I^g(y, v)$. It follows from (4.23), (1.5), (4.28), and (4.3) that for almost every $t \in \Omega_0$, $f(t, y(t), v(t)) \le M_3$. Together with (4.28), (4.23), (1.5), (4.13), (4.12), (4.1), (4.10), and (1.10) this inequality implies that for almost every $t \in \Omega_0$,

$$g(t, y(t), v(t)) \le f(t, y(t), v(t)) + \epsilon_0 < f(t, y(t), v(t)) + 1 \le M_3 + 1. \tag{4.29}$$

In view of (4.18), (1.8), and the monotonicity of ϕ for every $t \in \Omega_0$,

$$g(t, x(t), u(t)) \ge c_2\phi(||u(t)||) - \lambda_1(t) \ge c_2\phi(K) - \lambda_1(t).$$

By this inequality, (4.29), and (4.10),

$$\int_{\Omega_0} g(t,x(t),u(t))dt - \int_{\Omega_0} g(t,y(t),v(t))dt$$

$$\geq c_2\phi(K)\mathrm{mes}(\Omega_0) - \int_{\Omega_0} \lambda_1(t)dt - (M_3+1)\mathrm{mes}(\Omega_0) \geq -\int_{\Omega_0} \lambda_1(t)dt. \tag{4.30}$$

Relations (4.21), (4.22), (4.10), and (4.9) imply that $\mathrm{mes}(\Omega_0) \leq d \leq \gamma\kappa < \epsilon_1$. Together with (4.5) this inequality implies that $\int_{\Omega_0} \lambda_1(t)dt \leq \epsilon/8$. Combined with (4.30) this inequality implies that

$$\int_{\Omega_0} g(t,x(t),u(t))dt - \int_{\Omega_0} g(t,y(t),v(t))dt \geq -\epsilon/8. \tag{4.31}$$

By (4.28), (4.20), (4.18), (4.10), (4.23), (4.13), (4.12), and (1.10) for almost every $t \in [T_1,T_2] \setminus \Omega_0$,

$$|f(t,y(t),v(t)) - g(t,y(t),v(t))| \leq \epsilon_0,\ |f(t,x(t),u(t)) - g(t,x(t),u(t))| \leq \epsilon_0. \tag{4.32}$$

It follows from (4.18), (4.10), (4.23), (4.28), (4.20), the choice of δ (see (4.7)), and (4.27) that for almost every $t \in \Omega_1$,

$$|f(t,x(t),u(t)) - f(t,y(t),v(t))| = |f(t,x(t),u(t)) - f(t,y(t),u(t))| \leq \epsilon_0.$$

Combined with (4.32) this relation implies that for almost every $t \in \Omega_1$,

$$|g(t,x(t),u(t)) - g(t,y(t),v(t))| \leq |f(t,x(t),u(t)) - f(t,y(t),v(t))| + 2\epsilon_0 \leq 3\epsilon_0.$$

Together with (4.1) this relation implies that

$$\left|\int_{\Omega_1} g(t,x(t),u(t))dt - \int_{\Omega_1} g(t,y(t),v(t))dt\right| \leq 3\epsilon_0(T_2-T_1) < \epsilon/8. \tag{4.33}$$

By (4.18), (4.23), (4.28), (4.20), the choice of $\Gamma, \bar{\delta}$ (see (4.6)), (4.8), (4.7), and (4.27) for almost every $t \in \Omega_2$,

$$|f(t,x(t),u(t)) - f(t,y(t),v(t))| = |f(t,x(t),u(t)) - f(t,y(t),u(t))|$$
$$\leq \epsilon_0(f(t,x(t),u(t)) + \psi(t)).$$

Combined with (4.32) this relation implies that for almost every $t \in \Omega_2$,

$$|g(t,x(t),u(t)) - g(t,y(t),v(t))| \leq |f(t,x(t),u(t)) - f(t,y(t),v(t))| + 2\epsilon_0$$
$$\leq \epsilon_0(f(t,x(t),u(t)) + \psi(t)) + 2\epsilon_0 \leq \epsilon_0(g(t,x(t),u(t)) + \psi(t) + 3).$$

It follows from this relation, (1.8), (4.19), and (4.1) that

$$\left|\int_{\Omega_2} g(t,x(t),u(t))dt - \int_{\Omega_2} g(t,y(t),v(t))dt\right| \leq \epsilon_0 \int_{\Omega_2} (g(t,x(t),u(t)) + \psi(t) + 3)dt$$

$$\leq \epsilon_0 \int_{T_1}^{T_2} (g(t,x(t),u(t)) + \lambda_1(t) + \psi(t) + 3)dt$$

$$\leq \epsilon_0 \left[(M+M_1)(T_2-T_1) + 1 + 3(T_2-T_1) + \int_{T_1}^{T_2} \lambda_1(t)dt + \int_{T_1}^{T_2} \psi(t)dt\right] < \epsilon/8.$$

By this inequality, (4.33), (4.31), and (4.18), $I^g(x,u) - I^g(y,v) \geq -3\epsilon/8$. Together with (4.16) this relation implies that $I^g(y,v) \leq \inf(I^g; X(G,D)) + 3\epsilon/4$. This completes the proof of Theorem 1.3. □

Proof of Theorem 1.4. Lemma 2.5, (A2), and Lemma 2.7 imply that $f \in \mathcal{L}_{A,M}$ for each $M > 0$. Theorem 1.4 now follows from Theorem 1.3. □

5. A density result. In this section we prove the following result.

PROPOSITION 5.1. *Let $f \in \mathcal{M}_A$ and let ϵ, N, M be positive numbers. Then there exists $g \in \mathcal{L}_{A,M}$ such that*

- $(f,g) \in \mathcal{E}_{As}(N,\epsilon)$;
- *if $f \in \mathcal{M}_A^l$ (respectively, $f \in \mathcal{M}_A^c$), then $g \in \mathcal{M}^l \cap \mathcal{L}_{A,M}$ (respectively, $g \in \mathcal{L}_{A,M} \cap \mathcal{M}_A^c$);*
- *if $\lambda_1(t) = c_3$ for all $t \in [T_1,T_2]$ and if $f \in \mathcal{M}_B$ (respectively, $f \in \mathcal{M}_B^a$, $f \in \mathcal{M}_B^l$, $f \in \mathcal{M}_B^c$), then $g \in \mathcal{L}_{B,M}$ (respectively, $g \in \mathcal{L}_{B,M} \cap \mathcal{M}_B^a$, $g \in \mathcal{L}_{B,M} \cap \mathcal{M}_B^l$, $g \in \mathcal{L}_{B,M} \cap \mathcal{M}_B^c$) and there are constants $M_0 > 0, M_1 > 2, M_2 > 0, L > 0$ such that* (B5)(i) *and* (B5)(iii) *hold for g and that for almost every $t \in [T_1,T_2]$,*

$$g(t,x,u) \leq M_1 - 2 \text{ for each } x \in B_E(M + M_0 + N_*) \text{ and each } u \in B_F(M + M_0 + N_*),$$

$$|g(t,x_1,u) - g(t,x_2,u)| \leq L||x_1 - x_2|| \text{ for each } x_1, x_2 \in B_E(M_2 + 1) \text{ and each } u \in F.$$

In the proof of Proposition 5.1 we use the following auxiliary result which is proved in a straightforward manner.

LEMMA 5.1. *Let $f_1, f_2 : [T_1,T_2] \times E \times F \to R^1$ be scalar functions. Then the following assertions hold:*

1. *If* (A1)–(A3) *hold for f_1, f_2, then* (A1)–(A3) *hold for $f_1 + f_2$, $f_1 f_2$.*
2. *If* (B1)–(B3) *hold for f_1, f_2, then* (B1)–(B3) *hold for $f_1 + f_2$, $f_1 f_2$.*

Proof of Proposition 5.1. Consider a function $\tilde{\phi} : [0,\infty) \to [0,\infty)$ such that for each integer $i \geq 0$,

(5.1)
$$\tilde{\phi}(i) = \phi(i+1),\ \tilde{\phi}(\alpha i + (1-\alpha)(i+1)) = \alpha\tilde{\phi}(i) + (1-\alpha)\tilde{\phi}(i+1) \text{ for all } \alpha \in [0,1].$$

Clearly the function $\tilde{\phi} : [0,\infty) \to [0,\infty)$ is increasing and Lipschitzian on all bounded subsets of $[0,\infty)$,

$$\tilde{\phi}(t) \geq \phi(t) \text{ for all } t \geq 0 \text{ and } \lim_{t\to\infty} \tilde{\phi}(t)/t = \infty. \tag{5.2}$$

By Lemma 2.5 there exists $M_0 > 0$ such that the following property holds:

(P1) For each $D \in \mathcal{D}_M(E)$ and each $(x,u) \in X(G,D)$ which satisfies $||u(t)|| \leq M + N_*$ for almost every $t \in [T_1,T_2]$ the inequality $||x(t)|| \leq M_0$ holds for all $t \in [T_1,T_2]$.

By (A2) there exists $M_1 > 2$ such that for almost every $t \in [T_1,T_2]$ the following inequality holds:

(5.3)
$$f(t,x,u) \leq M_1 - 2 \text{ for each } x \in B_E(M + M_0 + N_*) \text{ and each } u \in B_F(M + M_0 + N_*).$$

By Lemma 2.7 there is $M_2 > 0$ such that the following property holds:

(P2) For each $g \in \mathcal{M}_A$, each $D \in \mathcal{D}_M(E)$, and each $(x,u) \in X(G,D)$ satisfying $I^g(x,u) \leq (M + M_1)(T_2 - T_1) + 1$ the inequality $||x(t)|| \leq M_2$ holds for all $t \in [T_1,T_2]$.

Choose natural numbers

$$N_1 > 4 + N + M + M_0 + N_*,\ N_2 > 4 + M_2 + N + M_0 + N_* + M. \tag{5.4}$$

For $i = 1, 2$ define

$$\begin{aligned} &\psi_i(t) = 1,\ t \in [0, N_i + 1], \\ &\psi_i(t) = 0,\ t \in [N_i + 2, \infty), \\ &\psi_i(t) = N_i + 2 - t,\ t \in (N_i + 1, N_i + 2). \end{aligned} \tag{5.5}$$

Set

$$d_0 = \sup\{||G(t, 0, 0)|| :\ t \in [T_1, T_2]\}. \tag{5.6}$$

It follows from (1.3), (1.4), and (5.6) that for each $(t, y, v) \in [T_1, T_2] \times B_E(N_1 + N_2) \times F$,

$$||G(t, y, v)|| \le ||G(t, 0, v)|| + c_0||y||\phi(||v||) \le ||G(t, 0, 0)|| + c_0||v||\phi(0)$$

$$+ c_0(N_1 + N_2)\phi(||v||) \le d_0 + c_0||v||\phi(0) + c_0(N_1 + N_2)\phi(||v||). \tag{5.7}$$

For each $(t, x, u) \in [T_1, T_2] \times E \times F$ define

$$\begin{aligned} h(t, x, u) = \psi_1(||u||)f(t, x, u) + (1 - \psi_1(||u||))[c_2\tilde{\phi}(||u||) \\ + c_1 d_0 + c_1 c_0||u||\phi(0) + c_1 c_0(N_1 + N_2)\tilde{\phi}(||u||)], \end{aligned} \tag{5.8}$$

$$g(t, x, u) = \psi_2(||x||)h(t, x, u) + (1 - \psi_2(||x||))f(t, x, u). \tag{5.9}$$

Clearly (A1) and (A2) hold for h, g.

It is not difficult to see that (B1), (B2), and (B3) hold for each of the following functions:

$$(t, x, u) \to \psi_1(||u||),\ (t, x, u) \to \psi_2(||x||),$$

$$(t, x, u) \to \tilde{\phi}(||u||),\ (t, x, u) \to ||u||,\ (t, x, u) \in [T_1, T_2] \times E \times F.$$

Together with Lemma 5.1 and the definition of g, h (see (5.8), (5.9)) this implies that (A3) holds for g, h, and if (B3) holds for f, then (B3) holds for g, h. Thus we have shown that g satisfies (A1)–(A3), and if $\lambda_1(t) = c_3$ for all $t \in [T_1, T_2]$ and if $f \in \mathcal{M}_B$, then g satisfies (B3). We will show that (A4) holds for g.

Let $(t, x, u) \in [T_1, T_2] \times E \times F$. There are two cases:

$$||x|| \ge N_2 + 2, \tag{5.10}$$

$$||x|| < N_2 + 2. \tag{5.11}$$

Assume that (5.10) is true. Then by (5.10), (5.9), (5.6), and (A4), which holds for f,

$$g(t, x, u) = f(t, x, u) \ge \max\{c_1||G(t, x, u)|| - \lambda_0(t),\ c_2\phi(||u||) - \lambda_1(t)\}. \tag{5.12}$$

Assume that (5.11) is true. It follows from (5.7), (5.11), (5.4), and (5.2) that

$$\max\{c_1||G(t, x, u)||, c_2\phi(||u||)\} \le [d_0 + c_0||u||\phi(0) + c_0(N_1 + N_2)\phi(||u||)]c_1 + c_0\phi(||u||)$$

$$\leq c_1[d_0 + c_0||u||\phi(0) + c_0(N_1 + N_2)\tilde{\phi}(||u||)] + c_2\tilde{\phi}(||u||).$$

Since (A4) holds for f it follows from this inequality, (5.8), and (5.5) that

$$h(t,x,u) \geq \max\{c_1||G(t,x,u)|| - \lambda_0(t),\ c_2\phi(||u||) - \lambda_1(t)\}.$$

Since (A4) holds for f it follows from this inequality, (5.5), and (5.9) that

$$g(t,x,u) \geq \max\{c_1||G(t,x,u)|| - \lambda_0(t),\ c_2\phi(||u||) - \lambda_1(t)\}.$$

Therefore this inequality holds in both cases, (A4) holds for g, and $g \in \mathcal{M}_A$.

It is clear that if $\lambda_1(t) = c_3$ for all $t \in [T_1, T_2]$ and if $f \in \mathcal{M}_B$, then (B4) holds for g and $g \in \mathcal{M}_B$. Thus we have shown that

(5.13) $g \in \mathcal{M}_A$ and if $\lambda_1(t) = c_3$ for all $t \in [T_1, T_2]$ and if $f \in \mathcal{M}_B$, then $g \in \mathcal{M}_B$.

Evidently

(5.14) if $f \in \mathcal{M}_A^l$ (respectively, $\mathcal{M}_A^c$), then $g \in \mathcal{M}_A^l$ (respectively, $\mathcal{M}_A^c$), if $\lambda_1(t) = c_3$ for all $t \in [T_1, T_2]$ and if $f \in \mathcal{M}_B^a$(respectively, $\mathcal{M}_B^l, \mathcal{M}_B^c$), then $g \in \mathcal{M}_B^a$ (respectively $\mathcal{M}_B^l, \mathcal{M}_B^c$).

We will show that (A5) holds for g. Note that we have already defined the constants $M_0 > 0, M_1 > 2, M_2 > 0$. By (5.9), (5.8), (5.5), and (5.4) for each $t \in [T_1, T_2]$, each $x \in B_E(M + M_0 + N_*)$, and each $u \in B_F(N + M_0 + N_*)$,

$$f(t,x,u) = g(t,x,u). \tag{5.15}$$

Properties (P1) and (P2) and (5.3) imply (A5)(i), (A5)(ii), and (A5)(iii) for g.

Let us show that (A5)(iv) holds.

Assume that

$$x_1, x_2 \in B_E(M_2 + 1),\ u \in F,\ ||u|| \geq N_1 + 2. \tag{5.16}$$

It follows from (5.9), (5.4), (5.16), and (5.5) that for $i = 1, 2$ and all $t \in [T_1, T_2]$,

$$g(t,x_i,u) = h(t,x_i,u) = c_2\tilde{\phi}(||u||) + c_1d_0 + c_1c_0||u||\phi(0) + c_1c_0(N_1 + N_2)\tilde{\phi}(||u||),$$

and $g(t,x_1,u) = g(t,x_2,u)$. Therefore (A5)(iv) holds for g, (A5) holds for g, and $g \in \mathcal{L}_{A,M}$.

Together with (5.14) this implies that

(5.17) $g \in \mathcal{L}_{A,M}$, and if $f \in \mathcal{M}_A^l$ (respectively, $\mathcal{M}_A^c$), then $g \in \mathcal{L}_{A,M} \cap \mathcal{M}_A^l$ (respectively, $\mathcal{L}_{A,M} \cap \mathcal{M}_A^c$).

It follows from (5.8), (5.9), (5.4), and (5.5) that for each $(t,x,u) \in [T_1, T_2] \times B_E(N) \times B_F(N)$,

$$g(t,x,u) = h(t,x,u) = f(t,x,u)$$

and

$$(f,g) \in \mathcal{E}_{As}(N,\epsilon).$$

Now assume that $\lambda_1(t) = c_3$ for all $t \in [T_1, T_2]$ and that $f \in \mathcal{M}_B$. We have already shown that $g \in \mathcal{M}_B$ (see (5.13)). We will show that (B5) holds for g. Note that we have already defined the constants $M_0 > 0, M_1 > 2, M_2 > 0$. Relation (5.15), properties (P1) and (P2), and (5.3) imply (B5)(i), (B5)(ii), and (B5)(iii) for g.

Let us show that (B5)(iv) holds. Since (B3) holds for g there exist $L > 0$ and a Lebesgue measurable set $\Omega \subset [T_1, T_2]$ such that $\mathrm{mes}(\Omega) = T_2 - T_1$ and

$$|g(t, x_1, u) - g(t, x_2, u)| \le L||x_1 - x_2|| \tag{5.18}$$

for each $t \in \Omega$, each $x_1, x_2 \in B_E(M_2 + N_1 + 2)$, and each $u \in B_F(M_2 + N_1 + 2)$.

Assume that

$$x_1, x_2 \in B_E(M_2 + 1),\ u \in F. \tag{5.19}$$

There are two cases:

$$||u|| \ge N_1 + 2, \tag{5.20}$$

$$||u|| < N_1 + 2. \tag{5.21}$$

By (5.9), (5.19), and (5.4) for all $t \in [T_1, T_2]$,

$$g(t, x_i, u) = h(t, x_i, u),\ i = 1, 2. \tag{5.22}$$

Assume that (5.20) holds. Then by (5.20), (5.5), and (5.8) for all $t \in [T_1, T_2]$ we have $h(t, x_1, u) = h(t, x_2, u)$. Together with (5.22) this implies that for all $t \in [T_1, T_2]$ we have $g(t, x_1, u) = g(t, x_2, u)$. Assume that (5.21) holds. Then in view of (5.19), (5.21), and the choice of L, inequality (5.18) is true for all $t \in \Omega$. Clearly, (5.18) holds in both cases for all $t \in \Omega$. Thus we have shown that (5.18) holds for each $u \in F$, each $x_1, x_2 \in B_E(M_2 + 1)$, and each $t \in \Omega$. Therefore (B5)(iv) holds for g and $g \in \mathcal{L}_{B,M}$. This completes the proof of Proposition 5.1. □

It is easy to see that Proposition 5.1 implies the following result.

PROPOSITION 5.2. *Let $M > 0$. Then the set $\mathcal{L}_{A,M}$ (respectively, $\mathcal{L}_{A,M} \cap \mathcal{M}_A^l$, $\mathcal{L}_{A,M} \cap \mathcal{M}_A^c$) is an everywhere dense subset of $\mathcal{M}_A$ (respectively, $\mathcal{M}_a^l$, $\mathcal{M}_A^c$) with the strong topology, and the set $\mathcal{L}_{B,M}$ (respectively, $\mathcal{L}_{B,M} \cap \mathcal{M}_B^a$, $\mathcal{L}_{B,M} \cap \mathcal{M}_B^l$, $\mathcal{L}_{B,M} \cap \mathcal{M}_B^c$) is an everywhere dense subset of $\mathcal{M}_B$ (respectively, $\mathcal{M}_B^a$, $\mathcal{M}_a^l$, $\mathcal{M}_A^c$) with the strong topology.*

6. Proofs of Theorems 1.5–1.7.

Proof of Theorem 1.5. Set $M_0 = 8M$. By Proposition 5.2, $\mathcal{L}_{B,M_0}$ (respectively, $\mathcal{L}_{B,M_0} \cap \mathcal{M}_B^a$, $\mathcal{L}_{B,M_0} \cap \mathcal{M}_B^l$, $\mathcal{L}_{B,M_0} \cap \mathcal{M}_B^c$) is an everywhere dense subset of $\mathcal{M}_B$ (respectively, $\mathcal{M}_B^a$, $\mathcal{M}_B^l$, $\mathcal{M}_B^c$) with the strong topology.

Let $f \in \mathcal{L}_{B,M_0}$. By Theorem 1.1 there is $K(f) > 0$ such that the following property holds:

(P3) For each $g \in \mathcal{M}_B$ satisfying $(f, g) \in \mathcal{E}_{Bw}(K(f), 8M)$, each $D \in \mathcal{D}_{8M}(E)$, and each $(x, u) \in X(G, D)$ satisfying $\mathrm{mes}(\{t \in [T_1, T_2] :\ ||u(t)|| > K(f)\}) > 0$, there exists $(y, v) \in X(G, D)$ such that $I^g(y, v) < I^g(x, u)$ and $||v(t)|| \le K(f)$ for almost every $t \in [T_1, T_2]$.

Denote by $\mathcal{U}(f)$ an open neighborhood of f in $\mathcal{M}_B$ with the weak topology such that

$$\{g \in \mathcal{M}_B :\ (f, g) \in \mathcal{E}_{Bw}(K(f), M)\} \subset \mathcal{U}(f) \subset \{g \in \mathcal{M}_B :\ (f, g) \in \mathcal{E}_{Bw}(K(f), 2M)\}. \tag{6.1}$$

Define $\mathcal{F} = \cup\{\mathcal{U}(f) : \ f \in \mathcal{L}_{B,M_0}\}$. Clearly, $\mathcal{F}$ is an open (in the weak topology) everywhere dense (in the strong topology) subset of $\mathcal{M}_B$, and $\mathcal{F} \subset \mathcal{M}_B^a$ (respectively, $\mathcal{F} \cap \mathcal{M}_B^l$, $\mathcal{F} \cap \mathcal{M}_B^c$) contains an open (in the weak topology) everywhere dense (in the strong topology) subset of $\mathcal{M}_B^a$ (respectively, $\mathcal{M}_B^l$, $\mathcal{M}_B^c$).

Let $g \in \mathcal{F}$. By the definition of $\mathcal{F}$ there is $f \in \mathcal{L}_{B,M_0}$ such that $g \in \mathcal{U}(f)$. Assume that

$$h \in \mathcal{M}_B, \ (g,h) \in \mathcal{E}_{Bw}(K(f), M), \ D \in \mathcal{D}_M(E), \ \ (x,u) \in X(G,D), \tag{6.2}$$

$$\mathrm{mes}(\{t \in [T_1,T_2] : \ ||u(t)|| > K(f)\}) > 0. \tag{6.3}$$

By (6.1), (6.2), and the inclusion $g \in \mathcal{U}(f)$ we have $(h,f) \in \mathcal{E}_{Bw}(K(f), 3M)$. It follows from this inclusion, property (P3), (6.2), and (6.3) that there is $(y,v) \in X(G,D)$ such that $I^h(y,v) < I^h(x,u)$ and $||v(t)|| \le K(f)$ for almost every $t \in [T_1,T_2]$. Theorem 1.5 is proved. □

Theorem 1.6 easily follows from Theorem 1.5.

Proof of Theorem 1.7. Let n be a natural number. By Proposition 5.2, $\mathcal{L}_{A,n}$ (respectively, $\mathcal{L}_{A,n} \cap \mathcal{M}_A^l$, $\mathcal{L}_{A,n} \cap \mathcal{M}_A^c$) is an everywhere dense subset of $\mathcal{M}_A$ (respectively, $\mathcal{M}_A^l$, $\mathcal{M}_A^c$) with the strong topology. By Theorem 1.3 for each $f \in \mathcal{L}_{A,n}$ there exist $K(f,n) > 0$ and an open neighborhood $\mathcal{U}(f,n)$ of f in $\mathcal{M}_A$ with the weak topology such that the following property holds:

(P4) For each $g \in \mathcal{U}(f,n)$ and each $D \in \mathcal{D}_n(E)$ there is $(x,u) \in X(G,D)$ such that $||u(t)|| \le K(f,n)$ for almost every $t \in [T_1,T_2]$ and $I^g(x,u) \le \inf(I^g; X(G,D)) + 1/n$.

Define $\mathcal{F}_A = \cap_{n=1}^{\infty} \cup \{\mathcal{U}(f,n) : \ f \in \mathcal{L}_{A,n}\}$. Clearly $\mathcal{F}_A$ is a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) subsets of $\mathcal{M}_A$, and $\mathcal{F}_A \cap \mathcal{M}_A^l$ (respectively, $\mathcal{F}_A \cap \mathcal{M}_A^c$) contains a countable intersection of open (in the weak topology) everywhere dense (in the strong topology) subsets of $\mathcal{M}_A^l$ (respectively, $\mathcal{M}_A^c$). Let

$$g \in \mathcal{F}_A, \ D \in \mathcal{D}(E), \ \epsilon > 0. \tag{6.4}$$

Choose a natural number m such that

$$D \in \mathcal{D}_m(E), \ \epsilon > 1/m. \tag{6.5}$$

By (6.4) and the definition of $\mathcal{F}_A$ there is $f_m \in \mathcal{L}_{A,m}$ such that $g \in \mathcal{U}(f_m, m)$. It follows from this inclusion, the property (P4), and (6.5) that there is $(x,u) \in X(G,D)$ such that

$$||u(t)|| \le K(f_m,m) \text{ for almost every } t \in [T_1,T_2], \ I^g(x,u) \le \inf(I^g : X(G,D)) + 1/m.$$

In view of these inequalities and (6.5) $\inf(I^g; X_L(G,D)) \le I^g(x,u) \le \inf(I^g; X(G,D)) + \epsilon$. Since ϵ is an arbitrary positive number we conclude that $\inf(I^g; X_L(G,D)) = \inf(I^g; X(G,D))$. This completes the proof of Theorem 1.7. □

7. Examples. We assume that $\phi : [0,\infty) \to [0,\infty)$ is an increasing function satisfying (1.2), $c_0, \tilde{c}_1, \tilde{c}_2$ are positive constants, and $G_1 : [T_1,T_2] \times E \to E$, $G_2 : [T_1,T_2] \times F \to E$ are continuous mappings such that

$$||G_1(t,y_1) - G_1(t,y_2)|| \le c_0||y_1 - y_2||\phi(0) \text{ for each } y_1, y_2 \in E \text{ and each } t \in [T_1,T_2],$$

$$||G_2(t,v_1) - G_2(t,v_2)|| \le c_0||v_1 - v_2||\phi(0) \text{ for each } v_1, v_2 \in F \text{ and each } t \in [T_1,T_2],$$

$$G(t,y,v) = G_1(t,y) + G_2(t,v),\ (t,y,v) \in [T_1,T_2] \times E \times F.$$

It is easy to see that (1.3) holds for each $t \in [T_1,T_2]$, each $y_1, y_2 \in E$, and each $v \in F$, and that (1.4) holds for each $t \in [T_1,T_2]$, each $y \in E$, and each $v_1, v_2 \in F$.

Example 1. Assume that $h : [T_1,T_2] \times E \to R^1$ and $g : [T_1,T_2] \times F \to R^1$ are nonnegative continuous functions which are bounded on bounded sets such that for each $M, \epsilon > 0$ there exists $\delta > 0$ such that $|h(t,x_1) - h(t,x_2)| \le \epsilon$ for each $t \in [T_1,T_2]$ and each $x_1, x_2 \in B_E(M)$ satisfying $||x_1 - x_2|| \le \delta$; that

$$h(t,x) \ge \tilde{c}_1||x||\phi(0) \text{ for each } (t,x) \in [T_1,T_2] \times E, \tag{7.1}$$

$$g(t,u) \ge \tilde{c}_1||u||\phi(0) \text{ for each } (t,u) \in [T_1,T_2] \times F;$$

and that

$$g(t,v) \ge \tilde{c}_2\phi(||v||) \text{ for each } (t,v) \in [T_1,T_2] \times E. \tag{7.2}$$

It is not difficult to see that the function $f(t,x,u) = h(t,x) + g(t,u)$, $(t,x,u) \in [T_1,T_2] \times E \times F$ satisfies (A1)–(A3), (A6), and (A4) with $c_2 = \tilde{c}_2$, $\lambda_1(t) = 0$ for all $t \in [T_1,T_2]$, $c_1 = \tilde{c}_1 c_0^{-1}$, $\lambda_0(t) = c_4$ for all $t \in [T_1,T_2]$, where c_4 is a sufficiently large positive constant.

Example 2. Assume that h and g are as in Example 1 and that the following condition holds:

• For each $M > 0$ there exists $L > 0$ such that

$$|h(t,x_1) - h(t,x_2)| \le L||x_1 - x_2|| \text{ for each } t \in [T_1,T_2] \text{ and each } x_1, x_2 \in B_E(M).$$

It is not difficult to see that the function $f(t,x,u) = h(t,x) + g(t,u)$, $(t,x,u) \in [T_1,T_2] \times E \times F$ satisfies (B1), (B2), (B3), (B6), and (B4) with $c_2 = \tilde{c}_2$, $c_3 = 0$, $c_1 = \tilde{c}_1 c_0^{-1}$, and $\lambda_0(t) = c_4$ for all $t \in [T_1,T_2]$, where c_4 is a sufficiently large positive number.

Example 3. Assume that h and g are as in Example 1 and that

$$\inf\{h(t,x) :\ t \in [T_1,T_2],\ x \in E\} > 0,\ \inf\{g(t,x) :\ t \in [T_1,T_2],\ v \in F\} > 0.$$

It is not difficult to see that the function $f(t,x,u) = h(t,x)g(t,u)$, $(t,x,u) \in [T_1,T_2] \times E \times F$ satisfies (A1), (A2), (A3), (A6), and (A4) with $\lambda_1(t) = 0$ for all $t \in [T_1,T_2]$, a sufficiently small positive constant c_2, and $\lambda_0(t) = c_4$ for all $t \in [T_1,T_2]$, where c_4 is a sufficiently large positive number.

Example 4. Assume that h, g are as in Example 3 and that the following condition holds:

• For each $M > 0$ there exists $L > 0$ such that

$$|h(t,x_1) - h(t,x_2)| \le L||x_1 - x_2|| \text{ for each } t \in [T_1,T_2] \text{ and each } x_1, x_2 \in B_E(M).$$

It is not difficult to see that the function $f(t,x,u) = h(t,x)g(t,u)$, $(t,x,u) \in [T_1,T_2] \times E \times F$ satisfies (B1), (B2), (B3), (B6), and (B4) with $c_3 = 0$, sufficiently small positive constants c_1, c_2, and a sufficiently large positive number c_4.

Example 5. Let us consider the following optimal control problem with $E = F = R^1$, $T_1 = 0$, $T_2 = 1$, and

$$\phi(t) = t^2 + 1 \text{ for all } t \in [0,\infty),\ c_0 = 100.$$

Set

$$G(t,y,v) = (3t+4y+1)|2t+v+2| + (t+v+4)|2t+5y+1|,\ (t,y,v) \in [0,1]\times R^1\times R^1.$$

Clearly the function $G : [0,1]\times R^1\times R^1 \to R^1$ is continuous. It is not difficult to verify that (1.2) holds, (1.3) holds for each $t \in [0,1]$ and each $y_2, y_2, v \in R^1$, and (1.4) holds for each $t \in [0,1]$ and each $y, v_1, v_2 \in R^1$.

For $t \in [0,1]$ set

$$U(t) = [-t^2 - 4t - 4, t^2 + 4t + 4].$$

Clearly (1.5) holds with $N_* = 1$ and $u_*(t) = 0$ for all $t \in [0,1]$.

For each $(t,x,u) \in [0,1]\times R^1\times R^1$ set

$$f(t,x,u) = (\ln(1+t)+1)(x^2+|x|+1)(e^u+2+t) + (u^2+|u|+2t+1)(e^{|x|}+t+1).$$

It is not difficult to see that the integrand f satisfies (B1), (B2), (B3), (B4), and (B6) with sufficiently small positive numbers c_1, c_2 and sufficiently large positive number c_3. Thus the assumptions of Theorem 1.2 hold and it can be applied to the optimal control problem

$$\int_0^1 (\ln(1+t)+1)((x(t))^2+|x(t)|+1)(e^{u(t)}+2+t)$$
$$+((u(t))^2+|u(t)|+2t+1)(e^{|x(t)|}+t+1)dt \to \min,$$

$$x'(t) = (3t+4x(t)+1)|2t+u(t)+2| + (t+u(t)+4)|2t+5x(t)+1|,\ t\in[0,1] \text{ a.e.},$$

$$x(0) = 1,\ u(t) \in [-t^2-4t-4, t^2+4t+4],\ t \in [0,1] \text{ a.e.}.$$

It is clear that this example is not covered by the results of [12], [18], [24].

8. Sarychev integrands. Let γ, σ, k, l, m, r be positive numbers which satisfy

$$l > k > 1,\ m \ge (1+kr)(1-k/l)^{-1}, \tag{8.1}$$

and let C^1-smooth functions $\eta : R^1 \to R^1$, $\Psi_0 : R^1 \to R^1$ and the C^2-smooth function $\psi_0 : R^2 \to R^1$ satisfy the following conditions:

$$\Psi_0(0) = 0,\ \Psi_0(z) \ge \gamma|z|^r \text{ for all } z \in R^1,\ \eta(v) \ge \sigma|v|^m \text{ for all } v \in R^1, \tag{8.2}$$

$$\psi_0(1,1) = 1,\ \psi_0(t,0) > 0 \text{ for all } t \in [0,1], \tag{8.3}$$

$$|\psi_0(t,x)| \le \Theta(x) \text{ for all } (t,x) \in [0,1]\times[0,\infty),$$

where

$$\lim_{x\to\infty} \Theta(x)x^{-1} = 0$$

and

$$(x\partial\psi_0/\partial x - l\psi_0) < 0 \text{ in } [0,1]\times[0,\infty). \tag{8.4}$$

Assume that $L(t,x,v)$ is a C^1-smooth integrand that admits the bounds

$$a \le L(t,x,v) \le b(t,x)(1+|v|^p),\ 0 \le p < l/(l-k), \tag{8.5}$$

with $a \in R^1$ and $b(t,x)$ being positive and bounded on bounded subsets of $[0,1] \times [0,\infty)$.

Sarychev [18, Theorem 2.1] established the existence of a positive number ϵ_0 such that for any $\epsilon \in [0,\epsilon_0]$, the variational problem

$$\int_0^1 \epsilon L(t,x(t),x'(t)) + \Psi_0((x(t))^l - t^k\psi_0(t,x(t)))\eta(x'(t))dt \to \min,$$

$$x(0) = 0,\ x(1) = 1$$

exhibits the Lavrentiev phenomenon.

In this section we prove the following result.

PROPOSITION 8.1. *For any $s \in R^1$, the function*

$$f(t,x,u) := sL(t,x,u) + \Psi_0(x^l - t^k\psi_0(t,x))\eta(u),\ (t,x,u) \in [0,1] \times R^1 \times R^1 \tag{8.6}$$

does not satisfy assumption (A6).

Proof. Let $s \in R^1$. Set

$$F(t,x) = x^l - t^k\psi_0(t,x),\ (t,x) \in R^1 \times R^1. \tag{8.7}$$

It is clear that $F \in C^2(R^2)$ and $F(1,1) = 0$. By (8.3) and (8.4),

$$(\partial F/\partial x)(1,1) = l - (\partial\psi_0/\partial x)(1,1) > 0.$$

By the implicit function theorem [5] there exist $\gamma_1 > 0$, $\gamma_2 > 0$, and a continuous differentiable function $\xi : (1-\gamma_2, 1+\gamma_2) \to R^1$ such that

$$\xi(1) = 1,$$

$$\text{if } t \in (1-\gamma_2, 1+\gamma_2),\ \text{then } |\xi(t) - 1| < \gamma_1 \text{ and } F(t,\xi(t)) = 0, \tag{8.8}$$

$$\text{for each } (t,x) \in (1-\gamma_2, 1+\gamma_2) \times (1-\gamma_1, 1+\gamma_1) \text{ the equality} \tag{8.9}$$

$$F(t,x) = 0 \text{ holds if and only if } x = \xi(t).$$

Assume that the integrand f satisfies assumption (A6) and let an integrable scalar nonnegative function $\psi(t)$, $t \in [0,1]$, be as guaranteed by (A6). In view of (A6) there exist $\Gamma, \delta > 0$ such that for almost every $t \in [0,1]$ the following property holds:

$$|f(t,x_1,u) - f(t,x_2,u)| \le f(t,x_2,u) + \psi(t) \tag{P5}$$

for all $x_1, x_2 \in [-2-\gamma_1, 2+\gamma_1]$ satisfying $|x_1 - x_2| \le \delta$ and each $u \in R^1$ satisfying $|u| \ge \Gamma$.

It follows from the definitions of Γ and δ that there exists

$$t_0 \in (1-\gamma_2/2, 1+\gamma_2/2) \cap (0,1) \tag{8.10}$$

such that $\psi(t_0)$ is finite and property (P5) holds with $t = t_0$.

Clearly $\xi(t_0)$ is well defined and

$$|\xi(t_0) - 1| < \gamma_1,\ F(t_0,\xi(t_0)) = 0. \tag{8.11}$$

Since property (P5) holds with $t = t_0$, it follows from (8.11) that

$$|f(t_0,x,u) - f(t_0,\xi(t_0),u)| \le f(t_0,\xi(t_0),u) + \psi(t_0) \tag{8.12}$$

for each

$$x \in [-2-\gamma_1, 2+\gamma_1] \cap [\xi(t_0)-\delta, \xi(t_0)+\delta] \text{ and each } u \in R^1 \setminus (-\Gamma, \Gamma). \tag{8.13}$$

Choose a number x_0 such that

$$x_0 \in (1-\gamma_1, 1+\gamma_1) \cap (\xi(t_0)-\delta, \xi(t_0)+\delta) \setminus \{\xi(t_0)\}. \tag{8.14}$$

It follows from (8.12), (8.13), and (8.14) that

$$|f(t_0, x_0, u) - f(t_0, \xi(t_0), u)| \le f(t_0, \xi(t_0), u) + \psi(t_0) \text{ for each } u \in R^1 \setminus (-\Gamma, \Gamma). \tag{8.15}$$

Let

$$u \in R^1 \setminus (-\Gamma, \Gamma). \tag{8.16}$$

By (8.10), (8.7), the choice of the function ξ, and (8.8),

$$(\xi(t_0))^l - t_0^k \psi_0(t_0, \xi(t_0)) = 0. \tag{8.17}$$

Together with (8.6) and (8.2) this implies that

$$f(t_0, \xi(t_0), u) = sL(t_0, \xi(t_0), u). \tag{8.18}$$

By (8.7), (8.10), (8.14), and (8.9),

$$|x_0^l - t_0^k \psi_0(t_0, x_0)| = |F(t_0, x_0)| > 0.$$

Together with (8.2) this inequality implies that

$$\Lambda := \Psi_0(x_0^l - t_0^k \psi_0(t_0, x_0)) > 0. \tag{8.19}$$

In view of (8.15), (8.18), (8.19), (8.6), and (8.5) for each $u \in R^1 \setminus (-\Gamma, \Gamma)$,

$$sL(t_0, \xi(t_0), u) + \psi(t_0) \ge f(t_0, x_0, u) - f(t_0, \xi(t_0), u)$$

$$= sL(t_0, x_0, u) + \Lambda\eta(u) - sL(t_0, \xi(t_0), u)$$

and

$$\Lambda\eta(u) \le \psi(t_0) + 2sL(t_0, \xi(t_0), u) - as \text{ for each } u \in R^1 \setminus (-\Gamma, \Gamma). \tag{8.20}$$

It follows from (8.2), (8.19), (8.20), and (8.5) that for each $u \in R^1 \setminus (-\Gamma, \Gamma)$,

$$|u|^m \le \sigma^{-1}\eta(u) \le \sigma^{-1}\Lambda^{-1}[\psi(t_0) + 2sL(t_0, \xi(t_0), u) - as]$$

$$\le \sigma^{-1}\Lambda^{-1}[\psi(t_0) + 2|s|a| + 2|s|b(t_0, \xi(t_0))(1+|u|^p) - as].$$

Since the inequality above holds for all $u \in R^1 \setminus (-\Gamma, \Gamma)$, we conclude that $m \le p$.

On the other hand, it follows from (8.1) and (8.5) that $m \ge (1+kr)l(l-k)^{-1} > p$. The contradiction we have reached proved that the integrand f does not satisfy (A6). Proposition 8.1 is proved. □

Acknowledgments. The author thanks the referees for useful comments and suggestions.

REFERENCES

[1] G. Alberti and F. Serra Cassano, *Non-occurrence of gap for one-dimensional autonomous functionals*, in Calculus of Variations, Homogenization and Continuum Mechanics (Marseille, 1993), Ser. Adv. Math. Appl. Sci. 18, World Scientific, River Edge, NJ, 1994, pp. 1–17.

[2] T. S. Angell, *A note on the approximation of optimal solutions of the calculus of variations*, Rend. Circ. Mat. Palermo, 2 (1979), pp. 258–272.

[3] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems* I. *The epigraphical distance*, Trans. Amer. Math. Soc., 3 (1991), pp. 695–729.

[4] H. Attouch and R. J.-B. Wets, *Quantitative stability of variational systems* II. *A framework for nonlinear conditioning*, SIAM J. Optim., 3 (1993), pp. 359–381.

[5] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley Interscience, New York, 1984.

[6] J. M. Ball and V. J. Mizel, *Singular minimizers for regular one-dimensional problems in the calculus of variations*, Bull. Amer. Math. Soc., 11 (1984), pp. 143–146.

[7] J. M. Ball and V. J. Mizel, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation*, Arch. Rational Mech Anal., 90 (1985), pp. 325–388.

[8] H. Brezis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[9] L. Cesari, *Optimization—Theory and Applications*, Springer-Verlag, Berlin, 1983.

[10] F. H. Clarke and R. B. Vinter, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985) pp. 73–98.

[11] F. H. Clarke and R. B. Vinter, *Regularity of solutions to variational problems with polynomial Lagrangians*, Bull. Polish Acad. Sci., 34 (1986), pp. 73–81.

[12] A. Ferriero, *The approximation of higher-order integrals of the calculus of variations and the Lavrentiev phenomenon*, SIAM J. Control Optim., 44 (2005), pp. 99–110.

[13] M. Lavrentiev, *Sur quelques problemes du calcul des variations*, Ann. Math. Pura Appl., 4 (1926), pp. 107–124.

[14] P. D. Loewen, *On the Lavrentiev phenomenon*, Canad. Math. Bull., 30 (1987), pp. 102–108.

[15] B. Mania, *Sopra un esempio di Lavrentieff*, Boll. Un. Mat. Ital., 13 (1934), pp. 146–153.

[16] V. J. Mizel, *New developments concerning the Lavrentiev phenomenon*, in Calculus of Variations and Differential Equations, Res. Notes in Math. 410, CRC Press, Boca Raton, FL, 2000, pp. 185–191.

[17] Z. Nitecki, *Differentiable Dynamics. An Introduction to the Orbit Structure of Diffeomorphisms*, The MIT Press, Cambridge, MA, 1971.

[18] A. V. Sarychev, *First-and second order integral functionals of the calculus of variations which exhibit the Lavrentiev phenomenon*, J. Dynam. Control Systems, 3 (1997), pp. 565–588.

[19] A. V. Sarychev and D. F. M. Torres, *Lipschitzian regularity of minimizers for optimal control problems with control-affine dynamics*, Appl. Math. Optim., 41 (2000), pp. 237–254.

[20] M. A. Sychev and V. J. Mizel, *A condition on the value function both necessary and sufficient for full regularity of minimizers of one-dimensional variational problems*, Trans. Amer. Math. Soc., 350 (1998), pp. 119–133.

[21] A. J. Zaslavski, *Generic well-posedness of nonconvex optimal control problems*, Nonlinear Anal., 59 (2004), pp. 1091–1124.

[22] A. J. Zaslavski, *Nonoccurrence of the Lavrentiev phenomenon for nonconvex variational problems*, Ann. Inst. H. Poincaré, Anal. Non Linéare, 22 (2005), pp. 579–596.

[23] A. J. Zaslavski, *Turnpike Properties in the Calculus of Variations and Optimal Control*, Springer, New York, 2006.

[24] A. J. Zaslavski, *Nonoccurrence of the Lavrentiev phenomenon for many nonconvex constrained variational problems*, Calc. Var. Partial Differential Equations, to appear.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1147–1164

ON UNIFORM SOLVABILITY OF PARAMETER-DEPENDENT LYAPUNOV INEQUALITIES AND APPLICATIONS TO VARIOUS PROBLEMS*

P. KRISHNAMURTHY† AND F. KHORRAMI†

Abstract. We consider the problem of finding a common quadratic Lyapunov function to demonstrate the stability of a not necessarily bounded family of matrices which incorporate design freedoms. Generically, this can be viewed as the problem of picking a family of controller (or observer) gains so that the family of closed-loop system matrices admits a common quadratic Lyapunov function. We provide several necessary and sufficient conditions for various structures of matrix families. Specifically, we consider matrix families which can be obtained through similarity transformations from a lower Hessenberg structure. Families of matrices containing a subset of diagonal matrices, which is invariant under the design freedoms, are also considered since they occur in many applications. The conditions for uniform solvability of the Lyapunov inequalities are explicitly given and involve inequalities regarding relative magnitudes of terms in the matrices. Various motivating applications of the obtained results to observer and controller designs for time-varying, switched, and nonlinear systems are highlighted.

Key words. coupled Lyapunov inequalities, state-dependent Lyapunov inequalities, nonlinear systems, switched systems

AMS subject classifications. 15A45, 93B51, 93D30

DOI. 10.1137/040619417

1. Introduction and problem statement.

We consider the following problem: Given families of matrices $\mathcal{A}(\theta) \in \mathcal{R}^{n\times n}$ and $\mathcal{C}(\theta) \in \mathcal{R}^{1\times n}$ indexed by θ ranging over some (not necessarily bounded) set Θ, is it possible to find a family of vectors $\mathcal{G}(\theta) \in \mathcal{R}^{n\times 1}$, a constant symmetric positive definite matrix $P \in \mathcal{R}^{n\times n}$, and a positive constant ν such that for all $\theta \in \Theta$, the Lyapunov inequality

$$P[\mathcal{A}(\theta) + \mathcal{G}(\theta)\mathcal{C}(\theta)] + [\mathcal{A}(\theta) + \mathcal{G}(\theta)\mathcal{C}(\theta)]^T P \leq -\nu I \tag{1.1}$$

is satisfied with I being the identity matrix of appropriate dimensions? If the answer is yes, provide an explicit construction.

Families of Lyapunov inequalities of the form (1.1) arise in many applications, some of which are outlined in section 5. In the literature, either the solvability of these inequalities is taken as an assumption and a numerical solution sought in specific cases, or the system class is restricted so that one of the approaches below can be applied. However, since the available approaches are predominantly analysis rather than design tools, this restriction could be quite conservative.

The problem statement above poses essentially a design problem since it entails the choosing of $\mathcal{G}(\theta)$ such that the family of matrices $\mathcal{A}(\theta)+\mathcal{G}(\theta)\mathcal{C}(\theta)$ admits a common quadratic Lyapunov function. In applications, this function is usually an observer or a controller gain vector. The associated analysis problem requires finding a matrix $P > 0$ and a positive constant ν to satisfy, for all $\theta \in \Theta$, the Lyapunov inequality

*Received by the editors November 22, 2004; accepted for publication (in revised form) March 28, 2006; published electronically September 12, 2006. An earlier version of this paper was presented at the American Control Conference, Boston, MA, June 2004.
http://www.siam.org/journals/sicon/45-4/61941.html
†Control/Robotics Research Laboratory (CRRL), Department of Electrical and Computer Engineering, Polytechnic University, 6 Metrotech Center, Brooklyn, NY 11201 (pk@crrl.poly.edu, khorrami@smart.poly.edu).

$P\mathcal{A}(\theta)+\mathcal{A}^T(\theta)P \leq -\nu I$. This analysis problem has been investigated extensively, and several results are available in the literature. These results are typically formulated as conditions on the eigenstructure of the matrices $\mathcal{A}(\theta)$, on the stability of associated matrix pencils, or on various geometric properties [17, 15, 13, 21]. Results involving algebraic conditions such as pairwise commutativity of the matrices $\mathcal{A}(\theta)$ are also available [16]. However, these results do not extend readily to the design problem above. Another class of results in the literature, which is inspired by the robust control problem, considers the set $\mathcal{A}(\theta)$ to have a *central* element with the radius of the set being bounded in some manner (for instance, by a bound on the singular values of the perturbation matrix, i.e., the difference between $\mathcal{A}(\theta)$ and the central element of the matrix set) so that the solution of the Lyapunov inequality with the central element also works for all the matrices in the set [18, 24, 23, 4, 2, 22]. This approach can be extended to the design problem to find $\mathcal{G}$ and a matrix $P > 0$ to satisfy (1.1). However, by the very nature of the approach, strong conditions such as compactness (with respect to the usual topology in $\mathcal{R}^{n\times n}$) are required on the matrix family $\mathcal{A}(\theta)$. For instance [23, 14, 20], under the assumptions that the elements of $\mathcal{A}(\theta)$ take values within a compact set, that each $\mathcal{A}(\theta)$ is in lower Hessenberg form with a uniform lower bound on the magnitudes of upper diagonal elements and with the sign of each upper diagonal element being independent of θ, and that $\mathcal{C}(\theta) = [1, 0, \ldots, 0]$, a constant vector $\mathcal{G}$ can be found such that the Lyapunov inequality (1.1) is satisfied for all $\theta \in \Theta$. A condition for the existence of a common similarity transformation to simultaneously transform all the elements of a matrix family $\mathcal{A}(\theta)$ into lower Hessenberg form can be formulated [13, 1] in terms of the solvability of the Lie algebra generated by the matrices in the family. However, these results rely crucially on compactness of the matrix family $\mathcal{A}(\theta)$. In contrast, the results in this paper do not require the matrix family $\mathcal{A}(\theta)$ to be compact.

The results given in this paper are therefore of particular interest in applications wherein the family $\mathcal{A}(\theta)$ is noncompact and wherein design freedoms $\mathcal{G}(\theta)$ are available and allowed to depend on the variable θ. This is the case, for instance, in time-varying systems (where $\theta = t$, the time), switched systems (where θ is a switching state), and nonlinear systems (where θ could be the state or some subset of the state). The obtained results provide extensions in each of these specific areas. We consider particularly the case wherein the matrix family $\mathcal{A}(\theta)$ contains matrices that are in lower Hessenberg form (or that are related to the lower Hessenberg form via a similarity transformation). For greater applicability, the matrix family is also allowed to contain a subset (obtained from diagonal matrices via a similarity transformation) which cannot be influenced by the design freedoms. The obtained results are new even in the case where the matrix family $\mathcal{A}(\theta)$ is compact since, unlike [23, 14, 20, 1], the results in this paper address both a family of the lower Hessenberg form *and* a family of diagonal matrices. Matrices of the lower Hessenberg form (or its transpose) appear in numerous applications, including control of nonlinear lower triangular (strict-feedback) and upper triangular (feedforward) systems. The results addressing the case wherein the matrix family $\mathcal{A}(\theta)$ contains a subset of diagonal matrices which are not influenced by design freedoms have applications in dynamic high-gain scaling–based control [9, 8, 10] and λ-tracking [3]. The applications of the obtained results are further elaborated on in section 5. The general framework of our results and some basic definitions and results are provided in section 2. Theorems giving sufficient conditions for uniform solvability of the Lyapunov inequalities are provided in section 3. The conditions are formulated as inequalities involving relative magnitudes of terms in the matrices. Necessity theorems are presented in section 4

and conditions under which the imposed assumptions are necessary and sufficient are identified.

2. Definitions and a basic theorem. We address the design problem formulated in the introduction for particular structures of $\mathcal{A}(\theta)$ and $\mathcal{C}(\theta)$. Specifically, we consider a partition of $\Theta = \Theta_1 \cup \Theta_2$ with $\mathcal{A}(\theta) = A(\theta)$, $\mathcal{G}(\theta) = G(\theta)$, $\mathcal{C}(\theta) = C(\theta)$ if $\theta \in \Theta_1$ and $\mathcal{A}(\theta) = -D(\theta)$, $\mathcal{G}(\theta) = 0$, $\mathcal{C}(\theta) = 0$ if $\theta \in \Theta_2$. This partition of Θ corresponds to some elements of the matrix family $\mathcal{A}(\theta)$ that cannot be modified by the design functions. The motivation for considering this problem is seen in the applications to scaling-based design and λ-tracking discussed in section 5. Equivalently, we consider the following problem: Given families of matrices $A(\theta) \in \mathcal{R}^{n\times n}$, $D(\theta) \in \mathcal{R}^{n\times n}$, and $C(\theta) \in \mathcal{R}^{1\times n}$, find $G(\theta) \in \mathcal{R}^{n\times 1}$, a constant symmetric positive definite matrix $P \in \mathcal{R}^{n\times n}$, and positive constants ν_1 and ν_2 such that for all $\theta \in \Theta$,

$$P[A(\theta) + G(\theta)C(\theta)] + [A(\theta) + G(\theta)C(\theta)]^T P \le -\nu_1 I, \tag{2.1}$$

$$PD(\theta) + D^T(\theta)P \ge \nu_2 I. \tag{2.2}$$

Definition 1. *A triple (A, C, D) which satisfies* (2.1) *and* (2.2) *with some choice of G, a symmetric positive definite matrix P, and positive constants ν_1 and ν_2 is said to have property $\mathcal{M}$ with the $\mathcal{M}$-associated* 4*-tuple (G, P, ν_1, ν_2).*

The problem posed above is in the "observer design" context, and the dual "controller design" problem can be formulated as follows: Given families of matrices $A(\theta) \in \mathcal{R}^{n\times n}$, $D(\theta) \in \mathcal{R}^{n\times n}$, and $B(\theta) \in \mathcal{R}^{n\times 1}$, find $H(\theta) \in \mathcal{R}^{1\times n}$, a constant symmetric positive definite matrix $P \in \mathcal{R}^{n\times n}$, and positive constants ν_1 and ν_2 such that for all $\theta \in \Theta$,

$$P[A(\theta) + B(\theta)H(\theta)] + [A(\theta) + B(\theta)H(\theta)]^T P \le -\nu_1 I, \tag{2.3}$$

$$PD(\theta) + D^T(\theta)P \ge \nu_2 I. \tag{2.4}$$

Definition 2. *A triple (A, B, D) which satisfies* (2.3) *and* (2.4) *with some choice of H, a symmetric positive definite matrix P, and positive constants ν_1 and ν_2 is said to have property $\mathcal{M}_d$ with the $\mathcal{M}_d$-associated* 4*-tuple (H, P, ν_1, ν_2).*

Theorem 1. *Let $Q \in \mathcal{R}^{n\times n}$ be a nonsingular matrix with $\lambda_{max}(QQ^T) = \overline{Q}$. If the triple (A, C, D) has property $\mathcal{M}$ with $\mathcal{M}$-associated* 4*-tuple (G, P, ν_1, ν_2), then*

(a) *the triple $(QAQ^{-1}, CQ^{-1}, QDQ^{-1})$ has property $\mathcal{M}$ with $\mathcal{M}$-associated* 4*-tuple $(QG, Q^{-T}PQ^{-1}, \frac{\nu_1}{\overline{Q}}, \frac{\nu_2}{\overline{Q}})$;*

(b) *the triple (A^T, C^T, D^T) has property $\mathcal{M}_d$ with $\mathcal{M}_d$-associated* 4*-tuple $(G^T, P^{-1}, \nu_1\lambda_{min}(P^{-2}), \nu_2\lambda_{min}(P^{-2}))$.*

If the triple (A, B, D) has property $\mathcal{M}_d$ with $\mathcal{M}_d$-associated 4*-tuple (H, P, ν_1, ν_2), then*

(c) *the triple (QAQ^{-1}, QB, QDQ^{-1}) has property $\mathcal{M}_d$ with $\mathcal{M}_d$-associated* 4*-tuple $(HQ^{-1}, Q^{-T}PQ^{-1}, \frac{\nu_1}{\overline{Q}}, \frac{\nu_2}{\overline{Q}})$;*

(d) *the triple (A^T, B^T, D^T) has property $\mathcal{M}$ with $\mathcal{M}$-associated* 4*-tuple $(H^T, P^{-1}, \nu_1\lambda_{min}(P^{-2}), \nu_2\lambda_{min}(P^{-2}))$.*

Proof of Theorem 1. The proof follows by standard matrix algebra. ∎

In sections 3 and 4, we consider particular structures of matrix families. By Theorem 1, the obtained results are applicable if the required structures are attained under any of the transformations considered in Theorem 1. Theorem 1 also allows theorems stated in the observer design context to be mapped to the dual controller design context and vice versa. In the rest of this paper, we primarily address the

observer design problem, i.e., solution of (2.1) and (2.2). Dual sufficiency results are summarized in Corollary 1. The duals of the necessity theorems can be obtained along the same lines as in the proof of Corollary 1.

3. Sufficiency theorems. In Theorem 2, conditions guaranteeing solvability of the coupled Lyapunov inequalities (2.1) and (2.2) are provided. Theorem 3 is a robust version of Theorem 2 wherein a perturbation term is introduced in the Lyapunov inequality (2.1). The duals of Theorems 2 and 3 in the controller context, i.e., coupled Lyapunov inequalities (2.3) and (2.4), are given in Corollary 1.

THEOREM 2. *Let $\phi_{(i,j)} : \Theta \mapsto \mathcal{R}$ and $D_i : \Theta \mapsto \mathcal{R}$, $i = 1, \ldots, n$, $j = 1, \ldots, i+1$. Let $A : \Theta \mapsto \mathcal{R}^{n\times n}$ be a matrix in the lower Hessenberg form,*

$$A(\theta) = \begin{bmatrix} \phi_{(1,1)}(\theta) & \phi_{(1,2)}(\theta) & 0 & 0 & \cdots & 0 \\ \phi_{(2,1)}(\theta) & \phi_{(2,2)}(\theta) & \phi_{(2,3)}(\theta) & 0 & \cdots & 0 \\ \phi_{(3,1)}(\theta) & \phi_{(3,2)}(\theta) & \phi_{(3,3)}(\theta) & \phi_{(3,4)}(\theta) & & 0 \\ & & & & \ddots & \\ \phi_{(n-1,1)}(\theta) & & \cdots & & \phi_{(n-1,n-1)}(\theta) & \phi_{(n-1,n)}(\theta) \\ \phi_{(n,1)}(\theta) & & \cdots & & \phi_{(n,n-1)}(\theta) & \phi_{(n,n)}(\theta) \end{bmatrix}, \tag{3.1}$$

and let $C(\theta) = C = [1, 0, \ldots, 0]$ and $D(\theta) = \operatorname{diag}(D_1(\theta), \ldots, D_n(\theta))$. If the sign of each $\phi_{(i,i+1)}(\theta)$ is independent of the argument θ, and if positive constants σ, ρ_{iA}, $i = 2, \ldots, n-1$, $\epsilon_{(i,j)}$, $j = 2, \ldots, i$, $i = 2, \ldots, n$, $\underline{\rho}_{iD}$, $i = 1, \ldots, n-1$, $\overline{\rho}_{iD}$, $i = 1, \ldots, n-1$, and $\underline{\rho}^o_{iD}$, $i = 1, \ldots, n$, exist such that for all $\theta \in \Theta$,

$$\begin{cases} |\phi_{(i,i+1)}(\theta)| \geq \sigma > 0, \quad 1 \leq i \leq n-1, \\ |\phi_{(i,i+1)}(\theta)| \leq \rho_{iA}|\phi_{(i-1,i)}(\theta)|, \quad 2 \leq i \leq n-1, \\ |\phi_{(i,j)}(\theta)| \leq \epsilon_{(i,j)}\sqrt{|\phi_{(i,i+1)}(\theta)||\phi_{(j-1,j)}(\theta)|}, \quad 2 \leq i \leq n-1,\ 2 \leq j \leq i, \\ |\phi_{(n,j)}(\theta)| \leq \epsilon_{(n,j)}\sqrt{|\phi_{(n-1,n)}(\theta)||\phi_{(j-1,j)}(\theta)|}, \quad 2 \leq j \leq n, \end{cases} \tag{3.2}$$

$$\begin{cases} D_i(\theta) \geq \underline{\rho}^o_{iD}, \quad i = 1, \ldots, n, \\ \underline{\rho}_{iD} D_i(\theta) \leq D_{i+1}(\theta) \leq \overline{\rho}_{iD} D_i(\theta), \quad i = 1, \ldots, n-1, \end{cases} \tag{3.3}$$

then the triple (A, C, D) has property $\mathcal{M}$.

Remark 1. Conditions (3.2) are a weaker version of the *cascading upper diagonal dominance* (CUDD) conditions introduced in [12, 7] which require the last two inequalities in (3.2) to be strengthened to

$$|\phi_{(i,j)}(\theta)| \leq \epsilon_{(i,j)}|\phi_{(i,i+1)}(\theta)|, \qquad 2 \leq i \leq n-1,\ 2 \leq j \leq i, \tag{3.4}$$

$$|\phi_{(n,j)}(\theta)| \leq \epsilon_{(n,j)}|\phi_{(n-1,n)}(\theta)|, \qquad 2 \leq j \leq n. \tag{3.5}$$

A matrix A of structure (3.1) and satisfying inequalities (3.2) is called a *weakly cascading upper diagonal dominant* (w-CUDD) matrix. The positive constants ρ_{iA}, $i = 2, \ldots, n-1$, and $\epsilon_{(i,j)}$, $i = 2, \ldots, n$, $j = 2, \ldots, i$, are referred to as the w-CUDD constants of A.

Remark 2. An interesting special case is that in which the elements of $A(\theta)$ and $D(\theta)$ vary within a compact set (or, more strongly, in which Θ is compact and the elements of $A(\theta)$ and $D(\theta)$ are continuous functions of θ). In this case, only the first equations in (3.2) and (3.3) need to be verified since the rest of the conditions in (3.2) and (3.3) are trivially satisfied. Hence, in the special case in which $\{\phi_{(i,j)}(\theta)|\theta \in \Theta\}$, $i = 1, \ldots, n$, $j = 1, \ldots, i+1$, and $\{D_i(\theta)|\theta \in \Theta\}$, $i = 1, \ldots, n$, are compact sets,

Theorem 2 states that a $G(\theta)$, a symmetric positive definite matrix P, and positive constants ν_1 and ν_2 can be found to satisfy the Lyapunov inequalities (2.1) and (2.2) if A, C, and D are of the form shown in Theorem 2 with the sign of each $\phi_{(i,i+1)}(\theta)$ being independent of θ and if positive constants σ and $\underline{\rho}^o_{iD}$, $i = 1, \ldots, n$, exist such that $|\phi_{(i,i+1)}(\theta)| \geq \sigma > 0$, $1 \leq i \leq n-1$, and $D_i(\theta) \geq \underline{\rho}^o_{iD}$, $i = 1, \ldots, n$. These conditions on $A(\theta)$ match those obtained in prior results [23, 14, 20] on uniform solvability of the family of matrix Lyapunov inequalities (2.1) with $A(\theta)$ being in lower Hessenberg form and belonging to a compact set. Note that, even in this special case, Theorem 2 is new since it addresses simultaneous solvability of both (2.1) *and* (2.2).

Theorem 2 is proved through a sequence of lemmas below. In Lemma 1, a family of matrices P that satisfy (2.2) is identified. In Lemma 2, the existence of a matrix in this family that also satisfies (2.1) is shown if the w-CUDD constants of A are small enough. Lemma 3 shows that a diagonal transformation can scale the w-CUDD constants. The proof of Theorem 2 is concluded by observing that D, being diagonal, is invariant under a diagonal transformation.

LEMMA 1. *Let*[1] $P = \hat{P}^T \hat{P}$, $\hat{P} = I - \text{lowerdiag}(k_2, \ldots, k_n)$, *with* $k_2, \ldots, k_n$ *being constants, and let* D *be as defined in Theorem* 2. *Assume that* (3.3) *is satisfied with positive constants* $\underline{\rho}_{iD}$, $\overline{\rho}_{iD}$, *and* $\underline{\rho}^o_{iD}$. *Then, a positive constant* κ *exists such that if* $k_2, \ldots, k_n$ *are smaller in magnitude than* κ, *then* (2.2) *holds with some positive constant* ν_2.

Proof of Lemma 1. See section A.1 in the appendix.

LEMMA 2. *Let* A *and* C *be as defined in Theorem* 2. *Given any constant* $\kappa > 0$, *a constant* $\overline{\rho}_A > 0$ *whose choice depends only on* κ *and* n *exists such that the following property is true: If* (3.2) *holds with some* $\sigma > 0$ *and with the w-CUDD constants* ρ_{iA} *and* $\epsilon_{(i,j)}$ *being smaller than* $\overline{\rho}_A$, *then constants* $k_2, \ldots, k_n$ *smaller in magnitude than* κ *and* $G(\theta) = [g_1(\theta), \ldots, g_n(\theta)]^T$ *exist such that* (2.1) *is satisfied for some positive* ν_1 *with* P *defined as in Lemma* 1.

Proof of Lemma 2. See section A.2 in the appendix.

LEMMA 3. *Let* A *be as defined in* (3.1) *and let* (3.2) *be satisfied with some positive constants* σ, ρ_{iA}, *and* $\epsilon_{(i,j)}$. *Given any positive constant* $\overline{\rho}_A$, *a constant diagonal matrix* T *exists such that the matrix* $\tilde{A} = TAT^{-1}$ *is a w-CUDD matrix with w-CUDD constants smaller than* $\overline{\rho}_A$.

Proof of Lemma 3. See section A.3 in the appendix.

Proof of Theorem 2. Given $A(\theta)$ and $D(\theta)$, obtain κ, $\overline{\rho}_A$, and T as in Lemmas 1, 2, and 3, respectively. Defining $\tilde{A} = TAT^{-1}$, $(\tilde{A}, C, D)$ has property $\mathcal{M}$ with $\mathcal{M}$-associated 4-tuple $(\tilde{G}, \tilde{P}, \tilde{\nu}_1, \tilde{\nu}_2)$ constructed as in the proofs of Lemmas 1 and 2. Using statement (a) in Theorem 1 and noting that $T^{-1}DT = D$, the triple (A, CT, D) has property $\mathcal{M}$ with $\mathcal{M}$-associated 4-tuple $(T^{-1}\tilde{G}, T\tilde{P}T, \tilde{\nu}_1\lambda_{min}(T^2), \tilde{\nu}_2\lambda_{min}(T^2))$. Observing that $CT = T_1 C$, where T_1 is the $(1,1)$th element of T, the triple (A, C, D) has property $\mathcal{M}$ with $\mathcal{M}$-associated 4-tuple $(T_1 T^{-1}\tilde{G}, T\tilde{P}T, \tilde{\nu}_1\lambda_{min}(T^2), \tilde{\nu}_2\lambda_{min}(T^2))$. ∎

THEOREM 3. *Let* A, C, *and* D *be as defined in Theorem* 2. *Let* $\Phi : \Theta \times \mathcal{R}^n \mapsto \mathcal{R}^n$ *be a function such that*[2] $|\Phi(\theta, \omega)|_e \leq_e \Gamma(\theta)|\omega|_e$ *for all* $\theta \in \Theta$ *and* $\omega \in \mathcal{R}^n$, *with* $\Gamma(\theta)$ *being a lower triangular* $n \times n$ *matrix function with nonnegative entries. If* A *and*

[1] lowerdiag$(k_2, \ldots, k_n)$ denotes the $n \times n$ matrix with lower diagonal elements $k_2, \ldots, k_n$ and zeros elsewhere.

[2] $|\beta|_e$ with β being a matrix denotes a matrix of the same dimension as β obtained by replacing each element of β with its absolute value. The relation $\leq_e$ between two vectors denotes an elementwise inequality.

$A+\Gamma$ are w-CUDD, then $G(\theta)$, a matrix $P>0$, and positive constants ν_1 and ν_2 exist such that for all $\theta\in\Theta$ and $\omega\in\mathcal{R}^n$,

$$\omega^T\{P[A(\theta)+G(\theta)C]+[A(\theta)+G(\theta)C]^TP\}\omega+2\omega^TP\Phi(\theta,\omega)\le-\nu_1|\omega|^2, \tag{3.6}$$

$$PD(\theta)+D(\theta)P\ge\nu_2 I. \tag{3.7}$$

Proof of Theorem 3. We have the inequality $\omega^TP\Phi\le|\omega^TP|_e|\Phi|_e\le|\omega^TP|_e\Gamma|\omega|_e=\omega^TPQ_1\Gamma Q_2\omega$, where Q_1 and Q_2 are appropriate diagonal matrices with 1 or -1 as each diagonal entry. Let $\mathcal{Q}:\{1,2,\dots,2^n\}\mapsto\mathcal{R}^{n\times n}$ be an enumeration of the 2^n diagonal matrices of dimension $n\times n$ with each diagonal entry 1 or -1. Consider $\theta'=(\theta,q_1,q_2)\in\Theta\times\{1,\dots,2^n\}\times\{1,\dots,2^n\}$. If $\tilde{A}(\theta')=A(\theta)+\mathcal{Q}(q_1)\Gamma(\theta)\mathcal{Q}(q_2)$ is w-CUDD, then using Theorem 2, the conclusion of Theorem 3 follows. Note that G can be taken to be a function of θ and not θ' since $|Q_1\Gamma Q_2|_e=|\Gamma|_e$, and the choice of G in (A.10) depends only on the upper diagonal elements of $\tilde{A}$ (which are equal to the upper diagonal elements of A) and upper bounds on absolute values of the other elements. Noting that $\mathcal{Q}(q_1)$ and $\mathcal{Q}(q_2)$ vary over all the 2^n diagonal matrices with each diagonal entry 1 or -1, it is seen that the necessary and sufficient condition for $\tilde{A}$ to be w-CUDD (i.e., that $\phi_{(i,j)}+\Gamma_{(i,j)}$ and $\phi_{(i,j)}-\Gamma_{(i,j)}$ should both satisfy the w-CUDD bounds) is that A and $A+\Gamma$ are w-CUDD. □

Corollary 1 (corollary to Theorems 2 and 3). *Let $A(\theta)$ be as given in* (3.1). *Let $D(\theta)=\mathrm{diag}(D_1(\theta),\dots,D_n(\theta))$ and $B(\theta)=B=[0,\dots,0,1]^T$. If positive constants σ, ρ_{iA}, $i=2,\dots,n-1$, $\epsilon_{(i,j)}$, $i=1,\dots,n-1$, $j=1,\dots,i$, $\underline{\rho}_{iD}$, $i=1,\dots,n-1$, $\overline{\rho}_{iD}$, $i=1,\dots,n-1$, and $\underline{\rho}^o_{iD}$, $i=1,\dots,n$, exist such that*[3] *for all $\theta\in\Theta$,*

$$\begin{cases}
|\phi_{(i,i+1)}(\theta)| \ge \sigma>0, & 1\le i\le n-1,\\
|\phi_{(i,i+1)}(\theta)| \ge \rho_{iA}|\phi_{(i-1,i)}(\theta)|, & 2\le i\le n-1,\\
|\phi_{(i,j)}(\theta)| \le \epsilon_{(i,j)}\sqrt{|\phi_{(i,i+1)}(\theta)||\phi_{(j-1,j)}(\theta)|}, & 2\le i\le n-1,\ 2\le j\le i,\\
|\phi_{(i,1)}(\theta)| \le \epsilon_{(i,1)}\sqrt{|\phi_{(i,i+1)}(\theta)||\phi_{(1,2)}(\theta)|}, & 1\le i\le n-1,
\end{cases} \tag{3.8}$$

and (3.3) *are satisfied, then the triple (A,B,D) has property $\mathcal{M}_d$. Furthermore, let $\Phi(\theta,\omega)$ be as defined in Theorem* 3. *If A and $A+\Gamma$ are dual w-CUDD, then $H(\theta)$, a matrix $P>0$, and positive constants ν_1 and ν_2 exist such that for all $\theta\in\Theta$ and $\omega\in\mathcal{R}^n$,*

$$\omega^T\{P[A(\theta)+BH(\theta)]+[A(\theta)+BH(\theta)]^TP\}\omega+2\omega^TP\Phi(\theta,\omega)\le-\nu_1|\omega|^2, \tag{3.9}$$

$$PD(\theta)+D(\theta)P\ge\nu_2 I. \tag{3.10}$$

Proof of Corollary 1. Define the matrices $\tilde{A}=QA^TQ$ and $\tilde{D}=QDQ$ with Q being the $n\times n$ matrix with 1's on the antidiagonal, i.e., $Q_{(i,n-i+1)}=1$, $i=1,\dots,n$, with zeros elsewhere. Premultiplication and postmultiplication of a matrix by Q reverses the order of the rows and columns, respectively. $\tilde{A}$ is of form (3.1) with (i,j)th element $\tilde{A}_{(i,j)}=A_{(n-j+1,n-i+1)}$, and is w-CUDD since, using (3.8),

$$\frac{|\tilde{A}_{(i,i+1)}|}{|\tilde{A}_{(i-1,i)}|}=\frac{|\phi_{(n-i,n-i+1)}|}{|\phi_{(n-i+1,n-i+2)}|}\le\frac{1}{\rho_{n-i+1,A}},\qquad i=2,\dots,n-1, \tag{3.11}$$

[3]Conditions (3.8) are dual to the w-CUDD conditions (3.2). Hence, a matrix of form (3.1) satisfying conditions (3.8) is referred to as a *dual w-CUDD* matrix with ρ_{iA} and $\epsilon_{(i,j)}$ called the dual w-CUDD constants.

$$\frac{|\tilde{A}_{(i,j)}|}{\sqrt{|\tilde{A}_{(i,i+1)}||\tilde{A}_{(j-1,j)}|}} = \frac{|\phi_{(n-j+1,n-i+1)}|}{\sqrt{|\phi_{(n-i,n-i+1)}||\phi_{(n-j+1,n-j+2)}|}}$$

$$\leq \epsilon_{(n-j+1,n-i+1)}, \qquad i = 2,\dots,n-1,\ j = 2,\dots,i, \tag{3.12}$$

$$\frac{|\tilde{A}_{(n,j)}|}{\sqrt{|\tilde{A}_{(n-1,n)}||\tilde{A}_{(j-1,j)}|}} = \frac{|\phi_{(n-j+1,1)}|}{\sqrt{|\phi_{(1,2)}||\phi_{(n-j+1,n-j+2)}|}}$$

$$\leq \epsilon_{(n-j+1,1)}, \qquad 2 \leq j \leq n. \tag{3.13}$$

$\tilde{D}$ is a diagonal matrix with (i,i)th element D_{n-i+1}. Using Theorem 2, the triple $(\tilde{A}, C, \tilde{D})$, where $C = [1, 0, \dots, 0]$, has property $\mathcal{M}$ with an $\mathcal{M}$-associated 4-tuple $(G, \tilde{P}, \tilde{\nu}_1, \tilde{\nu}_2)$. Using statements (b) and (c) of Theorem 1, and noting that $Q^{-1} = Q$ and $QC^T = B$, the triple (A, B, D) has property $\mathcal{M}_d$ with $\mathcal{M}_d$-associated 4-tuple $(G^TQ, Q\tilde{P}^{-1}Q, \tilde{\nu}_1\lambda_{min}(\tilde{P}^{-2}), \tilde{\nu}_2\lambda_{min}(\tilde{P}^{-2}))$. The second assertion in Corollary 1, i.e., solvability of (3.9) and (3.10), can be proved similarly using duality and the arguments in the proof of Theorem 3. □

4. Necessity theorems. In this section, the conservativeness of the assumptions required in Theorem 2 is evaluated by formulating necessary conditions for solvability of the coupled Lyapunov inequalities (2.1) and (2.2). In Theorem 4, negative definite submatrices of $P(A+GC)+(A+GC)^TP$ which are invariant under the design freedoms G are identified. This fundamental result in Theorem 4 is applied in Theorem 5 to prove that a necessary condition for solvability of (2.1) is that if the elements of A off the upper diagonal are smaller than the upper diagonal elements in a sense similar to the last two inequalities in (3.2), then the upper diagonal elements must satisfy the first two inequalities of (3.2). Lemma 4 further shows that the invariance of the signs of $\phi_{(i,i+1)}(\theta)$ is also necessary if the elements of A off the upper diagonal are zero. Theorem 6 shows that the assumed conditions on D are necessary for solvability of (2.2). Corollary 2 combines these results to show that the imposed assumptions are necessary and sufficient if A has nonzero entries only on the upper diagonal.

THEOREM 4. *Let $C = [1, 0, \dots, 0]$. If a matrix A of form (3.1) satisfies (2.1) with some $G(\theta)$, a constant matrix $P > 0$ and a constant $\nu_1 > 0$, then the following is true for every $i \in \{1, \dots, n\}$: Let A_i be the matrix obtained by deleting from A the first i rows and the first i columns. Let $C_i = [1, 0, \dots, 0]^T$ be a $1 \times (n-i)$ vector. Then, a constant vector $K_i = [k_{1_i}, \dots, k_{n-i_i}]^T$, a constant $(n-i) \times (n-i)$ matrix $P_i > 0$, and a positive constant $\tilde{\nu}_i$ exist such that for all $\theta \in \Theta$,*

$$P_i[A_i(\theta) + K_iC_i\phi_{(i,i+1)}(\theta)] + [A_i(\theta) + K_iC_i\phi_{(i,i+1)}(\theta)]^TP_i \leq -\tilde{\nu}_iI. \tag{4.1}$$

Proof of Theorem 4. Assume that (2.1) is satisfied with some G, P, and ν_1. Consider partitions

$$A + GC = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_i \end{bmatrix}, \quad P = \begin{bmatrix} P_{11} & P_{21}^T \\ P_{21} & P_i \end{bmatrix}. \tag{4.2}$$

Note that $A + GC$ differs from A only in the first column. The $(2,2)$ block of $P(A+GC)+(A+GC)^TP$ is $P_iA_i + A_i^TP_i + P_{21}A_{12} + A_{12}^TP_{21}^T$. The $i \times (n-i)$ matrix A_{12} has all elements zero except the $(i,1)$ element, which is equal to $\phi_{(i,i+1)}$. Hence, $P_i^{-1}P_{21}A_{12} = K_iC_i\phi_{(i,i+1)}$ with K_i being a constant $(n-i)\times 1$ vector. Noting that $P_iA_i + A_i^TP_i + P_{21}A_{12} + A_{12}^TP_{21}^T = P_i[A_i + K_iC_i\phi_{(i,i+1)}] + [A_i + K_iC_i\phi_{(i,i+1)}]^TP_i$

and that the $(2,2)$ block being on the principal diagonal must be negative definite for (2.1) to hold, the theorem follows. □

THEOREM 5. *Let A be of form (3.1) and $C = [1, 0, \ldots, 0]$. Let positive constants $\epsilon_{(i,j)}$, $i = 2, \ldots, n$, $j = 2, \ldots, i$, and $\bar{\epsilon}_{(i,j)}$, $i = 3, \ldots, n$, $j = 3, \ldots, i$, exist such that for all $\theta \in \Theta$,*

$$\begin{cases} |\phi_{(i,j)}(\theta)| \leq \epsilon_{(i,j)} \sqrt{|\phi_{(i,i+1)}(\theta)||\phi_{(j-1,j)}(\theta)|}, & 2 \leq i \leq n-1,\ 2 \leq j \leq i, \\ |\phi_{(n,j)}(\theta)| \leq \epsilon_{(n,j)} \sqrt{|\phi_{(n-1,n)}(\theta)||\phi_{(j-1,j)}(\theta)|}, & 2 \leq j \leq n, \\ |\phi_{(i,j)}(\theta)| \leq \bar{\epsilon}_{(i,j)} \sqrt{|\phi_{(i,i+1)}(\theta)||\phi_{(j-2,j-1)}(\theta)|}, & 3 \leq i \leq n-1,\ 3 \leq j \leq i, \\ |\phi_{(n,j)}(\theta)| \leq \bar{\epsilon}_{(n,j)} \sqrt{|\phi_{(n-1,n)}(\theta)||\phi_{(j-2,j-1)}(\theta)|}, & 3 \leq j \leq n. \end{cases} \tag{4.3}$$

If (2.1) is satisfied with some $G(\theta)$, a constant positive definite matrix P, and a positive constant ν_1, then positive constants σ and ρ_{iA}, $i = 2, \ldots, n-1$, exist such that

$$\begin{aligned} &|\phi_{(i,i+1)}(\theta)| \geq \sigma > 0, \quad 1 \leq i \leq n-1, \\ &|\phi_{(i,i+1)}(\theta)| \leq \rho_{iA} |\phi_{(i-1,i)}(\theta)|, \quad 2 \leq i \leq n-1. \end{aligned} \tag{4.4}$$

Proof of Theorem 5. Applying Theorem 4 with $i = n-1$, the existence of positive constants P_{n-1} and $\tilde{\nu}_{n-1}$ and a constant $k_{1_{n-1}}$ are inferred such that $2P_{n-1}[\phi_{(n,n)} + k_{1_{n-1}}\phi_{(n-1,n)}] \leq -\tilde{\nu}_{n-1}$. Noting that, by assumption, $|\phi_{(n,n)}| \leq \epsilon_{(n,n)}|\phi_{(n-1,n)}|$, it is seen that a positive constant σ_n exists such that $|\phi_{(n-1,n)}| \geq \sigma_n$. Applying Theorem 4 with $i = n-2$, a 2×2 matrix $P_{n-2} > 0$, a 2×1 vector K_{n-2}, and a positive constant $\tilde{\nu}_{n-2}$ are obtained such that $P_{n-2}[A_{n-2} + K_{n-2}C_{n-2}\phi_{(n-2,n-1)}] + [A_{n-2} + K_{n-2}C_{n-2}\phi_{(n-2,n-1)}]^T P_{n-2} \leq -\tilde{\nu}_{n-2} I$. Denote the (j,k)th element of P_i by $P_{(j,k)_i}$ and the jth element of vector K_i by k_{j_i}. Since the determinant of $P_{n-2}[A_{n-2} + K_{n-2}C_{n-2}\phi_{(n-2,n-1)}] + [A_{n-2} + K_{n-2}C_{n-2}\phi_{(n-2,n-1)}]^T P_{n-2}$ must be positive, we have

$$\begin{aligned} 0 < &-\Big[P_{(1,2)_{n-2}}\phi_{(n-1,n-1)} + P_{(2,2)_{n-2}}\phi_{(n,n-1)} \\ &\quad + \Big(P_{(1,2)_{n-2}}k_{1_{n-2}} + P_{(2,2)_{n-2}}k_{2_{n-2}}\Big)\phi_{(n-2,n-1)} \\ &\quad + P_{(1,1)_{n-2}}\phi_{(n-1,n)} + P_{(1,2)_{n-2}}\phi_{(n,n)}\Big]^2 \\ &+ 4\Big[P_{(1,1)_{n-2}}\phi_{(n-1,n-1)} + P_{(1,2)_{n-2}}\phi_{(n,n-1)} \\ &\quad + (P_{(1,1)_{n-2}}k_{1_{n-2}} + P_{(1,2)_{n-2}}k_{2_{n-2}})\phi_{(n-2,n-1)}\Big] \\ &\times \Big[P_{(1,2)_{n-2}}\phi_{(n-1,n)} + P_{(2,2)_{n-2}}\phi_{(n,n)}\Big]. \end{aligned} \tag{4.5}$$

Using (4.3) and (4.5), positive constants a_1, a_2, and a_3 exist such that

$$\begin{aligned} P^2_{(1,1)_{n-2}}\phi^2_{(n-1,n)} < &|\phi_{(n-1,n)}|\Big[a_1\sqrt{|\phi_{(n-1,n)}||\phi_{(n-2,n-1)}|} + a_2|\phi_{(n-2,n-1)}|\Big] \\ &+ a_3\phi^2_{(n-2,n-1)}. \end{aligned} \tag{4.6}$$

Since $|\phi_{(n-1,n)}(\theta)| > \sigma_n > 0$, $\phi_{(n-2,n-1)}(\theta)$ cannot be zero for any $\theta \in \Theta$. Dividing both sides of (4.6) by $\phi^2_{(n-2,n-1)}$, it is inferred that $\sup_{\theta\in\Theta} \frac{|\phi_{(n-1,n)}(\theta)|}{|\phi_{(n-2,n-1)}(\theta)|} < \infty$. Hence, a positive constant σ_{n-1} exists such that $|\phi_{(n-2,n-1)}(\theta)| \geq \sigma_{n-1}$ for all $\theta \in \Theta$.

We now proceed by induction. Assume that $\sup_{\theta\in\Theta}\frac{|\phi_{(j,j+1)}(\theta)|}{|\phi_{(j-1,j)}(\theta)|}<\infty$ and $\sigma_{j+1}\stackrel{\triangle}{=}\inf_{\theta\in\Theta}|\phi_{(j,j+1)}|>0$, $j=i+2,\dots,n-1$. Using Theorem 4, a matrix $P_i>0$, a vector K_i, and a positive constant $\tilde{\nu}_i$ are obtained to satisfy (4.1). Noting that the leading 2×2 minor of the left-hand side of (4.1) must have positive determinant,

$$\begin{aligned}0<{}&4\sum_{j=1}^{n-i}P_{(1,j)_i}(\phi_{(i+j,i+1)}+k_{j_i}\phi_{(i,i+1)})\sum_{j=1}^{n-i}P_{(2,j)_i}\phi_{(i+j,i+2)}\\&-\left[\sum_{j=1}^{n-i}P_{(1,j)_i}\phi_{(i+j,i+2)}+\sum_{j=1}^{n-i}P_{(2,j)_i}(\phi_{(i+j,i+1)}+k_{j_i}\phi_{(i,i+1)})\right]^2.\end{aligned}\tag{4.7}$$

Using (4.3) and (4.7), positive constants a_{1j},a_{2j},a_{3j}, $i+1\le j\le n-1$, a_4, and a_5 exist such that

$$\begin{aligned}P^2_{(1,1)_i}\phi^2_{(i+1,i+2)}<{}&\left[a_4|\phi_{(i,i+1)}|+\sum_{j=i+1}^{n-1}a_{1j}\sqrt{|\phi_{(j,j+1)}||\phi_{(i,i+1)}|}\right]\\&\times\left[\sum_{j=i+1}^{n-1}a_{2j}\sqrt{|\phi_{(j,j+1)}||\phi_{(i+1,i+2)}|}\right]\\&+\left[a_5|\phi_{(i,i+1)}|+\sum_{j=i+1}^{n-1}a_{3j}\sqrt{|\phi_{(j,j+1)}||\phi_{(i,i+1)}|}\right]^2.\end{aligned}\tag{4.8}$$

By the induction hypothesis, $\frac{|\phi_{(j,j+1)}(\theta)|}{|\phi_{(j-1,j)}(\theta)|}$, $j=i+2,\dots,n-1$, are bounded. Using (4.8), $\frac{|\phi_{(i+1,i+2)}(\theta)|}{|\phi_{(i,i+1)}(\theta)|}$ is bounded and $\sigma_{i+1}\stackrel{\triangle}{=}\inf_{\theta\in\Theta}|\phi_{(i,i+1)}(\theta)|>0$. The proof is complete by induction. □

LEMMA 4. *Let A be of form* (3.1) *and $C=[1,0,\dots,0]$. For each $i\in\{2,\dots,n\}$, let $\inf_{\theta\in\Theta}\sum_{j=i}^{n}|\phi_{(j,i)}(\theta)|=0$. Let* (2.1) *be satisfied with some $G(\theta)$, a constant matrix $P>0$, and a constant $\nu_1>0$. Then, the elements of P on the upper and lower diagonals, i.e., $P_{(i,i+1)}$, $i=1,\dots,n-1$, are nonzero. Furthermore, if $\phi_{(i,j)}\equiv 0$, $i=2,\dots,n$, $j=2,\dots,i$, then each $\phi_{(i,i+1)}(\theta)$, $i=1,\dots,n-1$, assumes the same sign for all arguments.*

Proof of Lemma 4. If (2.1) holds, the diagonal entries of $A_P\stackrel{\triangle}{=}P[A+GC]+[A+GC]^TP$ must be bounded above by a negative constant. The ith diagonal entry, $i\ge 2$, is given by $2\sum_{j=i-1}^{n}P_{(i,j)}\phi_{(j,i)}$. By assumption, $\inf_{\theta\in\Theta}\sum_{j=i}^{n}|\phi_{(j,i)}(\theta)|=0$. Therefore, the ith diagonal entry of A_P can be bounded above by a negative constant only if $P_{(i,i-1)}=P_{(i-1,i)}\neq 0$. If the entries of A off the upper diagonal are zero, then the ith diagonal entry of A_P is $2P_{(i,i-1)}\phi_{(i-1,i)}$ implying that $\phi_{(i-1,i)}(\theta)$ must assume the same sign for all arguments. □

THEOREM 6. *Let $D(\theta)=\operatorname{diag}(D_1(\theta),\dots,D_n(\theta))$. If a matrix $P>0$ exists such that* (2.2) *is satisfied with $\nu_2>0$, then $\inf_{\theta\in\Theta}D_i(\theta)>0$. Furthermore, if the entries of P on the upper and lower diagonals, i.e., $P_{(i,i+1)}$, $i=1,\dots,n-1$, are nonzero, then positive constants $\underline{\rho}_{iD}$ and $\overline{\rho}_{iD}$, $i=1,\dots,n-1$, exist such that $\underline{\rho}_{iD}D_i(\theta)\le D_{i+1}(\theta)\le\overline{\rho}_{iD}D_i(\theta)$, $i=1,\dots,n-1$.*

Proof of Theorem 6. The ith diagonal element of $PD+DP$ given by $2P_{(i,i)}D_i$ must be bounded below by a positive constant to satisfy (2.2). Hence, D_i must be bounded below by a positive constant. Furthermore, to satisfy (2.2), the 2×2 matrices

on the principal diagonal of $PD + DP$ must have positive determinant, i.e.,

$$4P_{(i,i)}P_{(i+1,i+1)}D_iD_{i+1} > P^2_{(i,i+1)}(D_i + D_{i+1})^2, \qquad i = 1, \ldots, n-1, \tag{4.9}$$

implying that $\frac{D_{i+1}(\theta)}{D_i(\theta)}$ and $\frac{D_i(\theta)}{D_{i+1}(\theta)}$ are upper and lower bounded by positive constants. ∎

COROLLARY 2. *Let $A(\theta)$ be a matrix of the form* (3.1) *with $\phi_{(i,j)} = 0$, $i = 2, \ldots, n$, $j = 2, \ldots, i$, i.e., A contains nonzero terms only on the upper diagonal and the first column. Let $C = [1, 0, \ldots, 0]$ and $D(\theta) = \mathrm{diag}(D_1(\theta), \ldots, D_n(\theta))$. A matrix $P > 0$, positive constants ν_1 and ν_2, and a $G(\theta)$ exist to satisfy* (2.1) *and* (2.2) *if and only if A is w-CUDD, each upper diagonal entry of A assumes the same sign for all arguments, and the elements of $D(\theta)$ satisfy* (3.3) *with $\underline{\rho}^o_{iD}$, $i = 1, \ldots, n$, $\underline{\rho}_{iD}$, and $\overline{\rho}_{iD}$, $i = 1, \ldots, n-1$, being positive constants. Furthermore, a matrix $P > 0$, a positive constant ν_1, and a $G(\theta)$ exist to satisfy* (2.1) *if and only if A is w-CUDD and each upper diagonal entry of A assumes the same sign for all arguments.*

Proof of Corollary 2. The proof is evident from Lemma 4 and Theorems 2, 5, and 6. ∎

5. Applications to controller and observer designs for nonlinear, time-varying, and switched systems. The problem stated in the introduction arises in several applications for nonlinear systems, time-varying systems, and switched systems, a few of which are outlined below.

1. *State-feedback*: Consider a system

$$\dot{x} = A_{l(t)}(x,t)x + B_{l(t)}(x,t)u + \Phi_{l(t)}(x,z,t) \tag{5.1}$$

with state $x \in \mathcal{R}^n$, input $u \in \mathcal{R}$, and disturbance $z \in \mathcal{R}^{n_z}$. $A_1(x,t), \ldots, A_N(x,t)$ are $n \times n$ matrices. $B_1(x,t), \ldots, B_N(x,t)$ and $\Phi_1(x,z,t), \ldots, \Phi_N(x,z,t)$ are $n \times 1$ vectors. $l(t)$ is a positive integer function of time taking values in the set $\{1, 2, \ldots, N\}$. Thus, (5.1) is a nonlinear time-varying switched system. If a matrix $P > 0$, a positive constant ν, and functions $K_i(x,t) \in \mathcal{R}^{1\times n}$, $i = 1, \ldots, N$, can be found to satisfy

$$\begin{aligned} x^T \Big\{ & P\Big[A_i(x,t) + B_i(x,t)K_i(x,t)\Big] \\ & + \Big[A_i(x,t) + B_i(x,t)K_i(x,t)\Big]^T P \Big\} x \\ & + 2x^T P \Phi_i(x,z,t) \le -\nu|x|^2 \\ & \qquad \forall x \in \mathcal{R}^n, z \in \mathcal{R}^{n_z}, i \in \{1, \ldots, N\}, \end{aligned} \tag{5.2}$$

then the state-feedback control law $u = K_{l(t)}(x,t)x$ makes the origin $x = 0$ a globally asymptotically stable equilibrium of system (5.1). The inequality (5.2) is of the form (3.9).

2. *Observer design* (*dual to state-feedback*): Given a system

$$\dot{x} = A_{l(t)}(y,t)x + B_{l(t)}(y,t)u; \quad y = C_{l(t)}(t)x, \tag{5.3}$$

with y being the output, a full-order state observer is given by

$$\dot{\hat{x}} = A_{l(t)}(y,t)\hat{x} + B_{l(t)}(y,t)u + G_{l(t)}(y,t)[C_{l(t)}(t)\hat{x} - y] \tag{5.4}$$

if a matrix $P > 0$, a positive constant ν, and functions $G_i \in \mathcal{R}^{n\times 1}$, $i = 1, \ldots, N$, can be found such that [12]

$$P\Big[A_i(y,t) + G_i(y,t)C_i(t)\Big] + \Big[A_i(y,t) + G_i(y,t)C_i(t)\Big]^T P \le -\nu I, \qquad i = 1, \ldots, N. \tag{5.5}$$

3. *Output-feedback*: A dynamic output-feedback controller can be constructed by combining the state-feedback and observer designs above as long as (5.2) and (5.5) can be satisfied (not necessarily with the same P).
4. *Dynamic high-gain scaling–based state-feedback/observer design/output-feedback*: Dynamic high-gain scaling [11, 5, 19, 6, 9, 8] provides a method for triangular systems to attenuate part of the system dynamics so that the system is approximated by a chain of nonlinear integrators; i.e., the system matrix is approximated by $A(x)$ with A having the form shown in (3.1) with only the upper diagonal entries being nonzero. The coupled inequalities (2.1) and (2.2) appear in observer design and their dual, (2.3) and (2.4), appear in controller design based on this technique. The approach is applicable to systems in both lower triangular (strict-feedback) [11, 5, 19, 8] and upper triangular (feedforward) [6, 9] forms and also some classes of nontriangular systems [10]. The approach is illustrated below for state-feedback controller design for a particular simple class of lower triangular systems.
Consider a system of the form

$$\dot{x}_i = \phi_{(i,i+1)}(x)x_{i+1} + \phi_i(x_0, \ldots, x_i, t), \qquad i = 0, \ldots, n, \tag{5.6}$$

where $x = [x_0, \ldots, x_n]^T$ is the system state and $x_{n+1} = u$ is the input. The functions $\phi_{(i,i+1)}$ are assumed to be continuous and lower bounded in magnitude by a positive constant σ, and the functions ϕ_i are assumed to be bounded as $|\phi_i(x_0, \ldots, x_i, t)| \le |\phi_{(0,1)}(x)|\gamma(x_0)\sum_{j=0}^{i}|x_j|$ with γ being a positive function. It is also assumed that $\phi_{(0,1)}(x) \le \overline{\phi}_{(0,1)}(x_0)|\phi_{(1,2)}(x)|$ for all $x \in \mathcal{R}^n$ with $\overline{\phi}_{(0,1)}$ being a positive function.
Define the scaling $\xi_1 = [x_1 + \zeta(x_0)]/r$, $\xi_i = x_i/r^i$, $i = 2, \ldots, n+1$, with $\zeta(x_0) = x_0\zeta_1(x_0)$ being a design function and r a dynamic scaling parameter. The dynamics of $\xi = [\xi_1, \ldots, \xi_n]^T$ can be written in matrix form as

$$\dot{\xi} = r\Big[A(x) + |\phi_{(1,2)}(x)|BK(x)\Big]\xi - \frac{\dot{r}}{r}D\xi + \Phi + H\xi_1 + \Xi \tag{5.7}$$

by picking the control input to be

$$u = \frac{r^{n+1}}{\phi_{(n,n+1)}}|\phi_{(1,2)}(x)|K(x)\xi, \tag{5.8}$$

where $K(x) \in \mathcal{R}^{1\times n}$, $A(x)$ is of the form in (3.1) with nonzero entries only on the upper diagonal, $B = [0, \ldots, 0, 1]^T$, $D = \mathrm{diag}(1, 2, \ldots, n)$, and

$$\Phi = \Big[\frac{\phi_1}{r}, \ldots, \frac{\phi_n}{r^n}\Big]^T, \tag{5.9}$$

$$H = \Big[\zeta'(x_0)\phi_{(0,1)}(x), 0, \ldots, 0\Big]^T, \tag{5.10}$$

$$\Xi = \frac{1}{r}\Big[(\phi_0 - \zeta(x_0)\phi_{(0,1)}(x))\zeta'(x_0), 0, \ldots, 0\Big]^T. \tag{5.11}$$

The dynamics of r are chosen to be the scalar Riccati equation

$$\dot{r} = |\phi_{(1,2)}(x)|r[-a_1(r-1) + a_2\gamma_1(x_0)] \tag{5.12}$$

with a_1 and a_2 being positive constants and γ_1 being a positive function. It will be shown below that the system is globally asymptotically stabilized with some choice of functions ζ_1 and γ_1 if the coupled Lyapunov inequalities

$$P[\hat{A}(x) + BK(x)] + [\hat{A}(x) + BK(x)]^T P \leq -\nu_1 I, \tag{5.13}$$

$$P[D - I/2] + [D - I/2]P \geq \nu_2 I, \tag{5.14}$$

where $\hat{A}(x) = A(x)/|\phi_{(1,2)}(x)|$, are satisfied with some $K(x)$, a matrix $P > 0$, and positive constants ν_1 and ν_2. The coupled Lyapunov inequalities (5.13) and (5.14) are of the forms (2.3) and (2.4), respectively. Since $\hat{A}$ has nonzero terms (which, by assumption, are continuous and lower bounded in magnitude by a positive constant) only on the upper diagonal, the dual of Corollary 2 gives the necessary and sufficient conditions for solvability of the coupled Lyapunov inequalities (5.13) and (5.14) to be that $|\phi_{(i,i+1)}(x)| \geq \rho_{iA}|\phi_{(i-1,i)}(x)|$, $i = 2, \ldots, n-1$, for all $x \in \mathcal{R}^{n+1}$ with ρ_{iA}, $i = 2, \ldots, n-1$, being positive constants.

The closed-loop stability can be analyzed using the Lyapunov function $V = \frac{1}{2}x_0^2 + r\xi^T P\xi$, which satisfies

$$\begin{aligned}\dot{V} = {} & x_0[\phi_{(0,1)}(x)x_1 + \phi_0] + 2r\xi^T P\Phi + 2r\xi^T PH\xi_1 + 2r\xi^T P\Xi \\ & + r^2|\phi_{(1,2)}(x)|\xi^T\Big[P[\hat{A}(x) + BK(x)] + [\hat{A}(x) + BK(x)]^T P\Big]\xi \\ & - \dot{r}\xi^T\Big[P[D - I/2] + [D - I/2]P\Big]\xi.\end{aligned} \tag{5.15}$$

Using the coupled Lyapunov inequalities (5.13) and (5.14) and the dynamics of r given in (5.12), (5.15) reduces to

$$\begin{aligned}\dot{V} = {} & x_0[\phi_{(0,1)}(x)x_1 + \phi_0] + 2r\xi^T P\Phi + 2r\xi^T PH\xi_1 + 2r\xi^T P\Xi \\ & - r^2|\phi_{(1,2)}(x)|\nu_1|\xi|^2 + a_1 r^2|\phi_{(1,2)}(x)|\overline{\nu}_2|\xi|^2 \\ & - r[a_1 + a_2\gamma_1(x_0)]|\phi_{(1,2)}(x)|\nu_2|\xi|^2,\end{aligned} \tag{5.16}$$

where $\overline{\nu}_2$ is a positive constant large enough to satisfy

$$P[D - I/2] + [D - I/2]P \leq \overline{\nu}_2 I. \tag{5.17}$$

With some algebraic manipulation involving mainly routine quadratic overbounding, it can be shown that if a_2 is chosen to be any positive constant and if a_1, ζ_1, and γ_1 are chosen to satisfy[4]

$$-\frac{3\nu_1}{4} + a_1\overline{\nu}_2 = -a^* < 0, \tag{5.18}$$

$$\zeta_1(x_0) = \frac{1}{\sigma}\operatorname{sign}(\phi_{(0,1)}(0))\left[\frac{\sigma\overline{\phi}_{(0,1)}(x_0)}{\nu_1} + \sigma\gamma(x_0) + 2\sigma + \zeta^*(x_0)\right], \tag{5.19}$$

[4]Since $\phi_{(0,1)}$ is continuous and bounded below in magnitude by a positive constant, it assumes the same sign for all arguments.

$$\gamma_1(x_0) \geq \frac{1}{a_2\nu_2}\Bigg[2\lambda_{max}(P)\overline{\phi}_{(0,1)}(x_0)\gamma(x_0)(n+1) + 2\lambda_{max}(P)|\zeta'(x_0)|\overline{\phi}_{(0,1)}(x_0) + \lambda^2_{max}(P)\overline{\phi}_{(0,1)}(x_0)\Big\{[\gamma(x_0)+|\zeta_1(x_0)|]^2[\zeta'(x_0)]^2 + 2(n+1)\gamma^2(x_0)[1+\zeta_1^2(x_0)]\Big\}\Bigg], \tag{5.20}$$

where ζ^* is a positive function bounded below in magnitude by a positive constant, then (5.16) simplifies to

$$\dot{V} \leq -\zeta^* x_0^2 - r^2 a^* \sigma |\xi|^2. \tag{5.21}$$

It can be inferred from (5.21) that $(r, x) = (1 + \frac{a_2}{a_1}\gamma_1(0), 0, \ldots, 0)$ is a globally asymptotically stable equilibrium of the closed-loop system formed by (5.6), (5.8), and (5.12). The key ingredient in attaining (5.21) is the use of the coupled Lyapunov inequalities (5.13) and (5.14) to obtain (5.16) from (5.15). This approach, which is based on dynamic high-gain scaling, can be applied to both state-feedback and output-feedback control of a variety of nonlinear systems, including strict-feedback systems, feedforward systems, and some classes of nontriangular systems, and can be extended to include adaptation to unknown parameters and robustness to uncertain system terms and appended input-to-state stable (ISS) dynamics [11, 5, 19, 6, 9, 8, 10].

5. *λ-tracking*: Inequalities (2.1) and (2.2), and their duals (2.3) and (2.4), also appear in the literature [3] on the λ-tracking problem for nonlinear systems, where the solvability of the coupled Lyapunov inequalities has been assumed so far. The control objective is to regulate the tracking error to a λ-neighborhood of zero.

6. Conclusion. We considered the problem of finding a uniform solution to a family of parameter-dependent Lyapunov inequalities with the matrices incorporating design freedoms. Generically, this problem can be viewed as picking a family of controller (or observer) gains so that the family of closed-loop system matrices admits a common Lyapunov function. Several sufficient and necessary conditions for specific structures of the matrix families were considered. The solvability conditions involve inequalities regarding relative magnitudes of terms in the matrices. The Lyapunov inequalities arising in the controller and observer contexts were seen to have a dual relation. The duals of some of our theorems were explicitly provided, and the duals of the other theorems can be similarly obtained. Several motivating applications were presented to highlight the importance of the considered structures of the matrix families. The obtained results provide extensions in controller and observer designs for time-varying, switched, and nonlinear systems.

Appendix. Proofs of Lemmas 1–3.

A.1. Proof of Lemma 1. The diagonal matrix D can be decomposed as $D = D_1(\theta)\hat{D}(\theta)$, where $\hat{D}(\theta) = \mathrm{diag}(\hat{D}_1(\theta), \hat{D}_2(\theta), \ldots, \hat{D}_n(\theta))$ with $\hat{D}_i(\theta) = \frac{D_i(\theta)}{D_1(\theta)}$. The inequality $P\hat{D} + \hat{D}P \geq \tilde{\nu}_2 I$ holds with some positive $\tilde{\nu}_2$ if $|k_i| < \kappa$ with small enough κ. This is seen by continuity since this inequality holds when $k_2 = \cdots = k_n = 0$. Note that the functions $\hat{D}_i(\theta)$ are bounded below and above by the positive constants

$\prod_{j=1}^{i-1} \underline{\rho}_{jD}$ and $\prod_{j=1}^{i-1} \overline{\rho}_{jD}$, respectively. An estimate (possibly conservative) for κ can be written, using the diagonal dominance condition, as

$$\kappa = \inf_{\theta \in \Theta} \min \left\{ \frac{2\hat{D}_i(\theta)}{\hat{D}_{i+1}(\theta) + 2\hat{D}_i(\theta) + \hat{D}_{i-1}(\theta)};\ 1 \le i \le n \right\} \tag{A.1}$$

with $\hat{D}_0 = -\hat{D}_1$ and $\hat{D}_{n+1} = -\hat{D}_n$. The inequality $PD + DP \ge \nu_2 I$ with $\nu_2 = \underline{\rho}^o_{1D} \tilde{\nu}_2$ follows if $|k_i| < \kappa$, $i = 2, \ldots, n$. □

A.2. Proof of Lemma 2. Consider the dynamical system $\dot{\omega} = A_c\omega$, where $\omega = [\omega_1, \ldots, \omega_n]^T \in \mathcal{R}^n$ and $A_c = A + GC$. Define the change of coordinates $\Omega = \hat{P}\omega$. We will prove below that, along the trajectories of the dynamical system $\dot{\omega} = A_c\omega$, the Lyapunov inequality $\frac{d}{dt}(\Omega^T\Omega) \le -\alpha\Omega^T\Omega$ is satisfied with α being a positive constant. Since $\Omega^T\Omega = \omega^T P\omega$ by construction, this will imply that (2.1) is satisfied with $\nu_1 = \alpha\lambda_{min}(P)$.

The inverse transformation relating ω to Ω is given by[5]

$$\omega_i = \sum_{l=1}^{i} \Omega_l \prod_{m=l+1}^{i} k_m = \sum_{l=1}^{i} p_k(l+1, i)\Omega_l, \tag{A.2}$$

where $p_k(n_1, n_2) \triangleq \prod_{m=n_1}^{n_2} k_m$. Note that $p_k(n_1, n_2) = 1$ if $n_1 > n_2$. For convenience, we introduce the dummy variables $\omega_{n+1} = \Omega_{n+1} = k_{n+1} = 0$ and $\phi_{(n+1,j)} = 0$, $1 \le j \le n+1$. Differentiating $\Omega_i = \omega_i - k_i\omega_{i-1}$, $2 \le i \le n$, we obtain

$$\begin{aligned}
\dot{\Omega}_i = {} & \Big(g_i - k_i g_{i-1} + \tilde{g}_i\Big)\Omega_1 + \phi_{(i,i+1)}\Omega_{i+1} \\
& + \Big(\phi_{(i,i)} - k_i\phi_{(i-1,i)} + k_{i+1}\phi_{(i,i+1)}\Big)\Omega_i \\
& + \Big(\phi_{(i,i)} - k_i\phi_{(i-1,i)} + k_{i+1}\phi_{(i,i+1)}\Big)\sum_{l=2}^{i-1} p_k(l+1, i)\Omega_l \\
& + \sum_{j=2}^{i-1}\left\{(\phi_{(i,j)} - k_i\phi_{(i-1,j)})\sum_{l=2}^{j} p_k(l+1, j)\Omega_l\right\},
\end{aligned} \tag{A.3}$$

$$\begin{aligned}
\tilde{g}_i = {} & \phi_{(i,1)} - k_i\phi_{(i-1,1)} + \Big(\phi_{(i,i)} - k_i\phi_{(i-1,i)} + k_{i+1}\phi_{(i,i+1)}\Big)p_k(2, i) \\
& + \sum_{j=2}^{i-1}(\phi_{(i,j)} - k_i\phi_{(i-1,j)})p_k(2, j).
\end{aligned} \tag{A.4}$$

Note that $|k_i|$, $i = 2, \ldots, n$, are required to be chosen smaller than κ if ρ_{iA} and $\epsilon_{(i,j)}$ are smaller than $\overline{\rho}_A$. The final construction of k_i and $\overline{\rho}_A$ will ensure these inequality conditions. We derive a bound below for $\sum_{i=2}^{n} \Omega_i\dot{\Omega}_i$ assuming these inequalities. Furthermore, since the positive constant $\overline{\rho}_A$ is free to be arbitrarily chosen, we assume $\overline{\rho}_A < 1$.

Using the w-CUDD conditions given in (3.2), for $2 \le i \le n$,

[5]We follow the convention that $\sum_{i=n_1}^{n_2} f_i = 0$ and $\prod_{i=n_1}^{n_2} f_i = 1$ if $n_1 > n_2$.

$$
\begin{aligned}
\Omega_i\dot{\Omega}_i \le{}& \frac{\kappa}{4}|\phi_{(i,i+1)}|\Omega_{i+1}^2 + \frac{1}{\kappa}|\phi_{(i,i+1)}|\Omega_i^2 \\
&+ \Big[-k_i\phi_{(i-1,i)} + \kappa\overline{\rho}_A|\phi_{(i-1,i)}| + \overline{\rho}_A|\phi_{(i-1,i)}|\Big]\Omega_i^2 \\
&+ \sum_{l=2}^{i-1}\frac{(i-2)(\phi_{(i,i)} - k_i\phi_{(i-1,i)} + k_{i+1}\phi_{(i,i+1)})^2}{\kappa|\phi_{(i-1,i)}|}[p_k(l+1,i)]^2\Omega_l^2 \\
&+ \frac{\kappa}{4}|\phi_{(i-1,i)}|\Omega_i^2 + \sum_{j=2}^{i-1}\sum_{l=2}^{j}\frac{(j-1)(i-2)(\phi_{(i,j)} - k_i\phi_{(i-1,j)})^2}{\kappa|\phi_{(i-1,i)}|}[p_k(l+1,j)]^2\Omega_l^2 \\
&+ \frac{\kappa}{4}|\phi_{(i-1,i)}|\Omega_i^2 + \frac{2}{\kappa}\frac{(g_i - k_ig_{i-1} + \tilde{g}_i)^2}{|\phi_{(i-1,i)}|}\Omega_1^2 + \frac{\kappa}{8}|\phi_{(i-1,i)}|\Omega_i^2 \\
={}& \frac{\kappa}{4}|\phi_{(i,i+1)}|\Omega_{i+1}^2 - k_i\phi_{(i-1,i)}\Omega_i^2 + \Big[\kappa\overline{\rho}_A + \overline{\rho}_A + \frac{\overline{\rho}_A}{\kappa} + \frac{5\kappa}{8}\Big]|\phi_{(i-1,i)}|\Omega_i^2 \\
&+ \sum_{l=2}^{i-1} q_{(i,l)}|\phi_{(l-1,l)}|\Omega_l^2 + \frac{2}{\kappa}\frac{(g_i - k_ig_{i-1} + \tilde{g}_i)^2}{|\phi_{(i-1,i)}|}\Omega_1^2, \qquad \text{(A.5)}
\end{aligned}
$$

where, for $i = 2, \dots, n$ and $l = 2, \dots, i-1$,

$$
\begin{aligned}
q(i,l) = \frac{1}{|\phi_{(l-1,l)}|}\Bigg[&\frac{(i-2)(\phi_{(i,i)} - k_i\phi_{(i-1,i)} + k_{i+1}\phi_{(i,i+1)})^2}{\kappa|\phi_{(i-1,i)}|}[p_k(l+1,i)]^2 \\
&+ \sum_{j=l}^{i-1}\frac{(j-1)(i-2)(\phi_{(i,j)} - k_i\phi_{(i-1,j)})^2}{\kappa|\phi_{(i-1,i)}|}[p_k(l+1,j)]^2\Bigg] \\
\le \frac{(n-2)}{\kappa}\Bigg[&(\epsilon_{(i,i)}\rho_{iA} + \rho_{iA}|k_{i+1}| + |k_i|)^2[p_k(l+1,i)]^2\prod_{r=l}^{i-1}\rho_{rA} \\
&+ (n-1)\sum_{j=l}^{i-1}(\epsilon_{(i,j)}\rho_{iA} + \epsilon_{(i-1,j)}|k_i|)^2[p_k(l+1,j)]^2\prod_{r=l}^{j-1}\rho_{rA}\Bigg] \\
\le \frac{(n-2)\overline{\rho}_A}{\kappa}&\Bigg[(1+2\kappa)^2\kappa^{2(i-l)} + (n-1)(1+\kappa)^2\frac{\kappa^{2(i-l)}-1}{\kappa^2-1}\Bigg]. \qquad \text{(A.6)}
\end{aligned}
$$

Choose any positive constant $\kappa^* < \frac{\kappa}{8}$ and choose $\overline{\rho}_A$ to be a positive constant satisfying

$$
\overline{\rho}_A \le \frac{\frac{\kappa}{8} - \kappa^*}{\kappa + 1 + \frac{1}{\kappa} + \frac{(n-2)}{\kappa}\sum_{l=i+1}^{n}\Big[(1+2\kappa)^2\kappa^{2(l-i)} + (n-1)(1+\kappa)^2\frac{\kappa^{2(l-i)}-1}{\kappa^2-1}\Big]}. \qquad \text{(A.7)}
$$

Note that $\overline{\rho}_A < 1$, implying that the inequalities in (A.5) and (A.6) are valid. Inequality (A.5) reduces to

$$
\begin{aligned}
\sum_{i=2}^{n}\Omega_i\dot{\Omega}_i \le{}& -\sum_{i=2}^{n}k_i\phi_{(i-1,i)}\Omega_i^2 + \sum_{i=2}^{n}(\kappa - \kappa^*)|\phi_{(i-1,i)}|\Omega_i^2 \\
&+ \frac{2}{\kappa}\sum_{i=2}^{n}\frac{(g_i - k_ig_{i-1} + \tilde{g}_i)^2}{|\phi_{(i-1,i)}|}\Omega_1^2. \qquad \text{(A.8)}
\end{aligned}
$$

The constants k_i, $i = 2, \dots, n$, are chosen as[6] $k_i = (\kappa - \kappa^*/4)\,\mathrm{sign}(\phi_{(i-1,i)}(0))$. Since

[6]Note that $\phi_{(i-1,i)}$ takes the same sign for all arguments by assumption.

$\Omega_1 = \omega_1$,

$$\dot{\Omega}_1 = (g_1 + \tilde{g}_1)\Omega_1 + \phi_{(1,2)}\Omega_2$$

$$\implies \Omega_1\dot{\Omega}_1 \leq \Big(g_1 + \tilde{g}_1 + \frac{1}{\kappa^*}|\phi_{(1,2)}|\Big)\Omega_1^2 + \frac{\kappa^*}{4}|\phi_{(1,2)}|\Omega_2^2, \tag{A.9}$$

where $\tilde{g}_1 = \phi_{(1,1)} + k_2\phi_{(1,2)}$. Defining G such that[7]

$$G = [g_1, \ldots, g_n]^T = -\hat{P}^{-1}\left[\frac{2}{\kappa}\sum_{i=2}^{n}\frac{\overline{g}_i^2}{|\phi_{(i-1,i)}|} + \overline{g}_1, 0, \ldots, 0\right]^T, \tag{A.10}$$

$$\overline{g}_1 \geq |\phi_{(1,1)}| + |k_2||\phi_{(1,2)}| + \frac{1}{\kappa^*}|\phi_{(1,2)}| + \frac{\kappa^*}{2}|\phi_{(1,2)}|, \tag{A.11}$$

$$\begin{aligned}\overline{g}_i \geq\ & |\phi_{(i,1)}| + |k_i||\phi_{(i-1,1)}| \\ & + \Big(|\phi_{(i,i)}| + |k_i||\phi_{(i-1,i)}| + |k_{i+1}||\phi_{(i,i+1)}|\Big)|p_k(2,i)| \\ & + \sum_{j=2}^{i-1}(|\phi_{(i,j)}| + |k_i||\phi_{(i-1,j)}|)|p_k(2,j)|, \qquad i = 2, \ldots, n,\end{aligned} \tag{A.12}$$

we finally obtain

$$\begin{aligned}\frac{d}{dt}\left(\frac{1}{2}\sum_{i=1}^{n}\Omega_i^2\right) &= \sum_{i=1}^{n}\Omega_i\dot{\Omega}_i \\ &\leq -\frac{\kappa^*}{2}\sum_{i=2}^{n}|\phi_{(i-1,i)}|\Omega_i^2 - \frac{\kappa^*}{2}|\phi_{(1,2)}|\Omega_1^2 \\ &\leq -\kappa^*\sigma\left(\frac{1}{2}\sum_{i=1}^{n}\Omega_i^2\right).\end{aligned} \tag{A.13}$$

Noting that $\sum_{i=1}^{n}\Omega_i^2 = \omega^T P\omega$, (2.1) is satisfied with $\nu_1 = \kappa^*\sigma\lambda_{min}(P)$. ∎

A.3. Proof of Lemma 3. If the w-CUDD constants of A are smaller than $\overline{\rho}_A$, T can be chosen to be the identity matrix. Otherwise, let

$$\rho_0 = \frac{\max\Big\{\max\{\rho_{iA}|i = 2, \ldots, n-1\}, \max\{\epsilon_{i,j}|i = 2, \ldots, n, j = 2, \ldots, i\}\Big\}}{\overline{\rho}_A}, \tag{A.14}$$

$$T = \mathrm{diag}(T_1, \ldots, T_n), \quad T_i = \rho_0^{n_i}, \quad n_i = (n-i)(n-i+1)/2, \quad i = 1, \ldots, n. \tag{A.15}$$

Note that ρ_0 as defined in (A.14) is greater than 1. The (i,j)th element of $\tilde{A}$ is $\tilde{A}_{(i,j)} = (T_i/T_j)\phi_{(i,j)}$, $i = 1, \ldots, n$, $j = 1, \ldots, i+1$, with zeros everywhere else. Using (A.14),

$$\begin{aligned}\frac{|\tilde{A}_{(i,i+1)}|}{|\tilde{A}_{(i-1,i)}|} = \frac{T_i^2}{T_{i-1}T_{i+1}}\frac{|\phi_{(i,i+1)}|}{|\phi_{(i-1,i)}|} &= \frac{1}{\rho_0}\frac{|\phi_{(i,i+1)}|}{|\phi_{(i-1,i)}|} \\ &\leq \frac{\rho_{iA}}{\rho_0} \leq \overline{\rho}_A,\end{aligned} \tag{A.16}$$

[7]This definition of $\overline{g}_i$ implies that the choice of G depends only on the upper diagonal entries and upper bounds on other elements of A.

$$\begin{aligned}\frac{|\tilde{A}_{(i,j)}|}{\sqrt{|\tilde{A}_{(i,i+1)}||\tilde{A}_{(j-1,j)}|}} &= \sqrt{\frac{T_iT_{i+1}}{T_{j-1}T_j}}\frac{|\phi_{(i,j)}|}{\sqrt{|\phi_{(i,i+1)}||\phi_{(j-1,j)}|}}\\ &= \rho_0^{\frac{(j-i-1)(2n-i-j+1)}{2}}\frac{|\phi_{(i,j)}|}{\sqrt{|\phi_{(i,i+1)}||\phi_{(j-1,j)}|}}\\ &\le \frac{\epsilon_{(i,j)}}{\rho_0} \le \overline{\rho}_A, \qquad i=2,\dots,n-1,\ j=2,\dots,i, \end{aligned} \tag{A.17}$$

$$\begin{aligned}\frac{|\tilde{A}_{(n,j)}|}{\sqrt{|\tilde{A}_{(n-1,n)}||\tilde{A}_{(j-1,j)}|}} &= \sqrt{\frac{T_n^3}{T_{j-1}T_jT_{n-1}}}\frac{|\phi_{(n,j)}|}{\sqrt{|\phi_{(n-1,n)}||\phi_{(j-1,j)}|}}\\ &= \frac{1}{\rho_0^{\frac{1+(n-j+1)^2}{2}}}\frac{|\phi_{(n,j)}|}{\sqrt{|\phi_{(n-1,n)}||\phi_{(j-1,j)}|}}\\ &\le \frac{\epsilon_{(n,j)}}{\rho_0} \le \overline{\rho}_A, \qquad j=2,\dots,n. \end{aligned} \tag{A.18}$$

Hence, $\tilde{A}$ is a w-CUDD matrix with w-CUDD constants smaller than $\overline{\rho}_A$. ∎

REFERENCES

[1] A. A. Agrachev and D. Liberzon, *Lie-algebraic stability criteria for switched systems*, SIAM J. Control Optim., 40 (2001), pp. 253–269.

[2] B. R. Barmish, *New Tools for Robustness of Linear Systems*, Macmillan, London, 1994.

[3] E. Bullinger and F. Allgöwer, *Adaptive λ-tracking for nonlinear systems with higher relative degree*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 4771–4776.

[4] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and H^∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[5] P. Krishnamurthy and F. Khorrami, *Generalized adaptive output-feedback form with unknown parameters multiplying high output relative-degree states*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 1503–1508.

[6] P. Krishnamurthy and F. Khorrami, *Global robust control of feedforward systems: State-feedback and output-feedback*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 6145–6150.

[7] P. Krishnamurthy and F. Khorrami, *Robust adaptive control for non-linear systems in generalized output-feedback canonical form*, Internat. J. Adapt. Control Signal Process., 17 (2003), pp. 285–311.

[8] P. Krishnamurthy and F. Khorrami, *Dynamic high-gain scaling: State and output feedback with application to systems with ISS appended dynamics driven by all states*, IEEE Trans. Automat. Control, 49 (2004), pp. 2219–2239.

[9] P. Krishnamurthy and F. Khorrami, *A high-gain scaling technique for adaptive output feedback control of feedforward systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 2286–2292.

[10] P. Krishnamurthy and F. Khorrami, *Generalized state scaling-based robust control of nonlinear systems and applications to triangular systems*, in Proceedings of the 24th IEEE American Control Conference, Portland, OR, 2005, pp. 3427–3432.

[11] P. Krishnamurthy, F. Khorrami, and R. S. Chandra, *Global high-gain-based observer and backstepping controller for generalized output-feedback canonical form*, IEEE Trans. Automat. Control, 48 (2003), pp. 2277–2284.

[12] P. Krishnamurthy, F. Khorrami, and Z. P. Jiang, *Global output feedback tracking for nonlinear systems in generalized output-feedback canonical form*, IEEE Trans. Automat. Control, 47 (2002), pp. 814–819.

[13] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Syst. Mag., 19 (1999), pp. 59–70.

[14] Y. Mori, T. Mori, and Y. Kuroe, *A solution to the common Lyapunov function problem for continuous-time systems*, in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997, pp. 3530–3531.

[15] Y. Mori, T. Mori, and Y. Kuroe, *On a class of linear constant systems which have a common quadratic Lyapunov function*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 2808–2809.

[16] K. S. Narendra and J. Balakrishnan, *A common Lyapunov function for stable LTI systems with commutating A-matrices*, IEEE Trans. Automat. Control, 39 (1994), pp. 2469–2471.

[17] T. Ooba and Y. Funahashi, *Two conditions concerning common quadratic Lyapunov functions for linear systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 719–722.

[18] R. Patel, M. Toda, and B. Sridhar, *Robustness of linear quadratic state feedback designs in the presence of system uncertainty*, IEEE Trans. Automat. Control, 22 (1977), pp. 945–949.

[19] L. Praly, *Asymptotic stabilization via output feedback for lower triangular systems with output dependent incremental rate*, IEEE Trans. Automat. Control, 48 (2003), pp. 1103–1108.

[20] R. N. Shorten and K. S. Narendra, *On the stability and existence of common Lyapunov functions for stable linear switching systems*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 3723–3724.

[21] R. N. Shorten, K. S. Narendra, and O. Mason, *A result on common quadratic Lyapunov functions*, IEEE Trans. Automat. Control, 48 (2003), pp. 110–113.

[22] K. Wang and A. N. Michel, *On the stability of a family of nonlinear time-varying systems*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 43 (1996), pp. 517–531.

[23] K. Wei, *Quadratic stabilizability of linear systems with structural independent time-varying uncertainties*, IEEE Trans. Automat. Control, 35 (1990), pp. 268–277.

[24] R. Yedavalli, *Improved measures of stability robustness for linear state space models*, IEEE Trans. Automat. Control, 30 (1985), pp. 577–579.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1165–1197

DOMAIN DEPENDENCE OF SOLUTIONS TO COMPRESSIBLE NAVIER–STOKES EQUATIONS*

P. I. PLOTNIKOV† AND J. SOKOLOWSKI‡

Abstract. The minimization of the drag functional for the stationary, isentropic, compressible Navier–Stokes equations (NSE) in three spatial dimensions is considered. In order to establish the existence of an optimal shape, the general result on compactness of families of generalized solutions to the NSE is established within in the framework of the modern theory of nonlinear PDEs [P. L. Lions, *Mathematical Topics in Fluid Dynamics. Vol.* 2. *Compressible Models*, Oxford University Press, Clarendon Press, New York, 1998], [E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, UK, 2004]. The family of generalized solutions to the NSE is constructed over a family of admissible domains $\mathcal{U}_{ad}$. Any admissible domain $\Omega = B \setminus S$ contains an obstacle S, e.g., a wing profile. Compactness properties of the family of admissible domains in the form of the condition $(\mathcal{H}_\Omega)$ is imposed. Roughly speaking, the condition $(\mathcal{H}_\Omega)$ is satisfied, provided that for any sequence of admissible domains $\{\Omega_n\} \subset \mathcal{U}_{ad}$ there is a subsequence convergent both in Hausdorff metrics and in the sense of Kuratowski and Mosco. The analysis is performed for the adiabatic constant $\gamma > 1$ in the pressure law $p(\rho) = \rho^\gamma$ and it is based on the technique used in [P. I. Plotnikov and J. Sokolowski, *Comm. Math. Phys.,* 258 (2005), pp. 567–608] in the case of discretized NSE. The result is a generalization to the stationary equations with $\gamma > 1$ of the results obtained in [E. Feireisl, A. H. Novotný, and H. Petzeltová, *Math. Methods Appl. Sci.,* 25 (2002), pp. 1045–1073], [E. Feireisl, *Appl. Math. Optim.,* 47 (2003), pp. 59–78] for evolution equations within the range $\gamma > 3/2$ for the adiabatic ratio.

Key words. Navier–Stokes equations, compressible fluids, shape optimization

AMS subject classifications. 35Q30, 49J20, 76N10

DOI. 10.1137/050635304

1. Introduction.

1.1. Shape optimization. In the present paper, a class of shape optimization problems for stationary isothermal compressible Navier–Stokes equations (NSE) is considered. Usually the mathematical analysis of such problems includes the following:

- Proof of the existence of optimal shapes for a sufficiently large class of admissible domains.
- Derivation of necessary optimality conditions, which characterize an optimal domain and can be used for the numerical solution of the shape optimization problem.
- The convergence proof for the numerical method, which can be used to evaluate an optimal shape.

To the best of our knowledge, only the first point is studied in the literature for the compressible NSE and drag minimization. We refer the reader to [10] for the existence results for shape optimization problems in the case of evolution equations for the adiabatic constant $\gamma > 3/2$; see also [9] for related results. In the present paper the case of $\gamma > 1$ is considered in three spatial dimensions. The technique used

*Received by the editors July 6, 2005; accepted for publication (in revised form) March 20, 2006; published electronically September 12, 2006. This paper was prepared in June 2005 during a visit of Pavel I. Plotnikov to the Institute Elie Cartan of the University Henri Poincaré Nancy 1.
http://www.siam.org/journals/sicon/45-4/63530.html

†Lavryentyev Institute of Hydrodynamics, Siberian Division of Russian Academy of Sciences, Lavryentyev pr. 15, Novosibirsk 630090, Russia (plotnikov@hydro.nsc.ru).

‡Institut Elie Cartan, Laboratoire de Mathématiques, Université Henri Poincaré Nancy 1, B.P. 239, 54506 Vandoeuvre lés Nancy Cedex, France (Jan.Sokolowski@iecn.u-nancy.fr).

is introduced in [12], [24], and [25]. The optimality conditions for drag minimization are derived in [27].

We restrict ourselves to the first issue on the above list of problems to be solved, which is already quite difficult, since there are no existence results for the PDE model itself in the range of parameters we are going to consider, i.e., for the adiabatic constant $\gamma > 1$ in the law $p(\rho) = \rho^\gamma$, which gives the pressure p of the fluid in the function of its density ρ in the isothermal regime. We also point out that the existence of generalized solutions for such equations with nonhomogeneous boundary conditions for a full range of parameters is an open problem; for some results in this direction we refer to [15], [16], [17] for two spatial dimensions and for special geometry of the flow region.

We show that, under appropriate assumptions, the convergence of the sequence of admissible domains implies the convergence of the associated sequence of shape functionals. To make our analysis representative we consider the drag functional, which is the standard choice in view of possible applications in the shape optimization of a wing.

A related shape optimization problem is considered in [10] for evolution equations, and with the adiabatic constant $\gamma > 3/2$. We consider the case of $\gamma > 1$ using the same technique as in [12], [25]. The novelty of our results, in comparison with paper [25], is the PDE model; i.e., we consider here the stationary problem and use a far-reaching approach to the problem based on the new kinetic formulation to the mass balance equation. In [25] the discretized problem introduced in [18] is considered; such a problem depends on the parameter $\alpha > 0$ of the time discretization of evolution equations [18], and the limit case of $\alpha = 0$ is the stationary problem.

The outline of the paper is the following. In section 1.2 an example of a shape optimization problem is presented and the existence result of optimal shapes is given. In section 1.3 the stationary NSE are introduced, and generalized solutions are defined. In section 1.4 the main result of the paper on compactness of generalized solutions and solvability of the drag minimization problem is presented. In section 2 local a priori estimates for generalized solutions are derived. In section 3 weak limits of sequences of generalized solutions are characterized. In section 4 the so-called effective viscous flux is defined and some properties of the viscous flux are obtained. In section 5 the oscillation defect measure is introduced for sequences of densities and its boundedness is proved for generalized solutions. In section 6 the kinetic formulation for the mass balance equation is derived in the form of a transport equation for the distribution function, which characterizes the Young measure associated with weakly convergent sequences of densities. In section 7 the renormalization of the kinetic formulation for the mass balance equation is performed, which shows the strong convergence of sequences of densities. Finally, in section 8 the proof of the main result of the paper is completed.

Since all of our presented results for the adiabatic ratio $\gamma > 1$ seem to be new, we provide the complete proofs for the convenience of the reader. To our knowledge such proofs are not known in the literature. The proof technique was introduced by Lions and by Feireisl, Novotný, and Petzeltová and it is adapted to the specific problem along the lines of [12], [24], and [25].

1.2. Example of the existence of an optimal shape. We use the notation introduced below. Let $(\mathbf{u}, \rho)$ be a generalized solution to the boundary value problem (1.3) posed in the geometrical domain $\Omega = B \setminus S$, where $B \subset \mathbb{R}^3$ is a fixed hold-all domain and S is an obstacle. The family of admissible domains $\Omega \in \mathcal{U}_{\rm ad}$ includes

domains with obstacles of the volume bounded from below, i.e., $\mathrm{meas}(S) \geq \mathrm{Vol}$ for all admissible domains $\Omega = B \setminus S$. Let us consider the problem of drag minimization, in which the shape functional $J(\Omega)$ takes the form

$$J(\Omega) := \mathbf{J}(\rho, \mathbf{u}, \Omega),$$

where the functional $\mathbf{J}(\rho, \mathbf{u}, \Omega)$ is defined in (1.5), over such a family of admissible domains, and let us assume that the set of generalized solutions to (1.3) over admissible domains $\Omega \in \mathcal{U}_{\mathrm{ad}}$ is nonempty. If the family $\mathcal{U}_{\mathrm{ad}}$ is compact with respect to the convergence defined by the condition $(\mathcal{H}_\Omega)$, then there is an optimal domain $\Omega^* = B \setminus S^*$ such that

$$J(\Omega^*) \leq J(\Omega)$$

for all $\Omega \in \mathcal{U}_{\mathrm{ad}}$. We prove the existence result without any regularity assumptions on the boundaries of obstacles. In fact, we prove a compactness result for the family of solutions to the NSE so that the existence of optimal shapes is a simple corollary of the theorem on the compactness of the set of generalized solutions.

1.3. The problem formulation. Suppose that compressible Newtonian fluid occupies the bounded region $\Omega \subset \mathbb{R}^3$. We will assume that $\Omega = B \setminus S$, where B is an open sufficiently large hold all with a sectionally smooth boundary containing a compact obstacle S. We could take, e.g., for B a ball of radius R, $B = \{x | |x| \leq R\}$. We do not impose restrictions on the topology of the flow region. The cases of S with a finite number of connected components or $S = \emptyset$ are taken into consideration. The fluid density $\rho : \Omega \mapsto \mathbb{R}^+$ and the velocity field $\mathbf{u} : \Omega \mapsto \mathbb{R}^3$ are governed by the NSE

$$-\nu\Delta\mathbf{u} - \xi\nabla\mathrm{div}\,\mathbf{u} + \rho\mathbf{u}\nabla\mathbf{u} + \nabla p(\rho) = \rho\mathbf{f}, \quad \mathrm{div}\ (\rho\mathbf{u}) = 0, \tag{1.1}$$

where ν, ξ are positive viscous coefficients and $\mathbf{f} : \Omega \mapsto \mathbb{R}^3$ is a given continuous vector field. We suppose that the flow is barotropic and $p(\rho) = \rho^\gamma$ with the adiabatic constant $\gamma > 1$. If the viscous stress tensor is defined by the equality

$$\Pi = \nu(\nabla\mathbf{u} + \nabla\mathbf{u}^\top) + (\xi - \nu)\mathrm{div}\,\mathbf{u}\,\mathbf{I}, \tag{1.2}$$

then the governing equations can be written in the equivalent divergence form

$$\mathrm{div}\,(\rho\mathbf{u} \otimes \mathbf{u}) + \nabla p(\rho) - \rho\mathbf{f} = \mathrm{div}\,\Pi, \quad \mathrm{div}\,(\rho\mathbf{u}) = 0 \ \text{ in } \ \Omega. \tag{1.3a}$$

In view of possible applications, e.g., to the shape optimization problem of a wing, it is assumed that the velocity field satisfies the nonhomogeneous boundary condition

$$\mathbf{u} = 0 \ \text{ on } \ \partial S, \quad \mathbf{u} = \mathbf{U}^\infty \ \text{ on } \ \partial B, \tag{1.3b}$$

and the density distribution is prescribed on the entrance set

$$\rho = \rho^\infty \ \text{ on } \ \Sigma^+ = \{x \in \partial B : \ \mathbf{U}^\infty \cdot \mathbf{n}(x) < 0\}. \tag{1.3c}$$

Here $\mathbf{n}$ is the outward unit normal vector to $\partial\Omega$. It is assumed that $\mathbf{U}^\infty \in \mathbb{R}^3$ is a given vector and ρ^∞ is a given nonnegative constant.

Boundary condition (1.3b) can be written in the form of the equality $\mathbf{u} = \mathbf{u}^\infty$ on $\partial\Omega$, where $\mathbf{u}^\infty(x)$ is a smooth function defined for any $x \in \mathbb{R}^3$, which vanishes in the vicinity of S and coincides with $\mathbf{U}^\infty$ in an open neighborhood of ∂B. The physical

quantities which characterize the flow include the total energy $\mathbf{E}$, the volume rate of energy dissipation $\mathbf{D}$, and the drag $\mathbf{J}$, and are defined by

$$\mathbf{E} = \int_\Omega \left(\frac{1}{2}\rho|\mathbf{u}|^2 + \frac{\rho^\gamma}{\gamma - 1} \right) dx, \quad \mathbf{D} = \int_\Omega (\nu|\nabla \mathbf{u}|^2 + \xi|\operatorname{div} \mathbf{u}|^2)\, dx, \tag{1.4}$$

$$\mathbf{J} = -\mathbf{U}^\infty \cdot \int_{\partial S} (\Pi - p(\rho)\mathbf{I}) \cdot \mathbf{n}\, dS.$$

The drag $\mathbf{J}$ accounts for the reaction of the surrounding fluid on the obstacle S. For our purposes, the formula for the drag can be written in the equivalent form

$$\mathbf{J}(\rho, \mathbf{u}, \Omega) = \int_\Omega (\Pi - \rho \mathbf{u} \otimes \mathbf{u} - p(\rho)\mathbf{I}) : \nabla \mathbf{u}^\infty dx + \int_\Omega (\mathbf{U}^\infty - \mathbf{u}^\infty) \cdot \mathbf{f} \rho dx. \tag{1.5}$$

We will consider the solutions to problems (1.3) for which the density is nonnegative and the total energies are bounded from above by some positive constant C_Ω. In what follows we will denote by c various constants depending only on $\mathbf{E}$, data $\|\mathbf{f}\|_{C(B)}$, ρ^∞, $\|\mathbf{u}^\infty\|_{C^1(B)}$, material constants γ, ν , ξ, and the domain B.

In this paper the standard notation is used for the function spaces. The space $H^{1,r}(\Omega)$ is the Sobolev space of functions integrable along with the first order generalized derivatives in $L_r(\Omega)$ equipped with its natural norm. For $r = 2$ we use the notation $H^{1,2}(\Omega)$ rather than $H^1(\Omega)$; the notation $H_0^{1,r}(\Omega)$ stands for the closure of $C_0^\infty(\Omega)$ in the norm of $H^{1,r}(\Omega)$.

DEFINITION 1.1. *For given $\mathbf{U}^\infty \in \mathbb{R}^3$ and $\mathbf{f} \in C(\Omega)^3$, a generalized solution to problem* (1.3) *is the pair $(\rho, \mathbf{u})$, where $\rho \in L^\gamma(\Omega)$ is a nonnegative function in Ω and $\mathbf{u} - \mathbf{u}^\infty \in H_0^{1,2}(\Omega)$, which satisfies the following conditions:*

(a) *The scalar function $\rho|\mathbf{u}|^2$ is integrable in Ω, i.e., the total energy $\mathbf{E}$ of the flow is finite. The mass density and the velocity field satisfy the energy inequality*

$$\nu\|\nabla(\mathbf{u} - \mathbf{u}^\infty)\|^2_{L^2(\Omega)} + \xi\|\operatorname{div}(\mathbf{u} - \mathbf{u}^\infty)\|^2_{L^2(\Omega)} + \int_\Omega \rho \mathbf{u} \otimes \mathbf{u} : \nabla \mathbf{u}^\infty\, dx$$

$$+ \int_\Omega p(\rho) \operatorname{div} \mathbf{u}^\infty\, dx - \int_\Omega \rho \mathbf{f} \cdot (\mathbf{u} - \mathbf{u}^\infty)\, dx + \frac{1}{\gamma - 1} \int_{\Gamma^+} (\rho^\infty)^\gamma (\mathbf{U}^\infty \cdot \mathbf{n})\, d\Gamma \le 0. \tag{1.6}$$

(b) *For all vector fields $\varphi \in C_0^1(\Omega)^3$,*

$$\int_\Omega \Big(\rho \mathbf{u} \otimes \mathbf{u} + p(\rho) \Big) : \nabla \varphi\, dx + \int_\Omega \rho \mathbf{f} \cdot \varphi\, dx = \int_\Omega \Pi : \nabla \varphi\, dx. \tag{1.7a}$$

(c) *The integral identity*

$$\int_\Omega \Big(G(\rho)\mathbf{u} \cdot \nabla \psi + \big(G(\rho) - G'(\rho)\rho\big)\psi \operatorname{div} \mathbf{u} \Big)\, dx - \int_{\Sigma^+} \psi G(\rho^\infty) \mathbf{U}^\infty \cdot \mathbf{n} d\Sigma = 0 \tag{1.7b}$$

holds for all functions $\psi \in C^1(\Omega)$ vanishing on $\Sigma^- = \partial B \setminus \Sigma^+$, and all functions $G \in C^1_{loc}[0, \infty)$ with the properties

$$\limsup_{r \to \infty} |G(r)|/r < \infty, \quad [0, \infty) \ni r \mapsto G(r) - G'(r)r \in \mathbb{R}$$

continuous and bounded.

Condition (c) of the above definition means that we consider the renormalized weak solutions of the stationary problem; see [11] for a discussion. Such a definition simplifies further analysis without any loss of generality.

Remark 1.1. Denote by $\Sigma^- = \partial B \setminus \Sigma^+$ the *exit* part of the boundary ∂B. It follows from the definition of generalized solutions that the extensions of the density and of the velocity vector field onto $\mathbb{R}^3$, given by the equalities

$$\begin{aligned} \rho(x) = 0 \ \text{ in } \ S, \qquad \rho(x) = \rho^\infty \ \text{ in } \ \mathbb{R}^3 \setminus (B \cup \Sigma^-), \\ \mathbf{u} = 0 \ \text{ in } \ S, \qquad \mathbf{u} = \mathbf{U}^\infty \ \text{ in } \ \mathbb{R}^3 \setminus B \end{aligned} \tag{1.8}$$

satisfy the integral identity

$$\int_{\mathbb{R}^3} \Big(G(\rho)\mathbf{u} \cdot \nabla\psi + \big(G(\rho) - G'(\rho)\rho\big)\psi \operatorname{div} \mathbf{u} \Big)\, dx = 0 \tag{1.9}$$

for any functions $\psi \in C_0^1(\mathbb{R}^3)$ vanishing near Σ^-.

1.4. Preliminaries and main results. The cost functional for shape optimization problems is the drag $\mathbf{J}(\rho, \mathbf{u}, \Omega)$ defined by formula (1.5). In applications, the drag is usually minimized within the class of admissible shapes. To the best of our knowledge there are no results on the shape optimization problem in the framework of generalized solutions to stationary problems for the adiabatic constant $\gamma > 1$; the case of evolution equations for the adiabatic constant $\gamma > 3/2$ is considered in [10].

The drag depends on the solution $(\rho, \mathbf{u})$ to problem (1.3); however, such a solution, if it does exist, is not in general unique. We point out that the existence of solutions for the adiabatic constant $\gamma \geq 1$ in three spatial dimensions is in general an open and difficult problem [19]. The case of discretized problems is considered in [25] for $\gamma \geq 1$; however, no dependence of solutions on geometrical domains is considered in [25]. On the other hand, the case of drag minimization in two spatial dimensions is studied in [24]. Furthermore, the drag depends on an admissible shape of the obstacle S. The dependence of the drag on the admissible shapes is twofold: first, it depends directly on Ω since the integrals in (1.5) are defined over Ω, and second, it depends on the generalized solutions to the NSE defined in Ω. The restrictions on the shapes of admissible obstacles S are defined in such a way that the set of admissible shapes and the set of the associated generalized solutions are compact. The precise conditions for admissible shapes are established below in the form of condition $(\mathcal{H}_\Omega)$. In the present paper we do not derive the necessary optimality conditions for the problem of drag minimization; we prove only the compactness of the set of solutions over the set of admissible shapes, which leads to the existence of an optimal shape, provided the set of generalized solutions is nonempty. The optimality conditions for the drag minimization problem will be given in forthcoming paper [27].

We are now in a position to formulate the main result of the paper.

Suppose that a sequence of flow domains $\Omega_n = B \setminus S_n$ satisfies the following three conditions, which we refer to collectively as condition $(\mathcal{H}_\Omega)$.

Condition $(\mathcal{H}_\Omega)$.

(a) There is a compact $K_S \Subset B$ such that $\cup_n S_n \subset K_S$.

(b) If a compact set $K \Subset \Omega$, then $K \Subset \Omega_n$ for all large n.

(c) If $w_n \to w$ weakly in $H^{1,2}(B)$ and $w_n \in H_0^{1,2}(\Omega_n)$, then $w \in H_0^{1,2}(\Omega)$.

In this case we will write

$$\Omega_n \overset{\mathcal{H}}{\to} \Omega.$$

It is easily seen that the convergence of the sequence Ω_n both in the Hausdorff metrics and in the sense of Kuratowski and Mosco implies all three conditions listed in $(\mathcal{H}_\Omega)$.

We deal with the complicated system of nonlinear PDEs. Therefore, we need stronger conditions on the convergence of geometrical domains, compared to the classical Kuratowski–Mosco convergence [14], which is adapted to the linear elliptic case. In particular, our conditions are more restrictive when compared to the Kuratowski–Mosco convergence of domains Ω_n.

The first main result of the paper is the following compactness theorem.

THEOREM 1.2. *Assume that there are a sequence (Ω_n) of domains which converges $\Omega_n \overset{\mathcal{H}}{\to} \Omega$ and a sequence of generalized solutions $\{\rho_n, \mathbf{u}_n\}$ to the compressible NSE*

$$\operatorname{div}(\rho_n \mathbf{u}_n \otimes \mathbf{u}_n) + \nabla p(\rho_n) - \rho \mathbf{f} = \operatorname{div} \Pi, \quad \operatorname{div}(\rho_n \mathbf{u}_n) = 0 \ \text{ in } \ \Omega_n, \tag{1.10a}$$

satisfying the boundary conditions

$$\mathbf{u}_n = 0 \ \text{ on } \ \partial S_n, \quad \mathbf{u}_n = \mathbf{U}^\infty \ \text{ on } \ \partial B, \quad \rho_n = \rho^\infty \ \text{ on } \ \Sigma^+. \tag{1.10b}$$

Suppose also that the total energies of the sequence $(\rho_n, \mathbf{u}_n)$ of generalized solutions to problem (1.10) *are uniformly bounded by a constant c_Ω,*

$$\mathbf{E}_n = \int_{\Omega_n} \left(\frac{1}{2} \rho |\mathbf{u}_n|^2 + \frac{1}{\gamma - 1} \rho_n^\gamma \right) dx \le c_\Omega. \tag{1.11}$$

Then there is a subsequence of the sequence $\{\rho_n, \mathbf{u}_n\}$, still denoted by $\{\rho_n, \mathbf{u}_n\}$, such that for any $r < \gamma$,

$$\rho_n \to \rho \ \text{ in } \ L^r(B), \quad p(\rho_n) \to p(\rho) \ \text{ in } \ L^1_{loc}(\Omega), \quad \mathbf{u}_n \to \mathbf{u} \ \text{ weakly in } \ H^{1,2}(B).$$

The pair of functions $(\rho, \mathbf{u})$ serves as a generalized solution to problem (1.3) *for the limit domain Ω; furthermore, the shape functionals converge for $n \to \infty$,*

$$\mathbf{J}(\rho_n, \mathbf{u}_n, \Omega_n) \to \mathbf{J}(\rho, \mathbf{u}, \Omega). \tag{1.12}$$

1.5. Shape optimization. We provide an existence result for the drag minimization. We also list, for the convenience of the reader, the related recent results on the compactness of the solutions to compressible NSE.

In order to compare our results with the recent results given in the literature on the subject, we note that for fixed S and $\gamma > 5/4$, compactness results were obtained in [12]. In the general case, the compactness of solutions to boundary value problems for NSE with $\gamma > 3/2$ was proved in [10] under the assumption that ∂S_n satisfy the "uniform thickness" condition.

Theorem 1.2 leads to the following result on the solvability of the shape optimization problem, which is the second main result of the paper. In order to formulate the result we introduce some notation. Choose an arbitrary continuous function $h : \mathbb{R}^3 \mapsto \mathbb{R}^+$ which is positive with a possible exception of a null capacity set. Choose also a positive constant V and a compact $K \Subset B$ with meas $K \ge V$. Denote by $\mathfrak{O}(h, K, V)$ the family of all compact obstacles $S \subset K$ with meas $K \ge V$ satisfying the following conditions.

Condition S. For each $x \in S$, there are $a^\pm(x) \in \mathbb{R}^3$ so that S contains the plane triangle $x + \Delta(x)$ with $\Delta(x) = \{ sa^- + ta^+ : s, t \ge 0, \, s + t \le 1 \}$ and

$$\min \{ |a^\pm(x)|, \ |a^+(x) \times a^-(x)| \} \ge h(x). \tag{1.13}$$

In particular, the angle $\vartheta(x) = \arcsin\big(|a^+(x) \times a^-(x)||a^-(x)|^{-1}|a^+(x)|^{-1}\big)$ satisfies the inequalities $0 < \vartheta(x) < \pi$ quasi everywhere (q.e.) in S. This condition is fulfilled, for example, if an obstacle admits a finite partition, such that every component of the partition is starlike with respect to two distinct points.

For positive C and compact $S \Subset B$, denote by $\mathcal{E}(C,S)$ the family of generalized solutions to problem (1.3) in domain $\Omega = B \setminus S$ with the total energies bounded from above by a constant, $\mathbf{E} \le C$. The following theorem gives sufficient conditions for the solvability of the drag minimization problem in the class $\mathfrak{O}(h,K,V)$.

THEOREM 1.3. *Suppose that there exist a compact $S \in \mathfrak{O}(h,K,V)$ and a generalized solution $\{\rho,\mathbf{u}\} \in \mathcal{E}(C,S)$ such that $\mathbf{J}(\rho,\mathbf{u},B\setminus S) < \infty$. Then there are $S^* \in \mathfrak{O}(h,K,V)$ and $\{\rho^*,\mathbf{u}^*\} \in \mathcal{E}(C,S^*)$ such that*

$$\mathbf{J}(\rho^*,\mathbf{u}^*,B\setminus S^*) = \inf_{S\in\mathfrak{O}(h,K,V)} \inf_{\{\rho,\mathbf{u}\}\in\mathcal{E}(C,S)} \mathbf{J}(\rho,\mathbf{u},B\setminus S). \tag{1.14}$$

Proof. Choose a minimizing sequences S_n, $(\rho_n,\mathbf{u}_n)$ so that

$$\mathbf{J}(\rho_n,\mathbf{u}_n,B\setminus S_n) \to \inf_{S\in\mathfrak{O}(h,K,V)} \inf_{\{\rho,\mathbf{u}\}\in\mathcal{E}(C,S)} \mathbf{J}(\rho,\mathbf{u},B\setminus S) \ \text{ as } \ n\to\infty.$$

Without any loss of generality, we can assume that the sequence S_n converges in the Hausdorff metric to some compact set $S^* \in \mathfrak{O}(h,K,V)$. By Theorem 1.2, it suffices to prove that

$$\Omega_n = B\setminus S_n \xrightarrow{\ \mathcal{H}\ } B\setminus S^* = \Omega^*.$$

Note that conditions (a) and (b) in $(\mathcal{H}_\Omega)$ follow from the definition of the class $\mathfrak{O}(h,K,V)$ and from the convergence of the sequence S_n in the Hausdorff metric. In order to prove (c) in $(\mathcal{H}_\Omega)$, let us consider an arbitrary sequence $w_n \in H^{1,2}_0(B)$ such that $w_n \to w$ weakly in $H^{1,2}(B)$ and $w_n \in H^{1,2}_0(\Omega_n)$. Without any loss of generality, we can assume that w_n vanishes on $\mathbb{R}^3\setminus\Omega_n$ and w is quasi continuous in $\mathbb{R}^3$. It is necessary to prove that $w \in H^{1,2}_0(\Omega)$. By the Hedberg approximation theorem (see [13]), it is sufficient to show that $w = 0$ q.e. on S^*.

To this end, we choose an arbitrary $x^* \in S^*$ with $h(x^*) > 0$ and note that $x^* = \lim_{n\to\infty} x_n \in S_n$. After passing to a subsequence we can assume that the sequence $a^\pm(x_n)$ converges to vectors $a^\pm(x^*)$ satisfying (1.13). Denote by $\mathcal{O}_n : \mathbb{R}^3 \to \mathbb{R}^3$ the orthogonal linear mapping, which maps $span\{a^\pm(x_n)\}$ onto the coordinate plane $\Pi = \{x\cdot\mathbf{e}_3 = 0\}$, and set

$$W_n(y) = w_n(x_n + \mathcal{O}_n^{-1}y), \quad b^\pm = \mathcal{O}a^\pm(x_n).$$

Since $\mathcal{O}_n \to \mathcal{O}^*$, the functions W_n converge weakly in $H^{1,2}(\mathbb{R}^3)$ and strongly in $L^2(\Pi)$ to $W(y) = w(x^* + \mathcal{O}^{*-1}y)$. Noting that W_n vanishes on the triangle $\mathcal{O}_n\Delta(x_n) \subset \Pi$ we can conclude from this that $W = 0$ on $\mathcal{O}^*\Delta(x^*)$, and hence $w = 0$ on $x^* + \Delta(x^*)$. Since the capacity of the plane disk of radius r is equal to $2r/\pi^2$, the inequality $\operatorname{cap}\,\Delta(x^*)\cap B(x^*,r) \ge r\vartheta(x^*)/\pi^3$ holds true for all $r < h(x^*)$, and hence

$$\liminf_{r\to 0} \frac{\operatorname{cap}\,\Delta(x^*)\cap B(0,r)}{\operatorname{cap}\,B(0,r)} \ge c(x^*) > 0.$$

From this and the above quasi-continuity property of w we can deduce $w(x^*) = 0$. Therefore, $w = 0$ q.e. in S and the proof of the theorem is completed. ∎

Remark 1.2. The minimal value of the drag in the right-hand side of (1.14) depends on V and a choice of a compact K and energy bound C. It increases in V and decreases in C and K.

Remark 1.3. A typical obstacle satisfying the conditions of Theorem 1.3 is a union of starlike bodies endowed with a system of plane wings. The case of curvilinear wings also can be included if we replace $\Delta(x)$ with $g(\Delta(x))$, where $g : \mathbb{R}^3 \to \mathbb{R}^3$ is an arbitrary bi-Lipschitz mapping with $g(0) = 0$.

To the best of our knowledge, there are no global results on the solvability of spatial stationary problems for compressible NSE with nonhomogeneous boundary conditions even for large γ. In this connection we note that there is a significant difference between the question on existence of solutions to boundary value problem (1.3) and the question on solvability of shape optimization problem (1.14). Thus, due to Remark 1.2, the conditions of Theorem 1.3 will be satisfied if there exists *at least one* obstacle S for which problem (1.3) has a finite energy solution. Such a solution can be constructed in the case when both B and S have a spherical symmetry or in the case when ∂S is a smooth surface sufficiently close to ∂B. This condition is also satisfied for sufficiently small data, since in such a case the existence of local solutions is proved at least for the homogeneous boundary conditions. We refer also to [15], [16], [17], [21] for the local existence results in two spatial dimensions with nonhomogeneous boundary conditions.

The remaining part of the paper is devoted to the proof of Theorem 1.2. We begin with the derivation of certain local a priori estimates for the generalized solutions to problem (1.3).

2. A priori estimates. In this section we prove the following theorem on local integrability of generalized solutions; cf. [12], [25], [26].

THEOREM 2.1. *Let $(\rho, \mathbf{u})$ be a generalized solution to problem* (1.3)*, and let Ω' be a subdomain of Ω with dist $(\Omega', \partial\Omega) > d > 0$. Then for $\kappa = 2(\gamma - 1)/(\gamma + 2) > 0$,*

$$\|\rho\mathbf{u}^2\|_{L^{1+\kappa}(\Omega')} \le cd^{-1}, \quad \|\rho\|_{L^{\gamma(1+\kappa)}(\Omega')} \le cN(\Omega'), \tag{2.1}$$

where the constant N depends only on Ω'.

The proof of Theorem 2.1 is divided into a sequence of lemmas. The first lemma gives the estimate of the rate of energy dissipation in terms of the total energy of the fluid.

LEMMA 2.2. *Under the assumptions of Theorem* 2.1*, the velocity vector field is bounded,*

$$\|\mathbf{u}\|_{H^{1,2}(\Omega)} \le c, \tag{2.2}$$

where the constant c depends only on the data of the boundary value problem.

Proof. The proof follows from condition (a) in the definition of generalized solutions to problem (1.3) and from the obvious inequalities

$$\|\mathbf{u}\|_{H^{1,2}(\Omega)} \le c\|\nabla(\mathbf{u} - \mathbf{u}^\infty)\|_{L^2(\Omega)} + \|\mathbf{u}^\infty\|_{C^1(\Omega)},$$

$$\int_\Omega \rho|\mathbf{f} \cdot (\mathbf{u} - \mathbf{u}^\infty)|\, dx \le c(\gamma, B)\|\mathbf{f}\|_{C(\Omega)}(\mathbf{E} + 1)(\|\mathbf{u}^\infty\|_{C^1(\Omega)} + 1). \qquad \square$$

Choose a domain Ω_0 with a smooth boundary so that

$$\Omega' \Subset \Omega_0 \Subset \Omega, \quad \text{dist } (\partial\Omega, \Omega_0) \ge d/3, \quad \text{dist } (\partial\Omega_0, \Omega')) \ge d/3\,.$$

The second lemma shows that the Newtonian potential of the pressure is uniformly bounded on Ω_0.

LEMMA 2.3. *Under the assumptions of Theorem* 2.1,

$$\operatorname*{ess\,sup}_{x\in\Omega_0}\int_\Omega \frac{\rho^\gamma(y)}{|x-y|}dy \le cd^{-1}\,. \tag{2.3}$$

Proof. Introduce the symmetric, nonnegative, matrix-valued function

$$\mathcal{E}=(\mathcal{E}_{i,j})_{3\times 3},\quad \mathcal{E}_{i,j}=\rho u_iu_j+p(\rho)\delta_{i,j},\quad \operatorname{Tr}\mathcal{E}=\rho|\mathbf{u}|^2+3p(\rho),$$

and rewrite integral identity (1.7a) in the form

$$\int_\Omega(\mathcal{E}-\Pi):\nabla\varphi=\int_\Omega\rho\mathbf{f}\cdot\varphi\,dy\quad\text{ for all }\ \varphi\in H_0^{1,2}(\Omega)^3\,. \tag{2.4}$$

Fix an arbitrary $x\in\Omega_0$ and set

$$\mathcal{M}_0(s)=\int_{B(x,s)}\mathbf{n}\otimes\mathbf{n}:\mathcal{E}(y)\,dy,\quad \mathcal{M}_1(s)=\int_{B(x,s)}\operatorname{Tr}\mathcal{E}(y)\,dy,$$
$$\text{where }\ \mathbf{n}=|y-x|^{-1}(y-x)\,.$$

Next choose positive $r<R\le d/3$, $\delta\in[0,R-r)$, $s\in[r,R-\delta)$, and a sequence of continuous functions $h_\delta(t)$ with $h_\delta(t)=1$ for $t\le s$, $h_\delta(s)=0$ for $t\ge s+\delta$, and $h_k'(t)=-1/\delta$ for $t\in(s,s+\delta)$. Substituting the vector function $\varphi(y)=h_\delta(|y-x|)(y-x)$ into (2.4) and taking into account that the functions $\mathcal{M}_j(s), j=0,1$, are absolutely continuous on the interval $[r,R]$, followed by the limit passage $\delta\to 0$, leads to

$$\mathcal{M}_1(s)-s\mathcal{M}_0'(s)=\int_{B(x,s)}\operatorname{Tr}\Pi\,dy-s\int_{\partial B(x,s)}\mathbf{n}\otimes\mathbf{n}:\Pi\,dS-\int_{B(x,s)}\rho(y-x)\cdot\mathbf{f}(y)dy$$

for almost everywhere (a.e.) $s\in(0,R]$. Multiplying both sides of this equality by s^{-2} and integrating the result over the interval $(r,R]$, we find that

$$\int_{(r,R]}\left(\frac{\mathcal{M}_1(s)}{s^2}-\frac{\mathcal{M}_0'(s)}{s}\right)ds=\frac{1}{r}\int_{B(x,r)}\operatorname{Tr}\Pi\,dy$$
$$-\frac{1}{R}\int_{B(x,R)}\operatorname{Tr}\Pi\,dy+\mathfrak{S}\Pi(x,r,R)-\int_{B(y,R)}k_{r,R}(y-x)\mathbf{f}(y)\rho(y)\,dy \tag{2.5}$$

with

$$k_{r,R}(x)=\min\{r^{-1},|x|^{-1}\}-R^{-1},$$
$$\mathfrak{S}\Pi(x,r,R)=\int_{r<|y|<R}\frac{1}{|x-y|}(\mathbf{I}-\mathbf{n}\otimes\mathbf{n}):\Pi(y)\,dy.$$

It is easily seen that

$$\int_{(r,R]}\Big(s^{-2}\mathcal{M}_1(s)-s^{-1}\mathcal{M}_0'(s)\Big)\,ds$$
$$=\int_{(r,R]}\Big(s^{-2}\mathcal{M}_1(s)-s^{-1}\mathcal{M}_1'(s)\Big)\,ds+\int_{(r,R]}s^{-1}d(\mathcal{M}_1(s)-\mathcal{M}_0(s)). \tag{2.6}$$

Integration by parts in the integrals on the right-hand side of (2.6) gives

(2.7)
$$\int_{(r,R]}\Big(s^{-2}\mathcal{M}_1(s)-s^{-1}\mathcal{M}'_1(s)\Big)\,ds=\frac{1}{r}\int_{B(y,r)}\operatorname{Tr}\mathcal{E}(y)\,dy-\frac{1}{R}\int_{B(y,R)}\operatorname{Tr}\mathcal{E}(y)\,dy,$$
$$\int_{(r,R]}s^{-1}d(\mathcal{M}_1(s)-\mathcal{M}_0(s))=\int_{r\le|x-y|\le R}\frac{1}{|x-y|}(\operatorname{Tr}\mathcal{E}(y)-\mathbf{n}\otimes\mathbf{n}:\mathcal{E}(y))\,dy.$$

Combining (2.5)–(2.7), we obtain

$$\frac{\mathcal{M}_1(x,r)}{r}+\int_{r\le|x-y|\le R}\frac{1}{|x-y|}(\operatorname{Tr}\mathcal{E}(y)-\mathbf{n}\otimes\mathbf{n}:\mathcal{E}(y))\,dy=\frac{\mathcal{M}_1(x,R)}{R}$$

(2.8)
$$-\frac{1}{R}\int_{B(x,R)}\operatorname{Tr}\,\Pi dy+\frac{1}{r}\int_{B(x,r)}\operatorname{Tr}\,\Pi dy+\mathfrak{S}\Pi(x,r,R)-\int_{B(y,R)}k_{r,R}(y-x)f(y)\rho(y)dy.$$

It follows from Lemma 2.2 that

$$\frac{1}{R}\int_{B(x,R)}|\operatorname{Tr}\,\Pi|\,dy+\frac{1}{r}\int_{B(x,r)}|\operatorname{Tr}\,\Pi|\,dy+|\mathfrak{S}\Pi(x,r,R)|\le c\int_{B(x,R)}\frac{|\Pi|}{|x-y|}\,dy$$
$$\le\Big(\int_\Omega\frac{dy}{|x-y|^2}\Big)^{1/2}\Big(\int_\Omega|\Pi|^2\,dy\Big)^{1/2}\le c\|\nabla\mathbf{u}\|_{L^2(\Omega)}\le c\,.$$

Passage to the limit $r\to 0$ in (2.8), in view of $\operatorname{Tr}\mathcal{E}(y)-\mathbf{n}\otimes\mathbf{n}:\mathcal{E}(y))\ge p(\rho)$, results in the inequality

$$\int_{B(x,R)}\frac{p\,dy}{|x-y|}\le\frac{c}{R}\int_{B(x,R)}(\rho|\mathbf{u}|^2+p)\,dy+c\|\nabla\mathbf{u}\|_{L^2(\Omega)}+c\|\mathbf{f}\|_{C(\Omega)}\le\frac{c}{R}\,,\tag{2.9}$$

which holds a.e. in Ω_0. It is sufficient to note that for $R=d/3$,

$$\int_\Omega\frac{p(\rho)}{|x-y|}\,dy\le\int_{B(x,d/3)}\frac{p(\rho)}{|x-y|}\,dy+d^{-1}\int_{\Omega\setminus B(x,d/3)}p(\rho)\,dy\le\int_{B(x,R)}\frac{p(\rho)}{|x-y|}\,dy+cd^{-1},$$

which along with (2.9) implies (2.3). The proof of Lemma 2.3 is completed. □

LEMMA 2.4. *Under the assumptions of Theorem* 2.1,

$$\int_{\Omega_0}\rho^\gamma|\mathbf{u}-\mathbf{u}^\infty|^2dx\le cd^{-1}.\tag{2.10}$$

Proof. Let us consider the sequence of functions $\mu_N:\mathbb{R}^3\mapsto\mathbb{R}^+$, $N\ge 2$, defined by

$$\mu_N(x)=\min\{\rho(x)^\gamma,N\}\ \text{ for }\ x\in\Omega_0,\qquad\mu_N(x)=0\ \text{ otherwise.}$$

Since spt μ_N, spt $(\mathbf{u}-\mathbf{u}^\infty)\subset$ cl B, the Adams embedding theorem (see [1], [2]) implies the inequality

$$\int_{\Omega_0}\mu_N|\mathbf{u}-\mathbf{u}^\infty|^2dx\le c\|\mathbf{u}-\mathbf{u}^\infty\|^2_{H^{1,2}(\Omega)}\|(-\Delta^{-1/2})\mu_N\|_{L^2(\mathbb{R}^3)},\tag{2.11}$$

in which the constant c depends only on B. Noting that $(-\Delta^{-1/2})\mu_N = c|x|^{-2} * \mu_N$, we obtain

$$\|(-\Delta^{-1/2})\mu_N\|^2_{L^2(\mathbb{R}^3)} = c\int_{\mathbb{R}^3}\left(\int_{\mathbb{R}^3}\frac{\mu_N(y)}{|x-y|^2}\,dy\right)^2 dx$$
$$= \int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\mu_N(y)\mu_N(z)}{|x-y|^2|x-z|^2}\,dxdydz.$$

Since

$$\int_{\mathbb{R}^3}\frac{dx}{|x-y|^2|x-z|^2} = c|y-z|^{-1},$$

we conclude from this that

$$\|(-\Delta^{-1/2})\mu_N\|^2_{L^2(\mathbb{R}^3)} = c\int_{\Omega_0}\mu_N(y)\left(\int_{\Omega_0}\frac{\mu_N(z)}{|y-z|}\,dz\right)dy$$
$$\le c\int_{\Omega}\rho^\gamma(y)dy\ \operatorname*{ess\,sup}_{x\in\Omega_0}\int_{\Omega_0}\frac{\mu_N(y)}{|y-x|}\,dy \le c\operatorname*{ess\,sup}_{x\in\Omega_0}\int_{\Omega}\frac{\rho^\gamma(y)}{|y-x|}\,dy.$$

Passage to the limit $N\to\infty$ in (2.11) results in the inequality

$$\int_{\Omega_0}\rho^\gamma|\mathbf{u}-\mathbf{u}^\infty|^2dx \le c\|\mathbf{u}-\mathbf{u}^\infty\|^2_{H_0^{1,2}(B)}\left\{\operatorname*{ess\,sup}_{x\in\Omega_0}\int_{\Omega}\frac{\rho^\gamma(y)}{|y-x|}\,dy\right\}^{1/2}.$$

Combining this result with (2.3), we obtain (2.10). The proof of Lemma 2.4 is completed. □

Proof. We are now in position to complete the proof of Theorem 2.1. Since $\mathbf{u}-\mathbf{u}^\infty \in H_0^{1,2}(B)$, Lemmas 2.4 and 2.2 imply

$$\int_{\Omega_0}\rho^\gamma|\mathbf{u}|^2\,dx \le 2\|\mathbf{u}^\infty\|^2_{C(B)}\mathbf{E} + 2\int_{\Omega_0}\rho^\gamma|\mathbf{u}-\mathbf{u}^\infty|^2dx \le cd^{-1}. \tag{2.12}$$

Applying the Hölder inequality, we obtain

$$\int_{\Omega_0}(\rho|\mathbf{u}|^2)^{1+\kappa}\,dx = \int_{\Omega_0}(\rho|\mathbf{u}|^{2/\gamma})^{1+\kappa}(|\mathbf{u}|^{2(\gamma-1)/\gamma})^{1+\kappa}\,dx$$
$$\le \left(\int_{\Omega_0}\rho^\gamma|\mathbf{u}|^2\,dx\right)^{(1+\kappa)/\gamma}\left(\int_{\Omega_0}|\mathbf{u}|^6\,dx\right)^{(\gamma-1-\kappa)/\gamma}$$

with $\kappa = 2(\gamma-1)/(\gamma+2) > 0$. From this, (2.12), and the boundedness of the embedding $H^{1,2}(B)\hookrightarrow L^6(B)$, we derive the inequality

$$\int_{\Omega_0}(\rho|\mathbf{u}|^2)^{1+\kappa}\,dx \le c\|\mathbf{u}\|^{6(\gamma-1-\kappa)/\gamma}_{H^{1,2}(\Omega)}d^{-(1+\kappa)/\gamma} \le cd^{-(1+\kappa)/\gamma}, \tag{2.13}$$

which along with Lemma 2.2 implies the first estimate from (2.1). In order to prove the second, we choose a function $\eta\in C_0^\infty(\mathbb{R}^3)$ such that

$$0\le\eta\le 1,\quad \eta(x)=1\ \text{ in }\ \Omega',\ \ \operatorname{spt}\eta\Subset\Omega_0,\quad |\nabla\eta|\le cd^{-1}.$$

From the moment equation, it follows that

$$\nabla(\eta p(\rho)) = \operatorname{div}\Big[\eta(\Pi - \rho\mathbf{u}\otimes\mathbf{u})\Big] + \Phi \in \mathcal{D}'(\Omega), \tag{2.14}$$

where $\Phi = \nabla\eta : (\rho\mathbf{u}\otimes\mathbf{u} - \Pi) + \eta\rho\mathbf{f}$. It is easily to see that $\|\Phi\|_{L^1(\Omega)} \le cd^{-1}$. Now choose $s = (1+\kappa)/\kappa > 3$. Recall (see [11]) that for a given function $q \in C_0^\infty(\Omega)$ with

$$[q] := \text{ meas } (\Omega_0)^{-1}\int_{\Omega_0} q\,dx = 0,$$

there exists the vector field $\varphi \in C^\infty(\Omega_0)^3$ such that

$$\operatorname{div}\varphi = q \ \text{ in } \ \Omega_0, \quad \varphi = 0 \ \text{ at } \ \partial\Omega_0, \quad \|\varphi\|_{H^{1,s}(\Omega_0)} \le N(\Omega_0)\|q\|_{L^s(\Omega_0)}. \tag{2.15}$$

Thus, we get

$$\int_{\Omega_0} q\eta p(\rho)\,dx = \int_{\Omega_0}\eta(\rho\mathbf{u}\otimes\mathbf{u} - \Pi) : \nabla\varphi\,dx + \int_{\Omega_0}\eta\varphi\Phi\,dx.$$

Since the embedding $H_0^{1,s}(\Omega_0) \hookrightarrow C(\Omega_0)$ is bounded, we have $|\varphi| \le c\|\varphi\|_{H^{1,s}(\Omega_0)} \le N(\Omega_0)\|q\|_{L^s(\Omega_0)}$, which yields

$$\begin{aligned}\left|\int_{\Omega_0} q\eta p(\rho)\,dx\right| &\le N(\Omega_0)(\|\rho|\mathbf{u}|^2\|_{L^{1+\kappa}(\Omega_0)} + \|\Pi\|_{L^{1+\kappa}(\Omega_0)})\|\varphi\|_{H^{1,s}(\Omega_0)}\\ &\quad + \|\Phi\|_{L^1(\Omega)}\|\varphi\|_{C(\Omega)} \le cN(\Omega_0)d^{-1}\|q\|_{L^s(\Omega_0)}.\end{aligned}$$

It follows from this that

$$\|\eta p(\rho) - [\eta p(\rho)]\|_{L^{s/(s-1)}(\Omega_0)} \le cN(\Omega_0).$$

Note that $s/(s-1) = 1+\kappa$, $\eta = 1$ on Ω', and $[\eta p(\rho)] \le c$. Hence $\|p\|_{L^{1+\kappa}(\Omega')} \le cN(\Omega')d^{-1}$, and the proof of Theorem 2.1 is completed. □

3. Weak convergence. Since the notion of weak limits plays a crucial role in our analysis, we begin with a short description of some basic facts concerning weak convergence and weak compactness. We refer the reader to, e.g., [5] for proofs of basic results.

Let A be an arbitrary bounded, measurable subset of $\mathbb{R}^3$ and $1 < r \le \infty$. Then for every bounded sequence $\{g_n\}_{n\ge1} \subset L^r(A)$, there exist a subsequence, still denoted by $\{g_n\}$, and a function $g \in L^r(A)$ such that for $n \to \infty$,

$$\int_A g_n(x)h(x)dx \to \int_A g(x)h(x)dx \ \text{ for all } \ h \in L^{r/(r-1)}(A).$$

We say the sequence converges $g_n \to g$ weakly in $L^r(A)$ for $r < \infty$ and converges star-weakly in $L^\infty(A)$ in the limit case of $r = \infty$. In the very special case of $r = 1$, it is known that the sequence of g_n contains a weakly convergent subsequence in $L^1(A)$ if and only if there is a continuous function $\Phi : \mathbb{R} \to \mathbb{R}^+$ such that

$$\lim_{s\to\infty}\Phi(s)/s = \infty \ \text{ and } \ \sup_{n\ge1}\|\Phi(g_n)\|_{L^1(A)} < \infty.$$

If the sequence of g_n is only bounded in $L^1(A)$ and A is open, then after passing to a subsequence we can assume that g_n converges star-weakly to a bounded Radon measure measure μ_g, i.e.,

$$\lim_{n\to\infty}\int_A g_n(x)h(x)dx=\int_A h(x)d\mu_g(x)\quad\text{for all compactly supported}\quad h\in C(A).$$

In what follows, the linear space of compactly supported functions on a set A is denoted by $C_0(A)$, and its dual is denoted by $C_0(A)^*$.

Ball's version [3] (see also [20], [23]) of the fundamental Tartar theorem on Young measures gives a simple and effective representation of weak limits in the form of integrals over families of probability measures. The following lemma is a consequence of Ball's theorem.

LEMMA 3.1. *Suppose that a sequence $\{g_n\}_{n\ge1}$ is bounded in $L^r(A)$, $1<r\le\infty$, where A is an open, bounded subset of $\mathbb{R}^3$. Then we have the following characterizations of weak limits.*

(i) *There exists a subsequence, still denoted by $\{g_n\}_{n\ge1}$, and a family of probability measures $\sigma_x\in C_0(\mathbb{R})^*$, $x\in A$, with a measurable distribution function $\Gamma(x,\lambda):=\sigma_x(-\infty,\lambda]$ so that the function $\lambda\mapsto\Gamma(x,\lambda)$ is monotone and continuous from the right and admits the limits $1,0$ for $\lambda\to\pm\infty$. Furthermore, for any continuous function $G:A\times\mathbb{R}\mapsto\mathbb{R}$ such that*

$$\lim_{|\lambda|\to\infty}\|G(\cdot,\lambda)\|_{C(A)}/|\lambda|^r=0\ \ \textit{for}\ \ r<\infty\ \ \textit{and}\ \sup_{|\lambda|}\|G(\cdot,\lambda)\|_{C(A)}<\infty\ \ \textit{for}\ \ r=\infty,$$

the sequence of $G(\cdot,g_n)$ converges weakly in $L^1(A)$ to a function

$$\overline{G}(x)=\int_{\mathbb{R}}G(x,\lambda)d_\lambda\Gamma(x,\lambda).\tag{3.1}$$

Moreover, the function

$$A\ni x\to\int_{\mathbb{R}}|\lambda|^r\,d_\lambda\Gamma(x,\lambda)\in\mathbb{R}$$

belongs to $L^1(A)$.

(ii) *If $G(x,\cdot)$ is convex and the sequence g_n converges weakly (star-weakly for $r=\infty$) to $g\in L^r(A)$, then $\overline{G}(x)\le G(x,g(x))$. If the functions g_n satisfy the inequalities $g_n\le M$ (resp., $g_n\ge m$), then $\Gamma(x,\lambda)=1$ for $\lambda\ge M$ (resp., $\Gamma(x,\lambda)=0$ for $\lambda<m$).*

(iii) *If $\Gamma(1-\Gamma)=0$ a.e. in A, then the sequence g_n converges to g in measure, and hence in $L^s(A)$, for positive $s<r$. Moreover, in this case $\Gamma(x,\lambda)=0$ for $\lambda<g(x)$ and $\Gamma(x,\lambda)=1$ for $\lambda\ge g(x)$*

Let us consider the sequence of generalized solutions $\{(\rho_n,\mathbf{u}_n)\}_{n\ge1}$ to problem (1.3) satisfying all assumptions of Theorem 1.2. Assume that the functions $(\rho_n,\mathbf{u}_n)$ are extended onto $\mathbb{R}^3$ by formulae (1.8), and fix an arbitrary bounded smooth domain D with $B\Subset D$. For such extended functions, by inequalities (1.11) and formulae (1.8), it follows that the sequence $(\rho_n,\mathbf{u}_n)$ contains a subsequence, still denoted by $(\rho_n,\mathbf{u}_n)$, such that

$$\begin{gathered}\rho_n\to\rho\ \text{weakly in}\ L^\gamma(D),\quad \mathbf{u}_n\to\mathbf{u}\ \text{weakly in}\ H^{1,2}(D),\\ \rho_n\to\rho\ \text{weakly in}\ L^{(1+\kappa)\gamma}(\Omega')\ \text{for all}\ \Omega'\Subset\Omega.\end{gathered}\tag{3.2}$$

The behavior of the functions $p(\rho_n)$ is more complicated. Since they are uniformly bounded in $L^1(D)$, we can assume, after passing to a subsequence if necessary, that $p(\rho_n)$ converge weakly to some finite Borel measure μ_p on D. On the other hand, the sequence $\{p(\rho_n)\}$ is bounded in $L^{1+\kappa}(K)$ for any compact $K \Subset \Omega$. Using the diagonal process, we obtain the existence of a subsequence which converges weakly on each compact $K \Subset \Omega$ to some function $\overline{p} \in L^{1+\kappa}_{loc}(\Omega)$. Since

$$\|\overline{p}\|_{L^1(K)} \le \liminf_{n\to\infty} \|p(\rho_n)\|_{L^1(K)} \le c$$

the function $\overline{p}$ is integrable over Ω. Next, we note that the functions $p(\rho_n) = (\rho^\infty)^\gamma$ are bounded and independent of n on $D\setminus B$. Hence the extended function $\overline{p}(x), x \in D$, defined as

$$\overline{p}(x) = \overline{p}(x) \ \text{ in } \ \Omega, \quad \overline{p}(x) = (\rho^\infty)^\gamma \ \text{ in } \ D \setminus B,$$

belongs to the class $L^1(D \setminus S) \cap L^{1+\kappa}_{loc}(\Omega) \cap L^\infty(D \setminus B)$, which implies the equality

$$\int_D h(x) d\mu_p = \int_{D\setminus S} h(x)\overline{p}(x)dx + \int_{\partial B} h(x) d\mu_p + \int_S h(x) d\mu_p \quad \text{for all} \quad h \in C_0(\mathbb{R}^3),$$

where the compact *obstacle* takes the form $S = B \setminus \Omega$. Applying Lemma 3.2 to the sequence of $g_n := \rho_n$ and to the sets $A = D, \Omega'$ leads to the following result on the representation of weak limits of a sequence of generalized solutions to problem (1.3).

LEMMA 3.2. *There exists a subsequence of the sequence* $\{\rho_n, \mathbf{u}_n\}$, *still denoted by* $\{\rho_n, \mathbf{u}_n\}$, *and a distribution function* $\Gamma : D \times \mathbb{R} \mapsto [0,1]$ *such that*

(i) $\Gamma(x,\lambda)$ *meets all requirements of Lemma* 3.1 *and satisfies the equalities*

$$\Gamma(x,\lambda) = 0 \ \ \textit{for} \ \ \lambda < 0 \ \ \textit{a.e. in} \ \ D,$$
$$\Gamma(x,\lambda) = 0 \ \ \textit{for} \ \ \lambda < \rho^\infty, \quad \Gamma(x,\lambda) = 1 \ \ \textit{for} \ \ \lambda \ge \rho^\infty \ \ \textit{a.e. in} \ \ D \setminus B\,.$$

(ii) *for any continuous function* $G : D\times\mathbb{R}$ *such that* $\lim_{\rho\to\infty} \rho^{-\gamma}\|G(\cdot,\rho)\|_{C(D)} = 0$, *the sequence* $G(\cdot,\rho_n)$ *converges weakly in* $L^1(D)$ *to the function*

$$\overline{G}(x) = \int_{[0,\infty)} G(x,\lambda)\, d_\lambda\Gamma(x,\lambda) \ \ \textit{a.e. in} \ \ D. \tag{3.3}$$

In particular, the weak limit of ρ_n *takes the form*

$$\rho(x) = \int_{[0,\infty)} \lambda\, d_\lambda\Gamma(x,\lambda) \equiv \int_{[0,\infty)} (1-\Gamma(x,\lambda))\, d\lambda \ \ \textit{a.e. in} \ \ D. \tag{3.4}$$

(iii) *the function* $\overline{p}$ *admits the representation*

$$\overline{p}(x) = \int_{[0,\infty)} \lambda^\gamma\, d_\lambda\Gamma(x,\lambda) \equiv \gamma \int_{[0,\infty)} \lambda^{\gamma-1}(1-\Gamma(x,\lambda))\, d\lambda \ \ \textit{a.e. in} \ \ D\setminus S\,. \tag{3.5}$$

Since the embedding $H^{1,2}_0(D) \hookrightarrow L^r(D)$ is compact for $r < 6$, we can expect that $\rho_n\mathbf{u}_n$ converge weakly to $\rho\mathbf{u}$. The corresponding result is given by the following lemma.

LEMMA 3.3. *For* $\iota = (\gamma-1)/(\gamma+1) > 0$ *and for any* $\Omega' \Subset \Omega$,

$$\rho_n\mathbf{u}_n \to \rho\mathbf{u} \ \textit{converges weakly in} \ L^{1+\iota}(D)^3,$$
$$\rho_n\mathbf{u}_n \otimes \mathbf{u}_n \to \rho\mathbf{u}\otimes\mathbf{u} \ \textit{weakly in} \ L^{1+\kappa}(\Omega')^9.$$

Proof. Since ρ_n and $\mathbf{u}_n$ are bounded and independent of n in the exterior of B, it suffices to show that $\rho_n\mathbf{u}_n \to \rho\mathbf{u}$ converges weakly in $L^{1+\iota}(B)$. We have

$$\begin{aligned}|\rho_n\mathbf{u}_n|^{1+\iota} &= (\rho_n)^{\frac{1+\iota}{2}}(\rho_n|\mathbf{u}_n|)^{\frac{1+\iota}{2}}\\ &\le \left(\rho_n^{(1+\iota)/2}\right)^{1+\gamma} + \left((\rho_n|\mathbf{u}_n|^2)^{(1+\iota)/2}\right)^{(1+\gamma)/\gamma} = \rho_n^\gamma + \rho_n|\mathbf{u}_n|^2,\end{aligned}$$

which implies the estimate

$$\int_B |\rho_n\mathbf{u}_n|^{1+\iota}\,dx \le \int_B (\rho^\gamma + \rho_n|\mathbf{u}_n|^2)dx \le c\mathbf{E}_n \le c.$$

Hence, it suffices to prove that $\rho_n\mathbf{u}_n$ converges to $\rho\mathbf{u}$ in $\mathcal{D}'$, i.e., in the sense of distributions. For any positive N, set $\rho_n^N = \min\{\rho_n, N\}$. Next note that the functions ρ_n^N converge star-weakly in $L^\infty(B)$ to some function $\overline{\rho^N}$. It is easily seen that the sequence of nonnegative functions $\rho - \overline{\rho^N}$ decreases for $N \to \infty$. Moreover, since

$$\int_B |\rho_n - \rho_n^N|\,dx \le \left(\int_B \rho_n^\gamma\,dx\right)^{1/\gamma} \text{meas}\,\{x : \rho(x)^\gamma \ge N^\gamma\}^{(\gamma-1)/\gamma} \le cN^{-(\gamma-1)}, \tag{3.6}$$

we can pass to the limit for $N \to \infty$,

$$\int_B (\rho - \overline{\rho^N})dx = \lim_{n\to\infty}\int_B (\rho_n - \rho_n^N)\,dx \to 0. \tag{3.7}$$

Since $\mathbf{u}_n$ converge strongly in $L^r(B)$, for $r < 6$, the sequence $\rho_n^N\mathbf{u}_n$ converges to $\overline{\rho^N}\mathbf{u}$ in $\mathcal{D}'(B)$. It follows from this that for any $h \in C_0^\infty(B)^3$,

$$\begin{aligned}&\limsup_{n\to\infty}\left|\int_B (\rho_n\mathbf{u}_n - \rho\mathbf{u})h\,dx\right|\\ &\le \limsup_{n\to\infty}\left|\int_B (\rho_n^N - \rho_n)\mathbf{u}_n h\,dx\right| + \left|\int_B (\overline{\rho^N} - \rho)\mathbf{u}h\,dx\right|.\end{aligned} \tag{3.8}$$

By the Cauchy–Schwarz inequality we have

$$\begin{aligned}&\left|\int_B (\rho_n^N - \rho_n)\mathbf{u}_n h\,dx\right|\\ &\le \|h\|_{C(B)}N^{(\gamma-1)/2}\int_B |\rho_n^N - \rho_n|\,dx + \|h\|_{C(B)}N^{-(\gamma-1)/2}\int_B \rho_n|\mathbf{u}_n|^2\,dx.\end{aligned} \tag{3.9}$$

Combining (3.6), (3.8), and (3.9) leads to the inequality

$$\limsup_{n\to\infty}\left|\int_B (\rho_n\mathbf{u}_n - \rho\mathbf{u})h\,dx\right| \le c\|h\|_{C(B)}N^{-(\gamma-1)/2} + \|h\|_{C(B)}\int_B (\rho - \overline{\rho^N})|\mathbf{u}|\,dx. \tag{3.10}$$

It follows from (3.7) that the sequence of nonnegative integrable functions $(\rho - \overline{\rho^N})|\mathbf{u}|$ decreases and converges to 0 a.e. in B for $N \to \infty$. From this, and by the Fatou theorem, we conclude that the right-hand side of (3.10) tends to 0 as $N \to \infty$, which yields the weak convergence of the sequence $\rho_n\mathbf{u}_n$.

It remains to prove the weak convergence of the sequence $\rho_n \mathbf{u}_n \otimes \mathbf{u}_n$. We begin with the observation that since this sequence is bounded in $L^{1+\kappa}(\Omega')$, it suffices to show that it converges in $\mathcal{D}'(\Omega')$. Arguing as before, we obtain

$$\limsup_{n\to\infty} \left| \int_B (\rho_n \mathbf{u}_n \otimes \mathbf{u}_n - \rho \mathbf{u} \otimes \mathbf{u}) : h \, dx \right| \tag{3.11}$$
$$\leq \limsup_{n\to\infty} \left| \int_B (\rho_n^N - \rho_n) \mathbf{u}_n \otimes \mathbf{u}_n : h \, dx \right| + \int_B (\rho - \overline{\rho^N}) |\mathbf{u}|^2 |h| \, dx$$

for all $h \in C_0^\infty(\Omega')^9$. Using the Young inequality and taking into account that $\rho_n \geq \rho_n^N$, we obtain

$$|(\rho_n - \rho_n^N)\mathbf{u}_n \otimes \mathbf{u}_n| \leq N^{(\gamma-1)/2}(\rho_n - \rho_n^N) + N^{-\kappa(\gamma-1)/2} \rho_n |\mathbf{u}_n|^{2(1+\kappa)}$$
$$\leq N^{(\gamma-1)/2}(\rho_n - \rho_n^N) + N^{-\kappa(\gamma-1)/2} \Big(|\mathbf{u}_n|^6 + (\rho_n |\mathbf{u}_n|^2)^{(1+\kappa)} \Big).$$

Since, by Lemma 2.2, $\|\mathbf{u}_n\|_{L^6(\Omega')} \leq c \|\mathbf{u}_n\|_{H^{1,2}(B)} \leq c$, we conclude from this and inequality (3.6) that

$$\limsup_{n\to\infty} \left| \int_B (\rho_n^N - \rho_n) \mathbf{u}_n \otimes \mathbf{u}_n : h \, dx \right|$$
$$\leq c \|h\|_{C(B)} (N^{-(\gamma-1)/2} + N^{-\kappa(\gamma-1)/2}) \to 0 \quad \text{for} \quad N \to \infty.$$

It is sufficient to observe that, by (3.7) and the Fatou theorem, the last term in the right-hand side of (3.11) tends to 0 for $n \to \infty$. The proof of Lemma 3.3 is completed. □

4. The effective viscous flux. Following [18], we introduce the quantity $V(\rho, \mathbf{u}) = p(\rho) - (\xi + \nu)\operatorname{div} \mathbf{u}$, which is called the effective viscous flux. As was shown in [18], [6], [7], [8], the effective viscous flux enjoys many remarkable properties. The most important is the multiplicative relation $\overline{\varphi(\rho)V} = \overline{\varphi(\rho)}\ \overline{V}$ for weak limits, which was proved in [18] for all $\gamma > 3/2$. The simple proof of this result, based on the new version of the compensated compactness principle, was given in papers [6], [8]. In our case, by Theorem 2.1, the critical estimate $\|\rho_n |\mathbf{u}_n|^2\|_{L^{(1+\kappa)}(\Omega')} \leq c(\Omega')$ holds for every $\Omega' \Subset \Omega$, which leads to the following local version of the compensated compactness result from [8]; see also [11], [22].

LEMMA 4.1. *Let there be given function $\Phi \in C^\infty(\Omega)$ so that $\Phi(x, \lambda)$ vanishes near $\partial\Omega$ and $\partial_\lambda \Phi(x, \lambda) = 0$ for sufficiently large λ. Then*

$$\int_\Omega \overline{\Phi V(\rho, \mathbf{u})} dx = \int_\Omega \overline{\Phi}\ \overline{V} dx, \quad \textit{where} \quad \overline{V} = \overline{p} - (2 + \nu)\operatorname{div} \mathbf{u}. \tag{4.1}$$

Proof. It suffice to prove the lemma for Φ in the form $\Phi = h(x)\varphi(\lambda)$ with $h \in C_0^\infty(\Omega)$, and $\varphi \in C^\infty(\mathbb{R})$ so that $\varphi'(\lambda) = 0$ for all large λ. Denote by $\mathbf{1}_\Omega$ the extension operator, for functions defined in Ω, such that $\mathbf{1}_\Omega u = u$ in Ω and $\mathbf{1}_\Omega u = 0$ in $\Omega^c = \mathbb{R}^3 \backslash \Omega$. The adjoint operator $\mathbf{1}_\Omega^*$ assigns to every function its restriction to Ω. Introduce the linear operators

$$A_i = \mathbf{1}_\Omega^* \Delta^{-1} \partial_{x_i} \mathbf{1}_\Omega, \quad R_{ij} = \mathbf{1}_\Omega^* \partial_{x_i} \Delta^{-1} \partial_{x_j} \mathbf{1}_\Omega, \quad 1 \leq i, \quad j \leq 3.$$

From the classical potential theory, it follows that they admit the following integral representations (the integrals for R_{ij}, R_{ii} are determined in the sense of Cauchy's

principal value):

$$A_i u(x) = \frac{1}{4\pi}\int_\Omega \frac{x_i - y_i}{|x-y|^3} u(y)\,dy,$$

$$R_{ij}u(x) = -\frac{3}{4\pi}\int_\Omega \frac{(x_i-y_i)(x_j-y_j)}{|x-y|^5}u(y)\,dy \quad \text{for} \quad i \neq j,$$

$$R_{ii}u(x) = \frac{1}{4\pi}\int_\Omega \frac{|x-y|^2 - 3|x_i-y_i|^2}{|x-y|^5}u(y)\,dy + \frac{1}{3}u(x).$$

We denote by A the vectorial operator with the entries A_i and by R the matrix operator with the entries $R_{ij}, i,j = 1,2,3$. Recall that the operators $R_{ij} : L^r(\Omega) \mapsto L^r(\Omega)$, $A_i : L^r(\Omega) \mapsto H^{1,r}(\Omega)$ are bounded for every $r > 1$. Since $\Omega_n \overset{\mathcal{H}}{\to} \Omega$, there is n_0 such that spt $h \Subset \Omega_n$ for all $n \geq n_0$.

Multiplying the moment balance equation for $(\rho_n, \mathbf{u}_n)$, $n \geq n_0$, by h we arrive at

$$\begin{aligned}\operatorname{div}\Big\{h\rho_n\mathbf{u}_n\otimes\mathbf{u}_n - \Pi(h\mathbf{u}_n) + hp(\rho_n)\mathbf{I} + \nu\mathbf{u}_n\otimes\nabla h + \nu\nabla h\otimes\mathbf{u}_n + (\xi-\nu)\nabla h\cdot\mathbf{u}_n\mathbf{I}\Big\}\\ -\Big\{\rho_n\mathbf{u}_n\otimes\mathbf{u}_n - \Pi(\mathbf{u}_n) + p(\rho_n)\mathbf{I}\Big\}\nabla h = h\rho_n\mathbf{f} \ \text{ in } \ \mathcal{D}'(\Omega).\end{aligned}$$

Next, apply to both sides of this identity the operator A to obtain

$$\begin{aligned}R : \{h\rho_n\mathbf{u}_n\otimes\mathbf{u}_n - \Pi(h\mathbf{u}_n) + hp(\rho_n)\mathbf{I} + 2\nu\mathbf{u}_n\otimes\nabla h + (\xi-\nu)\nabla h\cdot\mathbf{u}_n\mathbf{I}\}\\ = A\cdot(h\rho_n\mathbf{f}) + A\cdot\Big\{\Big(\rho_n\mathbf{u}_n\otimes\mathbf{u}_n - \Pi(\mathbf{u}_n) + p(\rho_n)\mathbf{I}\Big)\nabla h\Big\} \ \text{ in } \ L^1(\Omega).\end{aligned} \tag{4.2}$$

Since h is compactly supported in Ω, we have

$$\begin{aligned}R : \{-\Pi(h\mathbf{u}_n) + hp(\rho_n)\mathbf{I}\} = hV(\rho_n,\mathbf{u}_n) + (\xi+\nu)\nabla h\cdot\mathbf{u}_n,\\ R : ((\nabla h\cdot\mathbf{u}_n)\mathbf{I}) = \nabla h\cdot\mathbf{u}_n.\end{aligned} \tag{4.3}$$

Multiplying both sides of (4.2) by $\varphi_n = \varphi(\rho_n)$, integrating the result over Ω, and using in addition relations (4.3), we obtain

$$\int_\Omega h\varphi_n V(\rho_n,\mathbf{u}_n)dx + \int_\Omega \varphi_n\big(\mathcal{P}_n + R : (h\rho_n\mathbf{u}_n\otimes\mathbf{u}_n)\big)dx = 0, \tag{4.4}$$

$$\mathcal{P}_n = 2\xi\nabla h\cdot\mathbf{u}_n + 2\nu R : (\mathbf{u}_n\otimes\nabla h) - A\cdot\big\{(\rho_n\mathbf{u}_n\otimes\mathbf{u}_n - \Pi(\mathbf{u}_n) + p_n\mathbf{I})\nabla h + \rho_n h\mathbf{f}\big\}.$$

On the other hand, multiplying both sides of the renormalized mass balance equation for $(\rho_n, \mathbf{u}_n)$ by h and setting $G = \varphi$, we get

$$\operatorname{div}(h\varphi_n\mathbf{u}_n) + h\Big(\varphi_n'(\rho_n)\rho_n - \varphi_n\Big)\operatorname{div}\mathbf{u}_n - \varphi_n\nabla h\cdot\mathbf{u}_n = 0. \tag{4.5}$$

Introduce the vector field

$$\mathbf{v}_n^{(m)} = \rho_n^{(m)}\mathbf{u}_n, \quad \text{where} \quad \rho_n^{(m)} = \min\{m,\rho_n\}, \ m \geq 1.$$

Applying the operator $\mathbf{v}_n^{(m)}\cdot A$ to both sides of (4.5) and integrating the result over Ω, we arrive at the equality

$$\int_\Omega \big\{\mathbf{v}_n^{(m)}\mathfrak{P}_n + \mathbf{v}_n^{(m)}R(h\varphi_n\mathbf{u}_n)\big\}dx = 0, \tag{4.6}$$

where we denote

$$\mathfrak{P}_n = A \cdot \Big\{ h(\varphi'(\rho_n)\rho_n - \varphi_n)\mathrm{div}\,\mathbf{u}_n - \varphi_n \nabla h \cdot \mathbf{u}_n \Big\}.$$

Combining (4.4), (4.6), and the equality

$$\int_\Omega \Big(\mathbf{v}_n^{(m)} R(h\varphi_n \mathbf{u}_n) - \varphi_n R : (h\rho_n \mathbf{u}_n \otimes \mathbf{u}_n) \Big) dx$$

$$= \int_\Omega \Big(\mathbf{v}_{n,i}^{(m)} R_{ij}(h\varphi_n \mathbf{u}_{n,j}) - \varphi_n R_{ij}(h\mathbf{v}_{ni}^{(m)} \mathbf{u}_{n,j}) \Big) dx + \int_\Omega \varphi_n R : \{(\rho_n^{(m)} - \rho_n) h\mathbf{u}_n \otimes \mathbf{u}_n\} dx$$

$$= \int_\Omega h u_{n,j} \Big(\varphi_n R_{ij} \mathbf{v}_{ni}^{(m)} - \mathbf{v}_{n,i}^{(m)} R_{ij} \varphi_n \Big) dx + \int_\Omega \varphi_n R : \{(\rho_n^{(m)} - \rho_n) h\mathbf{u}_n \otimes \mathbf{u}_n\} dx,$$

we obtain the equality

$$\int_\Omega h\varphi_n V(\rho_n, \mathbf{u}_n) dx = \int_\Omega \Big(\mathbf{v}_n^{(m)} (\mathfrak{P}_n - \varphi_n \mathcal{P}_n + \mathfrak{R}_n \Big) dx + \varpi_n^{(m)}, \tag{4.7}$$

in which components of the vector $\mathfrak{R}_n$ and the scalar $\varpi_n^{(m)}$ are defined by

$$\mathfrak{R}_{n,i} = h\mathbf{u}_{n,j} \Big(\varphi_n R_{ij} \mathbf{v}_{n,i}^{(m)} - \mathbf{v}_{n,i}^{(m)} R_{ij} \varphi_n \Big), \tag{4.8}$$

$$\varpi_n^{(m)} = \int_\Omega \varphi_n R : \{(\rho_n^{(m)} - \rho_n) h\mathbf{u}_n \otimes \mathbf{u}_n\} dx, \tag{4.9}$$

respectively. Recall that $\rho_n \mathbf{u}_n \otimes \mathbf{u}_n \nabla h \to \rho \mathbf{u} \otimes \mathbf{u} \nabla h$ and $p_n \nabla h \to \overline{p} \nabla h$ weakly in $L^{1+\kappa}(\Omega)$ for $n \to \infty$. Hence, there are limits for $n \to \infty$,

$$\mathcal{P}_n \to \overline{\mathcal{P}} \ \text{ in } \ L^r(\Omega) \ \text{ for some } \ r > 1, \tag{4.10}$$

where

$$\overline{\mathcal{P}} = 2\xi \nabla h \mathbf{u} + 2\nu R : (\mathbf{u} \otimes \nabla h) - A \cdot \{ \big(\rho \mathbf{u} \otimes \mathbf{u} - \Pi(\mathbf{u}) + \overline{p} \mathbf{I} \big) \nabla h + h\rho \mathbf{f} \},$$

furthermore,

$$\mathfrak{P}_n \to \overline{\mathfrak{P}} \equiv A \cdot \{ h\overline{(\varphi'\rho - \varphi)\mathrm{div}\,\mathbf{u}} - \overline{\varphi} \nabla h \cdot \mathbf{u} \} \ \text{ in } \ L^2(\Omega)\ . \tag{4.11}$$

Since the sequences φ_n and $\mathbf{v}_n^{(m)}$ are bounded in $L^\infty(\Omega)$ and $L^6(\Omega)$, respectively, it follows from the compensated compactness Lemma from [8] that

$$\varphi_n R_{ij} \mathbf{v}_{n,i}^{(m)} - \mathbf{v}_{n,i}^{(m)} R_{ij} \varphi_n \to \overline{\varphi} R_{ij} \overline{\mathbf{v}^{(m)}}_i - \overline{\mathbf{v}^{(m)}}_i R_{ij} \overline{\varphi} \ \text{ weakly in } \ L^2(\Omega).$$

Therefore, $\mathfrak{R}_n$ converges weakly in $L^{3/2}(\Omega)$ to $\overline{\mathfrak{R}} = \{ \mathbf{u}_j (\overline{\varphi} R_{ij} \overline{\mathbf{v}^{(m)}}_i - \overline{\mathbf{v}^{(m)}}_i R_{ij} \overline{\varphi}) \}$. Passing to the limit in (4.7) and using (4.10)–(4.11), we obtain

$$\int_\Omega h\overline{\varphi V} dx = \int_\Omega (\overline{\mathbf{v}} \overline{\mathfrak{P}} - \overline{\varphi}\overline{\mathcal{P}} + \overline{\mathfrak{R}}) dx + \overline{\varpi^{(m)}} \tag{4.12}$$

with $|\overline{\varpi^{(m)}}| \le \limsup |\varpi_n^{(m)}|$. On the other hand, passage to the limit in equalities (4.2) and (4.5) gives

$$\mathrm{div} \Big\{ h\rho \mathbf{u} \otimes \mathbf{u} - \Pi(h\mathbf{u}) + h\overline{p} I + \nu \mathbf{u} \otimes \nabla h + \nu \nabla h \otimes \mathbf{u} + (\xi - \nu) \nabla h \cdot \mathbf{u} \Big\}$$

$$- \big(\rho \mathbf{u} \otimes \mathbf{u} - \Pi(\mathbf{u}) + \overline{p} I \big) \nabla h = h\rho \mathbf{f}, \tag{4.13}$$

$$\mathrm{div}\,(h\overline{\varphi}\mathbf{u}) + h\overline{(\varphi'\rho - \varphi)\mathrm{div}\,\mathbf{u}} - \overline{\varphi} \nabla h \cdot \mathbf{u} = 0. \tag{4.14}$$

Applying the operators $\overline{\varphi}A$ and $\overline{\mathbf{v}}\cdot A$, respectively, to both sides of the moment and the renormalized mass balance equations and arguing as before, we obtain

$$\int_\Omega h\overline{\varphi}\overline{V}dx = \int_\Omega \big(\overline{\mathbf{v}^{(m)}}\overline{\mathfrak{P}} - \overline{\varphi}\overline{\mathcal{P}} + \overline{\mathfrak{R}}\big)dx + \varpi^{(m)}, \tag{4.15}$$
$$\text{where } \ \varpi^{(m)} = \int_\Omega \overline{\varphi}R : [h(\overline{\mathbf{v}} - \rho\mathbf{u})\otimes\mathbf{u}]dx.$$

Combining (4.12) and (4.15), we finally obtain

$$\int_\Omega h\overline{\varphi}\overline{V}dx - \int_\Omega h\overline{\varphi V}dx = \varpi^{(m)} - \overline{\varpi^{(m)}}. \tag{4.16}$$

By Lemma 3.3 the sequence $h\rho_n\mathbf{u}_n\otimes\mathbf{u}_n$ converges weakly in $L^{1+\kappa}(\Omega)$ to $h\rho\mathbf{u}\otimes\mathbf{u}$; obviously $\rho_n^{(m)}\mathbf{u}_n = \mathbf{v}_n^{(m)}$ converges weakly in $L^6(\Omega)$ to $\overline{\mathbf{v}^{(m)}}$. From this we conclude that

$$\begin{aligned}|\varpi^{(m)}| &\le \|R : [h(\overline{\mathbf{v}^{(m)}} - \rho\mathbf{u})\otimes\mathbf{u}]\|_{L^1(\Omega)} \le c\|h(\overline{\mathbf{v}^{(m)}} - \rho\mathbf{u})\otimes\mathbf{u}\|_{L^{1+\kappa/2}(\Omega)}\\ &\le c\limsup_{n\to\infty}\|h(\mathbf{v}_n - \rho_n\mathbf{u}_n)\otimes\mathbf{u}_n\|_{L^{1+\kappa/2}(\text{spt } h)}.\end{aligned}$$

Similar arguments can be used to obtain the same bound for $\limsup_{n\to\infty}|\varpi_n^{(m)}|$. Since the sequence $\rho_n\mathbf{u}_n\otimes\mathbf{u}_n$ is bounded in $L^{1+\kappa}_{loc}(\Omega)$ and the sequence ρ_n is bounded in $L^1(\Omega)$, we have

$$\begin{aligned}\|h(\mathbf{v}_n^{(m)} - \rho_n\mathbf{u}_n)\otimes\mathbf{u}_n\|_{L^{1+\kappa/2}(\text{spt } h)} &\le c(h)\|(\rho_n^{(m)} - \rho_n)\mathbf{u}_n\otimes\mathbf{u}_n\|_{L^{1+\kappa/2}(\text{spt } h)}\\ &\le c\Big(\int_{\{\rho_n>m\}\cap\text{spt } h}(\rho_n|\mathbf{u}_n|^2)^{1+\kappa/2}dx\Big)^{2/(2+\kappa)}\\ &\le c\|\rho_n|\mathbf{u}_n|^2\|_{L^{1+\kappa}(\text{spt } h)}\ \text{mes}\ \{\rho_n > m\}^\beta \le cm^{-\beta},\end{aligned}$$

where $\beta = \kappa(1+\kappa)^{-1}(2+\kappa)^{-1} > 0$. Hence $|\varpi^{(m)}| + |\overline{\varpi^{(m)}}| \to 0$ for $m\to\infty$. It remains to note that the left-hand side of (4.16) does not depend on n, and the proof of Lemma 4.1 is completed. □

COROLLARY 4.2. *Assume that the function $\lambda\mapsto\varphi(\lambda)$ belongs to the class $C^\infty(\mathbb{R})$ and vanishes for sufficiently large λ. Let $\overline{\varphi p}\in L^\infty(D)$ be an L^∞-star-weak limit of the sequence $\{\varphi(\rho_n)p(\rho_n)\}$, $\overline{\varphi\text{div}\,\mathbf{u}}\in L^2(D)$ be L^2- weak limit of the sequence $\{\varphi(\rho_n)\text{div}\,\mathbf{u}_n\}$. Then*

$$\frac{1}{\xi+\nu}\overline{\varphi p} - \overline{\varphi\text{div}\,\mathbf{u}} = \frac{1}{\xi+\nu}\overline{\varphi p} - \overline{\varphi}\text{div}\,\mathbf{u} \quad \textit{in} \quad D\setminus S, \tag{4.17}$$

where $\overline{\varphi}$ and $\overline{p}$ are given by Lemma 3.2.

Proof. Recall that $D\setminus S = (D\setminus B)\cup\Omega$. Since the characteristic function of any compact subset of Ω is a pointwise limit of a sequence of smooth functions compactly supported in Ω, relation (4.17) holds true on each compact subset of Ω. It remains to note that both sides of (4.17) belong to $L^1(D\setminus S)$ and vanish outside of B. □

5. The oscillation defect measure. The notion of the oscillation defect measure was introduced in [6] to justify the existence theory for isentropic flows with *small* values of the adiabatic constant γ. Following [6], [11] the r-oscillation defect measure associated with the sequence $\{\rho_n\}_{n\le1}$ is defined as

$$\mathbf{osc}_r[\rho_n\to\rho\,](K) := \sup_{k\ge1}\limsup_{n\to\infty}\|T_k(\rho_n) - T_k(\rho)\|^r_{L^r(K)},$$

where $T_k(z) = kT(z/k)$, with $T(z)$ a smooth concave function, is equal to z for $z \le 1$ and is a constant for $z \ge 3$. The smoothness properties of T_k are not important and we can take the simplest form $T_k(z) = \min\{z, k\}$. Note that the total energy estimates provide the boundedness of the γ-oscillation defect measure on the whole domain D. The unexpected result was obtained by Feireisl [6] and Feireisl, Novotný, and Petzeltová [8], who showed that the $(1+\gamma)$-oscillation defect measure associated with the sequence $\{\rho_n\}$ is uniformly bounded on all compact subsets of Ω.

Note that in the shape optimization problem we cannot replace the compact subsets $K \Subset \Omega$ with the domain Ω itself, since the oscillation defect measure is not any *regular* set additive function on the family of compact subsets of Ω, i.e., it is not any measure in the sense of measure theory. To bypass this difficulty we observe that the finiteness of the oscillation defect measure on compacts gives some additional information on the properties of the distribution function Γ. Our task is to first extract this information and then to use it in the proof of Theorem 1.2. In order to formulate the appropriate auxiliary result, we define the function $\mathcal{T}_\vartheta(x)$ by the equality

$$\mathcal{T}_\vartheta(x) = \overline{\min\{\rho, \vartheta\}}(x) - \min\{\rho(x), \vartheta(x)\} \quad \text{for each} \quad \vartheta \in C(\Omega)\,.$$

LEMMA 5.1. *Under the assumptions of Theorem* 1.2 *and Lemma* 3.2*, there is a constant c independent of ϑ and K such that the inequalities*

$$\|\mathcal{T}_\vartheta\|^{1+\gamma}_{L^{1+\gamma}(K)} \le \lim_{n\to\infty} \int_\Omega |\min\{\rho_n(x), \vartheta(x)\} - \min\{\rho(x), \vartheta(x)\}|^{1+\gamma} dx \le c \tag{5.1}$$

hold for all $\vartheta \in C(\Omega)$ and $K \Subset \Omega$. We point out that the limit in (5.1) *does exist by the choice of the sequence ρ_n.*

Proof. The proof imitates the proof of Lemma 4.3 from [8]. It can be easily seen that

$$\|\mathcal{T}_\vartheta\|^{1+\gamma}_{L^{1+\gamma}(K)} \le \limsup_{n\to\infty} \int_K |\min\{\rho_n(x), \vartheta(x)\} - \min\{\rho(x), \vartheta(x)\}|^{1+\gamma}\, dx. \tag{5.2}$$

Hence it suffices to show that the right-hand side of this inequality admits a bound independent of ϑ. From the properties of $\min\{\cdot,\cdot\}$, it follows that

$$|\min\{s', \vartheta\} - \min\{s'', \vartheta\}|^{1+\gamma} \le (\min\{s', \vartheta\} - \min\{s'', \vartheta\})(s'^\gamma - s''^\gamma) \quad \text{for all} \quad s', s'' \in \mathbb{R}^+;$$

furthermore, for the weak limits we have the inequalities $\overline{\rho^\gamma} \ge \rho^\gamma$ and $\overline{\min\{\rho, \vartheta\}} \le \min\{\rho, \vartheta\}$, and therefore, for any compactly supported, nonnegative function $h \in C(\Omega)$, we get

$$\begin{aligned}
&\lim_{n\to\infty} \int_\Omega h |\min\{\rho_n, \vartheta\} - \min\{\rho, \vartheta\}|^{1+\gamma}\, dx \\
&\le \lim_{n\to\infty} \int_\Omega h(\min\{\rho_n, \vartheta\} - \min\{\rho, \vartheta\})(\rho_n^\gamma - \rho^\gamma)\, dx \\
&\le \lim_{n\to\infty} \int_\Omega h(\min\{\rho_n, \vartheta\} - \min\{\rho, \vartheta\})(\rho_n^\gamma - \rho^\gamma)\, dx \\
&\quad + \int_\Omega (\overline{\rho^\gamma} - \rho^\gamma)(\min\{\rho, \vartheta\} - \overline{\min\{\rho, \vartheta\}})\, dx \\
&= \lim_{n\to\infty} \int_\Omega h(\rho_n^\gamma \min\{\rho_n, \vartheta\} - \overline{\rho^\gamma}\,\overline{\min\{\rho, \vartheta\}})\, dx \\
&= \lim_{n\to\infty} \int_\Omega h(p(\rho_n) \min\{\rho_n, \vartheta\} - \overline{p}\,\overline{\min\{\rho, \vartheta\}})\, dx.
\end{aligned} \tag{5.3}$$

By Lemma 4.1 with $\Phi(\rho,x)=h(x)\min\{\rho,\vartheta(x)\}$, the right-hand side of (5.3), divided by $(\xi+\nu)$, is equal to

$$
\begin{aligned}
(5.4)\qquad &\lim_{n\to\infty}\int_\Omega h(\min\{\rho_n,\vartheta\}\operatorname{div}\mathbf{u}_n-\overline{\min\{\rho,\vartheta\}}\operatorname{div}\mathbf{u})\,dx\\
&=\lim_{n\to\infty}\int_\Omega h(\min\{\rho_n,\vartheta\}-\min\{\rho,\vartheta\})\operatorname{div}\mathbf{u}_n dx\\
&\quad-\lim_{n\to\infty}\int_\Omega h(\min\{\rho_n,\vartheta\}-\min\{\rho,\vartheta\})\operatorname{div}\mathbf{u}\,dx\\
&\le\delta\limsup_{n\to\infty}\int_\Omega h|\min\{\rho_n,\vartheta\}-\min\{\rho,\vartheta\}|^{1+\gamma}\,dx\\
&\quad+\delta^{-\gamma}\lim_{n\to\infty}\int_\Omega h(|\operatorname{div}\mathbf{u}_n|+|\operatorname{div}\mathbf{u}|)^{(1+\gamma)/\gamma}\\
&\le\delta\lim_{n\to\infty}\int_\Omega h|\min\{\rho_n,\vartheta\}-\min\{\rho,\vartheta\}|^{1+\gamma}\,dx+c\delta^{-\gamma}\|h\|_{C(\Omega)}.
\end{aligned}
$$

Combining (5.4) and (5.3), choosing $h=1$ on K, and choosing $\delta>0$ with δ sufficiently small, we obtain (5.1). □

We reformulate this result in terms of the distribution function Γ. Recall that the functions $\min\{\rho_n,\lambda\}$ are uniformly bounded in $\mathbb{R}^3$ and that $\min\{\rho_n,\lambda\}\operatorname{div}\mathbf{u}_n$ converges weakly in $L^2(D)$ for all nonnegative λ. Introduce the functions

$$
\begin{aligned}
(5.5)\qquad \mathcal{V}_\lambda&=\overline{\Big(\min\{\rho,\lambda\}\operatorname{div}\mathbf{u}\Big)}-\overline{\min\{\rho,\lambda\}}\operatorname{div}\mathbf{u}\in L^2(D),\\
\mathfrak{H}(x)&=\int_{[0,\infty)}\Gamma(x,s)(1-\Gamma(x,s))\,ds,\qquad \mathfrak{H}\in L^\gamma(D).
\end{aligned}
$$

Lemma 5.2. *There is a constant c independent of λ such that*

$$
(5.6)\qquad \|\mathfrak{H}\|_{L^{1+\gamma}(D\setminus S)}+\sup_\lambda\|\mathcal{V}_\lambda\|_{L^1(D\setminus S)}\le c.
$$

Proof. Recall that $D\setminus S=(D\setminus B)\cup\Omega$ and that $H=\mathcal{V}_\lambda=0$ on $D\setminus B$. Hence it is sufficient to prove that for all compacts $K\Subset\Omega$, we have

$$
(5.7)\qquad \|\mathfrak{H}\|_{L^{1+\gamma}(K)}+\sup_\lambda\|\mathcal{V}_\lambda\|_{L^1(K)}\le c,
$$

with the constant c independent of K. We begin with the observation that by Lemma 3.2,

$$
(5.8)\qquad \mathcal{T}_\vartheta(x)=\int_{[0,\infty)}\min\{\lambda,\vartheta(x)\}\,d_\lambda\Gamma(x,\lambda)-\min\left\{\int_{[0,\infty)}\lambda\,d_\lambda\Gamma(x,\lambda),\vartheta(x)\right\}
$$

for all functions $\vartheta\in C(D)$. From this and the identity $\rho(x)=\int_{[0,\infty)}(1-\Gamma(x,\lambda))\,d\lambda$, we conclude that

$$
(5.9)\qquad \mathcal{T}_\vartheta(x)=\int_0^{\vartheta(x)}\Gamma(x,s)\,ds\ \text{ for }\vartheta(x)\ge\rho(x)\ \text{ and }\ \mathcal{T}_\vartheta(x)=\int_{\vartheta(x)}^\infty(1-\Gamma(x,s))\,ds\ \text{ otherwise.}
$$

Next, choose a sequence of continuous nonnegative functions $\{\vartheta_k\}_{k\geq 1}$ which converge for $k \to \infty$ to the density $\vartheta_k \to \rho$ a.e. in $D \setminus S$. By Lemma 5.1 the norms in $L^{1+\gamma}(K)$ of functions $\mathcal{T}_{\vartheta_k}$ are uniformly bounded by a constant independent of k and K. Moreover, $\mathcal{T}_{\vartheta_k}$ converges a.e. in K to the function

$$\mathcal{T}_\rho(x) = \int_0^{\rho(x)} \Gamma(x,s)\,ds = \int_{\rho(x)}^{\infty} (1-\Gamma(x,s))\,ds,$$

which yields the estimates $\|\mathcal{T}_\rho\|_{L^{1+\gamma}(K)} \leq c$ with the constant c independent of K. It remains to note that estimate (5.7) for $\mathfrak{H}$ obviously follows from the inequality $\mathfrak{H} \leq 2\mathcal{T}_\rho$.

To estimate $\mathcal{V}_\lambda$ note that

$$\mathcal{V}_\lambda = \text{w-}\lim_{n\to\infty}\Big((\min\{\rho_n,\lambda\} - \min\{\rho,\lambda\})\operatorname{div}\mathbf{u}_n\Big) - \Big(\text{w-}\lim_{n\to\infty}\min\{\rho_n,\lambda\} - \min\{\rho,\lambda\}\Big)\operatorname{div}\mathbf{u},$$

where w-lim is denotes the weak limit in $L^1(D\setminus S)$. From this and the boundedness of norms $\|\operatorname{div}\mathbf{u}_n\|_{L^2(D)}$, we obtain

$$\begin{aligned}\|\mathcal{V}_\lambda\|_{L^1(K)} \leq \limsup_{n\to\infty}&(\|\operatorname{div}\mathbf{u}_n\|_{L^2(K)} + \|\operatorname{div}\mathbf{u}\|_{L^2(K)})\\ &\times \|\min\{\rho_n(x),\lambda\} - \min\{\rho(x),\lambda\}\|_{L^2(K)},\end{aligned}$$

which along with (5.1) implies (5.7). The proof of Lemma 5.2 is completed. □

6. Kinetic formulation of the mass balance equation. In this section we show that the distribution function $\Gamma(x,\lambda)$ of the Young measure, associated with a given sequence of solutions to problem (1.3), satisfies some integrodifferential transport equation, which is called the kinetic equation. This result is given by the following lemma. Fix an arbitrary function $\zeta(x,\lambda)$ satisfying the conditions

$$\zeta \in C_0^\infty(D\times\mathbb{R}), \qquad \operatorname{spt}\zeta \Subset D\setminus(\Sigma^-\cup S). \tag{6.1}$$

We use the notation ∂_λ for the partial derivatives with respect to the variable λ, e.g., $\partial_\lambda\zeta := \frac{\partial\zeta}{\partial\lambda}$. The absolutely continuous measure is denoted by $d_\lambda\zeta := \partial_\lambda\zeta d\lambda$.

Recall that the compact obstacle is of the form $S = B\setminus\Omega$ and that $\Sigma^- \subset \partial B$ is the *exit* set.

LEMMA 6.1. *Suppose that all assumptions of Theorem* 1.2 *are satisfied and* Γ *is a distribution function of the Young measure associated with a given sequence* $\{\rho_n\}$ *of solutions to problem* (1.3). *Then*

$$\int_{(D\setminus S)\times\mathbb{R}} \Gamma(x,\lambda)\nabla_{x,\lambda}\zeta\cdot\mathbf{w}\,d\lambda\,dx + \int_{(D\setminus S)\times\mathbb{R}} \lambda\mathcal{M}(x,\lambda)\,d_\lambda\zeta dx = 0. \tag{6.2}$$

Here $\mathbf{w}$ *is the solenoidal vector field of the form* $\mathbf{w}(x,\lambda) = (\mathbf{u}(x), -\lambda\operatorname{div}\mathbf{u})$, *and the function* $\mathcal{M}$ *is defined by the equalities*

$$\mathcal{M}(x,\lambda) = -\frac{1}{\xi+\nu}\int_{(-\infty,\lambda)}(s^\gamma - \overline{\rho})\,d_s\Gamma(x,s) = \frac{1}{\xi+\nu}\int_{[\lambda,\infty)}(s^\gamma-\overline{\rho})\,d_s\Gamma(x,s), \tag{6.3}$$

in which the weak limit for the pressure $\overline{\rho}(x) = \int_{\mathbb{R}}\lambda^\gamma d_\lambda\Gamma(x,\lambda)$ *is defined as in Lemma* 3.2. *Integral identity* (6.2) *is equivalent, in the sense of distributions, to the kinetic equation*

$$\begin{aligned}&\frac{\partial}{\partial\lambda}\big[\lambda\operatorname{div}\mathbf{u}(x)\Gamma(x,\lambda)\big] - \operatorname{div}\big(\Gamma(x,\lambda)\mathbf{u}(x)\big) - \frac{\partial}{\partial\lambda}[\lambda\mathcal{M}(x,\lambda)] = 0\\ &\quad \text{in}\ \ \mathcal{D}'\big((D\setminus(S\cup\Sigma^-))\times\mathbb{R}\big).\end{aligned} \tag{6.4}$$

Remark 6.1. Since the kinetic equation is understood in the sense of distributions, the equation remains valid if we replace the intervals of integration $(-\infty,\lambda)$ and $[\lambda,\infty)$ in formulae (6.3) with $(-\infty,\lambda]$ and (λ,∞) respectively; however, this creates some discomfort. To avoid such ambiguity, we observe that (6.2) also holds true if we replace the function $\mathcal{M}$ with its invariant form

$$\mathfrak{M}(x,\lambda) := \frac{1}{2}\Big(\lim_{s\to\lambda+0}\mathcal{M}(x,s)+\lim_{s\to\lambda-0}\mathcal{M}(x,s)\Big). \tag{6.5}$$

Proof. Choose an arbitrary $\varphi\in C_0^\infty(\mathbb{R})$, and note that for all $h\in C^\infty(D)$ with $\operatorname{spt} h \Subset D\setminus\big(\Sigma^-\cup S\big)$,

$$\int_{D\setminus S}\Big\{h\big(\varphi(\rho_n)-\varphi'(\rho_n)\rho_n\big)\operatorname{div}\mathbf{u}_n+\varphi(\rho_n)\nabla h\cdot\mathbf{u}_n\Big\}dx=0.$$

Taking the limit $n\to\infty$, we obtain

$$\int_{D\setminus S}\Big\{h\overline{\big(\varphi-\varphi'\rho\big)\operatorname{div}\mathbf{u}}+\overline{\varphi}\nabla h\cdot\mathbf{u}\Big\}dx=0,$$

which along with (4.17) gives the integral identity for weak limits,

$$\int_{D\setminus S}\Big[h\overline{\big(\varphi-\varphi'\rho\big)}\operatorname{div}\mathbf{u}+\overline{\varphi}\nabla h\cdot\mathbf{u}+\frac{1}{\xi+\nu}\big(\overline{(\varphi-\varphi'\rho)p}-\overline{(\varphi-\varphi'\rho)}\,\overline{p}\big)\Big]dx=0. \tag{6.6}$$

Next choose an arbitrary smooth function $\eta(\lambda)$ which vanishes near $+\infty$ and set $\varphi(\lambda)=\int_\lambda^\infty\eta(s)ds$. Identity (3.3) from Lemma 3.2 combined with the formula of integration by parts for Stieltjes integrals yields the representations

$$\Big(\overline{\varphi-\varphi'\rho}\Big)(x)=\int_{[0,\infty)}\eta(\lambda)\Gamma(x,\lambda)\,d\lambda+\int_{[0,\infty)}\lambda\eta(\lambda)d_\lambda\Gamma(x,\lambda)=\int_{[0,\infty)}\eta\,d_\lambda(\lambda\Gamma)\,,$$

which are substituted into (6.6) and lead to the integral identity

$$\begin{aligned}\int_{(D\setminus S)\times\mathbb{R}}&\Big\{h\operatorname{div}\mathbf{u}\,d_\lambda\big(\lambda\Gamma(x,\lambda)\big)+\Gamma\nabla h\mathbf{u}d\lambda\Big\}\eta\,dx\\&+\frac{1}{\xi+\nu}\int_{D\setminus S}h\Big\{\overline{(\varphi-\varphi'\rho)p}-\overline{(\varphi-\varphi'\rho)}\,\overline{p}\Big\}\,dx=0.\end{aligned} \tag{6.7}$$

Recall that Γ vanishes for $\lambda<0$, the function $G(x,\lambda)=(\varphi(\lambda)-\varphi'(\lambda)\lambda)h(x)$ satisfies all conditions of Lemma 3.2, and h is compactly supported in $D\setminus S$. It follows from this and Lemma 3.2 that

$$\int_{D\setminus S}h\Big(\overline{(\varphi-\varphi'\rho)p}-\overline{(\varphi-\varphi'\rho)}\,\overline{p}\Big)dx=\int_{(D\setminus S)\times\mathbb{R}}h\eta\lambda(\lambda^\gamma-\overline{p})d_\lambda\Gamma(x,\lambda)dx$$

$$+\int_{(D\setminus S)\times\mathbb{R}}h\Big\{\int_\lambda^\infty\eta(s)ds\Big\}(\lambda^\gamma-\overline{p})d_\lambda\Gamma(x,\lambda)dx=-(\xi+\nu)\int_{(D\setminus S)\times\mathbb{R}}h\eta\;d_\lambda\big(\lambda\mathcal{M}(x,\lambda)\big)dx.$$

By substitution of this identity into (6.7), since the linear hull of the set of functions in the form $h\eta$ is dense in $C_0^\infty((D\setminus S)\times\mathbb{R})$, we obtain (6.2), which completes the proof of Lemma 6.1. □

The next lemma describes the basic properties of the function $\mathcal{M}(x,\lambda)$, which are important for further analysis.

LEMMA 6.2. *For a.e.* $x \in D \setminus S$,

(i) $\mathcal{M}(x,\cdot)$ *is nonnegative and vanishes on* $\mathbb{R}^-$. *Moreover, if the Borel function* $\mathfrak{M}(x,.)$ *given by* (6.5) *vanishes* σ_x*-a.e. on the interval* (ω,∞) *with* $\omega = \overline{p}(x)^{1/\gamma}$, *then* $\sigma_x = d_\lambda\Gamma(x,\cdot)$ *is a Dirac measure and*

$$\Gamma(x,\lambda) = 0 \quad \textit{for} \quad \lambda < \overline{p}(x)^{1/\gamma}, \quad \Gamma(x,\lambda) = 1 \quad \textit{for} \quad \lambda \geq \overline{p}(x)^{1/\gamma}.$$

(ii) *for all* $g \in C_0^\infty(0,\infty)$,

$$\int_{\mathbb{R}_\lambda} g(\lambda)\mathcal{M}(x,\lambda)d\lambda = -\int_{[0,\infty)} g'(\lambda)\mathcal{V}_\lambda(x)d\lambda, \tag{6.8}$$

where $\mathcal{V}_\lambda$ *is defined by* (5.5).

Proof. Let θ be a standard mollifying kernel in $\mathbb{R}$,

$$\theta \in \mathcal{D}_+(\mathbb{R}), \quad \int_{\mathbb{R}} \theta(t)\,dt = 1, \quad \text{spt } \theta \Subset \{|t| \leq 1\},$$

with the corresponding mollifier

$$(\mathbf{S}_k\Gamma)(x,\lambda) := k\int_{\mathbb{R}} \theta(k(\lambda - t))\Gamma(x,t)dt.$$

We will write simply Γ_k instead of $\mathbf{S}_k\Gamma$. The mollified distribution function $\Gamma_k(x,\cdot)$ belongs to the class $C^\infty(\mathbb{R})$ and generates the absolutely continuous Stieltjes measure σ_{kx} of the form $d\sigma_{kx} = \partial_\lambda\Gamma_k d\lambda$. It is easily seen that for $k \to \infty$ the sequence of measures σ_{kx} converges star-weakly to the measure $\sigma_x = d_\lambda\Gamma$ in the space of Radon's measures on $\mathbb{R}$. In particular, for all λ with $\sigma_x\{\lambda\} := \lim_{s\to\lambda+0}\Gamma(x,s) - \lim_{s\to\lambda-0}\Gamma(x,s) = 0$, we can pass to the limit to obtain

$$\int_{[0,\lambda)} (t^\gamma - \overline{p})\partial_t\Gamma_k(x,t)\,dt \to \int_{[0,\lambda)} (t^\gamma - \overline{p})\,d_t\Gamma(x,t) \quad \text{for} \quad k \to \infty. \tag{6.9}$$

In other words, relation (6.9) holds true for all λ, except possibly for some countable set. Since $\partial_\lambda\Gamma_k \geq 0$, the function on the left-hand side of (6.9) increases on $(-\infty,\omega)$ and decreases on (ω,∞). From this and (6.9) we conclude that $\mathcal{M}(x,\cdot)$ does not decrease for $\lambda < \omega$ and does not increase for $\lambda > \omega$, which along with the obvious relations $\lim_{\lambda\to\pm\infty}\mathcal{M}(x,\lambda) = 0$ yields the nonnegativity of $\mathcal{M}$.

To prove the second part of (i), note that $\mathfrak{M}(x,\lambda) = \lim_{k\to\infty}\mathbf{S}_k\mathcal{M}(x,\lambda)$ belongs to the first Baire class and hence is measurable in σ_x. It follows from the monotonicity of $\mathfrak{M}(x,\cdot)$ on the interval (ω,∞) that if $\mathfrak{M}(x,\alpha) = 0$ for some $\alpha > \omega$, then $\mathfrak{M}(x,\lambda) = 0$ and $\Gamma(x,\lambda) = 1$ on (α,∞). Assume that $\mathfrak{M}(x,\cdot)$ vanishes σ_x-a.e. on (ω,∞), and consider the set

$$\mathcal{O} = \Big\{\alpha > \omega : \quad \sigma_x(\omega,\alpha) \equiv \lim_{s\to\alpha-0}\Gamma(x,s) - \lim_{s\to\omega+0}\Gamma(x,s) = 0\Big\}.$$

Let us prove that $\mathcal{O} = (\omega,\infty)$. If the set $\mathcal{O}$ is empty, then there is a sequence of points $\lambda_k \searrow \omega$ with $\mathfrak{M}(x,\lambda_k) = 0$, which yields $\Gamma(x,\cdot) = 1$ on (ω,∞), and thus $\mathcal{O} = (\omega,\infty)$. Hence $\mathcal{O} \neq \emptyset$. If $m = \sup\mathcal{O} < \infty$, then there is a sequence $\lambda_k \searrow m$ with $\mathfrak{M}(x,\lambda_k) = 0$, which yields $\Gamma(x,\cdot) = 1$ on (m,∞). By construction, $\Gamma(x,\lambda) = c =$ constant on (ω,m). In other words, restriction of σ_x to (ω,∞) is the monoatomic measure $(1-c)\delta(\cdot - m)$. Hence $\mathfrak{M}(x,m) = 2^{-1}(1-c)(m^\gamma - \omega^\gamma) = 0$, which yields

$c = 1$. From this we can conclude that $\Gamma(x,\cdot) = 1$ on (ω,∞), and σ_x is a probability measure concentrated on $[0,\omega]$. Recalling that $\omega^\gamma = \overline{p}(x)$, we obtain

$$(\xi+\nu)\mathcal{M}(x,0) = \int_{[0,\omega]} (\lambda^\gamma - \omega^\gamma)\, d_\lambda\Gamma(x,\lambda) \geq 0.$$

Hence $d_\lambda\Gamma(x,\lambda)$ is the Dirac measure concentrated at ω, which implies (i).

The proof of (ii) is straightforward. It is easily seen that

$$\begin{aligned}
-(\xi+\nu)\int_{\mathbb{R}} g(\lambda)\mathcal{M}(x,\lambda)\,d\lambda &= \int_{[0,\infty)} \left(\int_{[\lambda,\infty)} g'(s)\,ds\right)\left(\int_{[\lambda,\infty)} (t^\gamma - \overline{p}) d_t\Gamma(x,t)\right) d\lambda \\
&= \int_{[0,\infty)} g'(s)\left(\int_{[0,s)} d\lambda \int_{[\lambda,\infty)} (t^\gamma - \overline{p})\, d_t\Gamma(x,t)\right) ds \\
&= \int_{[0,\infty)} g'(s)\left(\int_{[0,\infty)} \min\{t,s\}(t^\gamma - \overline{p})\, d_t\Gamma(x,t)\right) ds \\
&= \int_{[0,\infty)} g'(s)(\overline{\min\{\rho,s\}p} - \overline{\min\{\rho,s\}}\overline{p})\, ds.
\end{aligned}$$

On the other hand, Lemma 4.1 yields $\overline{\min\{\rho,\lambda\}p} - \overline{\min\{\rho,\lambda\}}\overline{p} = (\xi+\nu)\mathcal{V}_\lambda(x)$. The proof of Lemma 6.2 is completed. □

7. Renormalization of the kinetic equation. The notion of a renormalized solution, introduced in the pioneering paper [4], plays an important role in the theory of compressible NSE developed by Lions and by E. Feireisl and coworkers. Moreover, the kinetic equation itself is a result of the renormalization procedure. Formally we can renormalize (6.4) by multiplying both sides by a function $\Psi'(\Gamma)$, which leads to the transport equation for the function $\Psi(\Gamma)$, but whether this construction is justified a delicate question. The corresponding result is given by the following lemma. Set $\Psi(\Gamma) = \Gamma(1-\Gamma)$.

LEMMA 7.1. *For all functions $h \in C_0^\infty(D\setminus S)$ with spt $h \Subset D\setminus(S\cup\Sigma^-)$ and for all functions $\eta \in C^\infty(\mathbb{R})$ vanishing near $+\infty$, we have the integral identity*

$$\int_{(D\setminus S)\times\mathbb{R}} \mathcal{F}(x,\lambda)dx\,d\lambda = 2\int_{(D\setminus S)} \left(\int_{[0,\infty)} \eta(\lambda)\lambda\mathfrak{M}(x,\lambda)d_\lambda\Gamma(x,\lambda)\right)h(x)\,dx, \tag{7.1}$$

where

$$\mathcal{F} \equiv \eta(\lambda)\Psi(\Gamma)\mathbf{u}(x)\nabla h(x) - \lambda h(x)\Psi(\Gamma)\eta'(\lambda)\operatorname{div}\mathbf{u}(x) + \lambda h(x)\Psi'(\Gamma)\mathcal{M}(x,\lambda)\eta'(\lambda).$$

In other words, the function $\Psi(\Gamma)$ satisfies the transport equation

$$\operatorname{div}_{\lambda,x}\big(\Psi(\Gamma)\mathbf{w}\big) + \frac{\partial}{\partial\lambda}\big(\lambda\Psi'(\Gamma)\mathcal{M}\big) + 2\lambda\mathfrak{M}\frac{\partial\Gamma}{\partial\lambda} = 0 \quad \textit{in} \quad \mathcal{D}'\Big(\big(D\setminus(\Sigma^-\cup S)\big)\times\mathbb{R}\Big).$$

Proof. Note that by assertion (i) of Lemma 3.2, the function Γ is well defined on each domain in $\mathbb{R}^4$, and hence on the whole space $\mathbb{R}^4$. The same conclusion can be drawn for the vector field $\mathbf{u} \in H^{1,2}_{loc}(\mathbb{R}^3)$. Let Θ be the regularizing kernel in $\mathbb{R}^3$, i.e.,

$$\Theta \in \mathcal{D}_+(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \Theta(x)dx = 1, \quad \text{spt } \Theta \Subset \{|x| \leq 1\},$$

and define the mollifier for scalar or vector functions,

$$\mathbf{T}_m u(x,\lambda) = m^3 \int_\Omega \Theta(m(x-y))u(y,\lambda)\,dy.$$

We will simply write $\Gamma_{k,m}$ and Γ_k instead of $\mathbf{S}_k\mathbf{T}_m\Gamma$ and $\mathbf{S}_k\Gamma$. Obviously, Γ_k are smooth functions of the variable λ and $|\partial^\alpha\Gamma_k(x,\lambda)| \le k^\alpha$. Recall that for any $f \in L^2(D')$ with $D \Subset D'$ and dist $(D', \partial D) > m^{-1}$,

$$\|\mathbf{T}_m f\|_{L^2(D)} \le \|f\|_{L^2(D')}, \quad \mathbf{T}_m f \to f \ \text{ in } \ L^2(D). \tag{7.2}$$

Substituting the test function

$$\zeta(x,\lambda) = m^3\Theta(m(x_0-x))\theta(k(\lambda_0-\lambda)), \quad \text{dist } (x_0,\partial D) > m^{-1}, \quad \lambda_0 \in \mathbb{R},$$

into (6.2), we arrive at the equality

$$\partial_\lambda[\lambda\Gamma_{k,m}\mathrm{div}\,\mathbf{u}] - \mathrm{div}\,(\Gamma_{k,m}\mathbf{u}) - \partial_\lambda\mathbf{S}_k(\lambda\mathbf{T}_m\mathcal{M}) \tag{7.3}$$

$$+\mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_3 = 0, \tag{7.4}$$

which holds true in any subdomain of $D \setminus (S \cup \Sigma^-) \times \mathbb{R}$ for sufficiently large m. Here the remainders are given by

$$\mathfrak{r}_1 = \partial_\lambda\{\mathbf{T}_m(\lambda\Gamma_k\mathrm{div}\,\mathbf{u}) - \lambda\mathrm{div}\,\mathbf{u}\mathbf{T}_m\Gamma_k\}, \quad \mathfrak{r}_2 = \mathbf{T}_m\Big[\mathrm{div}\,\mathbf{u}\partial_\lambda\big(\mathbf{S}_k(\lambda\Gamma) - \lambda\mathbf{S}_k\Gamma\big)\Big],$$
$$\mathfrak{r}_3 = \mathrm{div}\,\big[(\mathbf{T}_m\Gamma_k)\mathbf{u} - \mathbf{T}_m(\Gamma_k\mathbf{u})\big].$$

Recall that $\Gamma_{k,m}$ are smooth functions of the variables x and λ. Multiplying both sides of (7.3) by $h\eta\Psi'(\Gamma_{k,n})$ and integrating the result over $D \setminus S$, we arrive at the equality

$$\int_{(D\setminus S)\times\mathbb{R}} \Big\{\eta(\lambda)\Psi(\Gamma_{k,m})\mathbf{u}(x)\nabla h(x) - \lambda h(x)\Psi(\Gamma_{k,m})\eta'(\lambda)\mathrm{div}\,\mathbf{u}(x)\Big\}dx\,d\lambda \tag{7.5}$$

$$+ J^{k,m} + \sum_{l=1}^{3} I_l^{k,m} = 0,$$

where

$$J^{k,m} = -\int_{(D\setminus S)\times\mathbb{R}} F^{k,m}\partial_\lambda\mathbf{S}_k(\lambda\mathbf{T}_m\mathcal{M}),$$

$$I_i^{k,m} = \int_{(D\setminus S)\times\mathbb{R}} F^{k,m}\mathfrak{r}_i \ \text{ with } \ F^{k,m} = h\eta\Psi'(\Gamma_{k,m})\,.$$

Our first task is to find the limits for $m \to \infty$ of the sequences $\{J^{k,m}\}$, $\{I_l^{k,m}\}$. It follows from inequality (7.2) that for $i = 1, 2$,

$$\|\mathfrak{r}_i\|_{L^2((D\setminus S)\times\mathcal{I})} \le k^c(\mathcal{I}) < \infty \ \text{ for all compact } \ \mathcal{I} \subset \mathbb{R}.$$

On the other hand, since η vanishes near ∞ and $\Gamma_{k,m}$ vanishes for $\lambda < 0$, the functions $F^{k,m}$ are uniformly compactly supported, and their absolute values are uniformly bounded. Hence for $i = 1, 2$,

$$\|F^{k,m}\mathfrak{r}_i\|_{L^2((\mathbf{D}\setminus S)\times\mathbb{R})} \le c < \infty \text{ for all } m \ge 1. \tag{7.6}$$

Noting that $\partial_\lambda\big[\lambda\Gamma_k(\cdot,\lambda)\big] \in L^\infty(D)$, taking into account the form of limits in (7.2), we obtain for fixed λ and k, and for $m \to \infty$,

$$\Gamma_{k,m} \to \Gamma_k, \quad \mathbf{T}_m\Big[\partial_\lambda(\lambda\Gamma_k)\mathrm{div}\,\mathbf{u}\Big] - \mathrm{div}\,\mathbf{u}\mathbf{T}_m\Big[\partial_\lambda(\lambda\Gamma_k)\Big] \to 0,$$
$$\mathbf{T}_m\Big[\mathrm{div}\,\mathbf{u}\partial_\lambda\big(\mathbf{S}_k(\lambda\Gamma) - \lambda\mathbf{S}_k\Gamma\big)\Big] \to \Big[\mathrm{div}\,\mathbf{u}\partial_\lambda\big(\mathbf{S}_k(\lambda\Gamma) - \lambda\mathbf{S}_k\Gamma\big)\Big],$$

a.e. in $D \setminus S$, which along with (7.6) yield for $m \to \infty$,

$$F^{k,m}\mathfrak{r}_1 \to 0, \quad F^{k,m}\mathfrak{r}_2 \to F^k\partial_\lambda\Big[\mathrm{div}\,\mathbf{u}\big(\mathbf{S}_k(\lambda\Gamma) - \lambda\mathbf{S}_k\Gamma\big)\Big] \ \text{ in } \ L^1((D\setminus S)\times\mathbb{R}). \tag{7.7}$$

The next evaluations are based on the following result from [4]: Let $\mathbf{u} \in H^{1,2}(D)$ and $g \in L^\infty\big(D \times (a,b)\big)$ with $b - a < \infty$. Then for all $D' \Subset D$,

$$\mathrm{div}\,\big[(\mathbf{T}_m g)\mathbf{u} - \mathbf{T}_m(g\mathbf{u})\big] \to 0 \ \text{ in } \ L^1(D'\times(a,b)).$$

Setting $g(x,\lambda) = \Gamma_k(x,\lambda)$ since $F^{k,m}$ are uniformly bounded and compactly supported in $(D\setminus S)\times\mathbb{R}$, we can conclude that for $m \to \infty$ the functions $F^{k,m}\mathfrak{r}_3$ tend to zero strongly in $L^1\big((D\setminus S)\times\mathbb{R}\big)$. Combining this result with (7.7), we obtain

$$\lim_{m\to\infty} I_1^{m,k} = \lim_{m\to\infty} I_3^{m,k} = 0,$$
$$\lim_{m\to\infty} I_2^{m,k} = I_2^k := \int_{(D\setminus S)\times\mathbb{R}} F^k\Big[\mathrm{div}\,\mathbf{u}\partial_\lambda\big(\mathbf{S}_k(\lambda\Gamma) - \lambda\mathbf{S}_k\Gamma\big)\Big]\,dx\,d\lambda.$$

Similar arguments show that

$$\lim_{m\to\infty} J^{m,k} = J^k := -\int_{(D\setminus S)\times\mathbb{R}} F^k\partial_\lambda\mathbf{S}_k(\lambda\mathcal{M})\,dx\,d\lambda \ \text{ with } \ F^k = h(x)\eta(\lambda)\Psi'(\Gamma_k). \tag{7.8}$$

Taking the limits, first for $m \to \infty$ and then for $k \to \infty$ in (7.5), we arrive at

$$\int_{(D\setminus S)\times\mathbb{R}} \Big\{\eta(\lambda)\Psi(\Gamma)\mathbf{u}(x)\nabla h(x) - \lambda h(x)\Psi(\Gamma)\eta'(\lambda)\mathrm{div}\,\mathbf{u}(x)\Big\}dx\,d\lambda + \lim_{k\to\infty} J^k + \lim_{k\to\infty} I_2^k = 0. \tag{7.9}$$

Let us show that the last term in the left-hand side equals 0. It is easily seen that

$$\partial_\lambda[(\lambda\Gamma)_k - \lambda\Gamma_k](x,\lambda) = k\int_{\mathbb{R}} \tilde\theta(k(\lambda-t))\Gamma(x,t)\,dt, \quad \tilde\theta(t) = t\theta'(t) + \theta(t).$$

Since $\tilde\theta$ is compactly supported and $\int_{\mathbb{R}} \tilde\theta(s)ds = 0$, the sequence of functions $\partial_\lambda[(\lambda\Gamma)_k - \lambda\Gamma_{k,}](x,\lambda)$ is bounded and converges to 0 a.e. in $(D\setminus S)\times\mathbb{R}$. Recalling that the functions F^k are uniformly supported, i.e., the supports spt F^k are included in the given compact subset of $(D\setminus S)\times\mathbb{R}$, and uniformly bounded in $(D\setminus S)\times\mathbb{R}$, we can conclude that $\lim_{k\to\infty} I_2^k = 0$. Next, integration by parts in the formula for J^k leads to

$$J^k = \int_{(D\setminus S)\times\mathbb{R}} h\eta'\Psi'(\Gamma_k)\mathbf{S}_k(\lambda\mathcal{M})dx\,d\lambda - 2W^k \tag{7.10}$$

with

$$W^k = \int_{(D\setminus S)\times\mathbb{R}} h\eta \mathbf{S}_k(\lambda\mathcal{M})\partial_\lambda\Gamma_k \, dx\, d\lambda. \tag{7.11}$$

Integration by part in the Stieltjes integral results in

$$\partial_\lambda\Gamma_k(x,\lambda) = k\partial_\lambda \int_{\mathbb{R}} \theta(k(\lambda - t)\Gamma(x,t)dt = k\int_{\mathbb{R}} \theta(k(\lambda - t)d_t\Gamma(x,t);$$

therefore,

$$\begin{aligned} W^k &= \int_{(D\setminus S)\times\mathbb{R}} h\mathbf{S}_k\Big[\eta\mathbf{S}_k(\lambda\mathcal{M})\Big]\, d_\lambda\Gamma(x,\lambda)\, dx \\ &\equiv \int_{D\setminus S} h(x)\left(\int_{\mathbb{R}} \eta\mathbf{S}_k^2(\lambda\mathcal{M})\, d_\lambda\Gamma(x,\lambda)\right) dx \\ &\quad + \int_{D\setminus S} h(x)\left(\int_{\mathbb{R}} [\mathbf{R}_k\mathbf{S}_k(\lambda\mathcal{M})](x,\lambda)\, d_\lambda\Gamma(x,\lambda)\right) dx \end{aligned} \tag{7.12}$$

with the linear operator $\mathbf{R}_k$ defined, for any $g \in L^\infty\big((D\setminus S)\times\mathbb{R}\big)$, by

$$\mathbf{R}_k g = \int_{\mathbb{R}} r_k(\lambda,t)g(x,t)dt, \quad \text{where} \quad r_k(\lambda,t) = k(\eta(t) - \eta(\lambda))\theta(k(\lambda - t)).$$

Since the function $\eta(\lambda)$ is smooth and vanishes for $\lambda > N = \sup\ \mathrm{spt}\ \eta$, the kernel $r_k(\lambda,t)$ vanishes for $\lambda > N+1$ and satisfies the inequality $\int_{\mathbb{R}} |r_k(\lambda,t)|\, dt \le ck^{-1}$. Thus, we get

$$|\mathbf{R}_k g| \le ck^{-1} \ \operatorname*{ess\,sup}_x \Big\{ \sup_{s\le N+1} |g(x,s)| \Big\},$$

which along with

$$|\mathbf{S}_k(\lambda\mathcal{M})|(x,\lambda) \le \sup_{s\le\lambda+1} s\mathcal{M}(x,s) \le (\lambda+1)\sup_{\mathbb{R}} \mathcal{M}(x,s) \le 2(\lambda+1)\overline{p}(x)$$

implies $|\mathbf{R}_k\mathbf{S}_k(\lambda\mathcal{M})| \le c(\eta)k^{-1}\overline{p}(x)$. Thus,

$$\left| \int_{\mathbb{R}} [\mathbf{R}_k\mathbf{S}_k(\lambda\mathcal{M})](x,\lambda)\, d_\lambda\Gamma(x,\lambda) \right| \le c(\eta)k^{-1}\overline{p}(x).$$

From this and the Lebesgue dominant convergence theorem we conclude that the last term in the right-hand side of (7.12) vanishes for $k \to \infty$. Recall that $\mathbf{S}_k^2$ is the standard mollifying operator with the kernel $\theta * \theta$. Hence for a.e. $x \in D\setminus S$ and for all $\lambda \in \mathbb{R}$, we have $\lim_{k\to\infty}[\mathbf{S}_k^2(\lambda\mathcal{M})](x,\lambda) = \lambda\mathfrak{M}(x,\lambda)$. From this and the Lebesgue dominant convergence theorem related to the measure $d_\lambda\Gamma(x,\cdot)$, we obtain that for a.e. $x \in D\setminus S$,

$$\lim_{k\to\infty} \int_{\mathbb{R}} \eta[\mathbf{S}_k^2(\lambda\mathcal{M})](x,\lambda)\, d_\lambda\Gamma(x,\lambda) = \int_{\mathbb{R}} \eta\lambda\mathfrak{M}\, d_\lambda\Gamma(x,\lambda).$$

Since the integrands in the right-hand side of (7.12) are uniformly bounded, we have

$$\lim_{k\to\infty} W^k = W := \int_{D\setminus S} h(x)\left(\int_{\mathbb{R}} \eta\lambda\mathfrak{M}\, d_\lambda\Gamma(x,\lambda)\right) dx.$$

On the other hand, for $k \to \infty$ the sequences Γ_k and $\mathbf{S}_k\mathcal{M}$ converge a.e. in $(D\setminus S)\times\mathbb{R}$ to Γ and $\mathcal{M}$, respectively; thus, by the limit passage $k \to \infty$ in (7.10), we obtain

$$\lim_{k\to\infty} J^k = \int_{(D\setminus S)\times\mathbb{R}} h\eta'\Psi'(\Gamma)\lambda\mathcal{M}\,dx\,d\lambda - 2W.$$

Substituting this equality into (7.9) and recalling that $\lim_{k\to\infty} I_2^k = 0$, we obtain (7.1). The proof of Lemma 7.1 is completed. □

8. The proof of Theorem 1.2. The theorem will be proved if we show that any sequence of generalized solutions to problem (1.3) satisfying the hypotheses of Theorem 1.2 and Lemmas 3.2 and 3.3 converges a.e. on $D \setminus S$. In light of Lemmas 3.2 and 3.3, it suffices to verify the equality $\Psi(\Gamma) = 0$ in $(D \setminus S) \times \mathbb{R}$.

We begin with proving that renormalized integral identity (7.1) after substituting $h = 1$ turns into the integral inequality

$$\begin{aligned}\int_{(D\setminus S)\times\mathbb{R}} \Big\{\lambda\Psi'(\Gamma)\mathcal{M}\eta' - \lambda\Psi(\Gamma)\eta'\operatorname{div}\mathbf{u}\Big\}dx\,d\lambda \\ \geq 2\int_{(D\setminus S)} \left(\int_{[0,\infty)} \eta(\lambda)\lambda\mathfrak{M}(x,\lambda)d_\lambda\Gamma(x,\lambda)\right)dx,\end{aligned} \tag{8.1}$$

which holds true for all nonnegative functions $\eta \in C^\infty(\mathbb{R})$ vanishing near $+\infty$. The proof is based on the following approximation result, which is shown by an application of the Hedberg approximation theorem [13].

LEMMA 8.1. *For each $k > 1$, there exist a function $\zeta_k \in C^\infty(D)$ and a constant c independent of k such that ζ_k vanishes in a vicinity of S and*

$$0 \leq \zeta_k \leq 1, \quad \text{meas } A_k + \int_D |\nabla\zeta_k\mathbf{u}|\,dx \leq 1/k, \tag{8.2}$$

where $A_k = \{x \in D \setminus S : \zeta_k(x) \leq 1 - 1/k\} \subset D \setminus S$.

Proof. Introduce the function $f(x) = |\mathbf{u}(x)| + \text{dist }(x, S)$. Since $\mathbf{u} - \mathbf{u}^\infty$ belongs to $H_0^{1,2}(D)$ and $\mathbf{u}^\infty$ vanishes near S, it follows from the general theory of Sobolev spaces (see [28]) that f has a quasi-continuous representative which vanishes everywhere on S. By the Hedberg approximation theorem (see [13]), for each $N > 0$ there is $\omega_N \in C_0^\infty(D)$ so that $\omega_N = 1$ in some vicinity of S, $0 \leq \omega_N \leq 1$, and $\|(1-\omega_N)f\|_{H^{1,2}(D)} \leq N^{-1}$. On the other hand, by the absolute continuity property of the integral, there exists a positive continuous function $\mu(s)$ such that, whenever meas $A \leq \delta$, then

$$\int_A (f^2 + |\nabla f|^2)dx \leq \mu(\delta) \qquad \text{and} \quad \mu(\delta) \searrow 0 \quad \text{for} \quad \delta \searrow 0.$$

For any $\delta > 0$, set $S_\delta = \{x : \text{ dist } (x, S) < \delta\}$. Since $\cap S_\delta = S$, the measure of the set $S_\delta \setminus S$ tends to 0 as $\delta \to 0$, which yields

$$\int_{S_\delta} (f^2 + |\nabla f|^2)dx = \int_{S_\delta\setminus S} (f^2 + |\nabla f|^2)dx \leq \mu\big(\text{meas }(S_\delta \setminus S)\big) := r(\delta) \searrow 0 \quad \text{as} \quad \delta \searrow 0.$$

Obviously, for $B_k = \{x \in D \setminus S : 1 - \omega_N(x) > 1/k\}$, we have

$$\frac{\delta^2}{k^2}\int_{B_k\setminus S_\delta} dx \leq \int_D |f|^2(1-\omega_N)^2\,dx \leq N^{-2},$$

which implies

$$\text{meas } \{B_k \setminus S_\delta\} \le k^2(\delta N)^{-2}, \quad \text{meas } B_k \le k^2(\delta N)^{-2} + \text{meas } (S \setminus S_\delta). \tag{8.3}$$

Next, note that

$$\begin{aligned}\int_D (1-\omega_N)^2 |\nabla f|^2\, dx &\le \frac{1}{k^2}\int_{D\setminus B_k} |\nabla f|^2 dx + \int_{B_k\setminus S_\delta} |\nabla f|^2\, dx + \int_{S_\delta} |\nabla f|^2\, dx \\ &\le ck^{-2} + \mu(k^2(\delta N)^{-2}) + r(\delta),\end{aligned}$$

which along with the identity $f\nabla\omega_N = (1-\omega_N)\nabla f - \nabla\big((1-\omega_N)f\big)$ yields the estimate

$$\begin{aligned}\int_D |\nabla\omega_N| f\, dx &\le \|\nabla\big((1-\omega_N)f\big)\|_{L^1(D)} + \int_D (1-\omega_N)|\nabla f|\, dx \\ &\le c\|(1-\omega_N)f\|_{H^{1,2}(D)} + c\left\{\int_D (1-\omega_N)^2|\nabla f|^2\, dx\right\}^{1/2} \\ &\le N^{-1} + ck^{-1} + c\mu^{1/2}\big(k^2(\delta N)^{-2}\big) + cr^{1/2}(\delta).\end{aligned}$$

Fix $\delta > 0$ sufficiently small such that $r^{1/2}(\delta) + \text{meas } (S \setminus S_\delta) \le k^{-1}$; next, choose N sufficiently large such that $\mu^{1/2}(k^2(\delta N)^{-2}) + k^2(\delta N)^{-2} < k^{-1}$. In view of (8.3) and taking into account that $|\mathbf{u}| \le f$, we arrive at

$$\int_D |\nabla\omega_N||\mathbf{u}|\, dx + \text{meas } B_k \le ck^{-1}.$$

Hence $\zeta_k = 1 - \omega_N$ with $A_k = B_k$ satisfies (8.2). The proof of Lemma 8.1 is completed. □

Following [10] let us consider the sequence of functions $\chi_k(x) = \chi(k \text{ dist } (x, \Sigma^- \cup \partial D))$ with an arbitrary, smooth, monotone function χ such that $\chi(z) = 0$ for $z \le 1/2$ and $\chi(z) = 1$ for $z \ge 1$. Since $\Psi(\Gamma)(\cdot,\lambda)$ vanishes outside of B, we have for all sufficiently large k,

$$\nabla\chi_k(\mathbf{u} - \mathbf{u}^\infty) \to 0 \ \text{ in } \ L^1(D\setminus S), \quad \Psi(\Gamma)(\cdot,\lambda)\nabla\chi_k\mathbf{u}^\infty \le 0 \text{ in } D. \tag{8.4}$$

Set $h_k = \zeta_k\chi_k$ and note that

$$\nabla h_k\mathbf{u} \le |\nabla\zeta_k\mathbf{u}| + |\nabla\chi_k(\mathbf{u} - \mathbf{u}^\infty)| + \zeta_k\nabla\chi_k\mathbf{u}^\infty.$$

Recalling the inequality $\eta \ge 0$ and relations (8.2), (8.4), we obtain

$$\limsup_{k\to\infty} \int_{(D\setminus S)\times\mathbb{R}} \Psi(\Gamma)\eta\nabla h_k\mathbf{u}\, dx\, d\lambda \le 0. \tag{8.5}$$

Moreover, the functions h_k converge to 1 in measure on $D \setminus S$. The functions h_k are Lipschitz continuous in $\mathbb{R}^3$ and vanish in the vicinity of $\Sigma^- \cup S$; therefore, they can be used as test functions in (7.1). Substitution of h_k into (7.1), followed by the limit passage $k \to \infty$ in the resulting integral identity, leads to desired inequality (8.1).

Next, we claim that the right-hand side of (8.1) equals zero. To this end, choose an arbitrary nonnegative function $\upsilon \in C^\infty(\mathbb{R})$ with spt $\upsilon \subset (-1,1)$ and $\int_{\mathbb{R}} \upsilon(\lambda)d\lambda = 1$. For fixed $t > 2$, set $\eta(\lambda) = \int_\lambda^\infty \upsilon(s-t)\, ds$. Since $\eta'(\lambda) = 0$ for $|\lambda - t| \ge 1$, and $\eta(\lambda) = 1$ for $\lambda \le t - 1$, we can use (8.1) to obtain

(8.6)
$$2\int_{(D\setminus S)}\Big(\int_{[0,t-1)}\lambda\mathfrak{M}d_\lambda\Gamma\Big)dx\le-(t+1)\int_{(D\setminus S)\times\mathbb{R}}\Big\{\mathcal{M}+\Psi(\Gamma)|\operatorname{div}\mathbf{u}(x)|\Big\}\eta'\,dx\,d\lambda.$$

Using identity (6.8) and the relation

$$\begin{aligned}&-\int_{(D\setminus S)\times\mathbb{R}}\Psi(\Gamma)\ \eta'(\lambda)\ |\operatorname{div}\mathbf{u}|\,dx\,d\lambda\\&\quad\equiv\int_{[0,\infty)}\eta''\Big\{\int_{D\setminus S}\Big[\int_{\Omega\times[0,\lambda)}\Psi(\Gamma(x,s))\,ds\Big]|\operatorname{div}\mathbf{u}|\,dx\Big\}d\lambda\end{aligned}$$

we can rewrite inequality (8.6) in the form

$$2\int_{(D\setminus S)}\Big(\int_{[0,t-1)}\lambda\mathfrak{M}d_\lambda\Gamma\Big)dx\le(1+t)\int_{[1,\infty)}\eta''(\lambda)\wp(\lambda)d\lambda\ ,\tag{8.7}$$

where the function $\wp:[0,\infty)\mapsto\mathbb{R}$ is given by

$$\wp(\lambda)=\int_{(D\setminus S)\times[0,\lambda)}\Psi(\Gamma(x,s))|\operatorname{div}\mathbf{u}(x)|\,dx\,ds+\int_\Omega\mathcal{V}_\lambda(x)dx.$$

Since

$$\int_{[0,\lambda)}\Psi(\Gamma(x,s))\,ds\le\int_{[0,\infty)}\Psi(\Gamma(x,s))\,ds=\mathfrak{H}(x),$$

Lemma 5.2 implies the boundedness of $\wp$ on $\mathbb{R}^+$,

$$|\wp(\lambda)|\le c\|\mathbf{u}\|_{H^{1,2}(D\setminus S)}\,\|\mathfrak{H}\|_{L^2(D\setminus S)}+\|\mathcal{V}_\lambda\|_{L^1(D\setminus S)}\le c.$$

Taking into account that $\eta''(\lambda)=\partial_t\upsilon(\lambda-t)$, inequality (8.7) can be rewritten in the form

$$2\int_{(D\setminus S)}\Big(\int_{[0,t-1)}\lambda\mathfrak{M}d\Gamma(x,\lambda)\Big)dx\le(1+t)\frac{d}{dt}(\upsilon*\wp)(t).\tag{8.8}$$

Since the smooth function $(\upsilon*\wp)(t)$ is uniformly bounded on $\mathbb{R}^+$, there is a sequence $t_k\to\infty$ such that $\lim_{k\to\infty}(t_k+1)\frac{d}{dt}(\upsilon*\wp)(t_k)\le0$. Substitution of $t=t_k$ into (8.8) followed by the limit passage $k\to\infty$ in (8.8) leads to

$$\int_{[0,t-1)}\lambda\mathfrak{M}d\Gamma(x,\lambda)=0\quad\text{for a.e. }\ x\in D\setminus S.$$

In other words, $\mathcal{M}(x,\cdot)$ vanishes σ_x-a.e. on $(0,\infty)$, which along with Lemma 6.2 implies the equality $\Gamma(1-\Gamma)=0$ a.e. in $(D\setminus S)\times\mathbb{R}$. Hence ρ_n converges a.e. in $D\setminus S$. Estimates (1.11) imply the strong convergence of the sequence ρ_n in $L^r(D\setminus S)$ for all $r<\gamma$. Since $(\rho_n,\mathbf{u}_n)$ satisfy all assumptions of Theorem 2.1, we can make use of estimate (2.1) from this theorem, which yields the strong convergence ρ_n in $L^r_{loc}(\Omega)$ for $r<\gamma(1+\kappa)$. In particular, the sequence $p(\rho_n)$ converges to $\overline{p}=p(\rho)$ in $L^1_{loc}(\Omega)$. After substituting $(\rho_n,\mathbf{u}_n)$ into integral identities (1.7a) and (1.7b), followed by the limit passage $n\to\infty$, by Lemma 3.3 we can conclude that the pair $(\rho,\mathbf{u})$ is a generalized solution to problem (1.3). Finally, since $\nabla\mathbf{u}^\infty$ is compactly supported in Ω, the limit passage in (1.12) for the sequence of drag functionals follows by Lemma 3.3 and by the strong convergence of the sequence $p(\rho_n)$ in $L^1_{loc}(\Omega)$.

REFERENCES

[1] D. Adams, *On the existence of capacitary strong type estimates in R^n*, Ark. Mat., 14 (1976), pp. 125–140.

[2] D. Adams and L. Hedberg, *Function Spaces and Potential Theory*, Springer-Verlag, Berlin, 1995.

[3] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDEs and Continuum Models of Phase Transitions, Lecture Notes in Phys. 344, Springer, Berlin, 1989, pp. 207–215.

[4] R. J. DiPerna and P. L. Lions, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 48 (1989), pp. 511–547.

[5] R. E. Edwards, *Functional Analysis. Theory and Applications*, Holt, Rinehart, and Wilson, New York, Toronto, London, 1965.

[6] E. Feireisl, *On compactness of solutions to the compressible isentropic Navier–Stokes equations when the density is not square integrable*, Comment. Math. Univ. Carolin., 42 (2001), pp. 83–98.

[7] E. Feireisl, Š. Matušů-Nečasová, H. Petzeltová, and I. Straškraba, *On the motion of a viscous compressible fluid driven by a time-periodic external force*, Arch. Ration. Mech. Anal., 149 (1999), pp. 69–96.

[8] E. Feireisl, A. H. Novotný, and H. Petzeltová, *On the existence of globally defined weak solutions to the Navier–Stokes equations* J. Math. Fluid Mech., 3 (2001), pp. 358–392.

[9] E. Feireisl, A. H. Novotný, and H. Petzeltová, *On the domain dependence of solutions to the compressible Navier–Stokes equations of a barotropic fluid*, Math. Methods Appl. Sci., 25 (2002), pp. 1045–1073.

[10] E. Feireisl, *Shape optimization in viscous compressible fluids*, Appl. Math. Optim., 47 (2003), pp. 59–78.

[11] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford University Press, Oxford, UK, 2004.

[12] J. Frehse, S. Goj, and M. Steinhauer, *L^p-estimates for the Navier–Stokes equations for steady compressible flow*, Manuscripta Math., 116 (2005), pp. 265–275.

[13] L. Hedberg, *Spectral synthesis in Sobolev spaces and uniqueness of solutions of the Dirichlet problem*, Acta Math., 147 (1981), pp. 237–264.

[14] B. Kawohl, O. Pironneau, L. Tartar, and J. Zolesio, *Optimal Shape Design*, Lecture Notes in Math. 1740, Springer-Verlag, Berlin, 2000.

[15] J. R. Kweon and R. B. Kellogg, *Regularity of solutions to the Navier–Stokes system for compressible flows on a polygon*, SIAM J. Math. Anal., 35 (2004), pp. 1451–1485.

[16] J. R. Kweon and R. B. Kellogg, *Regularity of solutions to the Navier–Stokes equations for compressible barotropic flows on a polygon*, Arch. Ration. Mech. Anal., 163 (2002), pp. 35–64.

[17] J. R. Kweon and R. B. Kellogg, *Compressible Stokes problem on nonconvex polygonal domains*, J. Differential Equations, 176 (2001), pp. 290–314.

[18] P. L. Lions, *Mathematical Topics in Fluid Dynamics. Vol.* 2. *Compressible Models*, Oxford University Press, Clarendon Press, New York, 1998.

[19] P. L. Lions, *On some challenging problems in nonlinear partial differential equations*, in Mathematics: Frontiers and Perspectives, V. Arnold et al., eds., AMS Providence, RI, 2000, pp. 121–135.

[20] J. Malek, J. Necas, M. Rokyta, and M. Ruzicka, *Weak and Measure-Valued Solutions to Evolutionary PDEs*, Chapman and Hall, London, 1996.

[21] A. Novotný and M. Padula, *Existence and uniqueness of stationary solutions for viscous compressible heat conductive fluid with large potential and small nonpotential external forces*, Siberian Math. J., 34 (1993), pp. 120–146.

[22] A. Novotný and I. Straškraba, *Introduction to the Mathematical Theory of Compressible Flow*, Oxford Lecture Ser. Math. Appl. 27, Oxford University Press, Oxford, UK, 2004.

[23] P. Pedregal, *Parametrized Measures and Variational Principles*, Progr. Nonlinear Differential Equations Appl. 30, Birkhäuser Verlag, Basel, 1997.

[24] P. I. Plotnikov and J. Sokolowski, *On compactness, domain dependence, and existence of steady state solutions to compressible isothermal Navier–Stokes equations*, J. Math. Fluid Mech., 7 (2005), pp. 529–573.

[25] P. I. Plotnikov and J. Sokolowski, *Concentrations of solutions to time-discretized compressible Navier–Stokes equations*, Comm. Math. Phys., 258 (2005), pp. 567–608.

[26] P. I. Plotnikov and J. Sokolowski, *Stationary boundary value problems for Navier–Stokes equations with adiabatic index* $\nu < 3/2$, Dokl. Math., 70 (2004), pp. 535–538. Translated from Doklady Akademii Nauk, Volume 397, $N.^0$, 1-6, 2004.

[27] P. I. Plotnikov and J. Sokolowski, *Shape sensitivity analysis for compressible Navier–Stokes equations*, in preparation.

[28] W. P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1198–1221

FEASIBLE AND NONINTERIOR PATH-FOLLOWING IN CONSTRAINED MINIMIZATION WITH LOW MULTIPLIER REGULARITY*

M. HINTERMÜLLER† AND K. KUNISCH†

Abstract. Primal-dual path-following methods for constrained minimization problems in function space with low multiplier regularity are introduced and analyzed. Regularity properties of the path are proved. The path structure allows us to define approximating models, which are used for controlling the path parameter in an iterative process for computing a solution of the original problem. The Moreau–Yosida regularized subproblems of the new path-following technique are solved efficiently by semismooth Newton methods. The overall algorithmic concept is provided, and numerical tests (including a comparison with primal-dual path-following interior point methods) for state constrained optimal control problems show the efficiency of the new concept.

Key words. active set strategy, Moreau–Yosida regularization, path-following methods, primal-dual methods, semismooth Newton methods

AMS subject classifications. 49M15, 49M37, 65K05, 90C33

DOI. 10.1137/050637480

1. Introduction.

The efficient numerical solution of inequality constrained minimization problems in function space with low Lagrange multiplier regularity is still a significant challenge. Prototype problems include optimal control problems with pointwise state constraints, boundary control problems in which the control has to satisfy pointwise constraints on the boundary, and classes of control problems with "solutions très faibles" (that is, "very weak solutions" in the sense of Lions). The common feature is low regularity of the Lagrange multipliers associated with the pointwise constraints. This has an immediate effect when characterizing optimality of solutions. Indeed, in the presence of inequality constraints, first order optimality conditions typically involve a so-called *complementarity system* (see, e.g., [10]), which in turn is influenced by the regularity of the Lagrange multiplier associated with the inequality constraint. To be specific, let us assume that $x_1 \in X_1$ has to satisfy the pointwise (almost everywhere) constraint

$$x_1 \le \psi, \tag{1.1}$$

where X_1 denotes a Hilbert space continuously embedded into $L^2(\omega)$ and $\omega \subset \mathbb{R}^m$ is a bounded domain. Further $\psi \in L^2(\omega)$, and $\le$ represents the natural ordering in $L^2(\omega)$. Let $\lambda \in X_1^*$ denote the Lagrange multiplier associated with constraint (1.1), where X_1^* is the topological dual space of X_1. First order optimality characterizations include the complementarity condition

$$x_1 \le \psi, \quad \langle \lambda, x_1 - \xi \rangle_{X_1^*, X_1} \ge 0 \quad \text{for all } \xi \le \psi,\ \xi \in X_1 \tag{1.2}$$

at an optimal solution x_1 with associated Lagrange multiplier λ. Without additional regularity, system (1.2) does not admit a pointwise interpretation, which is frequently crucial for numerical algorithms.

*Received by the editors August 3, 2005; accepted for publication (in revised form) April 26, 2006; published electronically September 12, 2006.
http://www.siam.org/journals/sicon/45-4/63748.html

†Department of Mathematics and Scientific Computing, University of Graz, Heinrichstraße 36, A-8010 Graz, Austria (michael.hintermueller@uni-graz.at, karl.kunisch@uni-graz.at).

In fact, solution techniques and their (local) convergence behavior often hinge on the multiplier regularity. Classical active set methods, for instance, require a pointwise (almost everywhere) interpretation of λ for the active set estimation. In the case of pointwise constraints, techniques like the projected gradient methods will not work without modification since the sum of the iteration variable and the gradient of the objective, which coincides with the negative multiplier, is needed for the update. Since the iteration variable and the gradient have different regularity properties, this is not feasible in general. An analogous comment applies for projected Newton techniques. Recently it was found that semi smooth Newton methods are highly efficient in solving certain classes of constrained optimization problems in function space [4, 5, 9, 10, 11, 21]. These methods rely on a pointwise almost everywhere interpretation of the complementarity system (1.2) and smoothing properties of the control-to-adjoint-state mapping. In fact, the pointwise interpretation allows us to express (1.2) equivalently as

$$\lambda - \max(0, \lambda + c(x_1 - \psi)) = 0, \tag{1.3}$$

for some arbitrarily fixed $c > 0$, and the smoothing of the control-to-adjoint-state operator typically implies that the mapping

$$\theta : x_1 \mapsto \lambda(x_1) + c(x_1 - \psi)$$

can be considered as $\theta : X_1 \to L^q(\omega)$ with $q > 2$. The norm gap between $L^q(\omega)$ and the space $L^2(\omega)$, in which inequality (1.1) is posed, is crucial in proving generalized differentiability of

$$x_1 \mapsto \max(0, \theta(x_1)),$$

and in arguing well-definedness and locally superlinear convergence of the generalized (semismooth) Newton method for solving the underlying nonsmooth first order optimality system; see [10] for details. Again, the low multiplier regularity may prevent the pointwise interpretation and/or the smoothing of the control-to-adjoint-state mapping.

An approach for solving state constrained optimal control problems that does not rely on the use of multipliers was introduced in [13]. This method operates with the interface (boundary) between the active set $\{x_1 = \psi\}$ and the inactive set $\{x_1 < \psi\}$ as the optimization variable, and the constrained minimization problem is transformed into a shape optimization problem. Since the interface allows for a unique identification of the inactive region, the multiplier itself is not an issue. While this technique is appealing due to its favorable analytical properties, the implementation is rather technical.

Based on recent work [12], in this paper we propose a primal-dual path-following concept for solving the aforementioned constrained minimization problems. It relies on a (generalized) Moreau–Yosida-type regularization of the max-operation involving a scalar parameter. The resulting regular subproblems can be solved efficiently by, e.g., semismooth Newton methods. The relaxation parameter induces a primal-dual path and a path value functional, for which good low-parametric models can be found based on the structure of the relaxation term. These models are subsequently used for driving the path parameter to its limit, i.e., to find a solution of the original (less regular) problem. This procedure has several analytical as well as numerical benefits as follows:

- sufficiently regular subproblems for which standard methods (like semismooth Newton algorithms) converge rapidly in function space setting;

- simple path structure such that one can find good approximating models for the primal-dual path value functional;
- controlled path-parameter updates based on model functions to avoid ill-conditioning;
- wide applicability.

Compared to other path-following concepts, like primal-dual path-following interior point methods (see, e.g., [6, 22, 25, 26] for finite dimensional versions and [20, 23] for function space treatments), the numerical implementation of our technique is rather user-friendly. In fact, in many cases (see, e.g., [1, 25]) competitive primal-dual path-following interior point methods require the addition of slack variables, which increases the problem size, and then appropriate pivot choices for reducing the indefinite system, which has to be solved in every iteration, in order to make the method feasible for large scale optimization. Additionally, within our path-following framework existing subproblem solvers can readily be used. In contrast to our approach, the successful analysis of interior point methods in function space [20, 23, 24] requires sufficient regularity of the multipliers, which prevents an immediate application of these concepts to the problem class considered in this paper. For an alternative regularization concept for state constrained optimal control problems, we refer to [19].

In the case of regular Lagrange multipliers, our earlier work [1, 10] indicates that semismooth Newton and primal-dual active set methods are superior to path-following strategies. This includes a wide class of pointwise control constraints in the optimal control of partial differential equations.

The rest of the paper is organized as follows. In the next section we introduce the problem class under consideration, specify the relaxed path-problems, and analyze properties of the primal-dual path. In section 3 we discuss several constrained optimal control problems which are covered by our model problem. Among these we consider state constrained optimal control problems and Dirichlet boundary control problems. The primal-dual path value function, its differentiability properties, and the definition of approximating models are the contents of section 4. The algorithm is introduced in section 5. A report on test runs, including a comparison of the new method with primal-dual path-following interior point and primal-dual active set techniques, is also given.

2. Problem formulation and properties of the path. Let X_1, X_2, and W be real Hilbert spaces with

$$X_1 \hookrightarrow L^2(\omega) \hookrightarrow X_1^*,$$

where X_1^* denotes the dual of X_1 and ω a bounded domain in $\mathbb{R}^m$. Further set $X = X_1 \times X_2$ and let $x = (x_1, x_2)$ denote a generic element in X.

Let $E \in \mathcal{L}(X, W)$, $f \in W$, and $\psi \in X_1$. Further let $J : X \to \mathbb{R}$ denote a quadratic functional satisfying the following assumption.

ASSUMPTION 2.1. *There exists a constant $\alpha > 0$ such that*

$$\langle J'(x) - J'(y), x - y\rangle_{X^*,X} \geq \alpha |x - y|_X^2 \tag{2.1}$$

whenever $E(x - y) = 0$ for $x, y \in X$.

Here $\langle\cdot,\cdot\rangle_{X^*,X}$, at times denoted by $\langle\cdot,\cdot\rangle$, stands for the duality pairing between X and X^*. We set

$$C := \begin{pmatrix} E \\ (I, 0) \end{pmatrix} : X \to W \times X_1$$

and assume that

$$C \text{ is surjective.} \tag{2.2}$$

The problem under consideration is

$$\begin{aligned} &\text{minimize } J(x) \text{ over } x \in X \\ &\text{subject to } Ex = f, \\ &\qquad x_1 \le \psi, \end{aligned} \tag{P}$$

where $\le$ denotes the ordering in $L^2(\omega)$. By Assumption 2.1 and (2.2) there exists a unique solution $x^* \in X$ to (P). Further by (2.2) there exists [17] a unique Lagrange multiplier pair $(p^*, \lambda^*) \in W^* \times X_1^*$ such that

$$\begin{aligned} &J'(x^*) + E^*p^* + (\lambda^*, 0) = 0 \text{ in } X^*, \\ &\lambda^* \ge 0, \quad x_1^* - \psi \le 0, \quad \langle \lambda^*, x_1^* - \psi \rangle_{X_1^*, X_1} = 0, \end{aligned} \tag{2.3}$$

where $\lambda^* \ge 0$ stands for $\langle \lambda^*, \phi \rangle_{X_1^*, X_1} \ge 0$ for all $\phi \in X_1$ with $\phi \ge 0$. In section 3 we give examples from optimal control with control or state constraints which are covered by our general framework.

We also consider the following problem without inequality constraints:

$$\begin{aligned} &\text{minimize } J(x) \text{ over } x \in X \\ &\text{subject to } Ex = f, \end{aligned} \tag{$\tilde{\text{P}}$}$$

as well as the regularized problems

$$\begin{aligned} &\text{minimize } J(x, \gamma) := J(x) + \frac{1}{2\gamma} \int_\omega |\left(\bar\lambda + \gamma(x_1 - \psi)\right)^+|^2 dw \text{ over } x \in X \\ &\text{subject to } Ex = f, \end{aligned} \tag{P_γ}$$

where $\gamma > 0$ represents a relaxation (or regularization) parameter and $\bar\lambda \in L^2(\omega)$ is an optional shift-parameter. For every $\gamma > 0$ there exists a unique solution $x_\gamma = (x_{1,\gamma}, x_{2,\gamma})$ to (P_γ) satisfying

$$\begin{aligned} &Ex_\gamma = f, \\ &J'(x_\gamma) + E^*p_\gamma + (\lambda_\gamma, 0) = 0 \text{ in } X^*, \\ &\lambda_\gamma = \left(\bar\lambda + \gamma(x_{1,\gamma} - \psi)\right)^+ \text{ in } L^2(\omega). \end{aligned} \tag{2.4}$$

In view of (1.3), the last equation in (2.4) suggests that $\bar\lambda$ acts as a regular approximation of the Lagrange multiplier corresponding to $x_1 \le \psi$. Further, in section 3 we find that in some cases, $\bar\lambda$ can be chosen such that $x_{1,\gamma} \le \psi$ for γ sufficiently large.

We refer to

$$\mathcal{C} = \{(x_\gamma, p_\gamma, \lambda_\gamma) \in X \times W^* \times X_1^* : \gamma \in (0, \infty)\}$$

as the *primal-dual path* associated with (P). For $r > 0$ we further set

$$\mathcal{C}_r = \{(x_\gamma, p_\gamma, \lambda_\gamma) : \gamma \in [r, \infty)\}.$$

PROPOSITION 2.1. *For every $r > 0$, the path $\mathcal{C}_r$ is bounded and, as $\gamma \to \infty$, we have $(x_\gamma, p_\gamma, \lambda_\gamma) \rightharpoonup (x^*, p^*, \lambda^*)$ in $X \times W^* \times X_1^*$. Moreover $x_\gamma \to x^*$ in X.*

Proof. From (2.4) we have

$$\langle J'(x_\gamma) - J'(x^*), x_\gamma - x^*\rangle_{X^*,X} + (\lambda_\gamma, x_{1,\gamma} - x_1^*)_{L^2} = \langle J'(x^*), x^* - x_\gamma\rangle_{X^*,X},$$

and by (2.1),

$$\alpha|x_\gamma - x^*|_X^2 + (\lambda_\gamma, x_{1,\gamma} - x_1^*)_{L^2} \le \langle J'(x^*), x^* - x_\gamma\rangle_{X^*,X}. \tag{2.5}$$

As in [14],

$$(\lambda_\gamma, x_{1,\gamma} - x_1^*)_{L^2} \ge \frac{1}{\gamma}|\lambda_\gamma|_{L^2}^2 - \frac{1}{\gamma}(\lambda_\gamma, \bar\lambda)_{L^2}, \tag{2.6}$$

and hence,

$$\alpha|x_\gamma - x^*|_X^2 + \frac{1}{2\gamma}|\lambda_\gamma|_{L^2}^2 \le |J'(x^*)|_{X^*}|x^* - x_\gamma|_X + \frac{1}{2\gamma}|\bar\lambda|_{L^2}^2. \tag{2.7}$$

In particular we have that $\{x_\gamma\}_{\gamma\ge r}$ is bounded for every $r > 0$. By (2.2),

$$C^*(p_\gamma, \lambda_\gamma) = -J'(x_\gamma),$$

and since C is surjective it follows that $\{(p_\gamma, \lambda_\gamma)\}$ is bounded in $W^* \times X_1^*$. Hence, there exists $(\hat x, \hat p, \hat\lambda)$ such that for a subsequence, $(x_\gamma, p_\gamma, \lambda_\gamma) \rightharpoonup (\hat x, \hat p, \hat\lambda)$. Since $\lambda_\gamma \ge 0$ for all $\gamma > 0$, we find $\langle\hat\lambda, \phi\rangle_{X_1^*,X_1} \ge 0$ for all $\phi \in X_1$ with $\phi \ge 0$. As in [14, section 3], one argues that $\hat x \le \psi$. By (2.3) and (2.4),

$$\langle J'(x_\gamma) - J'(x^*), x_\gamma - x^*\rangle_{X^*,X} + \langle\lambda_\gamma - \lambda^*, x_{1,\gamma} - x_1^*\rangle_{X_1^*,X_1} = 0, \tag{2.8}$$

and from (2.6) we have $(\lambda_\gamma, x_{1,\gamma} - x_1^*) \ge -\frac{1}{2\gamma}|\bar\lambda|_{L^2}^2$. Further, from (2.8) with Assumption 2.1 holding we arrive at

$$0 \le \alpha|x_\gamma - x^*|_X^2 \le \frac{1}{2\gamma}|\bar\lambda|_{L^2}^2 + \langle\lambda^*, x_{1,\gamma} - x^*\rangle_{X_1^*,X_1}.$$

Passing to the limit as $\gamma \to \infty$ yields

$$0 \le \alpha \limsup_{\gamma\to\infty} |x_\gamma - x^*|_X^2 = \langle\lambda^*, \hat x_1 - x_1^*\rangle_{X_1^*,X_1} = \langle\lambda^*, \hat x_1 - \psi\rangle_{X_1^*,X_1} \le 0,$$

and therefore $\lim_{\gamma\to\infty} x_\gamma = x^*$. Taking the limit in (2.4), we find

$$\langle J'(x^*), \phi\rangle_{X^*,X} + \langle E^*\hat p, \phi\rangle_{X^*,X} + \langle\hat\lambda, \phi_1\rangle_{X_1^*,X_1} = 0$$

for all $\phi = (\phi_1, \phi_2) \in X$. This equation also holds with $(\hat p, \hat\lambda)$ replaced by (p^*, λ^*). Since C^* is injective, we have $(\hat p, \hat\lambda) = (p^*, \lambda^*)$. Since the accumulation point of every weakly convergent subsequence of $\{(x_\gamma, p_\gamma, \lambda_\gamma)\}$ is (x^*, p^*, λ^*), the whole sequence converges weakly. ∎

We point out that condition (2.2) is a convenient sufficient condition, which is used for the existence of a Lagrange multiplier and for the a priori estimate just below (2.7). It holds for a reasonably wide class of interesting applications, as will be shown in section 3. In more involved applications, these two consequences of (2.2) must be argued by utilizing the specific properties of the underlying problem.

Next we treat the case $\bar{\lambda} = 0$ and $\gamma \to 0^+$, i.e., convergence towards the problem without inequality constraints.

PROPOSITION 2.2. *If $\bar{\lambda} = 0$, then the path $\mathcal{C}_0$ is bounded in $X \times W^* \times X_1^*$ and $\lim_{\gamma \to 0^+}(x_\gamma, p_\gamma, \lambda_\gamma) = (\tilde{x}, \tilde{p}, 0)$, where $\tilde{x}$ is the solution to $(\tilde{\mathrm{P}})$ with Lagrange multiplier $\tilde{p}$ associated with the constraint $Ex = f$.*

Proof. By (2.4) and (2.7), the family $\{(x_\gamma, p_\gamma, \lambda_\gamma)\}_{\gamma>0}$ is bounded. Moreover $\lim_{\gamma \to 0^+} \lambda_\gamma = 0$ in $L^2(\omega)$ by (2.7). Taking the limit $\gamma \to 0^+$ in (2.4), we find that each accumulation point of $\{(x_\gamma, p_\gamma, \lambda_\gamma)\}_{\gamma>0}$ must satisfy

$$J'(\tilde{x}) + E^*\tilde{p} = 0, \quad E\tilde{x} = f.$$

This implies the claim. ∎

Now we study smoothness properties of the primal-dual path.

PROPOSITION 2.3. *The primal-dual path $\mathcal{C}_r$ is globally Lipschitz continuous for every $r > 0$, and $\gamma \mapsto \lambda_\gamma$ is locally Lipschitz continuous in $L^2(\omega)$. If $\bar{\lambda} = 0$, then $\mathcal{C}_0$ is globally Lipschitz continuous.*

Proof. Let $\gamma, \bar{\gamma} \in [r, \infty)$. Then

(2.9)
$$J'(x_\gamma) - J'(x_{\bar{\gamma}}) + E^*(p_\gamma - p_{\bar{\gamma}}) + \big(\bar{\lambda} + \gamma(x_{1,\gamma} - \psi)\big)^+ - \big(\bar{\lambda} + \bar{\gamma}(x_{1,\bar{\gamma}} - \psi)\big)^+ = 0,$$

and hence with Assumption 2.1 holding,

$$\alpha|x_\gamma - \bar{x}_\gamma|_X^2 + \Big(\big(\bar{\lambda} + \gamma(x_{1,\gamma} - \psi)\big)^+ - \big(\bar{\lambda} + \bar{\gamma}(x_{1,\gamma} - \psi)\big)^+ + \big(\bar{\lambda} + \bar{\gamma}(x_{1,\gamma} - \psi)\big)^+ - \big(\bar{\lambda} + \bar{\gamma}(x_{1,\bar{\gamma}} - \psi)\big)^+, x_{1,\gamma} - x_{1,\bar{\gamma}} \Big)_{L^2} \le 0.$$

Since $x \mapsto (z + x)^+$ is monotone, we have

$$\alpha|x_\gamma - x_{\bar{\gamma}}|_X^2 \le |\gamma - \bar{\gamma}|\,|x_{1,\gamma} - \psi|_{L^2}|x_{1,\gamma} - x_{1,\bar{\gamma}}|_{L^2}.$$

Boundedness of $\{x_{1,\gamma}\}_{\gamma \ge r}$ implies Lipschitz continuity of $\gamma \mapsto x_\gamma$ for $\gamma \ge r$.

From (2.9) we deduce that

$$C^*(p_\gamma - p_{\bar{\gamma}}, \lambda_\gamma - \lambda_{\bar{\gamma}}) = J'(x_{\bar{\gamma}}) - J'(x_\gamma).$$

Therefore surjectivity of C implies that $\gamma \mapsto (p_\gamma, \lambda_\gamma)$ is Lipschitz continuous on $[r, \infty)$. For $\bar{\lambda} = 0$ we have that $\{x_\gamma\}_{\gamma>0}$ is bounded, and global Lipschitz continuity of $\mathcal{C}_0$ follows as before.

Local Lipschitz continuity of $\gamma \mapsto \lambda_\gamma$ in $L^2(\omega)$ results from

$$\begin{aligned} |\lambda_\gamma - \lambda_{\bar{\gamma}}|_{L^2} &= \Big| \big(\bar{\lambda} + \gamma(x_{1,\gamma} - \psi)\big)^+ - \big(\bar{\lambda} + \bar{\gamma}(x_{1,\bar{\gamma}} - \psi)\big)^+ \Big|_{L^2} \\ &\le |(\gamma - \bar{\gamma})x_{1,\gamma}|_{L^2} + |\bar{\gamma}(x_{1,\gamma} - x_{1,\bar{\gamma}})|_{L^2} + |\gamma - \bar{\gamma}||\psi|_{L^2}. \end{aligned}$$ ∎

Henceforth we set for $\gamma > 0$

$$S_\gamma = \{z \in \omega : \bar{\lambda}(z) + \gamma(x_{1,\gamma} - \psi)(z) > 0\}$$

and

$$g(\gamma) = \bar{\lambda} + \gamma(x_{1,\gamma} - \psi).$$

Below we shall assume that X_1 embeds compactly into $L^2(\omega)$. Since $\gamma \mapsto (x_\gamma, p_\gamma)$ is Lipschitz continuous,

$$\left(\frac{1}{\bar\gamma-\gamma}(x_{\bar\gamma}-x_\gamma), \frac{1}{\bar\gamma-\gamma}(p_{\bar\gamma}-p_\gamma)\right)$$

admits a weak accumulation point $(\dot x_\gamma^+, \dot p_\gamma^+)$ in $X \times W^*$ as $\bar\gamma \to \gamma$. Further,

$$\frac{1}{\bar\gamma-\gamma}\left(g(\bar\gamma)-g(\gamma)\right)$$

has $\dot g(\gamma) = x_{1,\gamma} - \psi + \gamma \dot x_{1,\gamma}^+$ as a strong accumulation point in $L^2(\omega)$.

Since J is quadratic, $J''(x_\gamma)$ is independent of x_γ and we henceforth simply use J'' for the bounded linear operator from X to X^*.

PROPOSITION 2.4. *Assume that $X_1 \hookrightarrow L^2(\omega)$ is compact, and set*

$$S_\gamma^+ = S_\gamma \cup \{z : \bar\lambda(z) + \gamma(x_{1,\gamma}-\psi)(z) = 0 \wedge \dot g(\gamma)(z) \geq 0\}.$$

Then $(\dot x_\gamma, \dot p_\gamma)$ satisfies

$$\langle J''(\dot x_\gamma), v\rangle_{X^*,X} + \langle E^*\dot p_\gamma, v\rangle_{X^*,X} + \left((x_{1,\gamma}-\psi+\gamma\dot x_{1,\gamma}^+)\chi_{S_\gamma^+}, v_1\right) = 0$$

for all $v = (v_1, v_2) \in X$.

For the proof we refer to [12]. We now set

$$S_\gamma^0 = \{z \in \omega : \bar\lambda(z) + \gamma(x_{1,\gamma}-\psi)(z) = 0\}.$$

COROLLARY 2.1. *If $\operatorname{meas}(S_\gamma^0) = 0$, then $\gamma \mapsto (x_\gamma, p_\gamma) \in X \times W^*$ is weakly differentiable at γ and $(\dot x_\gamma, \dot p_\gamma)$ satisfies*

$$\langle J''(\dot x_\gamma), v\rangle_{X^*,X} + \langle E^*\dot p_\gamma, v\rangle_{X^*,X} + \left((x_{1,\gamma}-\psi+\gamma\dot x_{1,\gamma})\chi_{S_\gamma}, v_1\right) = 0 \tag{2.10}$$

for all $v = (v_1, v_2) \in X$. Moreover $\gamma \mapsto x_\gamma$ is strongly differentiable at γ.

Proof. Every accumulation point $(\dot x_\gamma, \dot p_\gamma)$ of

$$\left(\frac{1}{\bar\gamma-\gamma}(x_{\bar\gamma}-x_\gamma), \frac{1}{\bar\gamma-\gamma}(p_{\bar\gamma}-p_\gamma)\right)$$

as $\bar\gamma \to \gamma$ satisfies the conclusion of Proposition 2.4. If there are two weak accumulation points, then their difference $(\bar x_\gamma, \bar p_\gamma)$ satisfies $E\bar x_\gamma = 0$ and

$$\langle J''(\bar x_\gamma), v\rangle_{X^*,X} + \langle \bar p_\gamma, Ev\rangle_{X^*,X} + \gamma(\bar x_{1,\gamma}\chi_{S_\gamma}, v_1)_{L^2} = 0 \tag{2.11}$$

for all $v = (v_1, v_2) \in X$. Setting $v = \bar x_\gamma$ in (2.11), we have

$$\langle J''(\bar x_\gamma), \bar x_\gamma\rangle_{X^*,X} + \gamma(\bar x_{1,\gamma}\chi_{S_\gamma}, \bar x_{1,\gamma}) = 0,$$

which implies that $\bar x_\gamma = 0$ by Assumption 2.1. Due to (2.2) and (2.11) we have $\bar p_\gamma = 0$, and weak differentiability follows. From (2.4) and (2.11) we have

$$\left\langle J''(\dot x_\gamma) - \frac{1}{\bar\gamma-\gamma}\left(J'(x_{\bar\gamma}) - J'(x_\gamma)\right), \dot x_\gamma - \frac{1}{\bar\gamma-\gamma}(x_{\bar\gamma}-x_\gamma)\right\rangle_{X^*,X} + \epsilon(\bar\gamma,\gamma) = 0,$$

where

$$\epsilon(\bar\gamma,\gamma)=\Big((x_{1,\gamma}-\psi+\gamma\dot x_{1,\gamma})\chi_{S_\gamma} \\ -\frac{1}{\bar\gamma-\gamma}\left[\big(\bar\lambda+\bar\gamma(x_{1,\bar\gamma}-\psi)\big)^+ \right. \\ \left. -\big(\bar\lambda+\gamma(x_{1,\gamma}-\psi)\big)^+\right],\dot x_{1,\gamma}-\frac{1}{\bar\gamma-\gamma}(x_{1,\bar\gamma}-x_{1,\gamma})\Big)_{L^2}.$$

Since

$$\lim_{\bar\gamma\to\gamma}\epsilon(\bar\gamma,\gamma)=0,$$

strong differentiability of $\gamma\mapsto x_\gamma$ at γ follows from Assumption 2.1. ∎

3. Applications. In this section we discuss several classes of constrained optimal control problems which are special instances of the general problem (P).

3.1. State constraints. The optimal control problem with state constraints,

$$\begin{aligned}&\text{minimize } J(x)=\frac{1}{2}|y-y_d|^2_{L^2(L^2)}+\frac{\beta}{2}|u|^2_{L^2(L^2)} \text{ over } x\in X\\ &\text{subject to } y_t-\Delta y=u \text{ in } (0,T)\times\Omega,\\ &\qquad y(0,\cdot)=0,\quad y=0 \text{ on } (0,T)\times\Gamma,\\ &\qquad y\le\psi \text{ a.e. in } (0,T)\times\Omega,\end{aligned}$$

with $\Gamma=\partial\Omega$ and $x=(y,u)$ is a special case of (P), where

$$\begin{aligned}X_1&=\{y\in L^2(0,T;H^1_0(\Omega)\cap H^2(\Omega)) : y_t\in L^2(0,T;L^2(\Omega))\},\\ X_2&=W=L^2(0,T;L^2(\Omega)),\\ \omega&=(0,T)\times\Omega,\\ E(y,u)&=y_t-\Delta y-u,\end{aligned}$$

$\beta>0$, $T>0$, $y_d\in W$, $\psi\in X_1$, and Ω a bounded domain in $\mathbb{R}^n$, is a special case of (P) with Assumption 2.1 and (2.2) satisfied. The general framework of section 2 provides dual variables $(p^*,\lambda^*)\in W\times X_1^*$. Additional regularity can be obtained by using the optimality conditions; see, e.g., [2, 3].

3.2. Dirichlet control with control constraints. Consider that

$$\begin{aligned}&\text{minimize } J(x)=\frac{1}{2}|y-y_d|^2_{L^2(\Omega)}+\frac{\beta}{2}|u|^2_{H^1(\Gamma)} \text{ over } x\in X\\ &\text{subject to } -\Delta y=f \text{ in } \Omega,\\ &\qquad y=u \text{ on } \Gamma,\\ &\qquad u\le\psi \text{ a.e. in } \Gamma,\end{aligned}\tag{3.1}$$

is a special case of (P) with $x=(u,y)$,

$$\begin{aligned}X&=H^1(\Gamma)\times H^1(\Omega),\quad W=H^{1/2}(\Gamma)\times H^{-1}(\Omega),\\ E(y,u)&=(y-u,-\Delta y-f),\end{aligned}$$

$\beta>0$, $y_d\in H^2(\Omega)$, $\psi\in C^2(\Gamma)$, and Ω a bounded domain in $\mathbb{R}^n$, $n\le 4$, with smooth boundary Γ. Since the variational solution $y=y(u)$ to

$$-\Delta y=f \text{ in } \Omega,\quad y=u \text{ on } \Gamma,$$

satisfies

$$|y(u)|_{H^1} \leq K(|u|_{H^{1/2}} + |f|_{H^{-1}})$$

for some constant K independent of $u \in H^{1/2}(\Gamma)$ and $f \in H^{-1}(\Omega)$, it follows that Assumption 2.1 is satisfied. To verify (2.2) let $(\tilde{u}_1, \tilde{y}, \tilde{u}_2) \in H^{1/2}(\Gamma) \times H^{-1}(\Omega) \times H^1(\Gamma)$. Choose $u = \tilde{u}_2$ and let $y \in H^1(\Omega)$ be the unique variational solution to

$$-\Delta y = \tilde{y} \text{ in } \Omega, \quad y = \tilde{u}_1 + u \text{ on } \Gamma.$$

Then

$$C(y,u) = \begin{pmatrix} y - u \\ -\Delta y \\ u \end{pmatrix} = \begin{pmatrix} \tilde{u}_1 \\ \tilde{y} \\ \tilde{u}_2 \end{pmatrix} \in H^{1/2}(\Gamma) \times H^{-1}(\Omega) \times H^1(\Gamma),$$

and thus C is surjective.

In this case the Lagrange multiplier associated with the inequality constraint $u \leq \psi$ is in $H^{-1}(\Gamma)$ and, hence, the primal-dual active set strategy without regularization cannot be defined.

The regularized problem is given by

$$\begin{aligned} &\text{minimize } \frac{1}{2}|y - y_d|^2_{L^2(\Omega)} + \frac{\beta}{2}|u|^2_{H^1(\Gamma)} + \frac{1}{2\gamma}|\left(\bar{\lambda} + \gamma(u - \psi)\right)^+|^2_{L^2(\Gamma)} \\ &\text{subject to } -\Delta y = f \text{ in } \Omega, \\ &\qquad\qquad y = u \text{ on } \Gamma, \end{aligned} \tag{3.2}$$

with $\bar{\lambda} \in C(\Gamma)$, $\bar{\lambda} \geq 0$. The optimality system for this problem is given by

$$\begin{aligned} &-\Delta y = f \text{ in } \Omega, \quad y = u \text{ on } \Gamma, \\ &-\Delta p = y_d - y \text{ in } \Omega, \quad p = 0 \text{ on } \Gamma, \\ &\beta(-\Delta_\Gamma + I)u + \frac{\partial p}{\partial n} + \left(\bar{\lambda} + \gamma(u - \psi)\right)^+ = 0 \text{ in } H^{-1}(\Gamma), \end{aligned} \tag{3.3}$$

where Δ_Γ denotes the Laplace–Beltrami operator on Γ; see [8]. Let $(y^*, u^*) = (y^*(u^*), u^*)$ denote the solution to (3.1). Then

$$\begin{aligned} &\frac{1}{2}|y_\gamma - y_d|^2_{L^2(\Omega)} + \frac{\beta}{2}|u_\gamma|^2_{H^1(\Gamma)} + \frac{1}{2\gamma}|\left(\bar{\lambda} + \gamma(u_\gamma - \psi)\right)^+|^2_{L^2(\Gamma)} \\ &\qquad \leq \frac{1}{2}|y^* - y_d|^2_{L^2(\Omega)} + \frac{\beta}{2}|u^*|^2_{H^1(\Gamma)} + \frac{1}{2\gamma}|\bar{\lambda}|^2_{L^2(\Gamma)}, \end{aligned}$$

and hence,

$$|u_\gamma|_{H^1} \leq K\left(1 + \frac{1}{\sqrt{\gamma}}|\bar{\lambda}|\right)_{L^2(\Gamma)}. \tag{3.4}$$

Here and below, K denotes a constant independent of $|\bar{\lambda}|_{L^2(\Gamma)}$ and γ. From well-known estimates [16, p. 188] for elliptic equations (3.3) and (3.4), we deduce that

$$\begin{aligned} |y_\gamma|_{H^{3/2}(\Omega)} &\leq K\left(1 + \frac{1}{\sqrt{\gamma}}|\bar{\lambda}|_{L^2(\Gamma)}\right), \\ |p_\gamma|_{H^{7/2}(\Omega)} &\leq K\left(1 + \frac{1}{\sqrt{\gamma}}|\bar{\lambda}|_{L^2(\Gamma)}\right), \\ \left|\frac{\partial p_\gamma}{\partial n}\right|_{H^2(\Gamma)} &\leq K\left(1 + \frac{1}{\sqrt{\gamma}}|\bar{\lambda}|_{L^2(\Gamma)}\right), \end{aligned}$$

and hence,

$$\left|\frac{\partial p_\gamma}{\partial n}\right|_{L^\infty(\Gamma)} \le K\left(1+\frac{1}{\sqrt{\gamma}}|\bar\lambda|_{L^2(\Gamma)}\right), \tag{3.5}$$

if $n \le 4$. From the last equation in (3.3), we deduce that

$$\beta\left\langle(-\Delta_\Gamma + I)\ (u_\gamma-\psi),(u_\gamma-\psi)^+\right\rangle_{X_1^*,X_1} + \left(\frac{\partial p_\gamma}{\partial n}+\bar\lambda+\gamma(u_\gamma-\psi),(u_\gamma-\psi)^+\right)_{L^2(\Gamma)} + \beta\left((-\Delta_\Gamma+I)\psi,(u_\gamma-\psi)^+\right)_{L^2(\Gamma)} = 0,$$

and thus,

$$\beta|(u_\gamma-\psi)^+|^2_{H^1(\Gamma)}+\gamma|(u_\gamma-\psi)^+|^2_{L^2(\Gamma)}+\left(\bar\lambda+\beta(-\Delta_\Gamma+I)\psi+\frac{\partial p_\gamma}{\partial n},(u_\gamma-\psi)^+\right)_{L^2(\Gamma)} \le 0.$$

From (3.5) it follows that

$$\beta|(u_\gamma-\psi)^+|^2_{H^1(\Gamma)}+\gamma|(u_\gamma-\psi)^+|^2_{L^2(\Gamma)} + \left(\bar\lambda+\beta(-\Delta_\Gamma+I)\psi-K\left(1+\frac{|\bar\lambda|}{\sqrt{\gamma}}\right),(u_\gamma-\psi)^+\right)_{L^2(\Gamma)} \le 0.$$

Hence, if $\bar\lambda$ is sufficiently large such that

$$\bar\lambda(x) > \beta(\Delta_\Gamma - I)\psi(x) + 2K \quad \text{for all } x\in\Gamma,$$

then there exists $\bar\gamma$ such that

$$u_\gamma \le \psi \text{ a.e. on } \Gamma \text{ for all } \gamma \ge \bar\gamma,$$

and thus u_γ is feasible for $\gamma \ge \bar\gamma$.

If $|u|^2_{H^1(\Gamma)}$ is replaced with $|u|^2_{H^{1/2}(\Gamma)}$ in (3.1) and accordingly $X = H^{1/2}(\Gamma)\times H^1(\Omega)$, then the Lagrange multiplier associated with $u\le\psi$ is in $H^{-1/2}(\Gamma)$, and again regularization is necessary to employ the primal-dual active set strategy. Following the above arguments, it can be shown that $u_\gamma \le \psi$ if $\bar\lambda \in C^1(\Gamma)$ and if γ is sufficiently large, and that $n \le 3$.

3.3. Dirichlet control with control constraints, revisited. Here we consider two further alternative formulations for Dirichlet boundary control and their treatment by the primal-dual active set strategy. First, we focus on

$$\begin{aligned} &\text{minimize } \frac{1}{2}|y-y_d|^2_{H^1(\Omega)}+\frac{\beta}{2}|u|^2_{L^2(\Gamma)} \text{ over } (y,u)\in H^1(\Omega)\times L^2(\Gamma)\\ &\text{subject to } -\Delta y = f \text{ in } \Omega,\\ &\qquad\qquad y = u \text{ on } \Gamma,\\ &\qquad\qquad u\le\psi \text{ a.e. on } \Gamma, \end{aligned} \tag{3.6}$$

where $y_d\in H^2(\Omega)$, $\beta>0$, $f\in L^2(\Omega)$, and $\psi\in H^2(\Gamma)$. Note that (3.6) admits a unique solution $(y^*,u^*)\in H^1(\Omega)\times H^{1/2}(\Gamma)$. The optimality condition for (3.6) is found to be (the variational form of)

$$\begin{aligned} &-\Delta y = f \text{ in } \Omega, \quad y = u \text{ on } \Gamma,\\ &-\Delta p = y_d - y - \Delta y_d - f \text{ in } \Omega, \quad p = 0 \text{ on } \Gamma,\\ &\left(\beta u+\frac{\partial p}{\partial n}-\frac{\partial y_d}{\partial n},h\right)_{L^2(\Gamma)} \ge 0 \quad \text{for all } h\in\mathcal{C}(u^*), \end{aligned} \tag{3.7}$$

where $(y,p) \in H^1(\Omega) \times H^1(\Omega)$, and

$$\mathcal{C}(u^*) = \{h = u - u^* : u \in L^2(\Gamma),\ u \le \psi\}.$$

Next consider p as a function of $u \in L^2(\Gamma)$ defined via the first four equations in (3.7). Note that the mapping $u \mapsto y(u)$ is continuous from $L^2(\Gamma)$ to $L^2(\Omega)$ (see [16]) and further that $u \mapsto \frac{\partial p(u)}{\partial n}$ is continuous from $L^2(\Gamma)$ to $H^{1/2}(\Gamma)$. Hence, the inequality in (3.7) can be equivalently expressed as

$$\begin{aligned} \beta u + \lambda &= \frac{\partial}{\partial n}(y_d - p), \\ \lambda &= (\lambda + c(u - \psi))^+ \end{aligned}$$

for any $c > 0$. Choosing $c = \beta$, this in turn is equivalent to

$$\beta u + \frac{\partial}{\partial n}(y_d - p) + \left(\frac{\partial}{\partial n}(y_d - p) - \beta\psi\right)^+ = 0. \tag{3.8}$$

As observed above, $u \mapsto \frac{\partial p(u)}{\partial n}$ is continuous from $L^2(\Gamma)$ to $H^{1/2}(\Gamma)$. Hence, for each n there exists $q > 2$ such that $u \mapsto \frac{\partial p(u)}{\partial n}$ is continuous from $L^2(\Gamma)$ to $L^q(\Gamma)$. Now standard techniques [10] can be applied to argue that the primal-dual active set strategy applied to (3.8) is locally superlinearly convergent in $L^2(\Gamma)$.

3.4. Optimal control problem with "solutions très faibles." Finally, we turn to the boundary control problem with solutions in the sense of "solutions très faibles"; e.g. [15, p.76],

$$\begin{aligned} &\text{minimize } \frac{1}{2}\left|\frac{\partial y}{\partial n} - y_d\right|^2_{H^{-1}(\Gamma)} + \frac{\beta}{2}|u|^2_{L^2(\Gamma)} \text{ over } (u,y) \in L^2(\Gamma) \times L^2(\Omega) \\ &\text{subject to } -(y, \Delta v)_{L^2(\Omega)} - (f, v)_{L^2(\Omega)} + \left(u, \frac{\partial v}{\partial n}\right)_{L^2(\Gamma)} = 0 \\ &\qquad\qquad \text{for all } v \in H^2(\Omega) \cap H^1_0(\Omega), \\ &u \le \psi, \end{aligned} \tag{3.9}$$

where $y_d \in H^{-1}(\Gamma)$, $f \in L^2(\Omega)$, $\psi \in L^2(\Gamma)$. We recall that $u \mapsto \frac{\partial y}{\partial n}$ is a continuous mapping from $L^2(\Gamma)$ to $H^{-1}(\Gamma)$; see [15, p. 78]. The adjoint equation for the optimal control problem (3.9) is given by

$$-\Delta p = 0 \text{ in } \Omega, \quad p = (-\Delta_\Gamma)^{-1}\left(y_d - \frac{\partial y}{\partial n}\right) \text{ on } \Gamma, \tag{3.10}$$

and it is known (see [15, p. 77] and [16]) that $p \in H^{3/2}(\Omega)$ and $\frac{\partial p}{\partial n} \in L^2(\Gamma)$. The optimality system for (3.9) consists of the variational equation in (3.9), the adjoint equation (3.10), and the optimality condition

$$\begin{aligned} \beta u + \lambda + \frac{\partial p}{\partial n} &= 0, \\ \lambda &= (\lambda + c(u - \psi))^+ \end{aligned} \tag{3.11}$$

for any $c > 0$. Choosing $c = \beta$ we find, as above, that (3.11) is equivalent to

$$\beta u + \frac{\partial p}{\partial n} + \left(\frac{\partial p}{\partial n} - \beta\psi\right)^+ = 0. \tag{3.12}$$

Note that $u \mapsto \frac{\partial p}{\partial n}$ is a continuous mapping from $L^2(\Gamma)$ into itself, which is not smoothing. Thus, the max-operation in (3.12) is not generalized differentiable in the sense of [10]. The latter property, however, is essential for a superlinear convergence of the primal-dual active set strategy, which we, hence, cannot expect in connection with (3.12).

Turning to the regularized version, we consider, for $\gamma > 0$,

$$\begin{aligned} &\text{minimize } \frac{1}{2}\left|\frac{\partial y}{\partial n} - y_d\right|^2_{H^{-1}(\Gamma)} + \frac{\beta}{2}|u|^2_{L^2(\Gamma)} + \frac{1}{2\gamma}\int_\Gamma \left|\left(\bar\lambda + \gamma(u-\psi)\right)^+\right|^2 \\ &\qquad \text{over } (u,y) \in L^2(\Gamma)\times L^2(\Omega) \\ &\text{subject to } -(y,\Delta v)_{L^2(\Omega)} - (f,v)_{L^2(\Omega)} + \left(u, \frac{\partial v}{\partial n}\right)_{L^2(\Gamma)} = 0 \\ &\qquad \text{for all } v \in H^2(\Omega)\cap H^1_0(\Omega). \end{aligned} \tag{3.13}$$

The optimality system consists of the primal equation, the adjoint equation (3.10), and the optimality condition

$$\beta u + \frac{\partial p}{\partial n} + \left(\bar\lambda + \gamma(u-\psi)\right)^+ = 0.$$

Consequently the solution u_γ to (3.13) is still in only $L^2(\Gamma)$, in general, and the primal-dual active set strategy for (3.13) may not be (superlinearly) convergent.

4. The value functional and its model. Next we introduce the optimal value functional of (P_γ), study its smoothness properties, and approximate it by low-parametric families of model functions. These model functions will be used in section 5, where we develop a path-following algorithm operating with a reliable γ-update strategy.

DEFINITION 4.1. *The functional*

$$\gamma \mapsto V(\gamma) = J(x_\gamma) + \frac{1}{2\gamma}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+|^2 dw$$

defined on $(0,\infty)$ *is called the* primal-dual-path value functional.

The smoothness and monotonicity properties of V provide useful information for tuning γ in an iterative procedure.

PROPOSITION 4.1. *The value functional V is differentiable with*

$$\dot V(\gamma) = -\frac{1}{2\gamma^2}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma}-\psi))^+|^2 + \frac{1}{\gamma}\int_\omega (\bar\lambda + \gamma(x_{1,\gamma}-\psi))^+(x_{1,\gamma}-\psi).$$

Proof. For $\bar\gamma, \gamma \in (0,\infty)$ we have from (4),

$$\langle J'(x_{\bar\gamma} + x_\gamma), x_{\bar\gamma} - x_\gamma\rangle_{X^*,X} + (\lambda_{\bar\gamma} + \lambda_\gamma, x_{1,\bar\gamma} - x_{1,\gamma})_{L^2} = 0,$$

and therefore

$$\begin{aligned}
V(\bar\gamma) - V(\gamma) &= J(x_{\bar\gamma}) - J(x_\gamma) + \frac{1}{2\bar\gamma}\int_\omega |(\bar\lambda + \bar\gamma(x_{1,\bar\gamma} - \psi))^+|^2 \\
&\quad - \frac{1}{2\gamma}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+|^2 \\
&= \langle J'(x_{\bar\gamma} + x_\gamma), x_{\bar\gamma} - x_\gamma\rangle_{\mathrm{X}^*,\mathrm{X}} + \frac{1}{2\bar\gamma}\int_\omega |(\bar\lambda + \bar\gamma(x_{1,\bar\gamma} - \psi))^+|^2 \\
&\quad - \frac{1}{2\bar\gamma}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+|^2 \\
&= \frac{1}{2\bar\gamma}\int_\omega |(\bar\lambda + \bar\gamma(x_{1,\bar\gamma} - \psi))^+|^2 - \frac{1}{2\gamma}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+|^2 \\
&\quad - \frac{1}{2}\int_\omega ((\bar\lambda + \bar\gamma(x_{1,\bar\gamma} - \psi))^+ + (\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+)(x_{1,\bar\gamma} - x_{1,\gamma}) \\
&= \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3.
\end{aligned}$$

It can now be argued just as in the proof of Proposition 4.1 of [12] that

$$\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma}|\mathcal{I}_2| = \lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma}|\mathcal{I}_3| = 0$$

and further,

$$\begin{aligned}
\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma}\mathcal{I}_1 = &- \frac{1}{2\gamma^2}\int_\omega ((\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+)^2 \\
&+ \frac{1}{\gamma}\int_\omega (\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+(x_{1,\gamma} - \psi),
\end{aligned}$$

as desired. □

It turns out that the monotonicity properties of V may depend on the choice of $\bar\lambda$. For the following discussion we assume that the solution $\tilde x$ to $(\tilde{\mathrm{P}})$ does not satisfy $\tilde x_1 \le \psi$. Let us start by first considering the case $\bar\lambda = 0$. We have

$$\dot V(\gamma) = \frac{1}{2}\int_\omega |(x_{1,\gamma} - \psi)^+|^2 > 0. \tag{4.1}$$

Indeed, if $\dot V(\gamma) = 0$ for some $\gamma > 0$, then $x_{1,\gamma} \le \psi$, i.e., $x_{1,\gamma}$ is feasible. Thus, $\lambda_\gamma = 0$ and, from (2.3) and (2.4), we find that $(x_\gamma, \lambda_\gamma) = (\tilde x, 0)$ is the solution to (2.3) with $\tilde x_1 \le \psi$ which was ruled out by assumption. In case there exist $\bar\lambda$ and $\bar\gamma > 0$ such that

$$x_{1,\gamma} \le \psi \quad \text{for all} \quad \gamma \ge \bar\gamma, \tag{4.2}$$

we have by Proposition 4.1 that $\dot V(\gamma) \le 0$ for $\gamma \ge \bar\gamma$ and, unless the solution $\tilde x$ to $(\tilde{\mathrm{P}})$ satisfies $\tilde x_1 \le \psi$, we obtain $\dot V(\gamma) < 0$ for $\gamma \ge \bar\gamma$. In fact, if $\dot V(\gamma) = 0$, then

$$0 = \dot V(\gamma) \le -\frac{1}{2\gamma^2}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+|^2 \le 0,$$

and hence $\lambda_\gamma = 0$. Therefore $(x_\gamma, \lambda_\gamma) = (\tilde x, 0)$ is the solution of (2.3), which is excluded. Recall that a specific case in which (4.2) holds was given in section 3.2.

Next we study the second order differentiability of V.

PROPOSITION 4.2. *Assume that $X_1 \hookrightarrow L^2(\omega)$ is compact and let $\dot{x}_\gamma$ denote an accumulation point of $\frac{1}{\bar\gamma-\gamma}(x_{1,\bar\gamma} - x_{1,\gamma})$ as $\bar\gamma \to \gamma$. Then for a subsequence γ_n realizing the accumulation point,*

$$\begin{aligned}\lim_{\bar\gamma_n\to\gamma} \frac{1}{\bar\gamma_n - \gamma}(\dot V(\bar\gamma_n) - \dot V(\gamma)) &= \frac{1}{\gamma^3}\int_\omega |(\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+|^2 \\ &- \frac{2}{\gamma^2}\int_\omega (\bar\lambda + \gamma(x_{1,\gamma} - \psi))^+(x_{1,\gamma} - \psi) \\ &+ \frac{1}{\gamma}\int_\omega (x_{1,\gamma} - \psi)(x_{1,\gamma} - \psi + \gamma\dot x_{1,\gamma})\chi_{S_\gamma^+}.\end{aligned} \tag{4.3}$$

If $\operatorname{meas}(S_\gamma^\circ) = 0$, *then $\gamma \to V(\gamma)$ is twice differentiable at γ, and the second derivative is given by the right-hand side in* (4.3) *with $\chi_{S_\gamma^+}$ replaced by χ_{S_γ}.*

Proof. The first claim follows from Proposition 4.1. If meas $(S_\gamma^\circ) = 0$, then accumulation points of $\frac{1}{\bar\gamma-\gamma}(x_{1,\bar\gamma} - x_{1,\gamma})$ as $\bar\gamma \to \gamma$ are unique, and (4.3) with $\chi_{S_\gamma^+}$ replaced by χ_{S_γ}, as well as existence of $\ddot V$ at γ, follow. ∎

We turn to proposing low-parametric model functions m for V which share some of the qualitative properties of V. Notice that, in general, V is not at our disposal quantitatively. However, we shall see that m can be obtained from solves of the regularized problem for different choices of γ. In the next section, these model functions will then be used as a guideline for updating γ; see (5.2). Throughout the following discussion, we assume that the solution $\tilde x$ to $(\tilde P)$ does not satisfy $\tilde x_1 \le \psi$.

4.1. Case $\bar\lambda = 0$. In this case, $\gamma \to V(\gamma)$ is strictly increasing with $V(0)$ equal to the value of the cost in $(\tilde{\mathrm{P}})$ and V^∞ the value of the cost in (P).

PROPOSITION 4.3. *The mapping $\gamma \to \dot V(\gamma)$ is monotonically decreasing.*

Proof. For $\bar\gamma > \gamma > 0$ we have

$$\begin{aligned}J(x_\gamma) + \frac{\gamma}{2}|(x_{1,\gamma} - \psi)^+|^2 &\le J(x_{\bar\gamma}) + \frac{\gamma}{2}|(x_{1,\bar\gamma} - \psi)^+|^2 \\ &\le J(x_{\bar\gamma}) + \frac{\bar\gamma}{2}|(x_{1,\bar\gamma} - \psi)^+|^2 \le J(x_\gamma) + \frac{\bar\gamma}{2}|(x_{1,\gamma} - \psi)^+|^2,\end{aligned}$$

and hence

$$J(x_\gamma) - J(x_{\bar\gamma}) \le \frac{\gamma}{2}(|(x_{1,\bar\gamma} - \psi)^+|^2 - |(x_{1,\gamma} - \psi)^+|^2),$$

and further

$$J(x_{\bar\gamma}) - J(x_\gamma) \le \frac{\bar\gamma}{2}(|(x_{1,\gamma} - \psi)^+|^2 - |(x_{1,\bar\gamma} - \psi)^+|^2),$$

which implies

$$0 \le (\bar\gamma - \gamma)(|(x_{1,\gamma} - \psi)^+|^2 - |(x_{1,\bar\gamma} - \psi)^+|^2) = (\bar\gamma - \gamma)(\dot V(\gamma) - \dot V(\bar\gamma)).$$ ∎

Note that Proposition 4.3 implies that $\ddot V(\gamma) \le 0$ whenever the second derivative of V exists at γ. This can also be derived from (2.9) and (4.3). A class of functions that satisfies the above properties of V is given by

$$m(\gamma) = C_1 - \frac{C_2}{(D+\gamma)^r}, \tag{4.4}$$

with $C_1 \in \mathbb{R},\ C_2 > 0,\ D > 0,\ r > 0$. In fact, $\dot{m} > 0$, $\ddot{m} < 0$ and $m(0)$, $m(\gamma)$ for $\gamma \to \infty$ are well defined. In our use of m for path-following algorithms, (C_1, C_2, D) will be treated differently from r. While r will be chosen as a fixed number, (C_1, C_2, D) will be updated in an iterative procedure. Let us further note that by a simple rescaling, we can always assume that $D = 1$.

Remark 4.1. To give a second motivation for the choice of the model function, we consider the state constrained problem

$$\text{minimize } J(x) = \frac{1}{2}|y - y_d|^2_{L^2(\Omega)} + \frac{\beta}{2}|u|^2_{L^2(\Omega)} \text{ over } x \in X$$

$$\begin{aligned} \text{subject to } \quad -\Delta y &= u \quad && \text{in } \Omega, \\ y &= 0 && \text{on } \partial\Omega, \\ y &\le \psi && \text{a.e. in } \Omega, \end{aligned}$$

where $\beta > 0$, $x = (y, u) \in X = (H^1_0(\Omega) \cap H^2(\Omega)) \times L^2(\Omega)$, $y_d \in L^2(\Omega)$, $\psi \in L^2(\Omega)$, and Ω is a bounded domain in $\mathbb{R}^n$. We utilize the sensitivity equation with respect to γ and assume that meas $\{z : (y - \psi)(z) = 0\} = \emptyset$. Then from Proposition 2.4 with $S_\gamma = \{z : (y - \psi)(z) > 0\}$, we have

$$-\Delta \dot{y} = \frac{1}{2}\dot{p}, \quad -\Delta \dot{p} + (y_\gamma - \psi)^+ + \gamma \dot{y} \chi_{S_\gamma} = -\dot{y},$$

and hence,

$$\alpha \Delta^2 \dot{y} + \dot{y} + \gamma \dot{y} \chi_{S_\gamma} + (y_\gamma - \psi)^+ = 0.$$

Taking the inner product in $L^2(\Omega)$ with $(y_\gamma - \psi)^+$, we have

$$(\alpha \Delta^2 \dot{y} + \dot{y},\ (y_\gamma - \psi)^+) + \gamma(\dot{y}, (y_\gamma - \psi)^+) + |(y_\gamma - \psi)^+|^2_{L^2} = 0.$$

Replacing $(\alpha \Delta^2 \dot{y},\ (y_\gamma - \psi)^+)$ with $(\hat{D}\, \dot{y}, (y_\gamma - \psi)^+)$, where $\hat{D}$ is a positive constant, we arrive at

$$((D + \gamma)\dot{y},\ (y_\gamma - \psi)^+) + |(y_\gamma - \psi)^+|^2_{L^2} = 0, \tag{4.5}$$

where we set $D = \hat{D} + 1$. Recall from Proposition 4.1 and (4.1) that $|(y_\gamma - \psi)^+|^2_{L^2} = 2\dot{V}(\gamma)$ and $|(\dot{y}, y_\gamma - \psi)^+| = \ddot{V}(\gamma)$. Since we replaced $\alpha\Delta^2$ with the constant $\hat{D}$, for the purpose of deriving this model, we replace V with its model m and obtain from (4.5)

$$(D + \gamma)\ddot{m}(\gamma) + 2\dot{m}(\gamma) = 0.$$

The solution to this ordinary differential equation is given by (4.4) with $r = 1$.

4.2. Case $\bar{\boldsymbol{\lambda}} \neq \mathbf{0}$. If (4.2) holds, then $\gamma \to V(\gamma)$ is strictly decreasing for $\gamma \ge \bar{\gamma}$, with $\lim_{\gamma \to \infty} V(\gamma)$ equal to the value of the objective of (P), and $\lim_{\gamma \to 0} V(\gamma) = \infty$.

PROPOSITION 4.4. *Assume that $X_1 \hookrightarrow L^2(\omega)$ is compact, that* meas$(S^\circ_\gamma) = 0$, *and that* (4.2) *holds. Then $\ddot{V}(\gamma) \ge 0$.*

Proof. From (4.3) with $S^+_\gamma = S_\gamma$, we have

$$\ddot{V}(\gamma) = \frac{1}{\gamma^3}\int_{S_\gamma} \bar{\lambda}^2 + \int_{S_\gamma} (x_{1,\gamma} - \psi)\dot{x}_{1,\gamma} \ge \frac{1}{\gamma}\int_{S_\gamma} (x_{1,\gamma} - \psi)^2 + \int_{S_\gamma} (x_{1,\gamma} - \psi)\dot{x}_{1,\gamma}.$$

From Proposition 2.4 and (2.1), we have

$$\gamma|\dot{x}_{1,\gamma}|_{L^2(S_\gamma)} \le |x_{1,\gamma} - \psi|_{L^2(S_\gamma)}.$$

This implies $\ddot{V}(\gamma) \ge 0$. □

A class of model functions, which satisfy the above properties of V, is given by

$$m(\gamma) = C_1 - \frac{C_2}{(D+\gamma)^r} + \frac{B}{\gamma^r}, \tag{4.6}$$

with $C_1 \in \mathbb{R},\ B \ge C_2 > 0,\ D > 0$, and $r \in (0, 1]$. In fact, $m(0) = \infty$, $\lim_{\gamma\to\infty} m(\gamma) = C_1$, $\dot{m}(\gamma) < 0$, and $\ddot{m}(\gamma) > 0$. A rescaling argument similar to the one for $\bar{\lambda} = 0$ allows us to henceforth assume $D = 1$.

5. Numerics. In this section we specify an inexact path-following method for the numerical solution of (P). In the inner loop, it utilizes a locally superlinearly convergent algorithm for solving the regularized path problem (P_γ). The outer loop employs a γ-update strategy based on our model functions (4.4) (respectively, (4.6)). The section ends with a report on test runs for the solution of some state constrained optimal control problems with distributed control and a linear elliptic PDE as the governing equation. We also compare our new algorithm with a primal-dual path-following interior point method [18, 27] adapted to PDE-constrained minimization as in [1] and with the primal-dual active set strategy [1, 2]. Our test problems include cases when the optimal solution lacks strict complementarity, i.e., when the set $\mathcal{S}^* = \{y^* = \psi \ \wedge\ \lambda^* = 0\}$ has positive measure. It is known that lack of strict complementarity may slow the convergence of numerical algorithms. This is due to the difficulty of detecting the correct active (respectively, inactive) set structure in the neighborhood of the solution.

5.1. Inner iteration: An algorithm for solving (P_γ). Here we adopt the primal-dual active set strategy as proposed in [12] for solving problems of the type (P_γ). The method is equivalent to a semismooth Newton algorithm, and, using the techniques in [10], it can be shown to converge locally at a q-superlinear rate.

ALGORITHM PDAS$_\gamma$ (primal-dual active set strategy for γ-regularized problem (P_γ)).

(i) Choose $\bar{\lambda} \ge 0$ and $x^0 \in X$; set $l = 0$.

(ii) Determine the active and inactive sets

$$\mathcal{A}^{l+1} := \{w \in \omega : \bar{\lambda}(w) + \gamma(x_1^l(w) - \psi(w)) > 0\},$$
$$\mathcal{I}^{l+1} := \Omega \setminus \mathcal{A}^{l+1}.$$

(iii) Compute the solution x^{l+1} with associated adjoint state p^{l+1} of

$$\text{minimize } \langle J'(x^l) + \frac{1}{2}\langle J''(x^l)(x - x^l), x - x^l\rangle_{X^*,X} + \frac{1}{2\gamma}|(\bar{\lambda} + \gamma(x_1 - \psi))^+|^2_{L^2(\mathcal{A}^{l+1})} \quad \text{over } x \in X$$
$$\text{subject to } Ex = f \text{ in } W.$$

(iv) Compute

$$\lambda^{l+1} = \begin{cases} 0 & \text{on } \mathcal{I}^{l+1}, \\ \bar{\lambda} + \gamma(x_1^{l+1} - \psi) & \text{on } \mathcal{A}^{l+1}, \end{cases}$$

set $l = l + 1$, and go to (ii).

The first order optimality system of the minimization problem in step (iii) is given by

$$\begin{aligned} &Ex^{l+1} = f \text{ in } W, \\ &J''(x^l)x^{l+1} + E^*p^{l+1} + \big(\gamma(x_1^{l+1} - \psi)\chi_{\mathcal{A}^{l+1}}, 0\big) \\ &\qquad = -J'(x^l) + J''(x^l)x^l - \big(\bar\lambda\chi_{\mathcal{A}^{l+1}}, 0\big) \text{ in } X^*. \end{aligned}$$

Note that this system corresponds to a linearization of (2.4) at x^l. In this context, the max-function

$$x_1 \mapsto \big(\bar\lambda + \gamma(x_1 - \psi)\big)^+$$

is linearized separately on the active and inactive sets determined in step (ii). This is equivalent to employing a generalized (or slant) derivative of the max-function in the spirit of [10, 12] in the process of linearization. Consequently, step (iii) is identical to the solution of the linear system within an iteration of a semismooth Newton method for solving (2.4).

5.2. Outer iteration: Inexact solutions and γ-update. For small γ there is no need for highly accurate solutions of the regularized problem (P_γ), since we expect x_γ to be only a coarse approximation of x^*. Rather we propose a procedure requiring approximate solutions of the path problem lying in some neighborhood of the path only. This is similar to the concept considered in [12] and to path-following for log-barrier-functions. For this purpose we introduce the residuals

$$\begin{aligned} r_x(x) &= \|Ex - f\|_W, \\ r_p(x, p, \lambda) &= \|J'(x) + E^*p + (\lambda, 0)\|_{X^*}, \\ r_\lambda(x_1, \lambda) &= \|\lambda - (\bar\lambda + \gamma(x_1 - \psi))^+\|_{X_1^*} \end{aligned}$$

and define the neighborhood

$$\mathcal{N}(\gamma, r) = \left\{(x, p, \lambda) \in Z : \max\{r_x(x), r_p(x, p, \lambda), r_\lambda(x_1, \lambda)\} \le \frac{\tau}{\gamma^r}\right\}$$

with $Z = X \times X^* \times L^2(\omega)$ for some fixed $\tau > 0$ and $r > 0$. In our implementation, we typically choose r in accordance with our model (4.4) or (4.6). In the subsequent algorithm, for fixed γ, we stop Algorithm PDAS_γ after the first occurrence of $(x^l, p^l, \lambda^l) \in \mathcal{N}(\gamma, r)$ for the first time.

Once an approximate solution of (P_γ) is obtained, we have to update γ. To this end, we introduce the feasibility measure ρ^F and the complementarity measure ρ^C as follows:

$$\begin{aligned} \rho^F(x_1) &:= \int_\omega (x_1 - \psi)^+ dz, \\ \rho^C(x_1) &:= \int_{\mathcal{I}(x_1)} (x_1 - \psi)^+ dz + \int_{\mathcal{A}(x_1)} (x_1 - \psi)^- dz, \end{aligned}$$

where $\mathcal{A}(x_1) = \{z \in \omega : \bar\lambda(z) + \gamma(x_1(z) - \psi(z)) > 0\}$, $\mathcal{I}(x_1) = \omega \setminus \mathcal{A}(x_1)$, and $(\cdot)^- = \min(0, \cdot)$. Whenever $x_1 = x_1^{k+1}$ and $\gamma = \gamma_k$, we write $\mathcal{A}^{k+1}$, $\mathcal{I}^{k+1}$ and ρ_{k+1}^F, ρ_{k+1}^C. For $\max(\rho_{k+1}^F, \rho_{k+1}^C) > 0$, we obtain a candidate γ_{k+1}^+ for γ_{k+1} as

$$\gamma_{k+1}^+ = \max\left(\gamma_k \max\left(\tau_1, \frac{\rho_{k+1}^F}{\rho_{k+1}^C}\right), \frac{1}{\max(\rho_{k+1}^F, \rho_{k+1}^C)^q}\right) \tag{5.1}$$

with $\tau_1 > 1$ and $q \geq 1$; otherwise we set $\gamma_{k+1}^+ = \tau_1 \gamma_k$. We include the quotient $\rho_{k+1}^F / \rho_{k+1}^C$ in order to significantly increase γ whenever $\rho_{k+1}^F \gg \rho_{k+1}^C$, i.e., when the iterates primarily lack feasibility rather than complementarity. The choice $q > 1$ induces certain growth rates for γ. Similar to [12], we also incorporate the following safeguard based on our model (4.4) (respectively, (4.6)): Unless $\gamma_{k+1} < \tau_2 \gamma_k$, with $\tau_2 > 1$, we reduce γ_{k+1}^+ until

$$|t_k(\gamma_{k+1}) - m_k(\gamma_{k+1})| \leq \tau_3 |J(x^{k+1}, \gamma_k) - J(x^k, \gamma_{k-1})|, \tag{5.2}$$

where $0 < \tau_3 < 1$ and $t_k(\gamma) = J(x^{k+1}, \gamma_k) + \frac{\partial J(x^k, \gamma_k)}{\partial \gamma}(\gamma - \gamma_k)$. In other words, the linearization of m_k at γ_{k+1} should not be farther away from m_k than the distance of the previous two objective values of the regularized problem. As soon as (5.2) is satisfied, we set $\gamma_{k+1} = \gamma_{k+1}^+$.

Notice that our safeguard involves the model function in iteration k, which we denote by m_k. To determine the parameters in our model, we use the actual approximate information on the value functional and its derivative as well as the value function at some reference point. In what follows, we argue only for the model (4.6). The case (4.4) is treated similarly. Given γ_k in iteration k, for fixing B_k, $C_{1,k}$, and $C_{2,k}$ in the model $m_k(\gamma)$, we use the conditions

$$m_k(\gamma_k) = J(x^k, \gamma_k), \quad \dot{m}_k(\gamma_k) = \frac{\partial J(x^k, \gamma_k)}{\partial \gamma}(x^k, \gamma_k), \quad m_k(\hat{\gamma}) = J(\hat{x}, \hat{\gamma}),$$

where $\hat{x}$ denotes an approximate solution of (P_γ) at a reference value $\gamma = \hat{\gamma}$.

Now we are able to outline our overall algorithm.

Algorithm IPF (inexact path-following).

(i) Initialized $\gamma_0 > 0$, select $r > 0$, and set $k := 0$.
(ii) Compute $(x^k, p^k, \lambda^k) \in \mathcal{N}(\gamma_k, r)$.
(iii) Update γ_k by (5.1) with safeguard (5.2) to obtain γ_{k+1}.
(iv) Set $k = k + 1$ and go to (ii).

In step (ii) we use PDAS_γ for performing the inner iteration. The convergence of Algorithm IPF follows immediately from the convergence of Algorithm PDAS_γ for every fixed γ and the fact that $\gamma_{k+1} \geq \tau_1 \gamma_k$ with $\tau_1 > 1$ for all k, the property that $\frac{\tau}{\gamma_k^r} \to 0$ for $\gamma_k \to \infty$ in the definition of the neighborhood, and from Proposition 2.1.

5.3. Numerical tests. In this section we report on numerical results for the solution of the following state constrained optimal control problem:

$$\begin{aligned} &\text{Minimize } J(y,u) = \frac{1}{2}|y - y_d|_{L^2}^2 + \frac{\beta}{2}|u|_{L^2}^2 \text{ over } (y,u) \in X \\ &\text{subject to } -\Delta y = u \text{ in } \Omega, \quad y = 0 \text{ on } \Gamma, \\ &\qquad\qquad y \leq \psi \text{ a.e. in } \Omega. \end{aligned}$$

Thus, we have $x = (y, u)$, $X_1 = H_0^1(\Omega)$, $X_2 = L^2(\Omega)$, and $\omega = \Omega = (0,1)^2$.

For the discretization of the Laplace operator in two dimensions we use a standard regular five-point finite difference stencil with mesh size h. Unless specified otherwise, the subsequent test runs of all algorithms are based on a nested iteration technique. For this purpose we define a grid hierarchy with mesh sizes $\{h_i\}_{i=2}^8$ and $h_i = 2^{-i}$. On every grid, we stop Algorithm IPF as soon as

$$\max(r_x(x^k), r_p(x^k, p^k, \lambda^k), r_d(x_1^k, \lambda^k)) \leq 0.1\, h_i^2, \tag{5.3}$$

where $r_d(x_1^k, \lambda^k) = \|\lambda^k - (\lambda^k + x_1^k - \psi)^+\|_{X_1^*}$. For $i = 2$ we initialize the algorithms by choosing $y^0 = \psi_i$, $u_i^0 = -\Delta_i y_i$, and $p_i^0 = \beta u_i^0$. Here and below, subscript i refers to discretized quantities on grid level i. The initial Lagrange multiplier is $\lambda_i^0 = \chi_{\{y_i^0 > \psi_i\}}(\bar\lambda_i + \gamma_0(y_i^0 - \psi_i))$ with $\bar\lambda_i = \max\left(y_{d,i} - (\beta\Delta_i^2 + \mathrm{id}_i)\psi_i, 0\right)$ in the case considered in section 4.2; otherwise $\bar\lambda_i = 0$. For $i > 2$, the interpolation of $(x_{i-1}, p_{i-1}, \lambda_{i-1})$, which is the approximate solution for h_{i-1}, is used as the initial value on the grid with mesh size h_i. The interpolation process is as follows. First, we smooth $(u_{i-1}, p_{i-1}, \lambda_{i-1})$ by applying $(-\Delta_{i-1})^{-1}$, the discrete Laplace operator with homogeneous Dirichlet boundary conditions for h_{i-1}, to the respective components. Then a nine-point-interpolation scheme is applied to obtain $(\tilde u_i, \tilde p_i, \tilde\lambda_i)$; see, e.g., [7]. Finally, we compute

$$u_i^0 = -\left(D_{w_1,i}(D^0_{w_1,i}\tilde u_i) + D_{w_2,i}(D^0_{w_2,i}\tilde u_i)\right),$$

where $D^0_{w_1,i}$ stands for symmetric differences with homogeneous Dirichlet boundary conditions in the w_1-direction, and $D_{w_1,i}$ represents symmetric differences with linear extrapolation on the boundary; similar definitions hold for $D_{w_2,i}$ and $D^0_{w_2,i}$. The adjoint state p_i^0 and the multiplier λ_i^0 are computed analogously.

Whenever a nonzero shift $\bar\lambda$ is used, then we choose

$$\bar\lambda = \max\left(y_d - (\beta\Delta^2 + \mathrm{id})\psi, 0\right).$$

This choice of $\bar\lambda$ is inspired by the reduction of the system consisting of state and adjoint equations, by the optimality condition with respect to u, and by setting $x_1 = \psi$.

Problem 1. The problem data are $y_d = 10(\sin(2\pi x_1) + x_2)$, $\psi \equiv 0.01$, $\beta = 0.1$. In Figure 1 we show the optimal state y_h^* (top left plot) and control u_h^* (far right) and the Lagrange multiplier λ_h^* (bottom) on a 128×128 grid. The behavior of λ_h^* along the boundary between the active and inactive sets at the discrete solution clearly suggests the measure-valuedness of λ^*.

For our model of the primal-dual value functional we use (4.6), i.e., we apply the $\bar\lambda$-shift, with $r = 0.2$. We use $\gamma_0 = 1\mathrm{E}2$ initially. First, we report on the behavior of the algorithms on a fixed grid, i.e., without utilizing the nested iteration concept. In Table 1 we show the iteration numbers for the primal-dual active set method for solving (P) (abbreviated by PDAS), the primal-dual path-following interior point method (PDIP), and our new path-following concept (IPF) for various mesh sizes h. We point out that the stopping rule for each algorithm is given by (5.3). For IPF we also specify the total number of inner iterations, i.e., the total number of iterations of Algorithm PDAS$_\gamma$. In all cases, the algorithms are initialized as in the case of $i = 2$ described above. Further, for Algorithm IPF we use $q = 1.25$ and $\tau_1 = 10$ in (5.1), $\tau_2 = 1.01$ and $\tau_3 = 0.6$ in (5.2), and $\tau = 100$ in $\mathcal{N}(\gamma, r)$. The results in Table 1 indicate that Algorithms IPF and PDIP are superior to PDAS, whereas IPF appears to be more efficient than PDIP. This reflects also our experience from further test runs for additional problems. We also point out that in contrast to PDAS and PDIP our Algorithm IPF admits a function space analysis. As a consequence, the number of (inner) iterations behaves in a rather mesh-independent way. For PDAS a strong dependence of the iteration numbers on the mesh size is observed. The stabilizing iteration numbers for PDIP in the case of mesh refinements suggest that a function space analysis might be possible. However, to the best of our knowledge, no such analysis is yet available for the problem class considered here that includes a regularization parameter tending to zero in the numerical method.

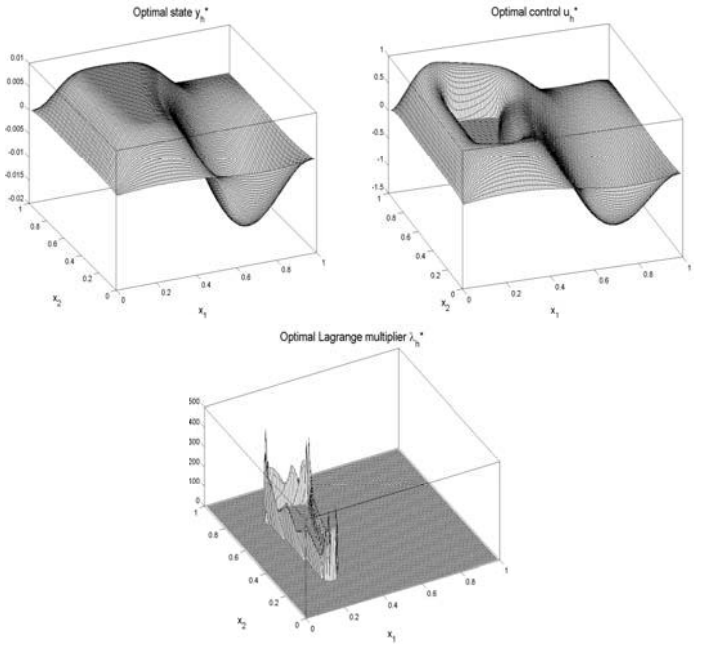


FIG. 1. *Optimal state (top left), optimal control (top right), and optimal multiplier (bottom) for problem* 1 *with* $h = 1/128$.

TABLE 1
Comparison of iteration numbers for different mesh sizes and methods.

Mesh size h	1/16	1/32	1/64	1/128	1/256
PDAS	14	27	54	113	226
PDIP	12	14	15	19	19
IPF	7(11)	9(15)	9(14)	7(13)	8(15)

Next we report on the outcome of our tests when employing the nested iteration concept and when using the interpolation of the approximate coarse grid solution as the initial point on the next finer grid. The results are summarized in Table 2. Note that for IPF we display only the total number of iterations on the respective grid. First observe that all algorithms experience a speed-up when using the nested iteration concept. PDAS especially, although lacking a function space convergence theory, performs remarkably well in this environment when compared to its variant on fixed grids. With respect to iteration numbers, our path concept is slightly faster than PDIP and PDAS. For a comparison of the CPU-times consumed by the respective algorithm, we define CPU-ratio(*algorithm*) = CPU-time(*algorithm*)/CPU-time(IPF). We have

$$\text{CPU-ratio(PDIP)} \approx 2 \quad \text{and} \quad \text{CPU-ratio(PDAS)} \approx 0.8;$$

i.e., PDIP requires twice as much CPU-time as IPF, while PDAS is slightly faster than IPF. This can be explained by the fact that PDAS performs system solves only on the currently inactive set, whereas IPF has to solve a system on the whole domain. However, we point out that the simple structure of the system matrix of IPF, i.e., $(-\Delta)_i + D_i$, with D_i a positive (semi)definite diagonal matrix on the respective grid

TABLE 2
Problem 1. *Comparison of iteration numbers for different mesh sizes and methods based on nested iteration.*

Mesh size h	1/4	1/8	1/16	1/32	1/64	1/128	1/256	total
PDAS	3	4	4	5	6	6	6	34
PDIP	3	2	4	4	5	6	7	31
IPF	4	3	3	4	5	5	5	29

with mesh size h_i, suggests that one can further speed up the solution process of the linear system in every iteration of Algorithm PDAS$_\gamma$. This, however, is beyond the scope of the present work.

Problem 2 *(lack of strict complementarity).* We construct a test problem for which the active set at the solution contains a subset, where strict complementarity does not hold. For this purpose define $\tilde{p}(w_1, w_2) = -w_1(1-w_1)w_2(1-w_2) \in H_0^1(\Omega) \cap C^\infty(\Omega)$ and $\ell(w_1, w_2) = -0.04 + 0.02w_1 - 0.03w_2$. The adjoint state at the solution is given by $p^* = \max(\tilde{p}, \ell)$. Further, we define the active set at the solution as $\mathcal{A}^* = \{\tilde{p} = \ell\}$ and the inactive set as its complement in Ω. We set

$$\tilde{\xi} = \begin{cases} \Delta\tilde{p} \text{ on } \mathcal{I}^*, \\ 0 \text{ on } \mathcal{A}^*, \end{cases}$$

$\xi = \Delta p$, and $\xi_L = \xi - \tilde{\xi}$. Next we decompose the Lagrange multiplier at the solution in a singular part λ_s^* and a regular part λ_r^*. The singular part consists of a line source contribution λ_L^* and an absolutely continuous part λ_a^*:

$$\lambda^* = \lambda_s^* + \lambda_r^* = \lambda_L^* + \lambda_a^* + \lambda_r^*.$$

We define

$$\lambda_L^* = \xi_L, \quad \lambda_a^* = \tilde{\xi}, \quad \lambda_r^*|_{\mathcal{I}^*} = -\tilde{\xi}, \quad \lambda_r^*|_{\mathcal{A}^*} = -100\tilde{p}|_{\mathcal{A}}.$$

Further, $u^* = p^*/\beta$, with $\beta = 0.01$, and $y^* = (-\Delta)^{-1}p^*/\beta$. The bound is given by $\psi|_{\mathcal{A}^*} = y^*|_{\mathcal{A}^*}$ and $\psi|_{\mathcal{I}^*} = (1/(1 + \mathrm{dist}(\mathcal{A}^*)^{0.1}))y^*|_{\mathcal{I}^*}$. Here $\mathrm{dist}(\mathcal{A}^*)$ denotes the distance function to the active set $\mathcal{A}^*$. The desired state is given by $y_d = y^* + \lambda_r^*$.

In Figure 2 we show the optimal state y_h^* (top left) and control u_h^* (top right) and the Lagrange multiplier λ_h^* (bottom) on a 128×128 grid.

Figure 3 provides the bound ψ (top left), the difference $\psi - y^*$ (top right), and the strongly active, weakly active and inactive sets (bottom). For the latter graph note that the weakly active set corresponds to the region where strict complementarity fails to hold (white region). The strongly active set, i.e., the set where $y^* = \psi$ and $\lambda^* > 0$, is displayed in gray. The inactive set is given by the black region.

The parameters in IPF had the values $r = 0.1$, $\tau_1 = 10$, $\tau_2 = 1.01$, $\tau_3 = 0.7$, and $\tau = 100$. We further set $\bar{\lambda} = 0$.

In Table 3 we report the result corresponding to Problem 2. The arrangement of the table is analogous to that of Table 2 for Problem 1. First we note that, compared to the previous problem, the iteration numbers for PDAS increase significantly as the mesh is refined. This is also reflected in the following CPU-time comparison for this example:

$$\text{CPU-ratio(PDIP)} \approx 1.2 \quad \text{and} \quad \text{CPU-ratio(PDAS)} \approx 1.8.$$

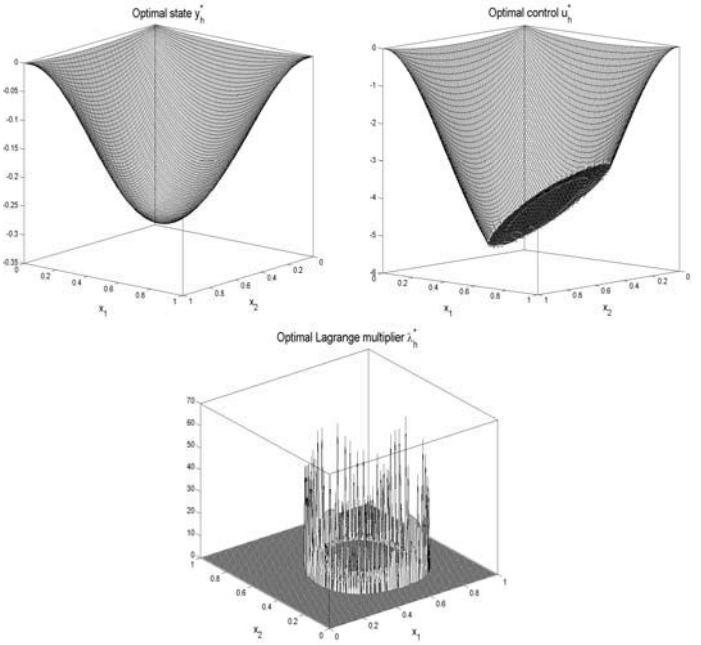


FIG. 2. *Problem* 2. *Optimal state (top left), optimal control (top right), and optimal multiplier (bottom) for* $h = 1/128$.

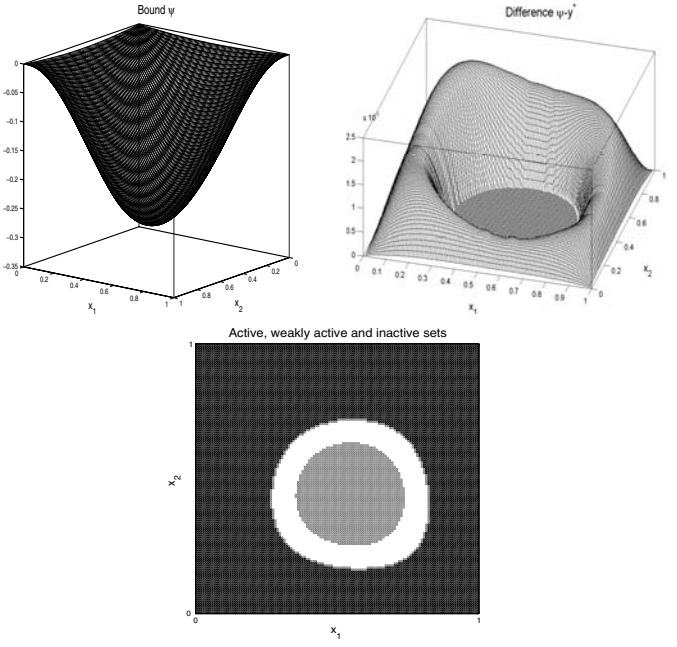


FIG. 3. *Problem* 2. *Bound* ψ *(top left), the difference* $\psi - y^*$ *(top right), and the active and inactive sets (bottom) for* $h = 1/128$.

TABLE 3
Problem 2. *Comparison of iteration numbers for different mesh sizes and methods based on nested iteration.*

Mesh size h	1/4	1/8	1/16	1/32	1/64	1/128	1/256	total
PDAS	2	4	5	9	10	21	40	91
PDIP	3	2	3	3	6	12	11	40
IPF	7	2	4	4	6	8	15	46

Further observe that PDIP needs the lowest total number of iterations, but it requires about 20% more CPU-time than IPF. This can be explained by the fact that interior point approaches typically enlarge the system size by adding slack variables. An a priori system reduction obtained by choosing specific pivots allows us to reduce the large indefinite system occurring in every iteration to a system of the same size as that for IPF. However, in order to compute a solution to the overall system, for PDIP several backward substitutions are required, in contrast to the efforts needed in IPF. Further, we point out that practically relevant versions of primal-dual path-following interior point methods, such as the Mehrotra predictor-corrector algorithm or the Mizuno–Todd–Ye variant (see [1] for an adaptation of these two variants to optimal control of PDEs and for further references), require additional system solves. Depending on the variant, one has either two system solves where the second system has a different right hand side but the same system matrix or two system solves where the second system has a different right hand side *and* a different system matrix as well. The triangular system solves for backward substitution, and the additional system solves mentioned above add significantly to the overall CPU-time. Let us further point out that the reduced systems of PDIP typically have a more involved structure when compared to the systems of IPF; for an account of this fact see, for instance, the systems (4.6) and (4.7) in [1], and the system in Algorithm PDAS$_\gamma$.

REFERENCES

[1] M. BERGOUNIOUX, M. HADDOU, M. HINTERMÜLLER, AND K. KUNISCH, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.

[2] M. BERGOUNIOUX AND K. KUNISCH, *Primal-Dual Strategy for State-Constrained Optimal Control Problems*, Comput. Optim. Appl., 22 (2002), pp. 193–224.

[3] E. CASAS, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[4] X. CHEN, Z. NASHED, AND L. QI, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.

[5] J.-C. DE LOS REYES AND K. KUNISCH, *A semi-smooth Newton method for control constrained boundary optimal control of the Navier–Stokes equations*, Nonlinear Anal., 62 (2005), pp. 1289–1316.

[6] A. FORSGREN, P. E. GILL, AND M. H. WRIGHT, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.

[7] W. HACKBUSCH, *Multigrid Methods and Applications*, Springer Ser. Computat. Math. 4, Springer-Verlag, Berlin, 1985.

[8] E. HEBEY, *Nonlinear Analysis on Manifolds: Sobolev Spaces and Inequalities*, Courant Lecture Notes in Math. 5, New York University, Courant Institute of Mathematical Sciences, New York, 1999.

[9] M. HINTERMÜLLER AND M. HINZE, *A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier–Stokes system subject to control constraints*, SIAM J. Optim., 16 (2006), pp. 1177–1200.

[10] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.

[11] M. Hintermüller, V. A. Kovtunenko, and K. Kunisch, *Generalized Newton methods for crack problems with nonpenetration condition*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 586–610.
[12] M. Hintermüller and K. Kunisch, *Path-following methods for a class of constrained minimization problems in function space*, SIAM J. Optim, 17 (2006), pp. 159–187.
[13] M. Hintermüller and W. Ring, *A level set approach for the solution of a state-constrained optimal control problem*, Numer. Math., 98 (2004), pp. 135–166.
[14] K. Ito and K. Kunisch, *Semi-smooth Newton methods for variational inequalities of the first kind*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 41–62.
[15] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170. Springer-Verlag, New York, 1971.
[16] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1, Travaux et Recherches Mathématiques, No. 17, Dunod, Paris, 1968.
[17] H. Maurer and J. Zowe, *First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
[18] S. Mehrotra, *On the implementation of a primal-dual interior point method*, SIAM J. Optim., 2 (1992), pp. 575–601.
[19] C. Meyer and F. Tröltzsch, *On an elliptic optimal control problem with pointwise mixed control-state constraints*, in Recent Advances in Optimization, A. Seeger, ed., Lecture Notes in Econom. Math. Systems 563, Springer, Berlin, 2006, pp. 187–204.
[20] U. Prüfert, F. Tröltzsch, and M. Weiser, *The Convergence of an Interior Point Method for an Elliptic Control Problem with Mixed Control-State Constraints*, Preprint 36-2004, Technische Universität, Berlin, 2004.
[21] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.
[22] R. J. Vanderbei, *Linear Programming Foundations and Extensions*, 2nd ed., Internat. Ser. Oper. Res. Sci. 37, Kluwer Academic Publishers, Boston, MA, 2001.
[23] M. Weiser, *Interior Point Methods in Function Space*, ZIB-Report 03-35, Konrad-Zuse Zentrum für Informationstechnik, Berlin, 2003.
[24] M. Weiser and P. Deuflhard, *The Central Path Towards the Numerical Solution of Optimal Control Problems*, ZIB-Report 01-12, Konrad-Zuse Zentrum für Informationstechnik, Berlin, 2001.
[25] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.
[26] Y. Ye, *Interior Point Algorithms. Theory and Analysis*, John Wiley & Sons Inc., New York, 1997.
[27] Y. Zhang and D. T. Zhang, *On polynomiality of the Mehrotra-type predictor-corrector interior-point algorithms*, Math. Programming, 68 (1995), pp. 303–318.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1222–1250

A NUMERICAL METHOD FOR A GENERALIZED ALGEBRAIC RICCATI EQUATION*

DELIN CHU†, WEN-WEI LIN‡, AND ROGER C. E. TAN†

Abstract. In this paper we develop a numerical method for computing the semistabilizing solution of a generalized algebraic Riccati equation (GARE). The semistabilizing solution of such a GARE has been used to characterize the solvability of the (J, J')-spectral factorization problem for general rational matrices which have poles and zeros on the extended imaginary axis. The main difficulty for solving such a GARE is that its associated skew-Hamiltonian/Hamiltonian pencil has eigenvalues on the extended imaginary axis; consequently, it is not clear which eigenspace of the associated skew-Hamiltonian/Hamiltonian pencil can characterize the desired semistabilizing solution; i.e., it is not clear which eigenvectors and principal vectors corresponding to the eigenvalues on the extended imaginary axis should be contained in the eigenspace that we wish to compute, and hence the well-known generalized eigenvalue approach for the classical algebraic Riccati equations cannot be directly employed for it. Our proposed method consists of computations of the eigendecomposition of the system pencil corresponding to the eigenvalues on the extended imaginary axis and the stable eigenspace of an augmented matrix pencil; hence, it is a generalization of the generalized eigenvalue approach for the classical algebraic Riccati equations.

Key words. generalized algebraic Riccati equation, semistabilizing solution, eigenvalues, stable eigenspace, extended imaginary axis

AMS subject classifications. 93B05, 93B40, 93B52, 65F35

DOI. 10.1137/050625035

1. Introduction.

Throughout this paper, $\mathbf{C}_-$ and $\mathbf{C}_0$ denote the open left half complex plane and the imaginary axis, respectively, and $\mathbf{C}_{0e} = \mathbf{C}_0 \cup \{\infty\}$.

The (J, J')-spectral factorization of rational matrices has many important applications in optimal Hankel-norm model reduction and H_∞ optimization [7, 19], transport theory [20], and stochastic filtering [13]. It is now well known that the algebraic Riccati equations (AREs) play important roles for solving various (J, J')-spectral factorization problems. In [17] the difficulties in numerically solving the classical spectral factorization (a special case of the (J, J')-spectral factorization) of rational matrix $G(s)$ are shown to be related to the existence of poles and zeros of $G(s)$ on $\mathbf{C}_{0e}$ and to the noninvertibility of $G^T(-s)G(s)$. If none of these elements are present, the computation of the classical spectral factorization problem reduces to solving a standard ARE for which numerically reliable algorithms are available [4, 8, 14, 24, 27, 28]. The canonical J-spectral factorization (the (J, J')-spectral factorization with the condition that $G(s)$ has no poles and zeros on $\mathbf{C}_{0e}$) has been solved in [16] in terms of the stabilizing solution of an indefinite ARE. The (J, J')-spectral factorization problem for general rational matrices with poles/zeros on $\mathbf{C}_{0e}$ has also been considered in [2, 3] based on the semistabilizing solution of a generalized

*Received by the editors February 22, 2005; accepted for publication (in revised form) March 7, 2006; published electronically September 15, 2006.
http://www.siam.org/journals/sicon/45-4/62503.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matchudl@nus.edu.sg, scitance@nus.edu.sg). The work of these authors was partially supported by Research grant R-146-000-047-112 of the National University of Singapore.

‡Department of Mathematics, National Tsing Hua University, Hsinchu 30043, Taiwan (wwlin@am.nthu.edu.tw).

algebraic Riccati equation (GARE) of the form

$$\begin{cases} A_a^T X_a + X_a^T A_a + (C_a^T J C_a - B_a J' B_a^T) - X_a^T B_a (J')^{-1} B_a^T X_a = 0, \\ E_a^T X_a = X_a^T E_a, \end{cases} \tag{1}$$

where

$$E_a = \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix}, \quad A_a = \begin{bmatrix} A & B \\ 0 & I \end{bmatrix}, \quad C_a = \begin{bmatrix} C & D \end{bmatrix}, \quad B_a = \begin{bmatrix} 0 \\ -I \end{bmatrix},$$

$E, A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{p\times n}$, $D \in \mathbf{R}^{p\times m}$, $J \in \mathbf{R}^{p\times p}$, $J' \in \mathbf{R}^{m\times m}$, E is singular, and J and J' are symmetric and nonsingular. The semistabilizing solution of the GARE (1) is defined as follows.

DEFINITION 1 (see [2, 3]). *A solution $X_a \in \mathbf{R}^{(n+m)\times(n+m)}$ of the GARE (1) is called a semistabilizing solution if*

(i) *the pencil $A_a - B_a(J')^{-1}B_a^T X_a - sE_a$ is regular and its eigenvalues lie in $\mathbf{C}_- \cup \mathbf{C}_{0e}$;*

(ii) *the matrix pair $(C_a, A_a - B_a(J')^{-1}B_a^T X_a - sE_a)$ has neither observable finite poles on $\mathbf{C}_0$ nor observable impulsive poles.*

Recently, a numerical method for computing the semistabilizing solution of the GARE (1) has been proposed in [2, 3] under the following assumptions:

(A1) (E, A, B) is finite dynamic stabilizable and impulse controllable, i.e.,

$$\operatorname{rank}\begin{bmatrix} -sE + A & B \end{bmatrix} = n \quad \forall s \in \mathbf{C}\backslash\mathbf{C}_-$$

and

$$\operatorname{rank}\begin{bmatrix} E & A & B \\ 0 & E & 0 \end{bmatrix} = n + \operatorname{rank}(E);$$

$$\max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} -sE + A & B \\ C & D \end{bmatrix} = n + m. \tag{A2}$$

A key step in this method is to compute a nonsingular solution of a nonsymmetric ARE of the form

$$A_{a,0}^T X_{a,0} + X_{a,0}^T A_{a,0} + Q_{a,0} + X_{a,0}^T R_{a,0} X_{a,0} = 0. \tag{2}$$

However, the nonsymmetric ARE (2) is much harder to solve than its symmetric version. Up to now, symmetric AREs have been studied extensively, and many numerically reliable methods are available [4, 8, 14, 24, 25, 27, 28]. On the contrary, these existing methods do not work for the nonsymmetric AREs of the form (2). Indeed, there is still a lack of numerically reliable methods for solving the nonsymmetric AREs of the form (2).

Denote

$$G(s) = D + C(sE - A)^{-1}B =: \left[\begin{array}{c|c} -sE + A & B \\ \hline C & D \end{array}\right]. \tag{3}$$

Regarding the assumption (A1), it implies that the descriptor system

$$E\dot{x} = Ax + Bu, \quad y = Cx + Du \tag{4}$$

is controllable at infinity. However, control of descriptor systems is of practical importance for constrained mechanical systems and mechatronics; there, controllability

at infinity can usually not be assumed. But, the uncontrollable part of the system (4) at infinity does not contribute to its transfer matrix $G(s)$. In fact, if the system (4) is not controllable at infinity, according to [11, 26], we can always compute orthogonal matrices P and $\mathcal{P}$ such that

$$\text{(5)}\qquad \begin{cases} \mathcal{P}(sE-A)P = \begin{matrix} \\ \left[\begin{array}{cc} s\mathcal{E}-\mathcal{A} & s\mathcal{E}_{12}-\mathcal{A}_{12} \\ 0 & s\mathcal{E}_{22}-\mathcal{A}_{22} \end{array}\right] \end{matrix} \begin{matrix} \mu & n-\mu \\ \} \mu \\ \} n-\mu \end{matrix}, \\ \mathcal{P}B = \left[\begin{array}{c} \mathcal{B} \\ 0 \end{array}\right] \begin{matrix} \} \mu \\ \} n-\mu \end{matrix}, \quad CP = \begin{matrix} \mu & n-\mu \\ \left[\begin{array}{cc} \mathcal{C} & \mathcal{C}_2 \end{array}\right] \end{matrix}, \end{cases}$$

such that $(\mathcal{E}, \mathcal{A}, \mathcal{B})$ is finite dynamic stabilizable and impulse controllable. Then we will have

$$G(s) = D + \mathcal{C}(s\mathcal{E} - \mathcal{A})^{-1}\mathcal{B}.$$

In addition, the assumption (A2) is necessary for the existence of the (J, J')-spectral factorization of $G(s)$ [2, 3]. Hence, in this paper we shall also assume that *the assumptions* (A1) *and* (A2) *hold*.

Due to the importance of the GARE (1) in solving the (J, J')-spectral factorization problem of general rational matrices, in this paper we study the numerical computation of the semistabilizing solution of the GARE (1) and extend the well-known generalized eigenvalue approach [8, 14, 27] for the classical AREs to the GARE (1). To the best of our knowledge, similar results do not exist in the literature.

The generalized eigenvalue approach [8, 14, 27] for solving the classical AREs is also called the Schur-type method or invariant subspace method and has been studied extensively in the literature (see [8, 14, 27] and the references therein).

Define

$$\text{(6)}\qquad H_a(s) := \left[\begin{array}{cc} A_a - sE_a & -B_a(J')^{-1}B_a^T \\ -C_a^T J C_a + B_a J' B_a^T & -A_a^T - sE_a^T \end{array}\right].$$

$H_a(s)$ is the skew-Hamiltonian/Hamiltonian pencil associated with the GARE (1). The following result is trivial.

COROLLARY 2. (i) *X_a is a solution of the GARE* (1) *if and only if*

$$H_a(s)\left[\begin{array}{c} I \\ X_a \end{array}\right] = \left[\begin{array}{c} I \\ X_a^T \end{array}\right](A_a - B_a(J')^{-1}B_a^T X_a - sE_a).$$

(ii) *The GARE* (1) *has a solution X_a such that the pencil $A_a - B_a(J')^{-1}B_a^T X_a - sE_a$ is regular, and all its eigenvalues are on $\mathbf{C}_- \cup \mathbf{C}_{0e}$ if and only if there exist matrices*

$$\begin{matrix} n+m \\ \left[\begin{array}{c} Y_1 \\ Y_2 \end{array}\right] \end{matrix} \begin{matrix} \\ \} n+m \\ \} n+m \end{matrix} \quad \textit{and} \quad \begin{matrix} n+m \\ \left[\begin{array}{c} Z_1 \\ Z_2 \end{array}\right] \end{matrix} \begin{matrix} \\ \} n+m \\ \} n+m \end{matrix}$$

such that

$$H_a(s)\left[\begin{array}{c} Y_1 \\ Y_2 \end{array}\right] = \left[\begin{array}{c} Z_1 \\ Z_2 \end{array}\right](\Phi - s\Theta), \quad \operatorname{rank}(Y_1) = \operatorname{rank}(Z_1) = n+m, \quad Y_1^T Z_2 = Y_2^T Z_1,$$

where the pencil $\Phi - s\Theta \in \mathbf{R}^{(n+m)\times(n+m)}$ is regular and all its eigenvalues are on $\mathbf{C}_- \cup \mathbf{C}_{0e}$. In this case, such a solution X_a is given by $X_a = Y_2 Y_1^{-1}$.

From Corollary 2, there exists a close connection between the semistabilizing solution of the GARE (1) and the eigenspace of the pencil $H_a(s)$ corresponding to the eigenvalues on $\mathbf{C}_-$ and $\mathbf{C}_{0e}$. However, the eigenstructure of the pencil $H_a(s)$ corresponding to the eigenvalues on $\mathbf{C}_{0e}$ is much more complicated than the structure of its stable eigenspace. This issue is easy to understand as follows: Let τ_1 and τ_2 denote the dimensions of the eigenspaces of the pencil $H_a(s)$ corresponding to the eigenvalues on $\mathbf{C}_-$ and $\mathbf{C}_{0e}$, respectively. Since E is singular, we have

$$\tau_1 < n,\ \tau_1 + \frac{1}{2}\tau_2 = n + m,$$

provided $H_a(s)$ is regular. This implies that there are many different eigenspaces with dimension $n+m$ corresponding to part of the eigenvalues on $\mathbf{C}_-$ and $\mathbf{C}_{0e}$. Hence, it is not possible to check whether one of such eigenspaces characterizes the existence of the semistabilizing solution of the GARE (1) without having some extra insight. Consequently, it is not clear which eigenvectors and principal vectors corresponding to the eigenvalues on $\mathbf{C}_{0e}$ should be contained in the eigenspace that we wish to compute. Therefore, Corollary 2 is not very useful for the numerical computation of the GARE (1), and the generalized eigenvalue approach [8, 14, 27] for solving the classical AREs cannot be directly employed for the GARE (1). In fact it is a challenge to extend the generalized eigenvalue approach for the classical AREs to the GARE (1).

The main contributions of this paper include numerically verifiable necessary and sufficient conditions for the existence of the semistabilizing solution of the GARE (1) and a new numerical method for computing such a semistabilizing solution. The main ingredients of our method include (i) the eigendecomposition of the system pencil corresponding to the eigenvalues on $\mathbf{C}_{0e}$; (ii) the stable eigenspace of an augmented matrix pencil. These two eigendecompositions can be computed by the numerically backward stable algorithms [6, 10, 26]. Hence, we extend the generalized eigenvalue approach for the classical AREs to the GARE (1).

2. Main results. In this section we present necessary and sufficient conditions for the existence of the semistabilizing solution of the GARE (1) and develop a numerically reliable method to compute such a semistabilizing solution. Since the main difficulty for computing the semistabilizing solution of the GARE (1) is that the skew-Hamiltonian/Hamiltonian pencil $H_a(s)$ defined in (6) has eigenvalues on $\mathbf{C}_{0e}$, and also because

$$\operatorname{rank}(H_a(s)) = \operatorname{rank}\left(\begin{bmatrix} A-sE & B & 0 & 0 \\ C & D & 0 & -J^{-1} \\ 0 & 0 & -A^T-sE^T & -C^T \\ 0 & 0 & -B^T & -D^T \end{bmatrix}\right) + m - p \quad \forall s \in \mathbf{C},$$

which implies that the eigenvalues of $H_a(s)$ on $\mathbf{C}_{0e}$ consist of the eigenvalues of the pencils

$$\begin{bmatrix} A-sE & B \\ C & D \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -A^T-sE^T & -C^T \\ -B^T & -D^T \end{bmatrix}$$

on $\mathbf{C}_{0e}$, considering that the pencils

$$\begin{bmatrix} A-sE & B \\ C & D \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -A^T-sE^T & -C^T \\ -B^T & -D^T \end{bmatrix}$$

have the same eigenvalues on $\mathbf{C}_{0e}$, we first isolate the eigenvalues of the pencil

$$\begin{bmatrix} A - sE & B \\ C & D \end{bmatrix}$$

on $\mathbf{C}_{0e}$ via the factorizations (7) and (8) in the following theorem.

THEOREM 3. *Let* $n_1 = \operatorname{rank}(E)$.

(i) *There exist orthogonal matrices* $U, V \in \mathbf{R}^{n\times n}$ *and* $\mathcal{V} \in \mathbf{R}^{(n-n_1+m)\times(n-n_1+m)}$ *such that*

$$U \begin{bmatrix} A - sE & B \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \mathcal{V} \end{bmatrix} = \begin{array}{c} \begin{matrix} n_1 & n-n_1 & m \end{matrix} \\ \begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \end{bmatrix} \begin{matrix} \}n_1 \\ \}n - n_1 \end{matrix} \end{array}, \tag{7}$$

where E_{11} *and* A_{22} *are nonsingular.*

(ii) *Denote*

$$\begin{bmatrix} C & D \end{bmatrix} \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \mathcal{V} \end{bmatrix} = \begin{array}{c} \begin{matrix} n_1 & n-n_1 & m \end{matrix} \\ \begin{bmatrix} C_1 & C_2 & \mathcal{D} \end{bmatrix}. \end{array}$$

Then the eigendecomposition of the pencil

$$\begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ C_1 & C_2 & \mathcal{D} \end{bmatrix}$$

corresponding to the eigenvalues on $\mathbf{C}_{0e}$ *is of the form*

$$S^T \begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ C_1 & C_2 & \mathcal{D} \end{bmatrix} Q$$

$$= \begin{array}{c} \begin{matrix} \mu_1 & n_1 - \mu_1 & m & n - n_1 \end{matrix} \\ \begin{bmatrix} \Phi_{11} - s\Theta_{11} & 0 & 0 & 0 \\ \Phi_{21} - s\Theta_{21} & \Phi_{22} - s\Theta_{22} & \Phi_{23} & \Phi_{24} - s\Theta_{24} \\ \Phi_{31} - s\Theta_{31} & \Phi_{32} & \Phi_{33} & \Phi_{34} \\ \Phi_{41} - s\Theta_{41} & 0 & 0 & \Phi_{44} \end{bmatrix} \begin{matrix} \}p + \mu_1 - m \\ \}n_1 - \mu_1 \\ \}m \\ \}n - n_1 \end{matrix} \end{array}, \tag{8}$$

where S *and* Q *are orthogonal matrices with partitioning*

$$Q = \begin{array}{c} \begin{matrix} \mu_1 & n_1 - \mu_1 & m & n - n_1 \end{matrix} \\ \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n - n_1 \\ \}m \end{matrix} \end{array},$$

$$S = \begin{array}{c} \begin{matrix} p + \mu_1 - m & n_1 - \mu_1 & m & n - n_1 \end{matrix} \\ \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & 0 & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n - n_1 \\ \}p \end{matrix} \end{array}, \tag{9}$$

and

$$\operatorname{rank}(\Theta_{11}) = \mu_1, \quad \operatorname{rank}(\Theta_{22}) = n_1 - \mu_1,$$

$$\operatorname{rank}(\Phi_{44}) = n - n_1, \quad \operatorname{rank}(S_{24}) = n - n_1, \tag{10}$$

$$\operatorname{rank}(\Phi_{11} - s\Theta_{11}) = \mu_1 \quad \forall s \in \mathbf{C}_{0e}, \tag{11}$$

$$\operatorname{rank}\begin{bmatrix} \Phi_{22} - s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \end{bmatrix} = (n_1 - \mu_1) + m \quad \forall s \in \mathbf{C}\backslash\mathbf{C}_0. \tag{12}$$

Proof. The decompositions (7) and (8) are constructed in the appendix. □

Remark 1. It is clear that all eigenvalues of

$$\begin{bmatrix} A - sE & B \\ C & D \end{bmatrix}$$

on $\mathbf{C}_{0e}$ are isolated to

$$\begin{bmatrix} \Phi_{22} - s\Theta_{22} & \Phi_{23} & \Phi_{24} - s\Theta_{24} \\ \Phi_{32} & \Phi_{33} & \Phi_{34} \\ 0 & 0 & \Phi_{44} \end{bmatrix}.$$

Let

$$M_a(s) := \begin{bmatrix} A_{11} - sE_{11} & A_{12} & 0 & 0 & B_1 \\ A_{21} & A_{22} & 0 & 0 & 0 \\ -C_1^T J C_1 & -C_1^T J C_2 & -A_{11}^T - sE_{11}^T & -A_{21}^T & -C_1^T J \mathcal{D} \\ -C_2^T J C_1 & -C_2^T J C_2 & -A_{12}^T & -A_{22}^T & -C_2^T J \mathcal{D} \\ -\mathcal{D}^T J C_1 & -\mathcal{D}^T J C_2 & -B_1^T & 0 & -\mathcal{D}^T J \mathcal{D} \end{bmatrix}. \tag{13}$$

Let ν be the dimension of the stable eigenspace of $M_a(s)$, and the columns of matrix

$$L = \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \overset{\nu}{} \\ \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \end{bmatrix} \end{matrix} \begin{matrix} \\ \}n_1 \\ \}n - n_1 \\ \}n_1 \\ \}n - n_1 \\ \}m \end{matrix}$$

form a basis for this stable eigenspace; consequently,

$$\begin{cases} \begin{bmatrix} A_{11} & A_{12} & 0 & 0 & B_1 \\ A_{21} & A_{22} & 0 & 0 & 0 \\ -C_1^T J C_1 & -C_1^T J C_2 & -A_{11}^T & -A_{21}^T & -C_1^T J \mathcal{D} \\ -C_2^T J C_1 & -C_2^T J C_2 & -A_{12}^T & -A_{22}^T & -C_2^T J \mathcal{D} \\ -\mathcal{D}^T J C_1 & -\mathcal{D}^T J C_2 & -B_1^T & 0 & -\mathcal{D}^T J \mathcal{D} \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \end{bmatrix} = \begin{bmatrix} E_{11} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & E_{11}^T & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \end{bmatrix} \Delta, \\ \Delta \in \mathbf{R}^{\nu \times \nu} \text{ is stable.} \end{cases} \tag{14}$$

It is known that if $n_1 = \nu = n$, then the stabilizing solution of the GARE (1) is determined by L. Analogously, a similar result holds true, as shown in the following theorem.

THEOREM 4. *Assume that the decomposition* (7) *and the eigendecompositions* (8) *and* (14) *have been determined. Then the GARE* (1) *has a semistabilizing solution* X_a *if and only if*

$$S_{33}^T J S_{33} = \mathcal{D}_0^T J' \mathcal{D}_0 \quad \textit{for some nonsingular matrix } \mathcal{D}_0 \in \mathbf{R}^{m\times m}, \tag{15}$$

$$(n_1 - \mu_1) + \nu = n_1 \ (\textit{i.e.,}\ \mu_1 = \nu),\ \begin{bmatrix} L_1 & Q_{12} \end{bmatrix} \textit{ is nonsingular.} \tag{16}$$

Furthermore, a semistabilizing solution X_a *of the GARE* (1) *is given by*

$$X_a = \begin{bmatrix} U^T & 0 \\ 0 & I \end{bmatrix} \hat{X}_a \begin{bmatrix} I & 0 \\ 0 & \mathcal{V}^T \end{bmatrix} \begin{bmatrix} V^T & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & J' \end{bmatrix}, \tag{17}$$

where $\hat{X}_a$ *is determined by the following linear system of equations:*

$$\begin{bmatrix} I & 0 & 0 \\ 0 & A_{22}^T & 0 \\ 0 & 0 & I_m \end{bmatrix} \hat{X}_a \begin{bmatrix} L_1 & Q_{12} & 0 \\ 0 & 0 & I_{n+m-n_1} \end{bmatrix}$$
$$= \begin{bmatrix} L_3 & 0 & 0 & 0 \\ A_{22}^T L_4 & -C_2^T J C_1 Q_{12} & -C_2^T J C_2/2 & -C_2^T J \mathcal{D} \\ 0 & -J'\mathcal{D}_0 \Phi_{32} & 0 & -J'\mathcal{D}_0 \Phi_{33} Q_{33}^T \end{bmatrix} \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ -L_2 & Q_{22} & I & 0 \\ -L_5 & 0 & 0 & I \end{bmatrix}. \tag{18}$$

Proof. The proof is given in section 3. □

Theorem 4 leads to the following algorithm for computing a semistabilizing solution of the GARE (1).

ALGORITHM 1.

Input: *Matrices* $A, E \in \mathbf{R}^{n\times n}, B, \in \mathbf{R}^{n\times m}, C \in \mathbf{R}^{p\times n}, D \in \mathbf{R}^{p\times m}$, $J \in \mathbf{R}^{p\times p}$, *and* $J' \in \mathbf{R}^{m\times m}$. *Here* J *and* J' *are symmetric.*

Output: *A semistabilizing solution* X_a *of the GARE* (1) *(if possible).*

Step 1: *Compute the decompositions* (7) *and* (8) *using the numerical procedure in the appendix.*

Step 2: *Verify the condition* (15). *If it is not true, print "The GARE* (1) *has no semistabilizing solutions" and then stop. Otherwise, compute the related matrix* $\mathcal{D}_0$.

Step 3: *Compute the eigendecomposition* (14).

Step 4: *Verify the singularity of* $\begin{bmatrix} L_1 & Q_{12} \end{bmatrix}$. *If it is singular, print "The GARE* (1) *has no semistabilizing solutions" and then stop. Otherwise, compute* $\hat{X}_a$ *by solving* (18) [6].

Step 5: *Compute* X_a *by* (17) *and then output it.*

Algorithm 1 is an extension of the generalized eigenvalue approach for the classical AREs, and its main ingredients are the decompositions (7), (8), and (14). All these decompositions are computed by existing numerically backward stable methods [6, 10, 26]. One possible source in Algorithm 1 for loss of accuracy in the numerical solution of the GARE (1) is the condition numbers of $\begin{bmatrix} L_1 & Q_{12} \end{bmatrix}$ and A_{22}. The similar problem is already known for the classical AREs, and there the problem can be resolved by proper scaling [18, 21]. It seems that the results in [18, 21] can be extended to the GARE (1). This is an interesting research topic and is worthy of further investigation.

In the following we present two examples[1] to illustrate Algorithm 1. All computations were performed using MATLAB 7.0 on a Linux system with IEEE standard double precision arithmetic and machine precision $\epsilon \approx 2.22 \times 10^{-16}$.

[1]These two examples were done by Dr. Bai Zhengjian.

Example 1 (see [2, 3]). Let

$$J = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad J' = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

and

$$G(s) = \begin{bmatrix} \frac{s}{(s+3)^2} & \frac{1}{s} \\ 0 & \frac{(s-1)(s^2+2)}{s} \\ 0 & 0 \end{bmatrix} = \left[\begin{array}{c|c} A - sE & B \\ \hline C & D \end{array}\right],$$

where

$$A - sE = \begin{bmatrix} -s & 1 & 0 & 0 & 0 & 0 \\ -9 & -s-6 & 0 & 0 & 0 & 0 \\ 0 & 0 & -s & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -s & 0 \\ 0 & 0 & 0 & 0 & 1 & -s \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -2 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}.$$

We compute a semistabilizing solutions X_a of the related GARE (1) by Algorithm 1 and obtain

$$X_a(:, 1:4) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -6.4358 \times 10^{-16} & 5.2500 \times 10^{-15} & 8.0000 \times 10^{0} & 0 \\ -3.2313 \times 10^{-16} & 2.6263 \times 10^{-15} & 4.0000 \times 10^{0} & 0 \\ 0 & 0 & 1.2546 \times 10^{-15} & 0 \\ 1.1684 \times 10^{-16} & 7.2990 \times 10^{-15} & 4.0000 \times 10^{0} & 2.0000 \times 10^{0} \\ -1.9971 \times 10^{-15} & 3.4206 \times 10^{-15} & 2.0000 \times 10^{0} & 1.0000 \times 10^{0} \\ -2.7210 \times 10^{-16} & 1.0000 \times 10^{0} & 1.0000 \times 10^{0} & -3.0761 \times 10^{-16} \end{bmatrix},$$

$$X_a(:, 5:8) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4.0000 \times 10^{0} & -1.6924 \times 10^{-15} & 0 & 0 \\ 2.0000 \times 10^{0} & -8.4131 \times 10^{-16} & 0 & 0 \\ 6.2731 \times 10^{-16} & 0 & 0 & 0 \\ 2.0000 \times 10^{0} & 2.0000 \times 10^{0} & 0 & 2.0000 \times 10^{0} \\ 1.0000 \times 10^{0} & 2.0000 \times 10^{0} & 1.0000 \times 10^{0} & 0 \\ -1.2835 \times 10^{-15} & 4.5762 \times 10^{-16} & 0 & -1.0000 \times 10^{0} \end{bmatrix}.$$

The accuracy of the GARE (1) can be measured by the following quantities:

$$\|A_a^T X_a + X_a^T A_a + (C_a^T J C_a - B_a J' B_a^T) - X_a^T B_a (J')^{-1} B_a^T X_a\|_2 = 1.4249 \times 10^{-14},$$

$$\frac{\|A_a^T X_a + X_a^T A_a + (C_a^T J C_a - B_a J' B_a^T) - X_a^T B_a (J')^{-1} B_a^T X_a\|_2}{\|A_a^T X_a\|_2 + \|X_a^T A_a\|_2 + \|C_a^T J C_a - B_a J' B_a^T\|_2 + \|X_a^T B_a (J')^{-1} B_a^T X_a\|_2} = 3.7333 \times 10^{-16}$$

and

$$\|E_a^T X_a - X_a^T E_a\|_2 = 8.6151 \times 10^{-15}, \quad \frac{\|E_a^T X_a - X_a^T E_a\|_2}{\|E_a^T X_a\|_2 + \|X_a^T E_a\|_2} = 4.3075 \times 10^{-16}.$$

Example 2. Example 5.5 in [9] is on a descriptor system describing an RLC electrical circuit with an independent loop containing capacitors and voltage sources only and an independent cutset with inductors and current sources only. The GARE (1) corresponding to this system has no semistabilizing solutions. Thus, we modify this system slightly and get

$$E = \begin{bmatrix} -1 & -1 & 0.005 & -0.005 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.005 & -0.005 & 0 & 0 & 0 & 0 \\ -0.001 & 0 & 0 & -0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -5 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & -0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.75 & 0 & 0.1 & -0.2 & -0.2 \end{bmatrix},$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 500 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ -1 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The matrix E is ill-conditioned since

$$\frac{\sigma_{max}(E)}{\sigma_{min}(E)} = 1.7897 \times 10^5,$$

where $\sigma_{max}(E)$ and $\sigma_{min}(E)$ denote the maximal amd mininal nonzero singular values of E. Let

$$J = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad J' = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

By performing Algorithm 1, we obtain a semistabilizing solution X_a of the GARE (1)

as follows:

$$X_a(:,1:5) = \begin{bmatrix} 7.5239\times10^{-15} & 7.5239\times10^{-15} & 3.1982\times10^{-14} & 1.9919\times10^{-15} & -6.7917\times10^{-16} \\ 1.1351\times10^{-15} & -4.8032\times10^{-16} & 1.0500\times10^{-2} & -4.9582\times10^{-4} & -2.2298\times10^{-16} \\ 0 & 0 & 4.6515\times10^{-16} & -1.7048\times10^{-16} & 0 \\ 0 & 0 & -1.4111\times10^{-14} & -3.3632\times10^{-16} & 2.9737\times10^{-16} \\ 1.2500\times10^{-1} & -1.2500\times10^{-1} & 2.5000\times10^{-1} & -4.8749\times10^{-1} & -5.2501\times10^{-3} \\ -4.0938\times10^{-15} & -4.1169\times10^{-15} & 5.0000\times10^{-1} & 2.5011\times10^{-2} & -1.0500\times10^{-2} \\ 2.7092\times10^{-15} & 2.6592\times10^{-15} & 4.7499\times10^{-1} & 2.5011\times10^{-2} & -1.0004\times10^{-2} \\ -1.2796\times10^{-15} & -1.2592\times10^{-15} & 1.0572\times10^{-13} & 5.7185\times10^{-15} & -2.2291\times10^{-15} \\ -9.1581\times10^{-1} & -1.0671\times10^{-1} & 4.0455\times10^{-1} & -1.9318\times10^{-1} & -8.4958\times10^{-3} \\ 2.9785\times10^{-1} & -5.1126\times10^{-1} & -4.0455\times10^{-1} & -1.0428\times10^{0} & 8.4958\times10^{-3} \end{bmatrix},$$

$$X_a(:,6:10) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -8.0583\times10^{-16} & -2.4175\times10^{-15} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 9.1926\times10^{-16} & -5.0668\times10^{-16} & -3.4015\times10^{-16} & -1.2500\times10^{-1} & -3.7500\times10^{-1} \\ -4.9531\times10^{-16} & 9.9060\times10^{-16} & 9.9060\times10^{-16} & 0 & 0 \\ -4.6191\times10^{-16} & 9.2380\times10^{-16} & 9.2380\times10^{-16} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -3.6153\times10^{-14} & 6.3089\times10^{-15} & 6.3228\times10^{-15} & -1.1067\times10^{0} & -1.0671\times10^{-1} \\ 6.4994\times10^{-14} & 6.1249\times10^{-15} & 6.1804\times10^{-15} & -5.1126\times10^{-1} & 4.8874\times10^{-1} \end{bmatrix}.$$

We have verified that

$$\|A_a^TX_a + X_a^TA_a + (C_a^TJC_a - B_aJ'B_a^T) - X_a^TB_a(J')^{-1}B_a^TX_a\|_2 = 4.0088\times10^{-13},$$

$$\frac{\|A_a^TX_a + X_a^TA_a + (C_a^TJC_a - B_aJ'B_a^T) - X_a^TB_a(J')^{-1}B_a^TX_a\|_2}{\|A_a^TX_a\|_2 + \|X_a^TA_a\|_2 + \|C_a^TJC_a - B_aJ'B_a^T\|_2 + \|X_a^TB_a(J')^{-1}B_a^TX_a\|_2} = 5.0409\times10^{-14},$$

and

$$\|E_a^TX_a - X_a^TE_a\|_2 = 1.1540\times10^{-13}, \quad \frac{\|E_a^TX_a - X_a^TE_a\|_2}{\|E_a^TX_a\|_2 + \|X_a^TE_a\|_2} = 2.4159\times10^{-14}.$$

Before closing this section, the following remark is in order.

Remark 2. (i) The factorization (7) can be related to the feedback regularization. By the assumption (A1), we can compute a feedback matrix F such that the pencil $A+BF-sE$ is regular and of index at most one [12] and then compute two orthogonal matrices $\hat{U}$ and $\hat{V}$ such that

$$\hat{U}(A+BF-sE)\hat{V} = \begin{array}{c} \begin{matrix} n_1 & n-n_1 \end{matrix} \\ \begin{bmatrix} \hat{A}_{11}-s\hat{E}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \end{array} \begin{matrix} \}n_1 \\ \}n-n_1 \end{matrix}, \tag{19}$$

where $\hat{E}_{11}$ and $\hat{A}_{22}$ are nonsingular. Let

$$\hat{U}B = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix} \begin{matrix} \}n_1 \\ \}n-n_1 \end{matrix}, \quad C\hat{V} = \begin{array}{c} \begin{matrix} n_1 & n-n_1 \end{matrix} \\ [\,\hat{C}_1 \quad \hat{C}_2\,] \end{array}. \tag{20}$$

Then

$$\hat{U}\begin{bmatrix} A-sE & B \end{bmatrix}\begin{bmatrix} \hat{V} & 0 \\ F\hat{V} & I \end{bmatrix}\begin{bmatrix} I & 0 & 0 \\ 0 & I & -\hat{A}_{22}^{-1}\hat{B}_2 \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} \hat{A}_{11}-s\hat{E}_{11} & \hat{A}_{12} & \hat{B}_1-\hat{A}_{12}\hat{A}_{22}^{-1}\hat{B}_2 \\ \hat{A}_{21} & \hat{A}_{22} & 0 \end{bmatrix}. \tag{21}$$

The decomposition (21) can play the same role as (7). But, if $\hat{A}_{22}$ is ill-conditioned, then the computation of (21) is not numerically reliable. Unfortunately, it is not easy to choose the feedback matrix F such that the resulting matrix $\hat{A}_{22}$ is well-conditioned. Therefore, (21) is not recommended, although it is based on the feedback regularization.

(ii) Let F be such that the pencil $A+BF-sE$ is regular and of index at most one. Define

$$G_F(s) := D + C(sE - A - BF)^{-1}B = \hat{D} + \hat{C}(sI-\hat{A})^{-1}\hat{B}.$$

If $\hat{D}$ is of full column rank, then we can consider the ARE

$$\hat{A}^T\hat{X} + \hat{X}\hat{A} - (\hat{X}\hat{B}^T + \hat{C}^TJ\hat{D})(\hat{D}^TJ\hat{D})^{-1}(\hat{B}^T\hat{X} + \hat{D}^TJ\hat{C}) = 0, \quad \hat{X} = \hat{X}^T. \tag{22}$$

Similar to Lemma 6 in section 3.2, we can show that the GARE (1) has a semistabilizing solution X_a if and only if the ARE (22) has a semistabilizing solution $\hat{X}$, i.e., a solution $\hat{X}$ satisfying that all eigenvalues of $\hat{A} - \hat{B}(\hat{D}^TJ\hat{D})^{-1}(\hat{B}^T\hat{X} + \hat{D}^TJ\hat{C})$ are on $\mathbf{C}_- \cup \mathbf{C}_0$. It seems that we can obtain a semistabilizing solution of the GARE (1) by computing a semistabilizing solution $\hat{X}$ of the ARE (22). However, two problems arise:

- the assumption (A2) cannot guarantee that $\hat{D}$ is of full column rank. If $\hat{D}$ is not of full column rank, then the ARE (22) is not well defined;
- the Hamiltonian matrix $\hat{H}$ associated with the ARE (22) is given by

$$\hat{H} = \begin{bmatrix} \hat{A} - \hat{B}(\hat{D}^TJ\hat{D})^{-1} & -\hat{B}(\hat{D}^TJ\hat{D})^{-1}\hat{B}^T \\ \hat{C}^T(\hat{D}^TJ\hat{D})^{-1}\hat{C} & -(\hat{A} - \hat{B}(\hat{D}^TJ\hat{D})^{-1})^T \end{bmatrix}.$$

  $\hat{H}$ has eigenvalues on $\mathbf{C}_0$, provided $H_a(s)$ has eigenvalues on $\mathbf{C}_0$. In this case, as analyzed in section 1, it is not clear which eigenspace we should compute in order to get a semistabilizing solution of the ARE (22).

Therefore, we cannot get a semistabilizing solution of the GARE (1) by simply applying the feedback regularization approach to the pair $(A-sE, B)$ and then solving the resulting ARE (22).

3. Proof of Theorem 4. Although Theorem 4 may look very simple, its proof is actually not trivial. It is motivated by the following results:

(i) the GARE (1) has a semistabilizing solution if and only if $G(s) = D + C(sE - A)^{-1}B$ has a (J, J')-spectral factorization, provided the pencil $A - sE$ is regular and stable [2, 3];

(ii) the (J, J')-spectral factorization of $G(s)$ is reduced to the canonical (J, J')-spectral factorization if $G(s)$ has no poles and zeros on $\mathbf{C}_{0e}$ [2, 3];

(iii) the canonical (J, J')-spectral factorization problem is solvable if and only a related ARE has a stabilizing solution [16];

(iv) the stabilizing solution of any ARE is characterized uniquely by the stable eigenspace of the associated Hamiltonian matrix [8, 14, 27].

We will prove Theorem 4 in this section outlined as follows:

- in section 3.1 we provide some important properties of eigendecompositions (8) and (14);
- in section 3.2 we reduce the GARE (1) to the GARE (28) which corresponds to the proper rational matrix $\mathcal{G}(s)$ defined in (27);
- in section 3.3, we connect the GARE (28) with the (J, J')-spectral factorization of the proper rational matrix $\mathcal{G}_{\mathcal{K}}(s)$ defined in (35);
- in section 3.4, we reduce the (J, J')-spectral factorization of $\mathcal{G}_{\mathcal{K}}(s)$ to the canonical (J, J')-spectral factorization of the proper rational matrix $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ defined in (37);
- in section 3.5, by characterizing the stable eigenspace of an associated Hamiltonian matrix we derive numerically verifiable necessary and sufficient conditions for the existence of a canonical (J, J')-spectral factorization of $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$;
- finally, in section 3.6, we prove Theorem 4 by using the results in sections 3.1–3.5.

3.1. Properties of eigendecompositions (8) and (14). In this section we present some basic properties of eigendecompositions (8) and (14). These properties are important in the development of sections 3.4–3.6.

LEMMA 5. (i) *In eigendecomposition* (8), *the orthogonal matrices* Q *and* S *with partitioning* (9) *satisfy*

$$E_{11}Q_{12} = S_{12}\Theta_{22}, \quad \begin{bmatrix} Q_{13} \\ Q_{23} \end{bmatrix} = 0, \quad \begin{bmatrix} Q_{31} & Q_{32} & Q_{34} \end{bmatrix} = 0, \quad \begin{bmatrix} S_{22} \\ S_{32} \end{bmatrix} = 0, \tag{23}$$

$$\operatorname{rank}\begin{bmatrix} Q_{11} & Q_{12} \end{bmatrix} = \operatorname{rank}\begin{bmatrix} S_{11} & S_{12} \end{bmatrix} = n_1, \quad Q_{33}^T Q_{33} = I, \quad \operatorname{rank}(S_{33}) = m. \tag{24}$$

(ii) *In eigendecomposition* (14),

$$L_1^T E_{11}^T L_3 = L_3^T E_{11} L_1, \quad Q_{12}^T E_{11}^T L_3 = 0, \quad L_3^T S_{12} = 0, \tag{25}$$

$$L_3^T S_{13} + (L_1^T C_1^T + L_2^T C_2^T + L_5^T D^T) J S_{33} = 0. \tag{26}$$

Proof. Part (i). From the eigendecomposition (8) it is obvious that

$$E_{11}Q_{12} = S_{12}\Theta_{22}, \quad E_{11}Q_{13} = 0, \quad \begin{bmatrix} S_{22} \\ S_{32} \end{bmatrix}\Theta_{22} = 0, \quad A_{22}Q_{23} = 0,$$

so the nonsingularity of E_{11}, A_{22}, and Θ_{22} gives

$$\begin{bmatrix} Q_{13} \\ Q_{23} \end{bmatrix} = 0, \quad \begin{bmatrix} S_{22} \\ S_{32} \end{bmatrix} = 0.$$

In addition, the orthogonality of Q and the relation

$$\begin{bmatrix} Q_{13} \\ Q_{23} \end{bmatrix} = 0$$

lead to

$$Q_{33}^T Q_{33} = I, \quad \begin{bmatrix} Q_{31} & Q_{32} & Q_{34} \end{bmatrix} = 0.$$

We also have

$$\begin{bmatrix} S_{11}^T \\ S_{12}^T \end{bmatrix} E_{11} \begin{bmatrix} Q_{11} & Q_{12} \end{bmatrix} = \begin{bmatrix} \Theta_{12} & 0 \\ \Theta_{21} & \Theta_{22} \end{bmatrix};$$

consequently,

$$\operatorname{rank}\left(\begin{bmatrix} S_{11}^T \\ S_{12}^T \end{bmatrix} E_{11} \begin{bmatrix} Q_{11} & Q_{12} \end{bmatrix} \right) = \mu_1 + (n_1 - \mu_1) = n_1,$$

and thus this implies that

$$\operatorname{rank} \begin{bmatrix} Q_{11} & Q_{12} \end{bmatrix} = \operatorname{rank} \begin{bmatrix} S_{11} & S_{12} \end{bmatrix} = n_1.$$

Note that S is orthogonal and $\begin{bmatrix} S_{11} & S_{12} \end{bmatrix}$ is of full row rank, as shown above, so

$$\begin{bmatrix} 0 & S_{24} \\ S_{33} & S_{34} \end{bmatrix}$$

is of full column rank, which yields that $\operatorname{rank}(S_{33}) = m$.

Part (ii). We have from the eigendecompositions (8) and (14) and part (i) that

$$\begin{bmatrix} -L_3^T & -L_4^T & L_1^T & L_2^T & L_5^T \end{bmatrix} M_a(s) = (\Delta^T + sI) \begin{bmatrix} L_3^T E_{11} & 0 & -L_1^T E_{11}^T & 0 & 0 \end{bmatrix},$$

$$M_a(s) \begin{bmatrix} Q_{12} & 0 \\ Q_{22} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & Q_{33} \end{bmatrix} = \begin{bmatrix} S_{12} & S_{13} \\ 0 & 0 \\ 0 & -C_1^T J S_{33} \\ 0 & -C_2^T J S_{33} \\ 0 & -\mathcal{D}^T J S_{33} \end{bmatrix} \begin{bmatrix} \Phi_{22} - s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \end{bmatrix},$$

which give that

$$\begin{aligned} (\Delta^T + sI)(L_3^T E_{11} L_1 - L_1^T E_{11}^T L_3) &= \begin{bmatrix} -L_3^T & -L_4^T & L_1^T & L_2^T & L_5^T \end{bmatrix} M_a(s) \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \end{bmatrix} \\ &= (L_1^T E_{11}^T L_3 - L_3^T E_{11} L_1)(\Delta - sI), \end{aligned}$$

$$(\Delta^T + sI) \begin{bmatrix} L_3^T E_{11} Q_{12} & 0 \end{bmatrix} = \begin{bmatrix} -L_3^T & -L_4^T & L_1^T & L_2^T & L_5^T \end{bmatrix} M_a(s) \begin{bmatrix} Q_{12} & 0 \\ Q_{22} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & Q_{33} \end{bmatrix}$$

$$= - \begin{bmatrix} L_3^T S_{12} & L_3^T S_{13} + (L_1^T C_1^T + L_2^T C_2^T + L_5^T \mathcal{D}^T) J S_{33} \end{bmatrix} \begin{bmatrix} \Phi_{22} - s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \end{bmatrix}.$$

In the above, $M_a(s)$ is defined by (13). Note that the pencils $\Delta^T + sI$ and $\Delta - sI$ have no common eigenvalues, and the pencils $\Delta^T + sI$ and

$$\begin{bmatrix} \Phi_{22} - s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \end{bmatrix}$$

also have no common eigenvalues; thus,

$$L_3^T E_{11} L_1 - L_1^T E_{11}^T L_3 = 0, \quad L_3^T E_{11} Q_{12} = 0,$$
$$L_3^T S_{12} = 0, \quad L_3^T S_{13} + (L_1^T C_1^T + L_2^T C_2^T + L_5^T \mathcal{D}^T) J S_{33} = 0.$$

Hence, part (ii) is proved. □

3.2. Reduction of the GARE (1) by eliminating the singularity of E. The singularity of the matrix E complicates the eigenstructure of $H_a(s)$ at infinity. To overcome this difficulty, we reduce the GARE (1) to an equivalent one by eliminating the singularity of E in this subsection. This reduction will significantly simplify the computation of the GARE (1).

The orthogonal matrices U, V, and $\mathcal{V}$ transform the pencil

$$\begin{bmatrix} A - sE & B \\ C & D \end{bmatrix}$$

into

$$\begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ C_1 & C_2 & \mathcal{D} \end{bmatrix};$$

as a result, the resulting rational matrix

$$\mathcal{G}(s) = \left[\begin{array}{cc|c} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ \hline C_1 & C_2 & \mathcal{D} \end{array}\right] = \left[\begin{array}{c|c} \mathcal{A}_{11} - sE_{11} & B_1 \\ \hline \mathcal{C}_1 & \mathcal{D} \end{array}\right] \tag{27}$$

is proper, although the original rational matrix $G(s)$ defined by (3) is nonproper. Here

$$\mathcal{A}_{11} = A_{11} - A_{12} A_{22}^{-1} A_{21}, \quad \mathcal{C}_1 = C_1 - C_2 A_{22}^{-1} A_{21}.$$

The GARE corresponding to the second realization of $\mathcal{G}(s)$ in (27) is

$$\begin{cases} \mathcal{A}_a^T \mathcal{X}_a + \mathcal{X}_a^T \mathcal{A}_a + (\mathcal{C}_a^T J \mathcal{C}_a - \mathcal{B}_a J' \mathcal{B}_a^T) - \mathcal{X}_a^T \mathcal{B}_a (J')^{-1} \mathcal{B}_a^T \mathcal{X}_a = 0, \\ \mathcal{E}_a^T \mathcal{X}_a = \mathcal{X}_a^T \mathcal{E}_a, \end{cases} \tag{28}$$

where

$$\mathcal{A}_a = \begin{bmatrix} \mathcal{A}_{11} & B_1 \\ 0 & I \end{bmatrix}, \quad \mathcal{E}_a = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{B}_a = \begin{bmatrix} 0 \\ -I_m \end{bmatrix}, \quad \mathcal{C}_a = \begin{bmatrix} \mathcal{C}_1 & \mathcal{D} \end{bmatrix}.$$

Let

$$\mathcal{H}_a(s) := \begin{bmatrix} \mathcal{A}_a - s\mathcal{E}_a & -\mathcal{B}_a (J')^{-1} \mathcal{B}_a^T \\ -\mathcal{C}_a^T J \mathcal{C}_a + \mathcal{B}_a J' \mathcal{B}_a^T & -\mathcal{A}_a^T - s\mathcal{E}_a^T \end{bmatrix}. \tag{29}$$

The GAREs (1) and (28) are related closely, as shown in the following lemma.

LEMMA 6. *The GARE* (1) *has a semistabilizing solution X_a if and only if the GARE* (28) *has a semistabilizing solution $\mathcal{X}_a$. Furthermore, if $\mathcal{X}_a$ with partitioning*

$$\mathcal{X}_a = \begin{matrix} \\ \end{matrix} \begin{matrix} n_1 & m \end{matrix} \\ \begin{bmatrix} \mathcal{X}_{11} & \mathcal{X}_{12} \\ \mathcal{X}_{21} & \mathcal{X}_{22} \end{bmatrix} \begin{matrix} \}n_1 \\ \}m \end{matrix} \tag{30}$$

is a semistabilizing solution of the GARE (28)*, then a semistabilizing solution* X_a *of the GARE* (1) *is*

$$X_a = \begin{bmatrix} U^T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \mathcal{X}_{11} & 0 & \mathcal{X}_{12} \\ \tilde{X}_{21} & \tilde{X}_{22} & \tilde{X}_{23} \\ \mathcal{X}_{21} & 0 & \mathcal{X}_{22} - J' \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \mathcal{V}^T \end{bmatrix} \begin{bmatrix} V^T & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & J' \end{bmatrix} \tag{31}$$

with

$$\begin{aligned} \tilde{X}_{21} &= -A_{22}^{-T}(C_2^T J C_1 + A_{12}^T \mathcal{X}_{11} - C_2^T J C_2 A_{22}^{-1} A_{21}/2), \\ \tilde{X}_{22} &= -A_{22}^{-T} C_2^T J C_2/2, \ \tilde{X}_{23} = -A_{22}^{-T} C_2^T J \mathcal{D}. \end{aligned} \tag{32}$$

Proof. First, let X_a with partitioning

$$X_a = \begin{matrix} \\ \end{matrix} \begin{matrix} n & m \end{matrix} \\ \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{matrix} \}n \\ \}m \end{matrix}$$

be a semistabilizing solution of the GARE (1). Define

$$\begin{aligned} Y_a &:= \begin{bmatrix} U & 0 \\ 0 & I \end{bmatrix} \left(X_a - \begin{bmatrix} 0 & 0 \\ 0 & J' \end{bmatrix} \right) \begin{bmatrix} V & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \mathcal{V} \end{bmatrix} \\ &= \begin{matrix} \\ \end{matrix} \begin{matrix} n_1 & n-n_1 & m \end{matrix} \\ \begin{bmatrix} Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ Y_{31} & Y_{32} & Y_{33} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n-n_1 \\ \}m \end{matrix} . \end{aligned}$$

Then we have by Corollary 2(i) and through a direct calculation that

$$\begin{aligned} &X_a \text{ is a solution of the GARE (1)} \\ &\Longrightarrow H_a(s) \begin{bmatrix} I \\ X_a \end{bmatrix} = \begin{bmatrix} I \\ X_a^T \end{bmatrix} (A_a - B_a (J')^{-1} B_a^T X_a - s E_a) \\ &\Longrightarrow \begin{bmatrix} A - sE & B & 0 \\ -C^T J C & -C^T J D & -A^T - sE^T \\ -D^T J C & -D^T J D & -B^T \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & I \\ X_{11} & X_{12} \end{bmatrix} \\ &\quad = \begin{bmatrix} I & 0 \\ X_{11}^T & X_{21}^T \\ X_{12}^T & (X_{22} - J')^T \end{bmatrix} \begin{bmatrix} A - sE & B \\ -(J')^{-1} X_{21} & -(J')^{-1}(X_{22} - J') \end{bmatrix} \\ &\Longrightarrow \begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 & 0 \\ -C_1^T J C_1 & -C_1^T J C_2 & -C_1^T J \mathcal{D} & -A_{11}^T - sE_{11}^T & -A_{21}^T \\ -C_2^T J C_1 & -C_2^T J C_2 & -C_2^T J \mathcal{D} & -A_{12}^T & -A_{22}^T \\ -\mathcal{D}^T J C_1 & -\mathcal{D}^T J C_2 & -\mathcal{D}^T J \mathcal{D} & -B_1^T & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ Y_{11} & Y_{12} & Y_{13} \\ Y_{21} & Y_{22} & Y_{23} \end{bmatrix} \\ &\quad = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ Y_{11}^T & Y_{21}^T & Y_{31}^T \\ Y_{12}^T & Y_{22}^T & Y_{32}^T \\ Y_{13}^T & Y_{23}^T & Y_{33}^T \end{bmatrix} \begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ -(J')^{-1} Y_{31} & -(J')^{-1} Y_{32} & -(J')^{-1} Y_{33} \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
&\Longrightarrow Y_{12}=0,\ Y_{13}=0,\ \text{and}\\
&\begin{bmatrix} \mathcal{A}_{11}-sE_{11} & B_1 & 0\\ -\mathcal{C}_1^TJ\mathcal{C}_1 & -\mathcal{C}_1^TJ\mathcal{D} & -\mathcal{A}_{11}^T-sE_{11}^T\\ -\mathcal{D}^TJ\mathcal{C}_1 & -\mathcal{D}^TJ\mathcal{D} & -B_1^T \end{bmatrix}\begin{bmatrix} I & 0\\ 0 & I\\ Y_{11} & 0 \end{bmatrix}\\
&=\begin{bmatrix} I & 0\\ Y_{11}^T & (Y_{31}-Y_{32}A_{22}^{-1}A_{21})^T\\ 0 & Y_{33}^T \end{bmatrix}\begin{bmatrix} \mathcal{A}_{11}-sE_{11} & B_1\\ -(J')^{-1}(Y_{31}-Y_{32}A_{22}^{-1}A_{21}) & -(J')^{-1}Y_{33} \end{bmatrix}\\
&\Longrightarrow \mathcal{H}_a(s)\begin{bmatrix} I\\ \mathcal{X}_a \end{bmatrix}=\begin{bmatrix} I\\ \mathcal{X}_a^T \end{bmatrix}(\mathcal{A}_a-\mathcal{B}_a(J')^{-1}\mathcal{B}_a^T\mathcal{X}_a-s\mathcal{E}_a),
\end{aligned}$$

where

$$\mathcal{X}_a=\begin{bmatrix} Y_{11} & 0\\ Y_{31}-Y_{32}A_{22}^{-1}A_{21} & Y_{33}+J' \end{bmatrix}.$$

Moreover, since the pencil $A_a-B_a(J')B_a^TX_a-sE_a$ is regular and all its eigenvalues lie in $\mathbf{C}_-\cup\mathbf{C}_{0e}$, and $(C_a,A_a-B_a(J')^{-1}B_a^TX_a-sE_a)$ has no observable finite poles on $\mathbf{C}_0$ and no observable impulsive poles, it is trivial to show that $\mathcal{A}_a-\mathcal{B}_a(J')\mathcal{B}_a^T\mathcal{X}_a-s\mathcal{E}_a$ is regular and all its eigenvalues lie in $\mathbf{C}_-\cup\mathbf{C}_{0e}$, and $(\mathcal{C}_a,\mathcal{A}_a-\mathcal{B}_a(J')^{-1}\mathcal{B}_a^T\mathcal{X}_a-s\mathcal{E}_a)$ has no observable finite poles on $\mathbf{C}_0$ and no observable impulsive poles. Hence, $\mathcal{X}_a$ is a semistabilizing solution of the GARE (28).

Conversely, let the GARE (28) have a semistabilizing solution $\mathcal{X}_a$ with partitioning (30). Then we have using Corollary 2(i) that

$$\begin{aligned}
&\mathcal{H}_a(s)\begin{bmatrix} I\\ \mathcal{X}_a \end{bmatrix}=\begin{bmatrix} I\\ \mathcal{X}_a^T \end{bmatrix}(\mathcal{A}_a-\mathcal{B}_a(J')^{-1}\mathcal{B}_a^T\mathcal{X}_a-s\mathcal{E}_a)\\
&\Longrightarrow \mathcal{X}_{12}=0 \text{ and}\\
&\begin{bmatrix} \mathcal{A}_{11}-sE_{11} & B_1 & 0\\ -\mathcal{C}_1^TJ\mathcal{C}_1 & -\mathcal{C}_1^TJ\mathcal{D} & -\mathcal{A}_{11}^T-sE_{11}^T\\ -\mathcal{D}^TJ\mathcal{C}_1 & -\mathcal{D}^TJ\mathcal{D}+J' & -B_1^T \end{bmatrix}\begin{bmatrix} I & 0\\ 0 & I\\ \mathcal{X}_{11} & 0 \end{bmatrix}\\
&=\begin{bmatrix} I & 0\\ \mathcal{X}_{11}^T & \mathcal{X}_{21}^T\\ 0 & (\mathcal{X}_{22}-J')^T \end{bmatrix}\begin{bmatrix} \mathcal{A}_{11}-sE_{11} & B_1\\ -(J')^{-1}\mathcal{X}_{21} & -(J')^{-1}(\mathcal{X}_{22}-J') \end{bmatrix}\\
&\Longrightarrow \begin{bmatrix} A_{11}-sE_{11} & A_{12} & B_1 & 0 & 0\\ A_{21} & A_{22} & 0 & 0 & 0\\ -C_1^TJC_1 & -C_1^TJC_2 & -C_1^TJ\mathcal{D} & -A_{11}^T-sE_{11}^T & -A_{21}^T\\ -C_2^TJC_1 & -C_2^TJC_2 & -C_2^TJ\mathcal{D} & -A_{12}^T & -A_{22}^T\\ -\mathcal{D}^TJC_1 & -\mathcal{D}^TJC_2 & -\mathcal{D}^TJ\mathcal{D} & -B_1^T & 0 \end{bmatrix}\begin{bmatrix} I & 0 & 0\\ 0 & I & 0\\ 0 & 0 & I\\ \mathcal{X}_{11} & 0 & 0\\ \tilde{X}_{21} & \tilde{X}_{22} & \tilde{X}_{23} \end{bmatrix}\\
&=\begin{bmatrix} I & 0 & 0\\ 0 & I & 0\\ \mathcal{X}_{11}^T & \tilde{X}_{21}^T & \mathcal{X}_{21}^T\\ 0 & \tilde{X}_{22}^T & 0\\ 0 & \tilde{X}_{23}^T & (\mathcal{X}_{22}-J')^T \end{bmatrix}\begin{bmatrix} A_{11}-sE_{11} & A_{12} & B_1\\ A_{21} & A_{22} & 0\\ -(J')^{-1}\mathcal{X}_{21} & 0 & -(J')^{-1}(\mathcal{X}_{22}-J') \end{bmatrix},
\end{aligned}$$

where

$$\tilde{X}_{21} = -A_{22}^{-T}(C_2^T J C_1 + A_{12}^T \mathcal{X}_{11} - C_2^T J C_2 A_{22}^{-1} A_{21}/2),$$
$$\tilde{X}_{22} = -A_{22}^{-T} C_2^T J C_2/2,\ \tilde{X}_{23} = -A_{22}^{-T} C_2^T J \mathcal{D}.$$

Hence, a direct verification using the relationship above and taking the fact that $\mathcal{X}_{12} = 0$ into account yields that X_a defined by (31) satisfies

$$H_a(s) \begin{bmatrix} I \\ X_a \end{bmatrix} = \begin{bmatrix} I \\ X_a^T \end{bmatrix} (A_a - B_a (J')^{-1} B_a^T X_a - sE_a).$$

By Corollary 2(i), X_a given by (31) with (32) is a solution of the GARE (1).

In addition, it is easy to see that the pencil $\mathcal{A}_a - \mathcal{B}_a (J')^{-1} \mathcal{B}_a^T \mathcal{X}_a - s\mathcal{E}_a$ is regular and all its eigenvalues lie in $\mathbf{C}_- \cup \mathbf{C}_{0e}$ if and only if the pencil $A_a - B_a (J')^{-1} B_a^T X_a - sE_a$ is regular and all its eigenvalues lie in $\mathbf{C}_- \cup \mathbf{C}_{0e}$, and the matrix pair $(\mathcal{C}_a, \mathcal{A}_a - \mathcal{B}_a (J')^{-1} \mathcal{B}_a^T \mathcal{X}_a - s\mathcal{E}_a)$ has neither observable finite poles on $\mathbf{C}_0$ nor observable impulsive poles if and only if the matrix pair $(C_a, A_a - B_a (J')^{-1} B_a^T X_a - sE_a)$ has neither observable finite poles on $\mathbf{C}_0$ nor observable impulsive poles. Therefore, X_a is a semistabilizing solution of the GARE (1). □

3.3. Connection of the GARE (28) with the (J, J')-spectral factorization of a proper rational matrix. Since Lemma 6 indicates clearly that we need to compute only a semistabilizing solution of the GARE (28) in order to get a semistabilizing solution of the GARE (1), and furthermore, that the GARE (28) is related closely to the (J, J')-spectral factorization problem [2, 3], naturally, we wish to find a connection between the GARE (28) and the (J, J')-spectral factorization of a proper rational matrix. This is the main purpose of this section.

Note that E_{11} is nonsingular and assumption (A1) implies that

$$\operatorname{rank} \begin{bmatrix} \mathcal{A}_{11} - sE_{11} & B_1 \end{bmatrix} = n_1 \quad \forall s \in \mathbf{C} \backslash \mathbf{C}_-;$$

i.e., $(E_{11}^{-1}\mathcal{A}_{11}, E_{11}^{-1}B_1)$ is stabilizable. We have the following lemma.

LEMMA 7 (see [23]). *Under assumption* (A1)*, there exists a* $\mathcal{K} \in \mathbf{R}^{n_1 \times m}$ *such that the pencil* $\mathcal{A}_{11} + B_1\mathcal{K} - sE_{11}$ *is stable.*

In the rest of this section, we shall assume that the matrix $\mathcal{K}$ has been determined such that the pencil $\mathcal{A}_{11} + B_1\mathcal{K} - sE_{11}$ is stable.

Consider the linear time-invariant system

$$E_{11}\dot{x} = \mathcal{A}_{11}x + B_1 u, \quad y = \mathcal{C}_1 x + \mathcal{D}u. \tag{33}$$

Let us introduce the state feedback $u = \mathcal{K}x + v$. Then the resulting closed-loop system is

$$E_{11}\dot{x} = (\mathcal{A}_{11} + B_1\mathcal{K})x + B_1 v, \quad y = (\mathcal{C}_1 + \mathcal{D}\mathcal{K})x + \mathcal{D}v. \tag{34}$$

The transfer matrix of system (34) from the input v to the output y is

$$\mathcal{G}_{\mathcal{K}}(s) := \left[\begin{array}{c|c} \mathcal{A}_{11} + B_1\mathcal{K} - sE_{11} & B_1 \\ \hline \mathcal{C}_1 + \mathcal{D}\mathcal{K} & \mathcal{D} \end{array}\right]. \tag{35}$$

The GARE associated with $\mathcal{G}_{\mathcal{K}}(s)$ is

$$\begin{cases} \mathcal{A}_{a,\mathcal{K}}^T \mathcal{X}_{a,\mathcal{K}} + \mathcal{X}_{a,\mathcal{K}}^T \mathcal{A}_{a,\mathcal{K}} + (\mathcal{C}_{a,\mathcal{K}}^T J \mathcal{C}_{a,\mathcal{K}} - \mathcal{B}_a J' \mathcal{B}_a^T) - \mathcal{X}_{a,\mathcal{K}}^T \mathcal{B}_a (J')^{-1} \mathcal{B}_a^T \mathcal{X}_{a,\mathcal{K}} = 0, \\ \mathcal{E}_a^T \mathcal{X}_{a,\mathcal{K}} = \mathcal{X}_{a,\mathcal{K}}^T \mathcal{E}_a, \end{cases} \tag{36}$$

where

$$\mathcal{A}_{a,\mathcal{K}} = \begin{bmatrix} \mathcal{A}_{11} + B_1\mathcal{K} & B_1 \\ 0 & I \end{bmatrix}, \quad \mathcal{C}_{a,\mathcal{K}} = \begin{bmatrix} \mathcal{C}_1 + \mathcal{D}\mathcal{K} & \mathcal{D} \end{bmatrix}.$$

The following lemma gives the relation between the GAREs (28) and (36).

LEMMA 8. (i) *The GARE* (28) *has a semistabilizing solution* $\mathcal{X}_a$ *if and only if the GARE* (36) *has a semistabilizing solution* $\mathcal{X}_{a,\mathcal{K}}$.

(ii) *The GARE* (36) *has a semistabilizing solution if and only if* $\mathcal{G}_{\mathcal{K}}(s)$ *has a* (J, J')*-spectral factorization; i.e., there exists an invertible rational matrix* $\Xi(s)(s)$ *such that the following conditions hold:*

(1) $\mathcal{G}_{\mathcal{K}}^T(-s)J\mathcal{G}_{\mathcal{K}}(s) = \Xi^T(-s)J'\Xi(s)$;

(2) *all poles and zeros of* $\Xi(s)$ *lie in* $\mathbf{C}_- \cup \mathbf{C}_{oe}$;

(3) $\mathcal{G}_{\mathcal{K}}(s)\Xi^{-1}(s)$ *is proper and has no poles on* $\mathbf{C}_0$.

Proof. Lemma 8(i) and (ii) follow directly from a simple calculation and Theorem 7.4.5 in [3], respectively. □

3.4. Elimination of zeros of $\mathcal{G}_{\mathcal{K}}(s)$ on $\mathbf{C}_{0e}$. According to Lemmas 6 and 8, we need to study the (J, J')-spectral factorization of $\mathcal{G}_{\mathcal{K}}(s)$. Note that $\mathcal{G}_{\mathcal{K}}(s)$ has zeros on $\mathbf{C}_{0e}$; in order to derive the necessary and sufficient conditions for the existence of a (J, J')-spectral factorization of $\mathcal{G}_{\mathcal{K}}(s)$, in the following we reduce the (J, J')-spectral factorization of $\mathcal{G}_{\mathcal{K}}(s)$ to the canonical (J, J')-spectral factorization of a related rational matrix without zeros on $\mathbf{C}_{0e}$.

Let

$$\mathcal{P}_{\mathcal{K}}(s) := \begin{bmatrix} \mathcal{A}_{11} + B_1\mathcal{K} - sE_{11} & B_1 \\ \mathcal{C}_1 + \mathcal{D}\mathcal{K} & \mathcal{D} \end{bmatrix}$$

and

$$\mathcal{M}_{a,\mathcal{K}}(s) := \begin{bmatrix} \mathcal{A}_{11} + B_1\mathcal{K} - sE_{11} & 0 & B_1 \\ -(\mathcal{C}_1 + \mathcal{D}\mathcal{K})^T J(\mathcal{C}_1 + \mathcal{D}\mathcal{K}) & -(\mathcal{A}_{11} + B_1\mathcal{K})^T - sE_{11}^T & -(\mathcal{C}_1 + \mathcal{D}\mathcal{K})^T J\mathcal{D} \\ -\mathcal{D}^T J(\mathcal{C}_1 + \mathcal{D}\mathcal{K}) & -B_1^T & -\mathcal{D}^T J\mathcal{D} \end{bmatrix}.$$

LEMMA 9. *If* $\mathcal{G}_{\mathcal{K}}(s)$ *has a* (J, J')*-spectral factorization, then*

$$\mu_1 = \nu, \quad \text{i.e.,} \quad (n_1 - \mu_1) + \nu = n_1.$$

Proof. We consider two cases.

Case 1: $p = m$. In this case, $\mathcal{P}_{\mathcal{K}}(s)$ is square and regular because of assumption (A1), and we know from the eigendecomposition (8) and Lemma 5 that the number of stable and antistable eigenvalues of $\mathcal{P}_{\mathcal{K}}(s)$ is

$$n_1 + m - [(n_1 - \mu_1) + m] = \mu_1.$$

Since

$$\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & -(\mathcal{C}_1+\mathcal{D}\mathcal{K})^T J & I & 0 \\ 0 & -\mathcal{D}^T J & 0 & I \\ 0 & I & 0 & 0 \end{bmatrix}$$
$$\times \begin{bmatrix} \mathcal{A}_{11}+B_1\mathcal{K}-sE_{11} & B_1 & 0 & 0 \\ \mathcal{C}_1+\mathcal{D}\mathcal{K} & \mathcal{D} & 0 & -J^{-1} \\ 0 & 0 & -(\mathcal{A}_{11}+B_1\mathcal{K})^T-sE_{11}^T & -(\mathcal{C}_1+\mathcal{D}\mathcal{K})^T \\ 0 & 0 & -B_1^T & -\mathcal{D}^T \end{bmatrix}$$
$$\times \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & I & 0 & 0 \\ J(\mathcal{C}_1+\mathcal{D}\mathcal{K}) & 0 & J\mathcal{D} & I \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{a,\mathcal{K}}(s) & 0 \\ 0 & -J^{-1} \end{bmatrix},$$

it is clear that the stable and antistable eigenvalues of $\mathcal{M}_{a,\mathcal{K}}(s)$ are the stable and antistable eigenvalues of $\mathcal{P}_{\mathcal{K}}(s)$ plus the stable and antistable eigenvalues of $-\mathcal{P}_{\mathcal{K}}(-s)$, so the number of such stable and antistable eigenvalues is $2\mu_1$. On the other hand, we have from the eigendecomposition (14) that the dimension of the stable eigenspace of $\mathcal{M}_{a,\mathcal{K}}(s)$ is ν, and it is a standard result [8, 14] that if λ is an eigenvalue of $\mathcal{M}_{a,\mathcal{K}}(s)$, then $-\bar{\lambda}$ is also an eigenvalue of $\mathcal{M}_{a,\mathcal{K}}(s)$. Therefore, the number of the stable and antistable eigenvalues of $\mathcal{M}_{a,\mathcal{K}}(s)$ is 2ν. Hence,

$$2\mu_1 = 2\nu, \quad \text{i.e.,} \quad \mu_1 = \nu, \quad (n_1-\mu_1)+\nu = n_1.$$

Case 2: $p > m$. Since $\mathcal{G}_{\mathcal{K}}(s)$ has a (J, J')-spectral factorization with a factorization factor $\Xi(s)$ satisfying

$$(G_{\mathcal{K}}(-s)\Xi^{-1}(-s))^T J(G_{\mathcal{K}}(s)\Xi^{-1}(s)) = J',$$

$G_{\mathcal{K}}(s)\Xi^{-1}(s)$ is proper and has no poles on $\mathbf{C}_0$; thus, $G_{\mathcal{K}}(s)\Xi^{-1}(s)$ has no finite zeros on $\mathbf{C}_0$ and nontrivial infinite zeros. Note that from $G_{\mathcal{K}}(s) = (G_{\mathcal{K}}(s)\Xi^{-1}(s))\Xi(s)$ we obtain that the zeros of $G_{\mathcal{K}}(s)$ on $\mathbf{C}_{0e}$ are the same as those of $\Xi(s)$; i.e., the dimension of the eigenspace of $\Xi(s)$ corresponding to the eigenvalues on $\mathbf{C}_{0e}$ is $(n_1-\mu_1)+m$. Therefore, from the above analysis in Case 1, $\Xi^T(-s)J'\Xi(s)$ has $2(n_1+m-\mu_1)$ zeros on $\mathbf{C}_{0e}$. Equivalently, $G_{\mathcal{K}}^T(-s)JG_{\mathcal{K}}(s)$ has $2(n_1+m-\mu_1)$ zeros on $\mathbf{C}_{0e}$. This means that $\mathcal{M}_{a,\mathcal{K}}(s)$ has $2(n_1+m-\mu_1)$ eigenvalues on $\mathbf{C}_{0e}$. The number of stable and antistable eigenvalues of $\mathcal{M}_{a,\mathcal{K}}(s)$ is $2\mu_1$. Hence, similar to Case 1, we must have

$$2\mu_1 = 2\nu, \quad \text{i.e.,} \quad \mu_1 = \nu, \quad (n_1-\mu_1)+\nu = n_1. \qquad \square$$

Since the eigendecomposition (8) and Lemma 5 yield that

$$\mathcal{P}_{\mathcal{K}}(s)\begin{bmatrix} Q_{12} & 0 \\ -\mathcal{K}Q_{12} & Q_{33} \end{bmatrix} = \begin{bmatrix} S_{12} & S_{13} \\ 0 & S_{33} \end{bmatrix}\begin{bmatrix} \Phi_{22}-s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \end{bmatrix},$$

we define

$$\mathcal{W}_{\mathcal{Z}}(s) := \left[\begin{array}{cc|c} \Phi_{22}-s\Theta_{22} & \Phi_{23} & \mathcal{Z} \\ \Phi_{32} & \Phi_{33} & I \\ \hline -\mathcal{K}Q_{12} & Q_{33} & 0 \end{array}\right].$$

LEMMA 10. (i) *For any* $\mathcal{Z} \in \mathbf{R}^{(n_1-\mu_1)\times m}$, $\mathcal{W}_{\mathcal{Z}}(s)$ *is invertible and, furthermore,*

$$\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s) := \mathcal{G}_{\mathcal{K}}(s)\mathcal{W}_{\mathcal{Z}}(s) = \left[\begin{array}{c|c} \mathcal{A}_{11}+B_1\mathcal{K}-sE_{11} & -S_{12}\mathcal{Z}-S_{13} \\ \hline \mathcal{C}_1+\mathcal{D}\mathcal{K} & -S_{33} \end{array}\right]. \tag{37}$$

(ii) *There exists* $\mathcal{Z}$ *such that all finite zeros of* $\mathcal{W}_{\mathcal{Z}}(s)$ *are on* $\mathbf{C}_-$. *For such a* $\mathcal{Z}$, $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ *is proper and stable and has no finite zeros on* $\mathbf{C}_0$ *and no nontrivial infinite zeros.*

(iii) *For a* $\mathcal{Z}$ *in* (ii), $\mathcal{G}_{\mathcal{K}}(s)$ *has the a* (J,J')*-spectral factorization if and only if* $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ *has a canonical* (J,J')*-spectral factorization; i.e, there exists an invertible rational matrix* $\Pi(s)$ *such that*

$$\mathcal{G}^T_{\mathcal{K},\mathcal{Z}}(-s)J\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s) = \Pi^T(-s)J'\Pi(s)$$

and all poles and zeros of $\Pi(s)$ *lie in* $\mathbf{C}_-$.

Proof. Note that Q_{33} is nonsingular (see Lemma 5), so $\mathcal{W}_{\mathcal{Z}}(s)$ is always invertible. In addition, (37) follows directly from a simple verification. So, we need to prove only (ii) and (iii) of Lemma 10.

Part (ii). Since S_{33} is of full column rank (see Lemma 5) and E_{11} is nonsingular, $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has no nontrivial infinite zeros. Furthermore, from Lemma 5,

$$\left[\begin{array}{c} S_{22} \\ S_{32} \end{array}\right] = 0$$

and $\operatorname{rank}(S_{33}) = m$, so

$$\left[\begin{array}{cc} S_{12} & S_{13} \\ 0 & S_{33} \end{array}\right]$$

is of full column rank; consequently, we have from (9) that

$$\begin{aligned} \operatorname{rank}\left[\begin{array}{cc} \Phi_{22}-s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \\ -\mathcal{K}Q_{12} & Q_{33} \end{array}\right] &= \operatorname{rank}\left(\left[\begin{array}{ccc} S_{12} & S_{13} & 0 \\ 0 & S_{33} & 0 \\ 0 & 0 & I \end{array}\right]\left[\begin{array}{cc} \Phi_{22}-s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \\ -\mathcal{K}Q_{12} & Q_{33} \end{array}\right]\right) \\ &= \operatorname{rank}\left[\begin{array}{cc} (\mathcal{A}_{11}+B_1\mathcal{K}-sE_{11})Q_{12} & 0 \\ (\mathcal{C}_1+\mathcal{D}\mathcal{K})Q_{12} & 0 \\ -\mathcal{K}Q_{12} & Q_{33} \end{array}\right] \quad \forall s \in \mathbf{C}. \end{aligned}$$

Since

$$\left[\begin{array}{cc} Q_{12} & 0 \\ 0 & Q_{33} \end{array}\right]$$

is of full column rank, $\mathcal{A}_{11}+B_1\mathcal{K}-sE_{11}$ is stable, and

$$\left[\begin{array}{cc} \Phi_{22}-s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \end{array}\right]$$

is nonsingular for any $s \in \mathbf{C}\backslash\mathbf{C}_0$, we get that

$$\left[\begin{array}{cc} \Phi_{22}-s\Theta_{22} & \Phi_{23} \\ \Phi_{32} & \Phi_{33} \\ -\mathcal{K}Q_{12} & Q_{33} \end{array}\right]$$

is of full column rank for any $s \in \mathbf{C}$; equivalently,

$$\begin{bmatrix} \Phi_{22} + \Phi_{23} Q_{33}^T \mathcal{K} Q_{12} - s\Theta_{22} \\ \Phi_{32} + \Phi_{33} Q_{33}^T \mathcal{K} Q_{12} \end{bmatrix}$$

is of full column rank for any $s \in \mathbf{C}$ (i.e., is observable). Consequently, there exists a $\mathcal{Z}$ such that the matrix

$$\Theta_{22}^{-1}[\Phi_{22} + \Phi_{23} Q_{33}^T \mathcal{K} Q_{12} - \mathcal{Z}(\Phi_{32} + \Phi_{33} Q_{33}^T \mathcal{K} Q_{12})]$$

is stable. Hence, part (ii) follows.

Part (iii). Assume that $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has a canonical (J, J')-spectral factorization:

$$\mathcal{G}_{\mathcal{K},\mathcal{Z}}^T(-s) J \mathcal{G}_{\mathcal{K},\mathcal{Z}}(s) = \Pi^T(-s) J' \Pi(s),$$

where all poles and zeros of $\Pi(s)$ lie in $\mathbf{C}_-$ and $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)\Pi^{-1}(s)$ is proper and has no poles on $\mathbf{C}_0$. Let $\Xi(s) = \Pi(s)\mathcal{W}_{\mathcal{Z}}^{-1}(s)$. Then, from (37), we immediately get

$$\mathcal{G}_{\mathcal{K}}^T(-s) J \mathcal{G}_{\mathcal{K}}(s) = \Xi^T(-s) J' \Xi(s).$$

Moreover, $\mathcal{G}_{\mathcal{K}}(s)\Xi^{-1}(s) = \mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)\Pi^{-1}(s)$ is proper and has no poles on $\mathbf{C}_0$. Furthermore, all poles and zeros of $\Pi(s)$ and $\mathcal{W}^{-1}(s)$ lie on $\mathbf{C}_- \cup \mathbf{C}_{0e}$, so all poles and zeros of $\Xi(s)$ also lie on $\mathbf{C}_- \cup \mathbf{C}_{0e}$. Hence $\mathcal{G}_{\mathcal{K}}(s)$ has a (J, J')-spectral factorization.

Conversely, assume that $\mathcal{G}_{\mathcal{K}}(s)$ has a (J, J')-spectral factorization:

$$\mathcal{G}_{\mathcal{K}}^T(-s) J \mathcal{G}_{\mathcal{K}}(s) = \Xi^T(-s) J' \Xi(s),$$

where all poles and zeros of $\Xi(s)$ lie in $\mathbf{C}_- \cup \mathbf{C}_{0e}$ and $\mathcal{G}_{\mathcal{K}}(s)\Xi^{-1}(s)$ is proper and has no poles on $\mathbf{C}_0$. Let $\Pi(s) = \Xi(s)\mathcal{W}_{\mathcal{Z}}(s)$. Then it can be verified that

$$\mathcal{G}_{\mathcal{K},\mathcal{Z}}^T(-s) J \mathcal{G}_{\mathcal{K},\mathcal{Z}}(s) = \Pi^T(-s) J' \Pi(s)$$

and $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)\Pi^{-1}(s) = (\mathcal{G}_{\mathcal{K}}(s)\mathcal{W}_{\mathcal{Z}}(s))(\Xi(s)\mathcal{W}_{\mathcal{Z}}(s))^{-1} = \mathcal{G}_{\mathcal{K}}(s)\Xi^{-1}(s)$ is proper and has no poles on $\mathbf{C}_0$.

Since $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ is stable and has no zeros on $\mathbf{C}_0$ and we have shown that $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)\Pi^{-1}(s)$ is proper and has no poles on $\mathbf{C}_0$, $\Pi^{-1}(s)$ cannot have poles and zeros on $\mathbf{C}_0$; i.e., $\Pi(s)$ cannot have poles and zeros on $\mathbf{C}_0$. Therefore, all poles and zeros of $\Pi(s)$ are on $\mathbf{C}_0$ because all poles and zeros of $\Xi(s)$ and $\mathcal{W}_{\mathcal{Z}}(s)$ are on $\mathbf{C}_0 \cup \mathbf{C}_{0e}$. Hence $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has a canonical (J, J')-spectral factorization. ∎

3.5. Solvability conditions for the canonical (J, J')-spectral factorization of $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$. Since $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ is proper and has no poles and zeros (finite zeros and nontrivial infinite zeros) on $\mathbf{C}_{0e}$, hence existing results [16] for the canonical (J, J')-spectral factorizations can be applied to $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ directly.

LEMMA 11. *Let $\mathcal{Z}$ be such that all finite zeros of $\mathcal{W}_{\mathcal{Z}}(s)$ are on $\mathbf{C}_-$. Then $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has a canonical (J, J')-spectral factorization if and only if the conditions* (15) *and* (16) *hold.*

Proof. It is well known [16] that $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has a canonical (J, J')-spectral factorization if and only if

$$S_{33}^T J S_{33} = \mathcal{D}_0^T J' \mathcal{D}_0 \quad \text{for some nonsingular } \mathcal{D}_0;$$

i.e., the condition (15) holds, and, furthermore, the ARE

$$\begin{aligned} &E_{11}^{-T}(\mathcal{A}_{11}+B_1\mathcal{K})^TP+P(\mathcal{A}_{11}+B_1\mathcal{K})E_{11}^{-1}+E_{11}^{-T}(\mathcal{C}_1+\mathcal{D}\mathcal{K})^TJ(\mathcal{C}_1+\mathcal{D}\mathcal{K})E_{11}^{-1}\\ &\quad-[P(S_{12}\mathcal{Z}+S_{13})+E_{11}^{-T}(\mathcal{C}_1+\mathcal{D}\mathcal{K})^TJS_{33}](S_{33}^TJS_{33})^{-1}\\ &\quad\times[(S_{12}\mathcal{Z}+S_{13})^TP+S_{33}^TJ(\mathcal{C}_1+\mathcal{D}\mathcal{K})E_{11}^{-1}]=0 \end{aligned} \tag{38}$$

has a stabilizing solution P; i.e., P is a solution of (38) and

$$(\mathcal{A}_{11}+B_1\mathcal{K})E_{11}^{-1}-(S_{12}\mathcal{Z}+S_{13})(S_{33}^TJS_{33})^{-1}[(S_{12}\mathcal{Z}+S_{13})^TP+S_{33}^TJ(\mathcal{C}_1+\mathcal{D}\mathcal{K})\mathcal{E}_{11}^{-1}] \tag{39}$$

is stable.

Define

$$\mathcal{A}_s:=\Phi_{22}+\Phi_{23}Q_{33}^T\mathcal{K}Q_{12}-\mathcal{Z}(\Phi_{32}+\Phi_{33}Q_{33}^T\mathcal{K}Q_{12}),$$
$$\mathcal{B}_s:=(\mathcal{Z}\Phi_{33}-\Phi_{32})Q_{33}^T,\quad \mathcal{C}_s:=\Phi_{32}+\Phi_{33}Q_{33}^T\mathcal{K}Q_{12},\quad \mathcal{D}_s:=-\Phi_{33}Q_{33}^T.$$

Then pencil $A_s-s\Theta_{22}$ is stable, as shown in the proof of Lemma 10. We also have after a simple calculation using the eigendecompositions (8) and (14) and Lemma 5 that

$$\begin{aligned} &\begin{bmatrix} \mathcal{A}_{11}+B_1\mathcal{K}-sE_{11} & 0 & -S_{12}\mathcal{Z}-S_{13}\\ -(\mathcal{C}_1+\mathcal{D}\mathcal{K})^TJ(\mathcal{C}_1+\mathcal{D}\mathcal{K}) & -(\mathcal{A}_{11}+B_1\mathcal{K})^T-sE_{11}^T & (\mathcal{C}_1+\mathcal{D}\mathcal{K})^TJS_{33}\\ S_{33}^TJ(\mathcal{C}_1+\mathcal{D}\mathcal{K}) & \mathcal{Z}^TS_{12}^T+S_{13}^T & -S_{33}^TJS_{33} \end{bmatrix}\\ &\quad\times\begin{bmatrix} L_1 & Q_{12}\\ L_3 & 0\\ \mathcal{D}_s(L_5-\mathcal{K}L_1) & \mathcal{C}_s \end{bmatrix}\\ &\quad=\begin{bmatrix} E_{11}L_1 & S_{12}\\ E_{11}^TL_3 & 0\\ 0 & 0 \end{bmatrix}\begin{bmatrix} \Delta-sI & 0\\ \mathcal{B}_s(L_5-\mathcal{K}L_1) & \mathcal{A}_s-s\Theta_{22} \end{bmatrix}. \end{aligned} \tag{40}$$

Since Lemmas 9 and 10(iii) hold, we have from the well-known relationship between the stabilizing solution of an ARE and its associated Hamiltonian matrix pencil [8, 14, 27] that $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has a canonical (J,J')-spectral factorization if and only if the condition (15) holds, $(n_1-\mu_1)+\nu=n_1$, and $\begin{bmatrix} L_1 & Q_{12}\end{bmatrix}$ is nonsingular; i.e., the conditions (15) and (16) hold. □

3.6. Proof of Theorem 4. Now we are ready to prove Theorem 4 based on the results in sections 3.1–3.5.

Proof. Let $\mathcal{K}$ and $\mathcal{Z}$ be such that $\mathcal{A}_{11}+B_1\mathcal{K}-sE_{11}$ is stable and all finite zeros of $\mathcal{W}_{\mathcal{Z}}(s)$ are on $\mathbf{C}_-$, respectively. We have

the GARE (1) has a semistabilizing solution X_a
$\iff$ the GARE (28) has a semistabilizing solution $\mathcal{X}_a$ (by Lemma 6)
$\iff$ $\mathcal{G}_{\mathcal{K}}(s)$ has a (J,J')-spectral factorization (by Lemma 8)
$\iff$ $\mathcal{G}_{\mathcal{K},\mathcal{Z}}(s)$ has a canonical (J,J')-spectral factorization (by Lemma 10)
$\iff$ the conditions (15) and (16) hold (by Lemma 11).

Furthermore, we have using Lemma 5 and the eigendecomposition (8) that

$$\mathcal{D}Q_{33}=S_{33}\Phi_{33},\quad C_1Q_{12}+C_2Q_{22}=S_{33}\Phi_{32},\quad A_{21}Q_{12}+A_{22}Q_{22}=0;$$

thus, under the conditions (15) and (16), we have using Lemma 5 and the eigendecompositions (8) and (14) that

$$\text{rank}\begin{bmatrix} L_1 & Q_{12} & 0 \\ L_5 & 0 & I \end{bmatrix} = \text{rank}\begin{bmatrix} E_{11}L_1 & S_{12} & S_{13} \\ 0 & 0 & \mathcal{D}_0 \end{bmatrix} = n_1 + m, \tag{41}$$

$$\mathcal{H}_a(s)\begin{bmatrix} L_1 & Q_{12} & 0 \\ L_5 & 0 & I \\ L_3 & 0 & 0 \\ J'L_5 & -J'\mathcal{D}_0\Phi_{32} & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T + J' \end{bmatrix}$$

$$= \begin{bmatrix} E_{11}L_1 & S_{12} & S_{13} \\ 0 & 0 & \mathcal{D}_0 \\ E_{11}^T L_3 & 0 & -\mathcal{C}_1^T J S_{33} \\ 0 & 0 & -\mathcal{D}^T J S_{33} + J'\mathcal{D}_0 \end{bmatrix} \begin{bmatrix} \Delta - sI & 0 & 0 \\ 0 & \Phi_{22} - s\Theta_{22} & \Phi_{23}Q_{33}^T \\ 0 & \Phi_{32} & \Phi_{33}Q_{33}^T \end{bmatrix}, \tag{42}$$

and

$$\begin{bmatrix} L_1 & Q_{12} & 0 \\ L_5 & 0 & I \end{bmatrix}^T \begin{bmatrix} E_{11}^T L_3 & 0 & -\mathcal{C}_1^T J S_{33} \\ 0 & 0 & -\mathcal{D}^T J S_{33} + J'\mathcal{D}_0 \end{bmatrix}$$

$$= \begin{bmatrix} L_3 & 0 & 0 \\ J'L_5 & -J'\mathcal{D}_0\Phi_{32} & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T + J' \end{bmatrix}^T \begin{bmatrix} E_{11}L_1 & S_{12} & S_{13} \\ 0 & 0 & \mathcal{D}_0 \end{bmatrix}. \tag{43}$$

So, Corollary 2(i) means that

$$\mathcal{X}_a = \begin{bmatrix} L_3 & 0 & 0 \\ J'L_5 & -J'\mathcal{D}_0\Phi_{32} & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T + J' \end{bmatrix} \begin{bmatrix} L_1 & Q_{12} & 0 \\ L_5 & 0 & I \end{bmatrix}^{-1} \tag{44}$$

is a semistabilizing solution of the GARE (28). Moreover, (42) yields that

$$\mathcal{A}_a - \mathcal{B}_a(J')^{-1}\mathcal{B}_a^T\mathcal{X}_a - s\mathcal{E}_a$$

$$= \begin{bmatrix} E_{11}L_1 & S_{12} & S_{13} \\ 0 & 0 & \mathcal{D}_0 \end{bmatrix}^{-1} \begin{bmatrix} \Delta - sI & 0 & 0 \\ 0 & \Phi_{22} - s\Theta_{22} & \Phi_{23}Q_{33}^T \\ 0 & \Phi_{32} & \Phi_{33}Q_{33}^T \end{bmatrix} \begin{bmatrix} L_1 & Q_{12} & 0 \\ L_5 & 0 & I \end{bmatrix}^{-1}$$

is regular and all its eigenvalues lie in $\mathbf{C}_- \cup \mathbf{C}_{0e}$. In addition, because

$$\begin{bmatrix} \mathcal{C}_1 & \mathcal{D} \end{bmatrix} \begin{bmatrix} L_1 & Q_{12} & 0 \\ L_5 & 0 & I \end{bmatrix} = \begin{bmatrix} \mathcal{C}_1 L_1 + \mathcal{D}L_5 & S_{33}\Phi_{32} & S_{33}\Phi_{33}Q_{33}^T \end{bmatrix},$$

it is easy to know using the fact that Δ is stable and $\begin{bmatrix} \Phi_{32} & \Phi_{33} \end{bmatrix}$ is of full row rank that the pair

$$\left(\begin{bmatrix} \mathcal{C}_1 L_1 + \mathcal{D}L_5 & S_{33}\Phi_{32} & S_{33}\Phi_{33}Q_{33}^T \end{bmatrix}, \begin{bmatrix} \Delta - sI & 0 & 0 \\ 0 & \Phi_{22} - s\Theta_{22} & \Phi_{23}Q_{33}^T \\ 0 & \Phi_{32} & \Phi_{33}Q_{33}^T \end{bmatrix} \right)$$

has neither observable finite poles on $\mathbf{C}_0$ nor observable impulsive poles; equivalently, the pair $(\mathcal{C}_a, \mathcal{A}_a - \mathcal{B}_a(J')^{-1}\mathcal{B}_a^T\mathcal{X}_a - s\mathcal{E}_a)$ has neither observable finite poles on $\mathbf{C}_0$ nor

observable impulsive poles. Therefore, $\mathcal{X}_a$ is a semistabilizing solution of the GARE (28).

Let $\tilde{X}_{21}$, $\tilde{X}_{22}$, and $\tilde{X}_{23}$ be defined in (32). We have

$$\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & 0 & I_m \\ 0 & I_{n_2} & 0 \end{bmatrix}\begin{bmatrix} \mathcal{X}_{11} & 0 & \mathcal{X}_{12} \\ \tilde{X}_{21} & \tilde{X}_{22} & \tilde{X}_{23} \\ \mathcal{X}_{21} & 0 & \mathcal{X}_{22}-J' \end{bmatrix}\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & 0 & I_{n_2} \\ 0 & I_m & 0 \end{bmatrix}\begin{bmatrix} L_1 & Q_{12} & 0 & 0 \\ L_5 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$$
$$= \begin{bmatrix} L_3 & 0 & 0 & 0 \\ 0 & -J'\mathcal{D}_0\Phi_{32} & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T & 0 \\ \tilde{X}_{21}L_1+\tilde{X}_{23}L_5 & \tilde{X}_{21}Q_{12} & \tilde{X}_{23} & \tilde{X}_{22} \end{bmatrix},$$

which yields

$$\begin{bmatrix} I & 0 & 0 \\ 0 & A_{22}^T & 0 \\ 0 & 0 & I_m \end{bmatrix}\begin{bmatrix} \mathcal{X}_{11} & 0 & \mathcal{X}_{12} \\ \tilde{X}_{21} & \tilde{X}_{22} & \tilde{X}_{23} \\ \mathcal{X}_{21} & 0 & \mathcal{X}_{22}-J' \end{bmatrix}\begin{bmatrix} L_1 & Q_{12} & 0 \\ 0 & 0 & I \end{bmatrix}$$
$$= \begin{bmatrix} L_3 & 0 & 0 & 0 \\ A_{22}^T\tilde{X}_{21}L_1 & A_{22}^T\tilde{X}_{21}Q_{12} & A_{22}^T\tilde{X}_{22} & A_{22}^T\tilde{X}_{23} \\ J'\mathcal{D}_0\Phi_{33}Q_{33}^TL_5 & -J'\mathcal{D}_0\Phi_{32} & 0 & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T \end{bmatrix}$$
$$= \begin{bmatrix} L_3 & 0 & 0 & 0 \\ A_{22}^T\tilde{X}_{21}L_1 & A_{22}^T\tilde{X}_{21}Q_{12} & -C_2^TJC_2/2 & -C_2^TJ\mathcal{D} \\ J'\mathcal{D}_0\Phi_{33}Q_{33}^TL_5 & -J'\mathcal{D}_0\Phi_{32} & 0 & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T \end{bmatrix}$$
$$= \begin{bmatrix} L_3 & 0 & 0 & 0 \\ A_{22}^T\tilde{X}_{21}L_1-C_2^TJC_2L_2/2-C_2^TJ\mathcal{D}L_5 & A_{22}^T\tilde{X}_{21}Q_{12}+C_2^TJC_2Q_{22}/2 & -C_2^TJC_2/2 & -C_2^TJ\mathcal{D} \\ 0 & -J'\mathcal{D}_0\Phi_{32} & 0 & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T \end{bmatrix}$$
$$\times\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ -L_2 & Q_{22} & I & 0 \\ -L_5 & 0 & 0 & I \end{bmatrix}.$$

Note that the eigendecompositions (8) and (14) and Lemma 5 give that

$$A_{21}Q_{12}+A_{22}Q_{22}=0, \quad A_{21}L_1+A_{22}L_2=0,$$
$$\mathcal{X}_{11}\begin{bmatrix} L_1 & Q_{12} \end{bmatrix}=\begin{bmatrix} L_3 & 0 \end{bmatrix},$$
$$-C_2^TJC_1L_1-C_2^TJC_2L_2-A_{12}^TL_3-C_2^TJ\mathcal{D}L_5=A_{22}^TL_4;$$

so

$$A_{22}^T\tilde{X}_{21}L_1-C_2^TJC_2L_2/2-C_2^TJ\mathcal{D}L_5=A_{22}^TL_4,$$
$$A_{22}^T\tilde{X}_{21}Q_{12}+C_2^TJC_2Q_{22}/2=-C_2^TJC_1Q_{12},$$

and thus

$$\begin{bmatrix} I & 0 & 0 \\ 0 & A_{22}^T & 0 \\ 0 & 0 & I_m \end{bmatrix}\begin{bmatrix} \mathcal{X}_{11} & 0 & \mathcal{X}_{12} \\ \tilde{X}_{21} & \tilde{X}_{22} & \tilde{X}_{23} \\ \mathcal{X}_{21} & 0 & \mathcal{X}_{22}-J' \end{bmatrix}\begin{bmatrix} L_1 & Q_{12} & 0 \\ 0 & 0 & I \end{bmatrix}$$
$$= \begin{bmatrix} L_3 & 0 & 0 & 0 \\ A_{22}^TL_4 & -C_2^TJC_1Q_{12} & -C_2^TJC_2/2 & -C_2^TJ\mathcal{D} \\ 0 & -J'\mathcal{D}_0\Phi_{32} & 0 & -J'\mathcal{D}_0\Phi_{33}Q_{33}^T \end{bmatrix}\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ -L_2 & Q_{22} & I & 0 \\ -L_5 & 0 & 0 & I \end{bmatrix}. \tag{45}$$

Hence, (17) and (18) follow directly from (45) and lemma 6. □

4. Conclusions. In this paper we have derived numerically verifiable necessary and sufficient conditions for the existence of the semistabilizing solution of the GARE (1) and developed Algorithm 1 for computing such a semistabilizing solution. The main ingredients of Algorithm 1 are the computations of eigendecompositions (8) and (14), which are computed by numerically backward stable algorithms in [6, 10, 26]. As a result, Algorithm 1 is a generalization of the well-known generalized eigenvalue approach for the classical AREs to the GARE (1).

There are still some interesting problems related to the GARE (1) that are yet to be solved. Such problems include (i) the conditioning analysis and (ii) the computation of

$$\begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_5 \end{bmatrix}$$

in (14) (i.e., the computation of the stable eigenspace of $M_a(s)$) based on the numerically stable and structure-preserving methods in [1, 4]. There are only very few results available for these problems in the existing literature.

Appendix. We construct the decompositions (7) and (8) as follows.

Construction of the decomposition (7).

Step 1. Compute the URV decomposition [6] of E to get orthogonal matrices U and V such that

$$UEV =: \begin{array}{c} \begin{matrix} n_1 & n-n_1 \end{matrix} \\ \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \}n_1 \\ \}n-n_1 \end{matrix} \end{array}, \quad \operatorname{rank}(E_{11}) = n_1.$$

Define

$$UAV =: \begin{array}{c} \begin{matrix} n_1 & n-n_1 \end{matrix} \\ \begin{bmatrix} A_{11} & A_{12}^{(1)} \\ A_{21} & A_{22}^{(1)} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n-n_1 \end{matrix} \end{array}, \quad UB =: \begin{array}{c} \\ \begin{bmatrix} B_1^{(1)} \\ B_2^{(1)} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n-n_1 \end{matrix} \end{array}.$$

Since assumption (A1) holds true, we have $\operatorname{rank}\begin{bmatrix} A_{22}^{(1)} & B_2^{(1)} \end{bmatrix} = n - n_1$.

Step 2. Compute the RQ decomposition [6] of $\begin{bmatrix} A_{22}^{(1)} & B_2^{(1)} \end{bmatrix}$ to get orthogonal matrix $\mathcal{V}$ such that

$$\begin{bmatrix} A_{22}^{(1)} & B_2^{(1)} \end{bmatrix} \mathcal{V} =: \begin{array}{c} \begin{matrix} n-n_1 & m \end{matrix} \\ \begin{bmatrix} A_{22} & 0 \end{bmatrix} \end{array}, \quad \operatorname{rank}(A_{22}) = n - n_1.$$

Set

$$\begin{bmatrix} A_{12}^{(1)} & B_1^{(1)} \end{bmatrix} \mathcal{V} =: \begin{array}{c} \begin{matrix} n-n_1 & m \end{matrix} \\ \begin{bmatrix} A_{12} & B_1 \end{bmatrix} \end{array}.$$

Then U, V, and $\mathcal{V}$ lead to the decomposition (7).

Construction of the eigendecomposition of (8).

Step 1. Compute the generalized upper triangular form [26, 10] of the pencil

$$\begin{bmatrix} A_{11} - sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ C_1 & C_2 & \mathcal{D} \end{bmatrix}$$

taking assumption (A2) into account to get orthogonal matrices S_1 and Q_1 with partitioning

$$S_1 = \begin{array}{c} \begin{matrix} p+\mu_1-m & n_1-\mu_1 & (n-n_1)+m \end{matrix} \\ \left[\begin{array}{ccc} S_{11} & S_{12} & S_{13}^{(1)} \\ S_{21} & S_{22} & S_{23}^{(1)} \\ S_{31} & S_{32} & S_{33}^{(1)} \end{array}\right] \end{array} \begin{matrix} \}n_1 \\ \}n-n_1 \\ \}p \end{matrix},$$

$$Q_1 = \begin{array}{c} \begin{matrix} \mu_1 & n_1-\mu_1 & (n-n_1)+m \end{matrix} \\ \left[\begin{array}{ccc} Q_{11} & Q_{12} & Q_{13}^{(1)} \\ Q_{21} & Q_{22} & Q_{23}^{(1)} \\ Q_{31} & Q_{32} & Q_{33}^{(1)} \end{array}\right] \end{array} \begin{matrix} \}n_1 \\ \}n-n_1 \\ \}m \end{matrix}$$

such that

$$S_1^T \left[\begin{array}{ccc} A_{11}-sE_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & 0 \\ C_1 & C_2 & \mathcal{D} \end{array}\right] Q_1$$

$$\begin{array}{c} \\ =: \end{array} \begin{array}{c} \begin{matrix} \mu_1 & n_1-\mu_1 & (n-n_1)+m \end{matrix} \\ \left[\begin{array}{ccc} \Phi_{11}-s\Theta_{11} & 0 & 0 \\ \Phi_{21}-s\Theta_{21} & \Phi_{22}^{(1)}-s\Theta_{22}^{(1)} & \Phi_{23}^{(1)} \\ \Phi_{31}^{(1)}-s\Theta_{31}^{(1)} & \Phi_{32}^{(1)} & \Phi_{33}^{(1)} \end{array}\right] \end{array} \begin{matrix} \}p+\mu_1-m \\ \}n_1-\mu_1 \\ \}(n-n_1)+m \end{matrix}, \tag{46}$$

where

$$\operatorname{rank}(\Theta_{11}) = \mu_1, \quad \operatorname{rank}(\Theta_{22}^{(1)}) = n_1-\mu_1,$$

$$\operatorname{rank}(\Phi_{11}-s\Theta_{11}) = \mu_1 \quad \forall s \in \mathbf{C}_{0e}, \tag{47}$$

$$\operatorname{rank}\left[\begin{array}{cc} \Phi_{22}^{(1)}-s\Theta_{22}^{(1)} & \Phi_{23}^{(1)} \\ \Phi_{32}^{(1)} & \Phi_{33}^{(1)} \end{array}\right] = m+n-\mu_1 \quad \forall s \in \mathbf{C}\backslash\mathbf{C}_0. \tag{48}$$

Note that (46) implies that $S_{22}\Theta_{22}^{(1)} = 0$ and $E_{11}Q_{13}^{(1)} = 0$, and so

$$S_{22} = 0, \quad Q_{13}^{(1)} = 0, \quad \left[\begin{array}{c} Q_{23}^{(1)} \\ Q_{33}^{(1)} \end{array}\right]$$

is nonsingular. We also have from (48) that

$$\operatorname{rank}\left[\begin{array}{cc} \Phi_{32}^{(1)} & \Phi_{33}^{(1)} \end{array}\right] = (n-n_1)+m.$$

Consequently, we have using (46) again that

$$n-n_1 = \operatorname{rank}(A_{22}) = \operatorname{rank}\left(\left[\begin{array}{cc} A_{22} & 0 \end{array}\right]\left[\begin{array}{c} Q_{23}^{(1)} \\ Q_{33}^{(1)} \end{array}\right]\right)$$

$$= \operatorname{rank}\left(\left[\begin{array}{cc} S_{22} & S_{23}^{(1)} \end{array}\right]\left[\begin{array}{c} \Phi_{23}^{(1)} \\ \Phi_{33}^{(1)} \end{array}\right]\right) = \operatorname{rank}(S_{23}^{(1)}\Phi_{33}^{(1)}),$$

which yields that $\operatorname{rank}(S_{23}^{(1)}) \geq n - n_1$. But it is clear that $\operatorname{rank}(S_{23}^{(1)}) \leq n - n_1$. Thus, we obtain that

$$\operatorname{rank}(S_{23}^{(1)}) = n - n_1.$$

Step 2. Compute the RQ factorization [6] of $S_{23}^{(1)}$ to get an orthogonal matrix S_2 such that

$$S_{23}^{(1)} S_2 =: \begin{array}{c} \begin{matrix} m & n-n_1 \end{matrix} \\ \left[\begin{array}{cc} 0 & S_{24} \end{array}\right] \end{array}, \quad \operatorname{rank}(S_{24}) = n - n_1.$$

Set

$$S = S_1 \begin{bmatrix} I & 0 \\ 0 & S_2 \end{bmatrix} = \begin{array}{c} \begin{matrix} p+\mu_1-m & n_1-\mu_1 & m & n-n_1 \end{matrix} \\ \left[\begin{array}{cccc} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & 0 & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \end{array}\right] \begin{matrix} \}n_1 \\ \}n-n_1 \\ \}p \end{matrix} \end{array},$$

$$S_2^T \begin{bmatrix} \Phi_{31}^{(1)} - s\Theta_{31}^{(1)} & \Phi_{32}^{(1)} & \Phi_{33}^{(1)} \end{bmatrix} =: \begin{array}{c} \begin{matrix} \mu_1 & n_1-\mu_1 & (n-n_1)+m \end{matrix} \\ \left[\begin{array}{ccc} \Phi_{31} - s\Theta_{31} & \Phi_{32}^{(2)} & \Phi_{33}^{(2)} \\ \Phi_{41} - s\Theta_{41} & \Phi_{42}^{(2)} & \Phi_{43}^{(2)} \end{array}\right] \begin{matrix} \}m \\ \}n-n_1 \end{matrix} \end{array}.$$

Then we have

$$n - n_1 = \operatorname{rank}(S_{23}^{(1)} \Phi_{33}^{(1)}) = \operatorname{rank}(S_{24} \Phi_{43}^{(2)}) = \operatorname{rank}(\Phi_{43}^{(2)}).$$

Step 3. Compute the RQ factorization [6] of $\Phi_{43}^{(2)}$ to get orthogonal matrix Q_2 such that

$$\Phi_{43}^{(2)} Q_2 = \begin{array}{c} \begin{matrix} m & n-n_1 \end{matrix} \\ \left[\begin{array}{cc} 0 & \Phi_{44}^{(3)} \end{array}\right] \end{array}, \quad \operatorname{rank}(\Phi_{44}^{(3)}) = n - n_1.$$

Set

$$\begin{bmatrix} \Phi_{23}^{(1)} \\ \Phi_{33}^{(2)} \end{bmatrix} Q_2 = \begin{array}{c} \begin{matrix} m & n-n_1 \end{matrix} \\ \left[\begin{array}{cc} \Phi_{23} & \Phi_{24}^{(3)} \\ \Phi_{33} & \Phi_{34}^{(3)} \end{array}\right] \begin{matrix} \}n_1-\mu_1 \\ \}m \end{matrix} \end{array}.$$

Step 4. Compute the RQ factorization [6] of $\begin{bmatrix} \Phi_{42}^{(2)} & \Phi_{44}^{(3)} \end{bmatrix}$ to get orthogonal matrix Q_3 with partitioning

$$Q_3 = \begin{array}{c} \begin{matrix} n_1-\mu_1 & n-n_1 \end{matrix} \\ \left[\begin{array}{cc} Q_{11}^{(3)} & Q_{12}^{(3)} \\ Q_{21}^{(3)} & Q_{22}^{(3)} \end{array}\right] \begin{matrix} \}n_1-\mu_1 \\ \}n-n_1 \end{matrix} \end{array}$$

such that

$$\begin{bmatrix} \Phi_{42}^{(2)} & \Phi_{44}^{(3)} \end{bmatrix} Q_3 = \begin{array}{c} \begin{matrix} n_1-\mu_1 & n-n_1 \end{matrix} \\ \left[\begin{array}{cc} 0 & \Phi_{44} \end{array}\right] \end{array}, \quad \operatorname{rank}(\Phi_{44}) = n - n_1.$$

Set

$$Q = Q_1 \begin{bmatrix} I & 0 \\ 0 & Q_2 \end{bmatrix} \begin{bmatrix} I_{\mu_1} & 0 & 0 & 0 \\ 0 & Q_{11}^{(3)} & 0 & Q_{12}^{(3)} \\ 0 & 0 & I & 0 \\ 0 & Q_{21}^{(3)} & 0 & Q_{22}^{(3)} \end{bmatrix} = \begin{matrix} \mu_1 \\ n_1 - \mu_1 \\ m \\ n - n_1 \end{matrix} \begin{matrix} \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} \end{bmatrix} \end{matrix} \begin{matrix} \}n_1 \\ \}n - n_1 \\ \}m \end{matrix},$$

$$\begin{bmatrix} \Phi_{22}^{(1)} - s\Theta_{22}^{(1)} & \Phi_{24}^{(3)} \\ \Phi_{32}^{(2)} & \Phi_{34}^{(3)} \end{bmatrix} Q_3 = \begin{matrix} n_1 - \mu_1 & n - n_1 \\ \begin{bmatrix} \Phi_{22} - s\Theta_{22} & \Phi_{24} - s\Theta_{24} \\ \Phi_{32} & \Phi_{34} \end{bmatrix} \end{matrix} \begin{matrix} \}n_1 - \mu_1 \\ \}m \end{matrix}.$$

Then a simple calculation gives that Θ_{22} is nonsingular, and (11) and (12) hold. Hence, S and Q give the eigendecomposition (8).

REFERENCES

[1] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Numerical computation of deflating subspaces of skew-Hamiltonian/Hamiltonian pencils*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 165–190.

[2] A. Kawamoto and T. Katayama, *The semi-stabilizing solution of generalized algebraic Riccati equation for descriptor systems*, Automatica, 38 (2002), pp. 1651–1662.

[3] A. Kawamoto, *Studies on Generalized Algebraic Riccati for Descriptor Systems and Applications*, Ph.D. thesis, Kyoto University, Kyoto, Japan, 1999.

[4] P. Benner, V. Mehrmann, and H. Xu, *A new method for computing the stable invariant subspace of a real Hamiltonian matrix*, J. Comput. Appl. Math., 86 (1997), pp. 17–43.

[5] P. Wangham and T. Mita, *Spectral Factorization of Singular Systems using Generalized Algebraic Riccati Equations*, in Proceedings of the 36th IEEE Conference on Decision and Control, 1997, pp. 4824–4829.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[7] M. Green and D. J. N. Limebeer, *Linear Robust Control*, Prentice–Hall, Englewood Cliffs, NJ, 1995.

[8] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.

[9] P. Misra, P. Van Dooren, and A. Varga, *Computation of structural invariants of generalized state space systems*, Automatica, 30 (1994), pp. 1921–1936.

[10] J. W. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. Part* I: *Theory and algorithms*, ACM Trans. Math. Software, 19 (1993), pp. 160–174.

[11] G. S. Miminis, *Deflation in eigenvalue assignment of descriptor systems using state feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 1322–1336.

[12] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Analysis Appl., 13 (1992), pp. 46–67.

[13] A. Lindquist and G. Picci, *A geometric approach to modelling and estimation of linear stochastic systems*, J. Math. Systems Estim. Control, 1 (1991), pp. 241–333.

[14] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem*, Springer-Verlag, Berlin, 1991.

[15] P. H. Petkov, N. D. Christov, and M. M. Konstantinov, *Computational Methods in Linear Control Systems*, Prentice–Hall, Hertfordshire, UK, 1991.

[16] M. Green, K. Glover, D. Limebeer, and J. C. Doyle, *A J-spectral factorization approach to H_∞ control*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.

[17] D. J. Clements and K. Glover, *Spectral factorization by Hermitian pencils*, Linear Algebra Appl., 122/124 (1989), pp. 797–846.

[18] C. Kenney, A. J. Laub, and M. Wette, *A stability-enhancing scaling procedure for Schur-Riccati solvers*, Systems Control Lett., 12 (1989), pp. 241–250.

[19] J. A. Ball and A. C. M. Ran, *Optimal Hankel norm model reduction and Wiener–Hopf factorization* I: *The canonical case*, SIAM J. Control Optim., 25 (1987), pp. 362–382.

[20] W. Greenberg, C. V. M. van der Mee, and V. Protopopescu, *Boundary Value Problems in Abstract Kinetic Theory*, Oper. Theory Adv. Appl. 23, Birkhäuser-Verlag, Basel, 1987.
[21] P. Hr. Petkov, N. D. Christov, and M. M. Konstantinov, *On the numerical properties of the Schur approach for solving the matrix Riccati equation*, Systems Control Lett., 9 (1987), pp. 197–201.
[22] H. Bart, I. Gohberg, and M. A. Kaashoek, *Constructive Methods of Wiener-Hopf Factorization*, Birkhäuser-Verlag, Basel, 1986.
[23] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.
[24] R. Byers, *Hamiltonian and Symplectic Algorithms for the Algebraic Riccati Equation*, Ph.D. thesis, Cornell University, Ithaca, NY, 1983.
[25] C. C. Paige and C. F. Van Loan, *Schur decomposition for Hamiltonian matrices*, Linear Algebra Appl., 41 (1981), pp. 11–32.
[26] P. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 26 (1981), pp. 111–129.
[27] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[28] A. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, 24 (1979), pp. 913–921.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1251–1278

ON THE GAP BETWEEN THE COMPLEX STRUCTURED SINGULAR VALUE AND ITS CONVEX UPPER BOUND*

I. M. JAIMOUKHA†, G. D. HALIKIAS‡, U. MALIK§, AND S. K. GUNGAH§

Abstract. The gap between the complex structured singular value of a complex matrix M and its convex upper bound is considered. New necessary and sufficient conditions for the existence of the gap are derived. It is shown that determining whether there exists such a gap is as difficult as evaluating a structured singular value of a reduced rank matrix (whose rank is equal to the multiplicity of the largest singular value of M). Furthermore, if an upper bound on this reduced rank problem can be obtained, it is shown that this provides an upper bound on the original problem that is lower than the convex relaxation upper bound. An example that illustrates our procedure is given. We also give the solution of several structured-approximation problems of independent interest.

Key words. structured uncertainty, convex relaxation, structured singular value, robust control, μ-analysis, distance to singularity, Toeplitz matrices

AMS subject classifications. 93B28, 93B35, 93B36, 93B40

DOI. 10.1137/040617340

1. Notation. $\mathcal{R}$ and $\mathcal{C}$ denote the sets of real and complex numbers, respectively. $\mathcal{C}^{n\times m}$ denotes the space of all $n\times m$ complex matrices. For $A\in\mathcal{C}^{n\times m}$, A' denotes the complex-conjugate transpose and $\sigma_i(A)$ denotes the ith largest singular value. The smallest and largest singular values are denoted by $\underline{\sigma}(A)$ and $\bar{\sigma}(A)$, respectively. The norm of A is defined as $\|A\|=\bar{\sigma}(A)$. For $A\in\mathcal{C}^{n\times n}$, $\lambda(A)$ is the set of all eigenvalues (or spectrum) of A and $\rho(A):=\max\{|\lambda_i|:\lambda_i\in\lambda(A)\}$ is the spectral radius of A. For Hermitian A, $\underline{\lambda}(A)$ denotes the smallest and $\bar{\lambda}(A)$ the largest eigenvalue of A, respectively. The m-dimensional identity matrix is denoted by I_m and the $m\times n$ null matrix is denoted by $0_{m,n}$. The subscripts of the identity and null matrices are dropped if the dimensions can be inferred from the context. All matrices are assumed to be complex unless otherwise stated. If m is a positive integer, then $\underline{m}$ is the set $\{i\in\mathcal{N}:1\le i\le m\}$, where $\mathcal{N}$ denotes the set of positive integers. If A and B are sets, then $A\backslash B=\{x\in A, x\notin B\}$.

A matrix Δ is called γ-unitary if $\Delta\Delta'=\Delta'\Delta=\gamma^2 I$. Let Δ_{11} be a complex matrix and let $\gamma\ge\|\Delta_{11}\|$. An embedding

$$\Delta=\begin{bmatrix}\Delta_{11} & \Delta_{12}\\ \Delta_{21} & \Delta_{22}\end{bmatrix} \tag{1.1}$$

is called a γ-completion of Δ_{11} if $\|\Delta\|=\gamma$. A γ-completion of Δ_{11} is called a γ-unitary completion if it is γ-unitary. If Δ is partitioned as (1.1) and U is a complex matrix such that $I-\Delta_{11}U$ is nonsingular, we define the upper linear fractional map

*Received by the editors October 20, 2004; accepted for publication (in revised form) April 24, 2006; published electronically September 19, 2006.
http://www.siam.org/journals/sicon/45-4/61734.html
†Control and Power Group, Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, UK (i.jaimouka@imperial.ac.uk).
‡School of Engineering and Mathematical Sciences, City University, Northampton Square, London EC1V 0HB, UK (G.Halikias@city.ac.uk).
§Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BT, UK (u.malik@imperial.ac.uk, s.gungah@ps.imperial.ac.uk).

as $\mathcal{F}_u(\Delta, U) = \Delta_{22} + \Delta_{21}U(I - \Delta_{11}U)^{-1}\Delta_{12}$. If $I - U\Delta_{22}$ is nonsingular, we define the lower linear fractional map as $\mathcal{F}_l(\Delta, U) = \Delta_{11} + \Delta_{12}U(I - \Delta_{22}U)^{-1}\Delta_{21}$.

2. Introduction. The structured singular value (μ) [3, 6, 7, 8, 10, 11, 12, 21, 27, 28] has proved to be an effective analytical and design tool in the area of robust control. As an analysis tool, it is used to establish necessary and sufficient conditions for the robust stability of feedback systems subject to structured uncertainty. Combined with $\mathcal{H}_\infty$ optimal control methods it serves as a synthesis tool for designing feedback systems subject to tight robust-stability and robust-performance specifications.

Unfortunately, deciding whether the structured singular value of a matrix is less than one (say) is an NP-hard problem [4, 5]. This suggests that given any algorithm to compute μ, there will be "worst-case" problems for which the algorithm will fail to find the answer in polynomial time. In practice, upper bounds are often used (giving sufficient robust stability and performance conditions), obtained via the so-called D-iteration procedure, which solves a convex optimization problem. The introduction of this method was initially motivated by the fact that the gap between μ and its convex upper bound is zero for certain simple uncertainty structures (specifically when $2S + F \leq 3$, where S is the number of scalar uncertainty blocks and F the number of full blocks). This class of problems, however, is too small for most practical applications. In general, the gap between μ and its convex upper bound can be arbitrarily large (but grows no faster than linearly in the number of uncertainty blocks) [19].

Despite these conclusions about the computational complexity of the general problem, solving special classes of μ problems can be relatively easy: The solution of the rank-one problem [25, 26] corresponds to the alignment of n complex vectors in two dimensions. Note that despite its simplicity, this class of problems is sufficiently general for carrying out Kharitonov-type analysis of uncertain polynomials. In [24], a low-complexity algorithm was given for solving a special class of rank-2 problems. Considering the Cauchy–Binet expansion of the determinant formula, it seems reasonable to conjecture that the computational complexity of the problem is intimately related to its rank. The main result of this work shows that the gap between μ and its convex upper bound may be breached if we can solve a reduced-rank μ problem. In this case, an upper bound tighter than the D-iteration bound can be obtained using a computationally simple procedure (the solution of an eigenvalue problem).

In our previous work [14] we obtained bounds on $\mu(M)$ by embedding the underlying block-structured uncertainty set within a larger set. This was constructed by imposing the least-conservative bound on the projection of the structured uncertainty in the direction defined by the singular vectors corresponding to the smallest singular value of M. The method has been used successfully for real and mixed-type structured uncertainty, resulting in algorithms with excellent computational performance [15, 17]. In its dynamic version, this technique was also used in [14] to identify the set of all maximally robust controllers which guarantee robust stability for the largest possible class of unstructured additive perturbations containing the uncertainty ball of maximum radius as a subset. (In this case, directionality arises from the Schmidt vectors of a Hankel operator related to the problem; these remain invariant for all maximally robust controllers.) As noted in [14], the approach followed in that work [14] suffered from the fundamental limitation that singular value multiplicity larger than one was not considered. As a result, the bounds on μ obtained by this method cannot outperform the D-iteration bounds (unless the largest singular value of the scaled matrix is simple, in which case the D-iteration bound is equal to μ). In this

paper we rectify this limitation by considering singular values of arbitrary multiplicity. In the following paragraphs we outline the general approach of this work more formally. Let $M \in \mathcal{C}^{n\times n}$ have a singular value decomposition

$$M = U\Sigma V' = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix}, \quad \Sigma_2 = \operatorname{diag}(\sigma_{m+1}, \ldots, \sigma_n), \tag{2.1}$$

with $U_1, V_1 \in \mathcal{C}^{n\times m}$. Assume that $1 > \sigma_{m+1} = \cdots = \sigma_{m+l} > \sigma_{m+l+1} \geq \cdots \geq \sigma_n$ and define $A = \Sigma^{-1} = \operatorname{diag}(A_1, A_2)$, where $A_1 = I_m$ and $A_2 = \operatorname{diag}(a_{m+1}, \ldots, a_n)$, with $1 < a_{m+1} = \cdots = a_{m+l} < a_{m+l+1} \leq \cdots \leq a_n$. Let $\mathbf{\Delta} \subseteq \mathcal{C}^{n\times n}$ denote a block diagonal structured uncertainty set and let $\mathcal{B}\mathbf{\Delta} = \{\Delta \in \mathbf{\Delta} \ : \ \|\Delta\| \leq 1\}$ denote its unit ball. The complex structured singular value of M is defined as $\mu_{\mathbf{\Delta}}(M) = r_{\mathbf{\Delta}}(M)^{-1}$, where

$$r_{\mathbf{\Delta}}(M) := \min_{\Delta\in\mathbf{\Delta}} \{\|\Delta\| : \det(I - \Delta M) = 0\} = \min_{\Delta\in\mathbf{\Delta}} \{\|\Delta\| : \det(A - V'\Delta U) = 0\} \tag{2.2}$$

is the structured stability radius (if $\det(I - \Delta M) \neq 0 \ \forall \Delta \in \mathbf{\Delta}$, define $\mu_{\mathbf{\Delta}}(M) = 0$). Let $\mathcal{D} = \{D \in \mathcal{C}^{n\times n} : D = D' \geq 0 : D\Delta = \Delta D \ \forall \Delta \in \mathbf{\Delta}\} \subseteq \mathcal{C}^{n\times n}$ denote a complementary block diagonal structure to $\mathbf{\Delta}$, and define $\mathcal{D}_+ = \{D \in \mathcal{D} : D > 0\}$. Then

$$\mu_{\mathbf{\Delta}}(M) \leq \inf \{\gamma : D \in \mathcal{D}_+, \ \gamma^2 D - M'DM \geq 0\} := \gamma_o, \tag{2.3}$$

where the infimum in the right-hand side is the so-called D-iteration upper bound. Note that an alternative form of this upper bound is

$$\gamma_o = \inf_{D\in\mathcal{D}_+} \|D^{1/2} M D^{-1/2}\|.$$

This follows readily from the equivalences:

$$\gamma^2 D - M'DM \geq 0 \Leftrightarrow \gamma^2 I - D^{-1/2} M'DMD^{-1/2} \geq 0 \Leftrightarrow \|D^{1/2} M D^{-1/2}\| \leq \gamma,$$

which hold for every $D \in \mathcal{D}_+$. In what follows, we assume that $D = I$ and $\gamma_o = 1$ are the minimizers and we investigate the gap between $\mu_{\mathbf{\Delta}}(M)$ and its upper bound $\|M\| = 1$.

Remark 2.1. We impose the above scaling on M because we intend to investigate in a single framework the case that M is arbitrary and also the case that M is D-optimally scaled. If M is arbitrary, redefining $M \leftarrow \|M\|^{-1} M$ gives M with norm one. Consider next the case arising at the end of the D-iteration: When the infimum in (2.3) is achieved by a $D_o \in \mathcal{D}_+$, the above scaling can be simply performed as $M \leftarrow \gamma_o^{-1} D_o^{1/2} M D_o^{-1/2}$ (we exclude the possibility that $\gamma_o = 0$ as in this case $\mu_\Delta(M) = 0$). Clearly, this is not defined when the infimum is not achieved, i.e., when there exists a sequence of D_i's in $\mathcal{D}_+$ such that $\lim_{i\to\infty} \|D_i^{1/2} M D_i^{-1/2}\| = \gamma_o$, but $\lim_{i\to\infty} D_i = D_o$ is singular. To avoid technical details with infimizing scaling-matrix sequences we restrict our attention throughout the paper to the class of problems for which the infimum in (2.3) is achieved. However, as we now argue, this is made only for simplicity of presentation and does not restrict the generality of our work.

To our knowledge, the question regarding the achievement of the infimum in (2.3) is only partially resolved. In [2], the authors study the limiting behavior of the D-iteration bound, in the equivalent formulation:

$$f_{\min}(M) = \{\|e^D M e^{-D}\| : D \in \mathcal{D}\}, \ \mathcal{D} = \{\mathcal{R}^{n\times n}, D = \operatorname{diag}(D), \operatorname{trace}(D) = 0\}. \tag{2.4}$$

Note that, although only simple diagonal scalings are considered, the results apply to general block diagonal/repeated scalings. The advantage of this formulation arises from the convexity of the function $f(M,D) = \|e^D M e^{-D}\|$. It is shown in [2] that the set of optimal scalings $\{D : D \in \mathcal{D}, \|e^D M e^{-D}\| = f_{\min}(M)\}$ is nonempty and bounded if M is irreducible (for a reducible M the infimum may or may not be attained). If M is reducible, there exists (by definition) a permutation P such that

$$PMP' = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix}$$

with $M_{11} \in \mathcal{C}^{r\times r}$, $r < n$. Since for any permutation matrix, $\|Pe^D P' PMP' Pe^{-D} P'\| = \|e^D M e^{-D}\|$, we can assume that every reducible matrix is of the above form (note that $Pe^D P'$ is diagonal and corresponds to the re-ordering of the diagonal structure). Now, $\mu_{\mathbf{\Delta}}^{-1}(M) = \min\{\|\Delta\| : \Delta \in \mathbf{\Delta}, \det(I - M\Delta) = 0\}$. Since $\det(I - \Delta M) = \det(I - \Delta_1 M_{11}) \det(I - \Delta_2 M_{22})$, where Δ_1 and Δ_2 denote, respectively, the first and second diagonal blocks of Δ, it follows that for every reducible matrix, $\mu_{\mathbf{\Delta}}(M) = \max\{\mu_{\mathbf{\Delta_1}}(M_{11}), \mu_{\mathbf{\Delta_2}}(M_{22})\}$, where $\mathbf{\Delta_1}$ and $\mathbf{\Delta_2}$ denote the two diagonal structures of dimension $r \times r$ and $(n-r) \times (n-r)$ inherited from $\mathbf{\Delta}$. Thus if M is reducible, the results of the paper can be applied separately to M_{11} and M_{22} (assumed irreducible) to obtain upper bounds $\mu_{\mathbf{\Delta_1}}(M_{11}) \le \bar{\mu}_{\mathbf{\Delta_1}}(M_{11})$ and $\mu_{\mathbf{\Delta_2}}(M_{22}) \le \bar{\mu}_{\mathbf{\Delta_2}}(M_{22})$, which are tighter (or at least no worse) than the two individual D-iteration upper bounds. We can then define $\bar{\mu}_{\mathbf{\Delta}}(M) := \max\{\bar{\mu}_{\mathbf{\Delta_1}}(M_{11}), \bar{\mu}_{\mathbf{\Delta_2}}(M_{22})\}$ as a bound on $\mu_{\mathbf{\Delta}}$. Since for any sequence (D^i) in $\mathcal{D}$ we have that

$$\left\| \begin{bmatrix} e^{D_1^i} & 0 \\ 0 & e^{D_2^i} \end{bmatrix} M \begin{bmatrix} e^{-D_1^i} & 0 \\ 0 & e^{-D_2^i} \end{bmatrix} \right\| \ge \max\{\|e^{D_1^i} M_{11} e^{-D_1^i}\|, \|e^{D_1^i} M_{11} e^{-D_1^i}\|\},$$

the bound $\bar{\mu}_{\mathbf{\Delta}}(M)$ can be no worse than the optimal D-iteration bound of M. If M_{11} or M_{22} (or both) are reducible, we can extend the process by partitioning these matrices further to obtain the same conclusion.

In what follows, extensive use is made of the following optimization problems associated with M and their solutions:

$$\max\{\rho(V_1' \Delta U_1) : \Delta \in \mathcal{B}\mathbf{\Delta}\} \tag{2.5}$$

and

$$\max\{\|V_1' \Delta U_1\| : \Delta \in \mathcal{B}\mathbf{\Delta}\}. \tag{2.6}$$

Remark 2.2. Let

$$M_0 = U\Sigma_0 V' := \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} = U_1 V_1', \tag{2.7}$$

so that M_0 is obtained from the singular value decomposition of M in (2.1) by setting $\Sigma_2 = 0$. Then the optimization in (2.5) becomes

$$\max_{\Delta\in\mathcal{B}\mathbf{\Delta}} \rho(V_1' \Delta U_1) = \max_{\Delta\in\mathcal{B}\mathbf{\Delta}} \rho(\Delta U_1 V_1') = \max_{\Delta\in\mathcal{B}\mathbf{\Delta}} \rho(\Delta M_0) =: \mu_{\mathbf{\Delta}}(M_0)$$

and is therefore a reduced rank μ problem. One of the contributions of this paper is to show that $\mu_{\mathbf{\Delta}}(M) = 1$ if and only if $\mu_{\mathbf{\Delta}}(M_0) = 1$.

The following result uses a semidefinite relaxation of the nonconvex optimization in (2.6) to derive an upper bound, shows that the gap is zero, and proves that the optimum is 1 if M is the (scaled) output of the D-iteration.

LEMMA 2.3. *With all variables as defined above,*

$$\max_{\Delta\in\mathcal{B}\mathbf{\Delta}} \|V_1'\Delta U_1\|^2 = \min\{\gamma : D\in\mathcal{D},\ D - V_1V_1' \geq 0,\ \gamma I - U_1'DU_1 \geq 0\}. \tag{2.8}$$

Furthermore, if the optimum in (2.3) *is* 1*, the optimum in* (2.6) *is* 1.

Proof. See the appendix. □

It is clear that $\|M\| = 1$ is an upper bound on $\mu_{\mathbf{\Delta}}(M)$ and on (2.5) and (2.6). Suppose that $\bar{\phi}_1$ and $\bar{\phi}_2$ are upper bounds on (2.5) and (2.6). The contribution of this paper is to show that if $\bar{\phi}_1$ or $\bar{\phi}_2$ is less than 1, then we can derive an upper bound on $\mu_{\mathbf{\Delta}}(M)$ which is also less than 1. More specifically, since $\rho(E_1'V'\Delta UE_1) \leq \bar{\phi}_1\|\Delta\|$ and $\|E_1'V'\Delta UE_1\| \leq \bar{\phi}_2\|\Delta\|\ \forall\ \Delta\in\mathbf{\Delta}$, where E_1 is the matrix of the first m columns of the $n\times n$ identity matrix, then

$$r_{\mathbf{\Delta}}(M) = \min_{\substack{\det(A-V'\Delta U)=0\\ \Delta\in\mathbf{\Delta}}} \|\Delta\| \geq \min_{\substack{\det(A-\Delta)=0\\ \rho(E_1'\Delta E_1)\leq\bar{\phi}_1\|\Delta\|\\ \|E_1'\Delta E_1\|\leq\bar{\phi}_2\|\Delta\|}} \|\Delta\| =: \underline{r}_{\mathbf{\Delta}}(M). \tag{2.9}$$

It follows that $[\underline{r}_{\mathbf{\Delta}}(M)]^{-1}$, evaluated in section 4, is an upper bound on $\mu_{\mathbf{\Delta}}(M)$.

Remark 2.4. It is implicit in the work in [21] that $\mu_{\mathbf{\Delta}}(M) = 1$ if $\mu_{\mathbf{\Delta}}(M_0) = 1$. It is also clear that the minimum on the right-hand sides of (2.9) is equal to 1 if $\bar{\phi}_1 = \bar{\phi}_2 = 1$. Since (2.6) is an upper bound on (2.5), we can always take $\bar{\phi}_1 \leq \bar{\phi}_2$. So in this work, we require, and therefore assume, that $\bar{\phi}_1 < 1$.

Remark 2.5. Our problem formulation assumes that M is square and nonsingular. This assumption involves no loss of generality and can be easily removed; see [14, Thm. 4.6] for details.

Our general approach in this work closely resembles the method used in [18] for breaching the duality gap between the quadratic integer programming problem and its semidefinite relaxation. Our results in that case involved considerably less technical detail than our present work, although the core idea is similar: The objective in both cases is to apply a convex relaxation to a computationally intractable (NP-hard) problem and, subsequently, to explore the properties of the optimal solution to the dual problem in order to reduce the duality gap. This is achieved in both cases by solving a problem of the same form as the primal, but of reduced complexity. In the following paragraphs we highlight the main results of [18] and its similarities with our present work. We hope that this will make the logic behind the sequence of arguments given in this paper easier to follow.

The quadratic integer programming (QIP) problem involves the maximization of the quadratic form $x'Qx$, in which $Q = Q' \in \mathcal{R}^{n\times n}$ and x is allowed to vary over $\{-1, 1\}^n$. Since this involves 2^{n-1} function evaluations, the computational complexity of the problem grows exponentially in n. Denote the maximum of the primal problem by γ. The convex relaxation upper bound $\bar{\gamma}$ of the problem involves the minimization of the trace of a diagonal matrix D such that $D - Q \geq 0$. It can be shown [18] that (i) the optimal D matrix, D_0 say, is unique; (ii) a simple sufficient condition for the duality gap $\gamma - \bar{\gamma}$ to be zero is that the null-space of $D_0 - Q$ is one-dimensional; (iii) a necessary and sufficient condition for the duality gap to be zero is that $\gamma_m = \frac{1}{n}\max\{x'VV'x\} = 1$, where $x \in \{-1,\ 1\}^n$ and V is an orthonormal matrix whose columns span the m-dimensional null-space of $D_0 - Q$. Let $V = [v_1, v_2, \ldots, v_m]$, where $\{v_i\}$ is an orthonormal basis of $\mathcal{N}(D_0 - Q)$. Introduce a row perturbation matrix

P so that $PV = [V_1' \; V_2']'$, where $V_1 \in \mathcal{R}^{m\times m}$ is nonsingular. Then, (iv) the duality gap is zero if and only if $V_2 V_1^{-1} z = \{-1,\ 1\}^{n-m}$ for some $z \in \{-1,\ 1\}^m$. Note that this gives a test which certifies if the duality gap is zero, requiring only 2^{m-1} evaluations. Finally, let $\underline{\lambda}_+$ be the smallest positive eigenvalue of $D_0 - Q$. Then, (v) $\gamma \leq \bar{\gamma} - n(1-\gamma_m)\underline{\lambda}_+ < \bar{\gamma}$. In other words the duality gap can be breached by evaluating γ_m. Now the maximization defining γ_m has the same form as the primal problem with the crucial difference that now $\text{rank}(VV') = m < n$ (potentially m is much smaller than n). It is shown in [18] that the evaluation of γ_m can be performed in polynomial time (in n) and is of complexity $O(n^{m-1})$. Geometrically the problem corresponds to the evaluation of the extreme points of a zonotope with n generating vectors in $m-1$ dimensions and can be solved by a variety of techniques. In [18] a reverse-enumeration algorithm was used, based on the work of [1], and was found to perform well in practice.

For the problem considered in this work, the primal problem is the evaluation of the structured singular value of an arbitrary matrix ($\mu(M)$), while the dual problem corresponds to its D-iteration upper bound (equation (2.3)). Since the constraint set of the dual problem is convex, the optimal solution can be easily calculated (e.g., via LMI techniques). Redefine M by absorbing the D-iteration matrix scalings and dividing by its norm so that $\|M\| = 1$. Let m be the multiplicity of the largest singular value of M (equal to one) and define U_1 and V_1 via the singular value decomposition of M in (2.1). It is well known [21] that $\mu(M) = 1$ if $m = 1$. In general, to test whether $\mu(M)$ is achieved at the end of the D-iteration we need to verify whether $\mu(M_0) = 1$, which is a reduced rank (m-rank) μ-problem. Breaching the duality gap also requires evaluating $\mu(M_0)$ (or at least obtaining an upper bound $\bar{\phi}_1$ less than 1) and solving an eigenvalue problem (see Theorem 4.1 and Remark 4.2). Note that in this case, since M is redefined at the end of the D-iteration procedure, we must have $\bar{\phi}_2 = 1$.

The similarities of the approach in the two problems are clear. The link is potentially important because it allows us to transfer intuition across different problem domains. We believe that this approach is sufficiently general and can be applied to a more general class of optimization problems when convex relaxations are used.

The structure of the paper is as follows. In section 3 we generalize an approach proposed in [16] to matrices with multiple largest singular value by solving a number of approximation problems involving the structured distance of a matrix to singularity, starting from the simplest case (zero uncertainty matrix subblock) and progressively moving to more general cases (uncertainty subblock is a fixed matrix; uncertainty subblock is free to vary over a compact set). In each case the minimum distance to singularity is obtained in a closed form and involves only straightforward numerical calculations (e.g., the solution of an eigenvalue problem); a complete parameterization of all optimal (minimum-norm) singularizing perturbations is also given in each case. The link to the structured singular value is obtained by specializing these results to a specific structured uncertainty set involving simultaneous spectral radius and norm constraints. This is obtained in Theorem 3.19, which gives the optimal solution in terms of a Toeplitz matrix (and its transformations). The link with μ is formalized in section 4 (Theorem 4.1), where the procedure for breaching the convex bound is outlined. An illustrative example is given in section 5. The results of the work are summarized in section 6. Finally, the main conclusions are drawn in section 7 and the proof of Lemma 2.3 is included in the appendix.

3. Structured distance to singularity. We first consider and generalize a structured uncertainty approach proposed in [16] and establish a connection with the structured singular value. We start by giving the following definition.

DEFINITION 3.1. *Let A be as defined in section 2 and let $E_1 = \begin{bmatrix} I_m & 0_{m\times(n-m)} \end{bmatrix}'$. For any set $\mathbf{\Delta_{11}} \subseteq \mathcal{C}^{m\times m}$ define the structured distance (of A) to singularity*

$$\gamma_{\mathbf{\Delta_{11}}} = \min \{\|\Delta\| : \det(A-\Delta) = 0,\ E_1{}'\Delta E_1 \in \mathbf{\Delta_{11}}\},$$

and the set of all optimal structured rank reducing perturbations

$$\mathcal{D}_{\mathbf{\Delta_{11}}} := \{\Delta \in \mathcal{C}^{n\times n} : \|\Delta\| = \gamma_{\mathbf{\Delta_{11}}}, \det(A-\Delta) = 0, E_1{}'\Delta E_1 \in \mathbf{\Delta_{11}}\}.$$

In this notation the (unstructured) distance to singularity will be denoted as

$$\gamma_{\mathcal{C}^{\mathbf{m}\times\mathbf{m}}} = \min \{\|\Delta\| : \det(A-\Delta)=0,\ E_1{}'\Delta E_1 \in \mathcal{C}^{\mathbf{m}\times\mathbf{m}}\} = \min \{\|\Delta\| : \det(A-\Delta)=0\}.$$

Recall that, unless specified otherwise, all matrices are assumed to be complex. It is thus implicit that $\Delta \in \mathcal{C}^{n\times n}$ in the definition of $\gamma_{\mathbf{\Delta_{11}}}$ above. For the remainder of this section, we evaluate $\gamma_{\mathbf{\Delta_{11}}}$ and $\mathcal{D}_{\mathbf{\Delta_{11}}}$ for several sets $\mathbf{\Delta_{11}}$.

3.1. $\mathbf{\Delta_{11}} = \mathcal{C}$. When $m = 1$, it is clear that the unstructured distance of A from singularity is

$$\gamma_{\mathcal{C}} = \min \{\|\Delta\| : \det(A-\Delta) = 0\} = \underline{\sigma}(A) = 1,$$

and the set of all optimal rank reducing perturbations is given by

$$\mathcal{D}_{\mathcal{C}} = \{\Delta = \operatorname{diag}(1, \Delta_{22}) \in \mathcal{C}^{n\times n}, \Delta_{22} \in \mathcal{C}^{(n-1)\times(n-1)}, \|\Delta_{22}\| \le 1\}.$$

3.2. $\mathbf{\Delta_{11}} = \{\delta \in \mathcal{C} : |\delta| \le \phi\}$, $\mathbf{0 \le \phi < 1}$. In [16] the (1,1) entry of the permissible perturbations was constrained to have modulus $\le \phi$, where $0 \le \phi < 1$ to obtain the following structured distance to singularity.

THEOREM 3.2 (see [16]). *Let $A = \operatorname{diag}(1, a_2, a_3, \ldots, a_n)$ with $1 < a_2 < \cdots < a_n$ be given and define $E_1 = \begin{bmatrix} 1 & 0_{1,n-1} \end{bmatrix}'$ and $\mathbf{\Delta}_\phi = \{\delta \in \mathcal{C} : |\delta| \le \phi\}$ for any $0 \le \phi \le 1$. Then*

$$\gamma_{\mathbf{\Delta}_\phi} := \min \{\|\Delta\| : \det(A-\Delta) = 0,\ E_1{}'\Delta E_1 \in \mathbf{\Delta}_\phi\} = \sqrt{a_2 - \phi(a_2-1)},$$

and for $\phi < 1$, the set of all optimal rank reducing perturbation is given by

$$\mathcal{D}_{\mathbf{\Delta}_\phi} = \left\{ \operatorname{diag}\left(\begin{bmatrix} \phi & \gamma_{\mathbf{\Delta}_\phi} e^{j\theta} \\ \gamma_{\mathbf{\Delta}_\phi} e^{-j\theta} & -\phi \end{bmatrix}, \Delta_{33} \right) : \theta \in \mathcal{R},\ \Delta_{33} \in \mathcal{C}^{(n-2)\times(n-2)}, \|\Delta_{33}\| \le \gamma_{\mathbf{\Delta}_\phi} \right\}.$$

Remark 3.3. Theorem 3.2 was used in [14] to derive an upper bound on μ in the case that m, the multiplicity of the largest singular value of M, is one. In Corollary 3.20 below, we give a generalization of Theorem 3.2 to the case $m \ge 1$. In section 4 this generalization is used to derive an upper bound on μ when $m \ge 1$.

3.3. $\mathbf{\Delta_{11}} = \mathcal{C}^{\mathbf{m}\times\mathbf{m}}$. To motivate the generalization of Theorem 3.2 for $m \ge 1$, we start with the unstructured multidimensional case.

LEMMA 3.4. *Let A and E_1 be as in Definition* 3.1. *Then*

$$\gamma_{\mathcal{C}^{\mathbf{m}\times\mathbf{m}}} = \min \{\|\Delta\| : \det(A-\Delta) = 0\} = 1,$$

and the set of all optimal rank reducing perturbations is given by

$$\mathcal{D}_{\mathcal{C}^{m\times m}} = \left\{ \begin{bmatrix} W & 0 \\ 0 & I_{n-m} \end{bmatrix} \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & \Delta_{22} & \Delta_{23} \\ \hline 0 & \Delta_{32} & \Delta_{33} \end{array}\right] \begin{bmatrix} W' & 0 \\ 0 & I_{n-m} \end{bmatrix} \in \mathcal{C}^{n\times n} : \right.$$

$$\left. WW' = W'W = I_m, \left\| \begin{bmatrix} \Delta_{22} & \Delta_{23} \\ \Delta_{32} & \Delta_{33} \end{bmatrix} \right\| \le 1 \right\}. \tag{3.1}$$

Proof. The proof that $\gamma_{\mathcal{C}^{m\times m}} = 1$ is well known. To prove that all minimizers are given by (3.1) let $\Delta \in \mathcal{C}^{n\times n}$ such that $\|\Delta\| = 1$ and $\det(A-\Delta) = 0$. Let $(A-\Delta)x = 0$ for some $x = \begin{bmatrix} x_1' & x_2' \end{bmatrix}'$ with $x_1 \in \mathcal{C}^m$ and such that $\|x\| = 1$. Then

$$(A-\Delta)x = 0 \Rightarrow x'A^2x = x'\Delta'\Delta x \Rightarrow x'\Delta'\Delta x = 1 \quad \text{and} \quad x_2 = 0 \Rightarrow {E_1}'\Delta E_1 x_1 = x_1$$
$$\Rightarrow 1 = \|{E_1}'\Delta E_1\| \;\&\; 1 \in \lambda({E_1}'\Delta E_1), \tag{3.2}$$

since $\underline{\sigma}(A) = 1$. It is straightforward to verify that all such Δ are captured by (3.1). ∎

3.4. $\boldsymbol{\Delta_{11} = \{0_{m\times m}\}}$. Thus a natural generalization of Theorem 3.2 is to impose constraints on ${E_1}'\Delta E_1$ so that (3.2) is violated. We start with the simplest constraint: ${E_1}'\Delta E_1 = 0$.

LEMMA 3.5. *Let $A = \operatorname{diag}(A_1, A_2) \in \mathcal{R}^{n\times n}$, with $A_1 = \operatorname{diag}(I_{m_1}, A_{22}) \in \mathcal{R}^{m\times m}$ and $A_2 = \operatorname{diag}(a_3 I_{m_3}, A_{44}) \in \mathcal{R}^{(n-m)\times(n-m)}$. Assume that $1 < \underline{\sigma}(A_{22})$, $0 < a_3 < \underline{\sigma}(A_{44})$, and $1 < a_3$. Then the structured distance to singularity is*

$$\gamma_{0_{m\times m}} := \min\{\|\Delta\| : \det(A-\Delta) = 0,\ {E_1}'\Delta E_1 = 0\} = \sqrt{a_3} =: \sqrt{\underline{\sigma}(A_1)\underline{\sigma}(A_2)}. \tag{3.3}$$

Furthermore all optimal rank reducing perturbations are generated by

$$W \left[\begin{array}{ccc||ccc} 0 & 0 & 0 & \sqrt{a_3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \Delta_{13} & \Delta_{14} \\ \hline 0 & 0 & 0 & 0 & \Delta_{23} & \Delta_{24} \\ \hline\hline \sqrt{a_3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \Delta_{31} & \Delta_{32} & 0 & \Delta_{33} & \Delta_{34} \\ \hline 0 & \Delta_{41} & \Delta_{42} & 0 & \Delta_{43} & \Delta_{44} \end{array}\right] W', \tag{3.4}$$

where $W = \operatorname{diag}(W_1, I_{m_2}, W_3, I_{m_4}) \in \mathcal{C}^{n\times n}$ is unitary and

$$\left\| \begin{bmatrix} 0 & 0 & \Delta_{13} & \Delta_{14} \\ 0 & 0 & \Delta_{23} & \Delta_{24} \\ \Delta_{31} & \Delta_{32} & \Delta_{33} & \Delta_{34} \\ \Delta_{41} & \Delta_{42} & \Delta_{43} & \Delta_{44} \end{bmatrix} \right\| \le \sqrt{a_3}. \tag{3.5}$$

Proof. A calculation verifies that all Δ of the form (3.4) have norm $\sqrt{a_3}$ and satisfy $\det(A-\Delta) = 0$. Thus $\sqrt{a_3}$ is an upper bound. Let $\Delta \in \mathcal{C}^{n\times n}$ be such that ${E_1}'\Delta E_1 = 0$ and $\det(A-\Delta) = 0$. Then there exists a nonzero $x = \begin{bmatrix} x_1' & x_2' \end{bmatrix}' \in \mathcal{C}^n$ with $x_1 \in \mathcal{C}^m$ such that $(A-\Delta)x = 0$. It is clear that $x_2 \neq 0$ since A_1 has full rank, and a little reflection will verify that $x_1 \neq 0$, since otherwise $\|\Delta\| \ge a_3 > \sqrt{a_3}$. Let $Z = \operatorname{diag}(z_1, z_2)$, where $z_1 = x_1/\|x_1\|$ and $z_2 = x_2/\|x_2\|$. Then

$$(A-\Delta)x = 0 \Rightarrow (A-\Delta)Z \begin{bmatrix} \|x_1\| \\ \|x_2\| \end{bmatrix} = 0 \Rightarrow \det\left(\begin{bmatrix} z_1'A_1z_1 & 0 \\ 0 & z_2'A_2z_2 \end{bmatrix} - Z'\Delta Z \right) = 0,$$

with the (1,1) entry of $Z'\Delta Z$ equal to zero. It follows from Theorem 3.2 (setting $\phi = 0$) and the fact that $\|Z\| = \|z_1\| = \|z_2\| = 1$ that

$$\|\Delta\| \ge \|Z'\Delta Z\| \ge \sqrt{|z_1'A_1z_1||z_2'A_2z_2|} \ge \sqrt{a_3},$$

with equality if and only if $z_1 = \begin{bmatrix} z_{11}' & 0 \end{bmatrix}'$ and $z_2 = \begin{bmatrix} z_{33}' & 0 \end{bmatrix}'$ with $z_{11} \in \mathcal{C}^{m_1}$ and $z_{33} \in \mathcal{C}^{m_3}$. Thus $\sqrt{a_3}$ is a lower bound and this proves (3.3).

To prove that (3.4) captures all optimal structured perturbations it suffices to show that any Δ which minimizes (3.3) can be written in the form of (3.4). Let

$$\Delta = \begin{bmatrix} O_{m,m} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{bmatrix} \in \mathcal{C}^{n\times n}$$

be any minimizer of (3.3). Also introduce the finer partitions,

$$\Delta_{12} = \begin{bmatrix} \hat{\Delta}_{11} & \hat{\Delta}_{12} \\ \hat{\Delta}_{21} & \hat{\Delta}_{22} \end{bmatrix}, \quad \Delta_{21} = \begin{bmatrix} \bar{\Delta}_{11} & \bar{\Delta}_{12} \\ \bar{\Delta}_{21} & \bar{\Delta}_{22} \end{bmatrix}, \quad \Delta_{22} = \begin{bmatrix} \tilde{\Delta}_{11} & \tilde{\Delta}_{12} \\ \tilde{\Delta}_{21} & \tilde{\Delta}_{22} \end{bmatrix},$$

where $\hat{\Delta}_{11} \in \mathcal{C}^{m_1\times m_3}$, $\bar{\Delta}_{11} \in \mathcal{C}^{m_3\times m_1}$, and $\tilde{\Delta}_{11} \in \mathcal{C}^{m_3\times m_3}$. Since Δ is optimal, a previous argument implies that

$$\det\left(\begin{bmatrix} z_1'A_1z_1 & 0 \\ 0 & z_2'A_2z_2 \end{bmatrix} - \begin{bmatrix} 0 & z_1'\Delta_{12}z_2 \\ z_2'\Delta_{21}z_1 & z_2'\Delta_{22}z_2 \end{bmatrix}\right) = 0, \tag{3.6}$$

and in fact the second matrix in the left-hand side of (3.6) is an optimal rank reducing approximation of the first matrix in the left-hand side of (3.6), with the constraint that the (1,1) entry is zero. Using the parameterization of Theorem 3.2 (with $\phi = 0$), and noting that $z_1'A_1z_1 = \|z_{11}\|^2 = 1$ and $z_2'A_2z_2 = a_3\|z_{33}\|^2 = a_3$, shows that

$$z_1'\Delta_{12}z_2 = z_{11}'\hat{\Delta}_{11}z_{33} = e^{j\theta}\sqrt{a_3}\ , \quad z_2'\Delta_{21}z_1 = z_{33}'\bar{\Delta}_{11}z_{33} = e^{-j\theta}\sqrt{a_3}\ , \tag{3.7}$$

and

$$z_2'\Delta_{22}z_2 = z_{33}'\tilde{\Delta}_{11}z_{33} = 0, \tag{3.8}$$

for some $\theta \in (-\pi, \pi]$. Next define any two unitary completions $U_\perp$ and $V_\perp$ of $e^{j\theta}z_{11}$ and z_{33}, respectively, to construct unitary matrices $W_1 = \begin{bmatrix} e^{j\theta}z_{11} & U_\perp \end{bmatrix}$ and $W_3 = \begin{bmatrix} z_{33} & V_\perp \end{bmatrix}$, and consider the product

$$\begin{bmatrix} W_1' & 0 \\ 0 & I_{m_2} \end{bmatrix} \Delta_{12} \begin{bmatrix} W_3 & 0 \\ 0 & I_{m_4} \end{bmatrix} = \begin{bmatrix} \psi z_{11}'\hat{\Delta}_{11}z_{33} & \psi z_{11}'\hat{\Delta}_{11}V_\perp & \psi z_{11}'\hat{\Delta}_{12} \\ U_\perp'\hat{\Delta}_{11}z_{33} & U_\perp'\hat{\Delta}_{11}V_\perp & U_\perp'\hat{\Delta}_{12} \\ \hat{\Delta}_{21}z_{33} & \hat{\Delta}_{21}V_\perp & \hat{\Delta}_{22} \end{bmatrix},$$

where $\psi = e^{-j\theta}$. Using the first equation in (3.7), the $(1,1)$ block in the right-hand side of the above equation may be written as $\sqrt{a_3}$. Now, since W_1 and W_3 are unitary and $\|\Delta_{12}\| \le \|\Delta\| = \sqrt{a_3}$, we conclude that $z_{11}'\hat{\Delta}_{11}V_\perp = 0$, $z_{11}'\hat{\Delta}_{12} = 0$, $U_\perp'\hat{\Delta}_{11}z_{33} = 0$, and $\hat{\Delta}_{21}z_{33} = 0$. Defining

$$\begin{bmatrix} U_\perp'\hat{\Delta}_{11}V_\perp & U_\perp'\hat{\Delta}_{12} \\ \hat{\Delta}_{21}V_\perp & \hat{\Delta}_{22} \end{bmatrix} = \begin{bmatrix} E_{13} & E_{14} \\ E_{23} & E_{24} \end{bmatrix}$$

thus gives

$$\begin{bmatrix} \hat{\Delta}_{11} & \hat{\Delta}_{12} \\ \hat{\Delta}_{21} & \hat{\Delta}_{22} \end{bmatrix} = \begin{bmatrix} W_1 & 0 \\ 0 & I_{m_2} \end{bmatrix} \begin{bmatrix} \sqrt{a_3} & 0 & 0 \\ 0 & E_{13} & E_{14} \\ 0 & E_{23} & E_{24} \end{bmatrix} \begin{bmatrix} W_3' & 0 \\ 0 & I_{m_4} \end{bmatrix}. \tag{3.9}$$

A similar argument using the second equation in (3.7) and (3.8) shows that

$$\begin{bmatrix} \bar{\Delta}_{11} & \bar{\Delta}_{12} \\ \bar{\Delta}_{21} & \bar{\Delta}_{22} \end{bmatrix} = \begin{bmatrix} W_3 & 0 \\ 0 & I_{m_4} \end{bmatrix} \begin{bmatrix} \sqrt{a_3} & 0 & 0 \\ 0 & E_{31} & E_{31} \\ 0 & E_{41} & E_{42} \end{bmatrix} \begin{bmatrix} W_1' & 0 \\ 0 & I_{m_2} \end{bmatrix} \tag{3.10}$$

and

$$\begin{bmatrix} \tilde{\Delta}_{11} & \tilde{\Delta}_{12} \\ \tilde{\Delta}_{21} & \tilde{\Delta}_{22} \end{bmatrix} = \begin{bmatrix} W_3 & 0 \\ 0 & I_{m_4} \end{bmatrix} \begin{bmatrix} 0 & \tilde{E}_{23} & \tilde{E}_{24} \\ \tilde{E}_{32} & E_{33} & E_{34} \\ \tilde{E}_{42} & E_{43} & E_{44} \end{bmatrix} \begin{bmatrix} W_3' & 0 \\ 0 & I_{m_4} \end{bmatrix}, \tag{3.11}$$

where again the E_{ij}'s and $\tilde{E}_{ij}$'s are well-defined matrices of the appropriate dimensions. Again using $\|\Delta\| = \sqrt{a_3}$ shows that all $\tilde{E}_{ij} = 0$. Writing Δ in full using the form of its partitions defined in (3.9), (3.10), and (3.11) and extracting matrix factors $W = \mathrm{diag}(W_1, I_{m_2}, W_3, I_{m_4})$ and W' from the left and the right respectively, shows that Δ agrees with the form given in (3.4). Finally, the required contractive property in (3.5) follows directly from $\|\Delta\| = \sqrt{a_3}$. ∎

3.5. $\boldsymbol{\Delta_{11} = \{\Delta_{11}\}, \Delta_{11} \in \mathcal{C}^{m\times m}, \|\Delta_{11}\| \le 1, \det(I - \Delta_{11}) \neq 0}$. Next, we impose the constraint ${E_1}'\Delta E_1 = \Delta_{11}$, where Δ_{11} is a given matrix such that (3.2) is violated. First, we need the following results, which can be deduced from [23], concerning properties of linear fractional maps.

LEMMA 3.6. *Let* $H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$ *and* U *be complex matrices and assume that* $(I - H_{11}U)^{-1}$, H_{12}^{-1}, *and* H_{21}^{-1} *exist.*

1. *If* H *is nonsingular, then* $\Phi = \mathcal{F}_u(H, U)$ *if and only if* $U = \mathcal{F}_l(H^{-1}, \Phi)$.
2. *If* H *and* U *are nonsingular, then* $[\mathcal{F}_u(H, U)]^{-1} = \mathcal{F}_u(H^{-1}, U^{-1})$.
3. *If* H *is* γ*-unitary for some* $\gamma > 0$, *then* $\|\mathcal{F}_u(H, U)\| = \gamma$ *if and only if* $\|U\| = \gamma^{-1}$ *and* $\|\mathcal{F}_u(H, U)\| < \gamma$ *if and only if* $\|U\| < \gamma^{-1}$.
4. *If* H *is* γ*-unitary for some* $\gamma > 1$, H_{11} *is square,* $I - H_{11}$ *is nonsingular, and* $\|H_{11}\| \le 1$, *then*

$$\|\mathcal{F}_u(H, I)\| = \|(\gamma^2 I - H_{11})(I - H_{11})^{-1}\| = \|(I - H_{11})^{-1}(\gamma^2 I - H_{11})\|.$$

Proof.

1. This follows from a simple calculation.
2. This follows from the definition of the linear fractional map.
3. This is a standard result and follows from the identity

$$\gamma^2 I - \mathcal{F}_u(H, U)'\mathcal{F}_u(H, U) = H_{12}'(I - U'H_{11}')^{-1}(I - \gamma^2 U'U)(I - H_{11}U)^{-1}H_{12}.$$

4. Define $E = (I - H_{11})^{-1}(\gamma^2 I - H_{11})$, $F = \mathcal{F}_u(H, I)$, and $a = \|F\|$. Then

$$\begin{aligned} a = \|F\| &\Leftrightarrow \underline{\lambda}(a^2 I - F'F) = 0 \Leftrightarrow \underline{\lambda}[(a^2 - \gamma^2)I + \gamma^2 I - F'F] = 0 \\ &\Leftrightarrow \underline{\lambda}[(a^2 - \gamma^2)I - (\gamma^2 - 1)H_{12}'(I - H_{11}')^{-1}(I - H_{11})^{-1}H_{12}] = 0 \\ &\Leftrightarrow \underline{\lambda}[(a^2 - \gamma^2)I - (\gamma^2 - 1)(I - H_{11})^{-1}H_{12}H_{12}'(I - H_{11}')^{-1}] = 0 \\ &\Leftrightarrow \underline{\lambda}[(a^2 - \gamma^2)I - (\gamma^2 - 1)(I - H_{11})^{-1}(\gamma^2 I - H_{11}H_{11}')(I - H_{11}')^{-1}] = 0, \end{aligned}$$

where we have used the identity in part 3. On the other hand,

$$\begin{aligned} a^2 I - EE' &= (a^2 - \gamma^2)I + \gamma^2 I - EE' \\ &= (a^2 - \gamma^2)I - (\gamma^2 - 1)(I - H_{11})^{-1}(\gamma^2 I - H_{11}H_{11}')(I - H_{11}')^{-1}, \end{aligned}$$

so that $\underline{\lambda}(a^2 I - EE') = 0$. Thus $a = \|E\|$, which proves the result. ∎

LEMMA 3.7. *Let* A *and* E_1 *be as in Definition* 3.1. *Let* $\Delta_{11} \in \mathcal{C}^{m\times m}$ *be given and assume that* $\|\Delta_{11}\| \le 1$ *and that* $I - \Delta_{11}$ *is nonsingular. Then*

$$\min_{\substack{\det(A-\Delta)=0 \\ {E_1}'\Delta E_1 = \Delta_{11}}} \|\Delta\| = \min_{\substack{\|(\gamma^2 I - \Delta_{11})(I - \Delta_{11})^{-1}\| = a_{m+1} \\ \gamma > 1}} \gamma. \tag{3.12}$$

Proof. We use a lifting procedure to reduce the problem to the case $\Delta_{11}=0$. Since $\|\Delta_{11}\|\le 1$, for any $\gamma>1$ there exist γ-unitary completions of Δ_{11} of the form

$$\Delta_1^\gamma=\begin{bmatrix}\Delta_{11} & \Delta_{13}^\gamma\\ \Delta_{31}^\gamma & \Delta_{33}\end{bmatrix}\in\mathcal{C}^{2m\times 2m},\quad \Delta_0^\gamma=\begin{bmatrix}\Delta_{11} & 0 & \Delta_{13}^\gamma & 0\\ 0 & 0 & 0 & \gamma I_{n-m}\\ \Delta_{31}^\gamma & 0 & \Delta_{33} & 0\\ 0 & \gamma I_{n-m} & 0 & 0\end{bmatrix}\in\mathcal{C}^{2n\times 2n}.$$

It is clear that a_{m+1} is an upper bound to (3.12) (take $\Delta=\operatorname{diag}(\Delta_{11},a_{m+1}I_{n-m})$), so we can assume that $\gamma\le a_{m+1}$. Define

$$X_A^\gamma=\mathcal{F}_u[(\Delta_0^\gamma)^{-1},A]=\operatorname{diag}\left(\mathcal{F}_u[(\Delta_1^\gamma)^{-1},I],\gamma^{-2}A_2\right)=:\operatorname{diag}\left(X_1^\gamma,X_2^\gamma\right).$$

Parts 2 and 3 of Lemma 3.6 prove that $(X_1^\gamma)^{-1}=\mathcal{F}_u(\Delta_1^\gamma,I)$ and $\|(X_1^\gamma)^{-1}\|>\gamma$, respectively, so that $\underline{\sigma}(X_1^\gamma)<\gamma^{-1}$. Also, $\underline{\sigma}(X_2^\gamma)=\gamma^{-2}a_{m+1}>\underline{\sigma}(X_1^\gamma)$ since $\gamma\le a_{m+1}$. For any $\Delta\in\mathcal{C}^{n\times n}$ define

$$\Phi_\Delta^\gamma:=\mathcal{F}_u[(\Delta_0^\gamma)^{-1},\Delta]. \tag{3.13}$$

It is not difficult to show that

$$\|\Delta\|=\gamma\Leftrightarrow\|\Phi_\Delta^\gamma\|=\gamma^{-1},$$
$$E_1{}'\Delta E_1=\Delta_{11}\Leftrightarrow E_1{}'\Phi_\Delta^\gamma E_1=0,$$

and

$$\det(A-\Delta)=0\Leftrightarrow\det(X_A^\gamma-\Phi_\Delta^\gamma)=0,$$

and so

$$\begin{aligned}\min_{\substack{\det(A-\Delta)=0\\ \gamma=\|\Delta\|\\ E_1{}'\Delta E_1=\Delta_{11}}}\gamma&=\min_{\substack{\det(X_A^\gamma-\Phi_\Delta^\gamma)=0\\ \|\Phi_\Delta^\gamma\|=\gamma^{-1}\\ E_1{}'\Phi_\Delta^\gamma E_1=0}}\gamma=\min_{\sqrt{\underline{\sigma}(X_1^\gamma)\underline{\sigma}(X_2^\gamma)}\le\gamma^{-1}}\gamma\\ &=\min_{\|(X_1^\gamma)^{-1}\|\ge a_{m+1}}\gamma,\end{aligned} \tag{3.14}$$

where the last two equalities follow from Lemma 3.5. Now $\|(X_1^\gamma)^{-1}\|$ is continuous in γ, $\|(X_1^1)^{-1}\|=1$ (since Δ_1^1 is unitary), and $\|(X_1^\gamma)^{-1}\|\ge\gamma$. It follows that the inequality constraint in (3.14) is an equality. Part 4 of Lemma 3.6 now establishes that

$$\|(X_1^\gamma)^{-1}\|=\|\mathcal{F}_u(\Delta_1^\gamma,I)\|=\|(\gamma^2I-\Delta_{11})(I-\Delta_{11})^{-1}\|,$$

and this proves (3.12). ∎

Remark 3.8. The optimization on the right-hand side of (3.12) is an eigenvalue problem, as the following argument shows. Setting $\gamma^2=1+\zeta,\zeta>0$, shows that

$$\min_{\substack{\|(\gamma^2I-\Delta_{11})(I-\Delta_{11})^{-1}\|=a_{m+1}\\ \gamma>1}}\gamma=\sqrt{1+\min_{\substack{\|I+\zeta(I-\Delta_{11})^{-1}\|=a_{m+1}\\ \zeta>0}}\zeta}. \tag{3.15}$$

Note that

$$\begin{aligned}a_{m+1}&=\|I+\zeta(I-\Delta_{11})^{-1}\|\\ &\Leftrightarrow\underline{\lambda}\{a_{m+1}^2I-[I+\zeta(I-\Delta_{11})^{-1}][I+\zeta(I-\Delta_{11}')^{-1}]\}=0\\ &\Leftrightarrow\underline{\lambda}\left\{\zeta^{-1}I-\frac{(I-\Delta_{11})^{-1}+(I-\Delta_{11}')^{-1}}{a_{m+1}^2-1}-\zeta\frac{(I-\Delta_{11})^{-1}(I-\Delta_{11}')^{-1}}{a_{m+1}^2-1}\right\}=0\\ &\Leftrightarrow\underline{\lambda}\left(\begin{bmatrix}\zeta^{-1}I-\frac{(I-\Delta_{11})^{-1}+(I-\Delta_{11}')^{-1}}{a_{m+1}^2-1} & -\frac{(I-\Delta_{11})^{-1}}{\sqrt{a_{m+1}^2-1}}\\ -\frac{(I-\Delta_{11}')^{-1}}{\sqrt{a_{m+1}^2-1}} & \zeta^{-1}I\end{bmatrix}\right)=0,\end{aligned}$$

and since $\zeta > 0$, it follows that the minimum ζ in (3.15) is given by ζ_o, where

$$\zeta_o^{-1} = \bar{\lambda}\left(\begin{bmatrix} \frac{(I-\Delta_{11})^{-1}+(I-\Delta_{11}')^{-1}}{a_{m+1}^2-1} & -\frac{(I-\Delta_{11})^{-1}}{\sqrt{a_{m+1}^2-1}} \\ -\frac{(I-\Delta_{11}')^{-1}}{\sqrt{a_{m+1}^2-1}} & 0 \end{bmatrix}\right).$$

Remark 3.9. Lemmas 3.5 and 3.7 generalize the results in [13] by allowing multiplicities of the singular values of A.

Remark 3.10. Let γ_o be the maximum in (3.12) and define $\Phi_\Delta^{\gamma_o}$ as in (3.13). Then it is clear from part 1 of Lemma 3.6 that $\Phi_\Delta^{\gamma_o} = \mathcal{F}_u[(\Delta_0^{\gamma_o})^{-1}, \Delta]$ if and only if $\Delta = \mathcal{F}_l(\Delta_0^{\gamma_o}, \Phi_\Delta^{\gamma_o})$. It follows that all structured rank reducing perturbations for Lemma 3.7 can be obtained via a suitable lower linear fractional map from all structured perturbations for Lemma 3.5. The details, although straightforward, are tedious and are hence omitted.

Lemma 3.7 shows that the structured distance to singularity, when the perturbation Δ is constrained such that ${E_1}'\Delta E_1 = \Delta_{11}$ for a given $\Delta_{11} \in \mathcal{C}^{m\times m}$ such that $\|\Delta_{11}\| \le 1$ and $I - \Delta_{11}$ is nonsingular, reduces to a simple eigenvalue evaluation.

3.6. $\boldsymbol{\Delta_{11} = \{\Delta_{11} \in \mathcal{C}^{m\times m} : \|\Delta_{11}\| \le 1,\ (I-\Delta_{11})^{-1}}$ exists and is bounded}. The following result allows us to evaluate the structured distance to singularity for a general class of constraints on ${E_1}'\Delta E_1$.

LEMMA 3.11. *Let A and E_1 be as in Definition* 3.1 *and let $\boldsymbol{\Delta_{11}}$ be any compact subset of $\mathcal{C}^{m\times m}$ such that $\|\Delta_{11}\| \le 1$ and that $(I-\Delta_{11})^{-1}$ exists and is bounded for all $\Delta_{11} \in \boldsymbol{\Delta_{11}}$. Then*

$$\min\{\|\Delta\| : \det(A-\Delta) = 0,\ {E_1}'\Delta E_1 \in \boldsymbol{\Delta_{11}}\} = \sqrt{1+\zeta_o}\ , \tag{3.16}$$

where

$$\zeta_o = \min\{\zeta : \|I + \zeta(I-\Delta_{11})^{-1}\| = a_{m+1},\ \zeta > 0,\ \Delta_{11} \in \boldsymbol{\Delta_{11}}\}, \tag{3.17}$$

and ζ_o is (strictly) increasing in a_{m+1}. Furthermore if $\Delta_o \in \boldsymbol{\Delta_{11}}$ is a minimizer for (3.17)*, then Δ_o is a maximizer for*

$$\max_{\Delta_{11}\in\boldsymbol{\Delta_{11}}} \|I + \zeta_o(I-\Delta_{11})^{-1}\|. \tag{3.18}$$

Proof. The first part follows from Lemma 3.7 and the compactness of $\boldsymbol{\Delta_{11}}$. For any $\Delta_{11} \in \boldsymbol{\Delta_{11}}$, define $f_{\Delta_{11}}(\zeta) = \|I + \zeta(I-\Delta_{11})^{-1}\|$ for $\zeta \ge 0$. Then $f_{\Delta_{11}}(\zeta)$ is continuous in ζ, $f_{\Delta_{11}}(0) = 1$, and $f_{\Delta_{11}}(\zeta) \to \infty$ as $\zeta \to \infty$. Let Δ_o be a minimizer for (3.17) so that $f_{\Delta_o}(\zeta_o) = a_{m+1}$. Then $f_{\Delta_o}(\zeta) < a_{m+1}\ \forall\ \zeta < \zeta_o$. Suppose that Δ_o is not a maximizer for (3.18) so that $f_{\Delta_1}(\zeta_o) > f_{\Delta_o}(\zeta_o) = a_{m+1}$ for some $\Delta_1 \in \boldsymbol{\Delta_{11}}$. Then by continuity, there exists $0 < \zeta_1 < \zeta_o$ such that $f_{\Delta_1}(\zeta_1) = a_{m+1}$, contradicting the minimality of ζ_o. That ζ_o is increasing in a_{m+1} also follows from continuity. ∎

Remark 3.12. Lemma 3.7 transforms the structured distance to singularity on the left-hand side of (3.16) to the minimization in (3.17), which in turn is transformed using Lemma 3.11 into the minimization in (3.18). Note, however, that this last maximization involves ζ_o, the minimum of (3.17). Furthermore, the lemma shows only that the minimizers of (3.17) are a subset of the maximizers of (3.18). For the class of structured constraints we consider below, we show that the maximizers of (3.18) are (i) independent of ζ_o and (ii) are also minimizers for (3.17).

3.7. $\boldsymbol{\Delta_{11} = \Delta^m_{\phi_1,\phi_2} := \{\Delta_{11} \in \mathcal{C}^{m\times m} : \rho(\Delta_{11}) \le \phi_1, \|\Delta_{11}\| \le \phi_2\},\ 0 \le \phi_1 \le \phi_2 \le 1}$. Finally, we evaluate the structured distance to singularity for a subset of $\boldsymbol{\Delta_{11}}$ defined in subsection 3.6.

PROBLEM 3.13. *Let A and E_1 be as in Definition* 3.1. *Then for any $0 \le \phi_1 \le \phi_2 \le 1$, find the structured distance to singularity*

$$\gamma_{\boldsymbol{\Delta}^m_{\phi_1,\phi_2}} = \min\{\|\Delta\| : \det(A-\Delta) = 0,\ E_1{}'\Delta E_1 \in \boldsymbol{\Delta}^m_{\phi_1,\phi_2}\}, \tag{3.19}$$

where

$$\boldsymbol{\Delta}^m_{\phi_1,\phi_2} = \left\{ \Delta_{11} = \begin{bmatrix} \delta_{11} & \cdots & \delta_{1,m} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_{mm} \end{bmatrix} \in \mathcal{C}^{m\times m} : |\delta_{ii}| \le \phi_1\ \forall i; \delta_{ij} = 0\ \forall i > j; \|\Delta_{11}\| \le \phi_2 \right\}, \tag{3.20}$$

and the set of all structured rank reducing perturbations

$$\mathcal{D}_{\boldsymbol{\Delta}^m_{\phi_1,\phi_2}} := \left\{ \Delta \in \mathcal{C}^{n\times n} : \|\Delta\| = \gamma_{\boldsymbol{\Delta}^m_{\phi_1,\phi_2}},\ \det(A-\Delta) = 0,\ E_1{}'\Delta E_1 \in \boldsymbol{\Delta}^m_{\phi_1,\phi_2} \right\}. \tag{3.21}$$

Remark 3.14. The structured distance to singularity in (3.19) is unchanged if Δ is replaced with $W'\Delta W$, where $W = \operatorname{diag}(W_1, I_{n-m})$ and $W_1 \in \mathcal{C}^{m\times m}$ is any unitary matrix. So (3.20) is the upper triangular Schur form of the structured constraint set

$$\{\Delta_{11} \in \mathcal{C}^{m\times m} : \rho(\Delta_{11}) \le \phi_1, \|\Delta_{11}\| \le \phi_2\}.$$

We opted for (3.20) for convenience.

Remark 3.15. Note that, unlike the block diagonal uncertainty set associated with the structured singular value, the perturbation set $\boldsymbol{\Delta}^m_{\phi_1,\phi_2}$ is not a vector space. However, if $\gamma > 0$, then $\boldsymbol{\Delta}^m_{\gamma\phi_1,\gamma\phi_2} = \gamma\boldsymbol{\Delta}^m_{\phi_1,\phi_2}$. In the next section, we use this observation to relate the results of this section to the structured singular value.

It follows from Lemmas 3.7 and 3.11 that the evaluation of the distance to singularity requires the solution of the maximization problem in (3.18). Theorem 3.19 below proves that all maximizers for $\max_{\Delta_{11}\in\boldsymbol{\Delta}^m_{\phi_1,\phi_2}} \|I + \zeta(I-\Delta_{11})^{-1}\|$ are independent of $\zeta\ \forall\ \zeta > 0$, are all minimizers for (3.17), and are given by

$$\boldsymbol{\Delta}^m_{\phi_1,\phi_2,\mathrm{all}} = \{T'\Delta^m_{\phi_1,\phi_2}T : T \in \mathcal{C}^{m\times m}, T'T = I, T \text{ is diagonal}\}, \tag{3.22}$$

where $\Delta^m_{\phi_1,\phi_2} \in \boldsymbol{\Delta}^m_{\phi_1,\phi_2}$ is the upper triangular Toeplitz matrix defined by

$$\Delta^m_{\phi_1,\phi_2,ij} = \left\{ \begin{array}{cc} 0, & j < i \\ \phi_1, & j = i \\ (-\frac{\phi_1}{\phi_2})^{j-i-1}\frac{\phi_2^2-\phi_1^2}{\phi_2}, & j > i \end{array} \right\}, \tag{3.23}$$

e.g. for $m = 5$,

$$\Delta^m_{\phi_1,\phi_2} = \begin{bmatrix} \phi_1 & \frac{\phi_2^2-\phi_1^2}{\phi_2} & -\frac{\phi_1}{\phi_2}\frac{\phi_2^2-\phi_1^2}{\phi_2} & (\frac{\phi_1}{\phi_2})^2\frac{\phi_2^2-\phi_1^2}{\phi_2} & -(\frac{\phi_1}{\phi_2})^3\frac{\phi_2^2-\phi_1^2}{\phi_2} \\ 0 & \phi_1 & \frac{\phi_2^2-\phi_1^2}{\phi_2} & -\frac{\phi_1}{\phi_2}\frac{\phi_2^2-\phi_1^2}{\phi_2} & (\frac{\phi_1}{\phi_2})^2\frac{\phi_2^2-\phi_1^2}{\phi_2} \\ 0 & 0 & \phi_1 & \frac{\phi_2^2-\phi_1^2}{\phi_2} & -\frac{\phi_1}{\phi_2}\frac{\phi_2^2-\phi_1^2}{\phi_2} \\ 0 & 0 & 0 & \phi_1 & \frac{\phi_2^2-\phi_1^2}{\phi_2} \\ 0 & 0 & 0 & 0 & \phi_1 \end{bmatrix}.$$

We first derive some properties of $\Delta^m_{\phi_1,\phi_2}$. The constraints in (3.20) can be written as

$$|\lambda_i(\Delta_{11})| \le \phi_1, \quad \sigma_i(\Delta_{11}) \le \phi_2 \ \forall i.$$

All the eigenvalues of $\Delta^m_{\phi_1,\phi_2}$ are equal to ϕ_1 so that the spectral radius constraint in $\mathbf{\Delta}^m_{\phi_1,\phi_2}$ is maximally achieved by $\Delta^m_{\phi_1,\phi_2}$. The next result gives the singular values of $\Delta^m_{\phi_1,\phi_2}$ and gives an idea of how the singular value constraints are also maximally achieved.

LEMMA 3.16. *For $m>1$, partition $\Delta^m_{\phi_1,\phi_2}$ as*

$$\Delta^m_{\phi_1,\phi_2} = \begin{bmatrix} \Delta^{m-1}_{\phi_1,\phi_2} & x_m \\ 0 & \phi_1 \end{bmatrix}.$$

Then

$$x_m x_m' = \frac{\phi_2^2-\phi_1^2}{\phi_2^2}[\phi_2^2 I_{m-1} - (\Delta^{m-1}_{\phi_1,\phi_2})(\Delta^{m-1}_{\phi_1,\phi_2})']. \tag{3.24}$$

Hence,

$$\lambda_i(\Delta^m_{\phi_1,\phi_2})=\phi_1 \ \forall i; \quad \sigma_i(\Delta^m_{\phi_1,\phi_2})=\phi_2, i\le m-1; \quad \sigma_m(\Delta^m_{\phi_1,\phi_2})=\phi_1\left(\frac{\phi_1}{\phi_2}\right)^{m-1}. \tag{3.25}$$

Finally, all $\Delta_{11} \in \mathbf{\Delta}^m_{\phi_1,\phi_2}$ such that (3.25) *is satisfied are given by $\mathbf{\Delta}^m_{\phi_1,\phi_2,\mathrm{all}}$ in* (3.22).

Proof. Note first that

$$x_{m+1} = \begin{bmatrix} -\frac{\phi_1}{\phi_2}x_m \\ \frac{\phi_2^2-\phi_1^2}{\phi_2} \end{bmatrix}.$$

The result is clearly true for $m=2$. Assume the result for m and let $\alpha = \frac{\phi_2^2-\phi_1^2}{\phi_2^2}$. Then,

$$\alpha[\phi_2^2 I - (\Delta^m_{\phi_1,\phi_2})(\Delta^m_{\phi_1,\phi_2})'] = \alpha \begin{bmatrix} \phi_2^2 I_{m-1} - (\Delta^{m-1}_{\phi_1,\phi_2})(\Delta^{m-1}_{\phi_1,\phi_2})' - x_m x_m' & -x_m\phi_1 \\ -\phi_1 x_m' & \phi_2^2-\phi_1^2 \end{bmatrix}$$
$$= \begin{bmatrix} \frac{\phi_1^2}{\phi_2^2}x_m x_m' & -\frac{\phi_1}{\phi_2}\frac{\phi_2^2-\phi_1^2}{\phi_2}x_m \\ -\frac{\phi_1}{\phi_2}\frac{\phi_2^2-\phi_1^2}{\phi_2}x_m' & (\frac{\phi_2^2-\phi_1^2}{\phi_2})^2 \end{bmatrix} = x_{m+1}x_{m+1}'.$$

This proves the first result and establishes that $[\phi_2^2 I - (\Delta^m_{\phi_1,\phi_2})(\Delta^m_{\phi_1,\phi_2})']$ is a nonnegative rank 1 matrix. It follows that $m-1$ singular values of $\Delta^m_{\phi_1,\phi_2}$ are equal to ϕ_2. Since $\Delta^m_{\phi_1,\phi_2}$ has m eigenvalues equal to ϕ_1, it follows that the smallest singular value is $\phi_1(\frac{\phi_1}{\phi_2})^{m-1}$, which proves (3.25). To prove the last part, let

$$\Delta_m = \begin{bmatrix} \Delta_{m-1} & x_m \\ 0 & \phi_1 \end{bmatrix} \in \mathbf{\Delta}^m_{\phi_1,\phi_2}$$

satisfy (3.25). Then $\phi_2^2 I - \Delta_m\Delta_m' \ge 0$ and has rank 1. Thus

$$\phi_2^2 I - \Delta_m\Delta_m' = \begin{bmatrix} \phi_2^2 I_{m-1} - \Delta_{m-1}\Delta_{m-1}' - x_m x_m' & -\phi_1 x_m \\ -\phi_1 x_m' & \phi_2^2-\phi_1^2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\begin{bmatrix} x_1' & \bar{x}_2 \end{bmatrix},$$

for some $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathcal{C}^{m\times 1}$. Solving for x_1 and x_2 gives

$$x_m x_m' = \frac{\phi_2^2-\phi_1^2}{\phi_2^2}(\phi_2^2 I_{m-1} - \Delta_{m-1}\Delta_{m-1}'), \tag{3.26}$$

so that (3.25) is satisfied for Δ_{m-1} with m replaced by $m-1$. Repeating the procedure with Δ_m replaced by Δ_{m-1} shows that (3.26) is satisfied. Since $\Delta^m_{\phi_1,\phi_2}$ satisfies (3.24), it follows that the only nonuniqueness is in the Cholesky factorizations in (3.26). It is a simple exercise to verify that this is covered by (3.22). ∎

Before we state our main results, we need the following two technical lemmas.

LEMMA 3.17. *Suppose that* $L = L_1 - L_2$ *with* $L_1 = L_1' > 0$, $L_2 = L_2' \geq 0$, $\operatorname{rank}(L_2) = 1$, *and* $L = L' \geq 0$. *Let* $Y = \beta L_2 - L_3$ *with* $\beta > 0$, $L_3 = L_3' \geq 0$, $\operatorname{rank}(L_3) = 1$, *and* $Y = Y' \geq 0$. *Then for any* $\epsilon > 0$,

$$\underline{\lambda}(L + \epsilon Y) = \underline{\lambda}(L) = 0 \Rightarrow Y = 0.$$

Proof. First write L_2 and L_3 in the forms $L_2 = l_2 l_2'$ and $L_3 = l_3 l_3'$, where l_2 and l_3 are column vectors, and decompose l_3 in directors parallel and orthogonal to l_2, i.e.,

$$l_3 = \frac{l_2' l_3}{\|l_2\|^2} l_2 + p, \quad p := l_3 - \frac{l_2' l_3}{\|l_2\|^2} l_2.$$

Clearly $l_2' p = 0$. Thus, $Y = \beta l_2 l_2' - l_3 l_3' \geq 0 \Rightarrow \beta p' l_2 l_2' p = 0 \geq p' l_3 l_3' p \Rightarrow l_3' p = 0$. Hence,

$$l_3' \left(l_3 - \frac{l_2' l_3}{\|l_2\|^2} l_2 \right) = 0 \quad \Rightarrow |l_2' l_3| = \|l_2\| \|l_3\|,$$

and hence $l_3 = \lambda l_2$ for some $\lambda \in \mathcal{C}$. Using again $Y \geq 0$ shows that $|\lambda| \leq \sqrt{\beta}$, and so

$$Y = \beta \left(1 - \frac{|\lambda|^2}{\beta} \right) L_2 := \alpha\beta L_2,$$

for some $0 \leq \alpha \leq 1$. Suppose that $\alpha > 0$, so that $Y \neq 0$. Pick a nonzero vector y such that $y'(L + \epsilon Y)y = 0$. Then

$$y'(L + \epsilon Y)y = 0 \Rightarrow y' L y = 0, y' Y y = 0 \Rightarrow y' L_2 y = 0 \Rightarrow y' L_1 y = 0,$$

contradicting the positive definiteness of L_1. ∎

LEMMA 3.18. *Let*

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} \\ 0 & \delta_{22} \end{bmatrix} \in \mathcal{C}^{2\times 2}.$$

There exists unitary W *such that*

$$W' \Delta W = \begin{bmatrix} \delta_{22} & \delta_{12} \\ 0 & \delta_{11} \end{bmatrix}.$$

Proof. If $\delta_{11} = \delta_{22}$, take $W = I_2$ and if $\delta_{12} = 0$, take W to be a permutation matrix. Otherwise take

$$W = \frac{1}{\sqrt{|\delta_{22} - \delta_{11}|^2 + |\delta_{12}|^2}} \begin{bmatrix} |\delta_{12}| & -(\bar{\delta}_{22} - \bar{\delta}_{11})|\delta_{12}|/\bar{\delta}_{12} \\ (\delta_{22} - \delta_{11})|\delta_{12}|/\delta_{12} & |\delta_{12}| \end{bmatrix}. \qquad ∎$$

The following is our main result for this section.

THEOREM 3.19. *Let $0 \le \phi_1 \le \phi_2 \le 1$, with $\phi_1 < 1$. Consider the optimization*

$$\max \{\|I + \zeta(I - \Delta_{11})^{-1}\| : \Delta_{11} \in \mathbf{\Delta}^m_{\phi_1,\phi_2}\}. \tag{3.27}$$

Then $\Delta^m_{\phi_1,\phi_2}$ is a maximizer for all $\zeta \ge 0$, where $\mathbf{\Delta}^m_{\phi_1,\phi_2}$ is defined in (3.20) *and $\Delta^m_{\phi_1,\phi_2}$ in* (3.23). *Furthermore, for all $\zeta > 0$, the set of all maximizers is given by $\mathbf{\Delta}^m_{\phi_1,\phi_2,\mathrm{all}}$ defined in* (3.22). *Finally, $\mathbf{\Delta}^m_{\phi_1,\phi_2,\mathrm{all}}$ is the set of all minimizers for* (3.17).

Proof. Let $\Delta^m_{11} \in \mathbf{\Delta}^m_{\phi_1,\phi_2}$ be partitioned as

$$\Delta^m_{11} = \begin{bmatrix} \Delta^{m-1}_{11} & x \\ 0 & \delta \end{bmatrix},$$

and define $X = I - \Delta^{m-1}_{11}$ and $F = I + \zeta X^{-1}$. Using a Schur complement argument,

$$\begin{aligned}
\|\Delta^m_{11}\| \le \phi_2 &\Leftrightarrow \phi_2^2 I - \Delta^m_{11}(\Delta^m_{11})' \ge 0 \\
&\Leftrightarrow \begin{bmatrix} \phi_2^2 I - \Delta^{m-1}_{11}(\Delta^{m-1}_{11})' - xx' & -\bar{\delta}x \\ -\delta x' & \phi_2^2 - |\delta|^2 \end{bmatrix} \ge 0 \\
&\Leftrightarrow \phi_2^2 I - \Delta^{m-1}_{11}(\Delta^{m-1}_{11})' - xx' - \frac{|\delta|^2}{\phi_2^2 - |\delta|^2} xx' \ge 0 \\
&\Leftrightarrow xx' \le \frac{\phi_2^2 - |\delta|^2}{\phi_2^2}(\phi_2^2 I - \Delta^{m-1}_{11}(\Delta^{m-1}_{11})') \\
&\Leftrightarrow X^{-1}xx'X^{-\prime} \le \frac{\phi_2^2 - |\delta|^2}{\phi_2^2}[X^{-1} + X^{-\prime} - I - (1 - \phi_2^2)X^{-1}X^{-\prime}] \\
&\Leftrightarrow X^{-1}xx'X^{-\prime} = \frac{\phi_2^2 - |\delta|^2}{\phi_2^2}[X^{-1} + X^{-\prime} - I - (1 - \phi_2^2)X^{-1}X^{-\prime}] - Y,
\end{aligned}$$

for some Y such that

$$0 \le Y = Y' \le \frac{\phi_2^2 - |\delta|^2}{\phi_2^2}[X^{-1} + X^{-\prime} - I - (1 - \phi_2^2)X^{-1}X^{-\prime}]. \tag{3.28}$$

Now,

$$\begin{aligned}
\gamma &= \|I + \zeta(I - \Delta^m_{11})^{-1}\| \\
&\Leftrightarrow \underline{\lambda}\left(\begin{bmatrix} \gamma^2 I - FF' - \frac{\zeta^2}{|1-\delta|^2}X^{-1}xx'X^{-\prime} & -\frac{\zeta}{1-\delta}(1 + \frac{\zeta}{1-\delta})X^{-1}x \\ -\frac{\zeta}{1-\delta}(1 + \frac{\zeta}{1-\delta})x'X^{-\prime} & \gamma^2 - |1 + \frac{\zeta}{1-\delta}|^2 \end{bmatrix}\right) = 0 \\
&\Leftrightarrow \underline{\lambda}\left[\gamma^2 I - FF' - \frac{\zeta^2\gamma^2}{|1-\delta|^2(\gamma^2 - |1 + \frac{\zeta}{1-\delta}|^2)}X^{-1}xx'X^{-\prime}\right] = 0 \\
&\Leftrightarrow \underline{\lambda}\left[L(\gamma^2) + \frac{\zeta^2\gamma^2}{|1-\delta|^2}Y\right] = 0,
\end{aligned}$$

where we have set

$$L(\gamma^2) = \left(\gamma^2 - |1 + \frac{\zeta}{1-\delta}|^2\right)(\gamma^2 I - FF') - \frac{\zeta^2\gamma^2}{|1-\delta|^2}\frac{\phi_2^2 - |\delta|^2}{\phi_2^2}[X^{-1} + X^{-\prime} - I - (1 - \phi_2^2)X^{-1}X^{-\prime}],$$

and where we have assumed that $\gamma > \max\{|1 + \frac{\zeta}{1-\delta}|^2, \|(I + \zeta X^{-1})\|\}$ since the right-hand side is clearly not the maximum γ. Thus (3.27) can be written as

$$\max\left\{\gamma : \underline{\lambda}\left[L(\gamma^2) + \frac{\zeta^2\gamma^2}{|1-\delta|^2}Y\right] = 0, |\delta| \le \phi_1, \Delta^{m-1}_{11} \in \mathbf{\Delta}^{m-1}_{\phi_1,\phi_2}, Y \text{ satisfies (3.28)}\right\}. \tag{3.29}$$

Next, we maximize (3.29) with respect to Y; that is, for given $\Delta_{11}^{m-1} \in \mathbf{\Delta}_{\phi_1,\phi_2}^{m-1}$ and $\delta \in \mathcal{C}$ such that $|\delta| \le \phi_1$, we solve

$$\max\left\{\gamma : \underline{\lambda}\left[L(\gamma^2) + \frac{\zeta^2\gamma^2}{|1-\delta|^2}Y\right] = 0, \quad Y \text{ satisfies (3.28)}\right\}. \tag{3.30}$$

Since $Y = Y' \ge 0$, $L(0) > 0$, and $L(\gamma^2) \sim \gamma^4 I > 0$ as $\gamma^2 \to \infty$, it follows that γ is maximized by $Y = 0$. Note that

$$\begin{aligned} Y = 0 &\Leftrightarrow xx' = \frac{\phi_2^2 - |\delta|^2}{\phi_2^2}(\phi_2^2 I - \Delta_{11}^{m-1}(\Delta_{11}^{m-1})') \\ &\Rightarrow \operatorname{rank}(\phi_2^2 I - \Delta_{11}^{m}(\Delta_{11}^{m})') = 1 \\ &\Leftrightarrow \sigma_i(\Delta_{11}^{m}) = \phi_2; i \le m-1 \end{aligned}$$

and is therefore feasible since Δ_{ϕ_1,ϕ_2}^m satisfies this property. That $Y = 0$ is the only maximizer for (3.30) follows from Lemma 3.17 with

$$\begin{aligned} &L_1 = (\gamma_o^2 - |1 + \tfrac{\zeta}{1-\delta}|^2)(\gamma_o^2 I - FF'), && \beta = \tfrac{|1-\delta|^2}{\zeta^2\gamma_o^2}, && \epsilon = \tfrac{\zeta^2\gamma_o^2}{|1-\delta|^2}, \\ &L_2 = \tfrac{\zeta^2\gamma_o^2}{|1-\delta|^2}\tfrac{\phi_2^2 - |\delta|^2}{\phi_2^2}[X^{-1} + X^{-\prime} - I - (1-\phi_2^2)X^{-1}X^{-\prime}], && L_3 = X^{-1}xx'X^{-\prime}, \end{aligned}$$

where γ_o is the maximum in (3.30). It follows that (3.29) can be written as

$$\max\left\{\gamma : \underline{\lambda}[L(\gamma^2)] = 0, |\delta| \le \phi_1, xx' = \frac{\phi_2^2 - |\delta|^2}{\phi_2^2}(\phi_2^2 I - \Delta_{11}^{m-1}(\Delta_{11}^{m-1})'), \Delta_{11}^{m-1} \in \mathbf{\Delta}_{\phi_1,\phi_2}^{m-1}\right\}.$$

Next, we maximize γ with respect to δ. Now,

$$\begin{aligned} \underline{\lambda}[L(\gamma^2)] = 0 &\Leftrightarrow \underline{\lambda}\left[\gamma^2 I - FF' - \frac{\zeta^2\gamma^2(\phi_2^2 - |\delta|^2)/\phi_2^2}{|1-\delta|^2(\gamma^2 - |1 + \frac{\zeta}{1-\delta}|^2)}BB'\right] = 0 \\ &\Leftrightarrow \gamma^2 - \left|1 + \frac{\zeta}{1-\delta}\right|^2 - \frac{\zeta^2\gamma^2(\phi_2^2 - |\delta|^2)}{\phi_2^2|1-\delta|^2}B'(\gamma^2 I - FF')^{-1}B = 0, \end{aligned}$$

where $BB' := X^{-1} + X^{-\prime} - I - (1-\phi_2^2)X^{-1}X^{-\prime}$. Setting $\delta = r\phi_1 e^{j\theta}$, with $0 \le r \le 1$ and $-\pi \le \theta \le \pi$, a simple calculation gives

$$\frac{\partial\gamma^2}{\partial\theta} = \frac{2r\phi_1(1 + \zeta - \phi_2^2\gamma^2)\sin\theta}{\phi_2^2(1 - 2r\phi_1\cos\theta + r^2\phi_1^2) + \zeta^2(\phi_2^2 - r^2\phi_1^2)B'(\gamma^2 I - FF')^{-1}FF'(\gamma^2 I - FF')^{-1}B},$$

which vanishes when δ is real. Setting δ real,

$$\frac{\partial\gamma^2}{\partial\delta} = \frac{2\zeta}{1-\delta}\,\frac{1 + \zeta - \delta + \gamma^2\zeta(\phi_2^2 - \delta)B'(\gamma^2 I - FF')^{-1}B}{\phi_2^2(1-\delta)^2 + \zeta^2(\phi_2^2 - \delta^2)B'(\gamma^2 I - FF')^{-1}FF'(\gamma^2 I - FF')^{-1}B},$$

which is positive for all δ, so γ is maximized by $\delta = \phi_1$. It follows from Lemma 3.18 that all diagonal entries of a maximizing Δ_{11}^m are equal to ϕ_1 since, if not, we can use a sequence of unitary similarity transformations that preserve the upper triangular structure of Δ_{11}^m to bring this diagonal entry to the (m, m) entry, contradicting the maximality of Δ_{11}^m. It follows from the last part of Lemma 3.16 that Δ_{ϕ_1,ϕ_2}^m is a maximizer and all maximizers are given by $\mathbf{\Delta}_{\phi_1,\phi_2,\text{all}}^m$. Note that the maximizers for

(3.27) are independent of ζ. Finally, Lemma 3.11 shows that all minimizers for (3.17) are a subset of all maximizers for (3.27). However, it is clear that the cost function in (3.17) is the same for all $\Delta_{11}^m \in \mathbf{\Delta}^m_{\phi_1,\phi_2,\text{all}}$. This proves the last part. ∎

The solution of Problem 3.13 is now summarized in the following result.

COROLLARY 3.20. *Let all variables be as in Problem* 3.13. *Then*

$$\begin{aligned}\gamma^2_{\mathbf{\Delta}^m_{\phi_1,\phi_2}} &:= \min_{\substack{\det(A-\Delta)=0\\ E_1{}'\Delta E_1 \in \mathbf{\Delta}^m_{\phi_1,\phi_2}}} \|\Delta\|^2 = \min_{\substack{\|(\gamma^2 I-\Delta_{11})(I-\Delta_{11})^{-1}\|=a_{m+1}\\ \gamma>1\\ \Delta_{11}\in\mathbf{\Delta}^m_{\phi_1,\phi_2}}} \gamma^2\\ &= \min_{\substack{\|(\gamma^2 I-\Delta^m_{\phi_1,\phi_2})(I-\Delta^m_{\phi_1,\phi_2})^{-1}\|=a_{m+1}\\ \gamma>1}} \gamma^2 = 1 + \min_{\substack{\|I+\zeta(I-\Delta_{11})^{-1}\|=a_{m+1}\\ \zeta>0\\ \Delta_{11}\in\mathbf{\Delta}^m_{\phi_1,\phi_2}}} \zeta\\ &= 1 + \min_{\substack{\|I+\zeta(I-\Delta^m_{\phi_1,\phi_2})^{-1}\|=a_{m+1}\\ \zeta>0}} \zeta = 1+\zeta_{\mathbf{\Delta}^m_{\phi_1,\phi_2}},\end{aligned}$$

where

$$\zeta^{-1}_{\mathbf{\Delta}^m_{\phi_1,\phi_2}} = \bar{\lambda}\left(\begin{bmatrix} \frac{(I-\Delta^m_{\phi_1,\phi_2})^{-1}+(I-\Delta^m_{\phi_1,\phi_2}{}')^{-1}}{a^2_{m+1}-1} & -\frac{(I-\Delta^m_{\phi_1,\phi_2})^{-1}}{\sqrt{a^2_{m+1}-1}}\\ -\frac{(I-\Delta^m_{\phi_1,\phi_2}{}')^{-1}}{\sqrt{a^2_{m+1}-1}} & 0\end{bmatrix}\right).$$

Remark 3.21. To characterize the set of all optimal structured perturbations in (3.21) we proceed as follows. Theorem 3.19 gives $\mathbf{\Delta}^m_{\phi_1,\phi_2,\text{all}}$ as the set of minimizers for (3.17). Remark 3.14 shows that we can absorb the unitary diagonal T in (3.22) in a unitary W_1. This gives all leading $m\times m$ blocks of (3.21) (it can be shown that if $\phi_1=\phi_2$, then this gives only a class of solutions). For the other blocks, we appeal to Remark 3.10 with Δ_{11} replaced by $\Delta^m_{\phi_1,\phi_2}$. We spare the readers the full details.

The following result gives an interpretation of $\Delta^m_{\phi_1,\phi_2}$. Since the solution of Theorem 3.19 is independent of ζ, letting $\zeta\to\infty$ suggests the following.

THEOREM 3.22. *Let* $0\le\phi_1\le\phi_2\le 1$ *be given. Consider the minimization*

$$\min\{\underline{\sigma}(I-\Delta_{11}) : \Delta_{11}\in\mathbf{\Delta}^m_{\phi_1,\phi_2}\} = \left(\max\{\|(I-\Delta_{11})^{-1}\| : \Delta\in\mathbf{\Delta}^m_{\phi_1,\phi_2}\}\right)^{-1}.$$

Then $\Delta_{11}=\Delta^m_{\phi_1,\phi_2}$ *is a minimizer and the set of all minimizers is given by* $\mathbf{\Delta}^m_{\phi_1,\phi_2,\text{all}}$.

Proof. The proof is almost identical to that of Theorem 3.19 and is omitted. ∎

Remark 3.23. Thus an interpretation of $\Delta^m_{\phi_1,\phi_2}$ is that it is the "most nearly rank reducing matrix" within $\mathbf{\Delta}^m_{\phi_1,\phi_2}$ to the identity matrix. This generalizes the unstructured as well as the structured case for $m=1$.

4. Upper bound on the structured singular value. In this section we use our previous results to derive an upper bound on the structured singular value.

THEOREM 4.1. *Let* $M\in\mathcal{C}^{n\times n}$ *be as defined in section* 2 *and let* $\mathbf{\Delta}\subseteq\mathcal{C}^{n\times n}$ *denote a block diagonal structured uncertainty set. Let* $\bar{\phi}_1$ *and* $\bar{\phi}_2$ *be upper bounds on the maximizations in* (2.5) *and* (2.6), *respectively, and assume that* $0\le\bar{\phi}_1\le\bar{\phi}_2\le 1$. *Then a lower bound on the structured stability radius* $r_{\mathbf{\Delta}}(M)$, *defined as in* (2.2), *is given by*

$$\begin{aligned}\underline{r}_{\mathbf{\Delta}}(M) &:= \min\{\|\Delta\| : \det(I-\Delta\Sigma)=0, \rho(E_1{}'\Delta E_1)\le\bar{\phi}_1\|\Delta\|, \|E_1{}'\Delta E_1\|\le\bar{\phi}_2\}\\ &= \min\{\gamma : \|(\gamma I-\Delta^m_{\bar{\phi}_1,\bar{\phi}_2})(\gamma^{-1}I-\Delta^m_{\bar{\phi}_1,\bar{\phi}_2})^{-1}\| = \sigma^{-1}_{m+1}, \gamma\ge 1\}\ge 1 \qquad (4.1)\end{aligned}$$

and is decreasing in σ_{m+1}, *where* $\Delta^m_{\bar{\phi}_1,\bar{\phi}_2}$ *is defined in* (3.23). *Furthermore,*

$$\bar{\phi}_1 < 1 \Rightarrow r_{\mathbf{\Delta}}(M) > 1 \quad \textit{and} \quad \underline{r}_{\mathbf{\Delta}}(M) > 1.$$

Finally, $r_{\mathbf{\Delta}}(M) = \underline{r}_{\mathbf{\Delta}}(M)$ *if and only if there exists* $\Delta \in \mathbf{\Delta}$ *such that*

$$V_1'\Delta U_1 = [\underline{r}_{\mathbf{\Delta}}(M)]^{-1} W' \Delta^m_{\bar{\phi}_1,\bar{\phi}_2} W,$$

for some unitary W.

Proof. The first part follows from (2.9), Remark 3.15, and Corollary 3.20. First note that

$$\begin{aligned}
\gamma_{\mathbf{\Delta}^m_{\bar{\phi}_1,\bar{\phi}_2}} &= \min\{\|\Delta\| : \det(A-\Delta) = 0, \rho(E_1{}'\Delta E_1) \le \bar{\phi}_1, \|E_1'\Delta E_1\| \le \bar{\phi}_2\} \\
&= \min\{\gamma : \|(\gamma^2 I - \Delta^m_{\bar{\phi}_1,\bar{\phi}_2})(I - \Delta^m_{\bar{\phi}_1,\bar{\phi}_2})^{-1}\| = a_{m+1}, \gamma > 1\} \\
&= \min\{\gamma : \|(\gamma I - \gamma^{-1}\Delta^m_{\bar{\phi}_1,\bar{\phi}_2})(\gamma^{-1} I - \gamma^{-1}\Delta^m_{\bar{\phi}_1,\bar{\phi}_2})^{-1}\| = a_{m+1}, \gamma > 1\} \\
&= \min\{\gamma : \|(\gamma I - \Delta_{\gamma^{-1}\bar{\phi}_1,\gamma^{-1}\bar{\phi}_2})(\gamma^{-1} I - \Delta_{\gamma^{-1}\bar{\phi}_1,\gamma^{-1}\bar{\phi}_2})^{-1}\| = a_{m+1}, \gamma > 1\}. \qquad (4.2)
\end{aligned}$$

To turn the "absolute" bounds $\rho(E_1'\Delta E_1) \le \bar{\phi}_1$ and $\|E_1'\Delta E_1\| \le \bar{\phi}_2$ into the "relative" bounds $\rho(E_1'\Delta E_1) \le \bar{\phi}_1\|\Delta\|$ and $\|E_1'\Delta E_1\| \le \bar{\phi}_2\|\Delta\|$, we simply write $\rho(E_1'\Delta E_1) \le \frac{\bar{\phi}_1}{\gamma}\gamma$ and $\|E_1'\Delta E_1\| \le \frac{\bar{\phi}_2}{\gamma}\gamma$; i.e., we re-define $\gamma^{-1}\bar{\phi}_1 \to \bar{\phi}_1$ and $\gamma^{-1}\bar{\phi}_2 \to \bar{\phi}_2$ in (4.2), which gives immediately the required expression for $\underline{r}_{\mathbf{\Delta}}(M)$. The fact that $\underline{r}_{\mathbf{\Delta}}(M)$ is (strictly) decreasing in σ_{m+1} follows from Lemma 3.11 since $a_{m+1} = \sigma_{m+1}^{-1}$.The fact that $r_{\mathbf{\Delta}}(M) > 1$ if $\bar{\phi}_1 < 1$ follows from Corollary 3.20. Finally, consider the two minimization problems in (2.9). The set of all minimizers in the right-hand side of (2.9) is given by all $\Delta \in \mathcal{C}^{n\times n}$ such that $\det(A-\Delta) = 0$, $\rho(E_1'\Delta E_1) \le \bar{\phi}_1 \underline{r}_{\mathbf{\Delta}}(M)$, and $\|E_1'\Delta E_1\| \le \bar{\phi}_2 \underline{r}_{\mathbf{\Delta}}(M)$, or equivalently,

$$E_1'\Delta E_1 = W \Delta^m_{\bar{\phi}_1 \underline{r}_{\mathbf{\Delta}}^{-1}(M), \bar{\phi}_2 \underline{r}_{\mathbf{\Delta}}^{-1}(M)} W' = \underline{r}_{\mathbf{\Delta}}^{-1}(M) W \Delta^m_{\bar{\phi}_1,\bar{\phi}_2} W',$$

where W is an arbitrary unitary matrix (see Remarks 3.14, 3.15, and 3.21). Now equality is possible in (2.9) if and only if there exists a minimizer of the left-hand side of (2.9), which is also a minimizer of the right-hand side of (2.9), i.e., if and only if there exists a $\Delta \in \mathbf{\Delta}$ such that $V_1'\Delta U_1 = [\underline{r}_{\mathbf{\Delta}}(M)]^{-1} W' \Delta^m_{\bar{\phi}_1,\bar{\phi}_2} W$. ∎

Remark 4.2. The evaluation of $\underline{r}_{\mathbf{\Delta}}(M)$ in (4.1) is a simple eigenvalue problem of dimension $4m \times 4m$, as the following development shows. Let $a = \sigma_{m+1}^{-1}$ and $\Psi = \Delta^m_{\bar{\phi}_1,\bar{\phi}_2}$. Then

$$\begin{aligned}
a &= \|(\gamma I - \Psi)(\gamma^{-1} I - \Psi)^{-1}\| \\
&\Rightarrow \det\{a^2 I - (\gamma^{-1} I - \Psi')^{-1}(\gamma I - \Psi')(\gamma I - \Psi)(\gamma^{-1} I - \Psi)^{-1}\} = 0 \\
&\Rightarrow \det\{\gamma^4 I - \gamma^3(\Psi + \Psi') - \gamma^2(a^2-1)\Psi'\Psi + \gamma a^2(\Psi + \Psi') - a^2 I\} = 0 \\
&\Rightarrow \det\left(\gamma I - \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ a^2 I & -a^2(\Psi+\Psi') & (a^2-1)\Psi'\Psi & \Psi + \Psi' \end{bmatrix}\right) = 0. \qquad (4.3)
\end{aligned}$$

It follows that γ is the smallest real eigenvalue larger than 1 of the $4m \times 4m$ block matrix in (4.3).

Remark 4.3. The work shows that if we can get an upper bound on $\mu_{\mathbf{\Delta}}(M_0)$, where M_0 is defined in (2.7) and is a reduced rank matrix (with the rank equal to the multiplicity of the largest singular value of M), then we can easily obtain a bound on M. This suggests that μ is "hard" for unitary M.

Remark 4.4. The theorem is a generalization of the result in [14] which solved the case $m = 1$. That result was limited since, at the end of the D-iteration, the largest singular value has multiplicity larger than one (otherwise it is equal to $\mu_{\mathbf{\Delta}}(M)$) [21].

In the final part of this section we outline the main steps of an algorithm which may be used to breach the D-iteration upper bound of M, provided a reduced rank μ problem can be solved.

ALGORITHM 4.5. *Given $M \in \mathcal{C}^{n\times n}$ and a complex block diagonal uncertainty structure $\mathbf{\Delta}$, the following algorithm obtains an upper bound of $\mu_{\mathbf{\Delta}}(M)$, $\bar{\mu}_{\mathbf{\Delta}}(M)$:*

1. *(Optional.) Solve the LMI optimization* (2.3) *to obtain the D-iteration upper bound of M. Normalize M by absorbing the D-iteration scalings obtained from the optimization.*
2. *If step* 1 *has been carried out and the multiplicity of the largest singular value of M is one, then $\mu_{\mathbf{\Delta}}(M) = \bar{\mu}_{\mathbf{\Delta}}(M) = \|M\|$ and the algorithm terminates.*
3. *Normalize $\|M\| = 1$ by dividing by its norm. Carry out the singular value decomposition of M given in* (2.1) *and let m be the multiplicity of the largest singular value of M (which is* 1*).*
4. *If step* 1 *(optional) has been carried out, set $\bar{\phi}_2 = 1$; else solve the LMI optimization* (2.3) *and set $\bar{\phi}_2 = \sqrt{\gamma}$.*
5. *From the singular value decomposition in step* 3*, define the (reduced rank) matrix M_0 as in Remark* 2.2*. Calculate the tightest possible upper bound of $\mu_{\mathbf{\Delta}}(M_0)$ and set it equal to $\bar{\phi}_1$.*
6. *If $\bar{\phi}_1 = 1$, set $\bar{\mu}_{\mathbf{\Delta}}(M) = \|M\|$ (after reversing the scaling in step* 3 *and/or* 1*) and exit. No tighter bound can be obtained.*
7. *Set $\Psi = \Delta^m_{\bar{\phi}_1,\bar{\phi}_2}$ defined in* (3.23) *and $a = \sigma^{-1}_{m+1}$, where σ_{m+1} is the $(m+1)$st largest singular value of M.*
8. *Solve the $4m \times 4m$ eigenvalue problem given in* (4.3) *and set $\underline{r}_{\mathbf{\Delta}}(M)$ as the smallest real eigenvalue larger than one. Then $\bar{\mu}_{\mathbf{\Delta}}(M) = \underline{r}^{-1}_{\mathbf{\Delta}}(M)$ (after reversing the scaling in step* 3 *and/or* 1*).*

5. Example. Consider a modified example from [21] with uncertainty structure $m_c = 0$ and $m_C = 4$. Let a matrix $M \in \mathcal{C}^{4\times 4}$ after D-iteration have the form

$$M = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} I_2 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix}, \quad U_1 = \begin{bmatrix} a & 0 \\ b & b \\ c & jc \\ d & f \end{bmatrix}, \quad V_1 = \begin{bmatrix} 0 & a \\ b & -b \\ c & -jc \\ e^{j\psi_1} f & e^{j\psi_2} d \end{bmatrix},$$

where $a = \sqrt{2/k}$, $b = c = 1/\sqrt{k}$, $d = -\sqrt{\beta/k}$, $f = (1+j)\sqrt{1/k\beta}$, $\psi_1 = -\pi/2$, and $\psi_2 = \pi$, with $k = 3 + \sqrt{3}$ and $\beta = \sqrt{3} - 1$. Also $\|\Sigma_2\| = \sigma_2$. Here, U_2 and V_2 are any unitary completions of U_1 and V_1, respectively. Let $M_0 = U_1 V_1'$. Then $\sigma_1(M_0) = 1$.

It is shown in [21] that $\mu_{\mathbf{\Delta}}(M_0) < 0.895 = \bar{\phi}_1$. (In fact, a rather involved calculation shows that

$$\mu_{\mathbf{\Delta}}(M_0) = \sqrt{\frac{1 + 2\sqrt{16 - 9\sqrt{3}}}{3}} = 0.872359\ldots,$$

although we use the value given in [21]). Figure 5.1 shows how the upper bound on $\mu_{\mathbf{\Delta}}(M) = \frac{1}{\underline{r}_{\mathbf{\Delta}}(M)}$ (calculated by finding the smallest root larger than 1 of the quartic in (4.3)) varies with σ_2. Note that the D-iteration upper bound is equal to 1 $\forall$ σ_2 such that $0 \le \sigma_2 \le 1$.

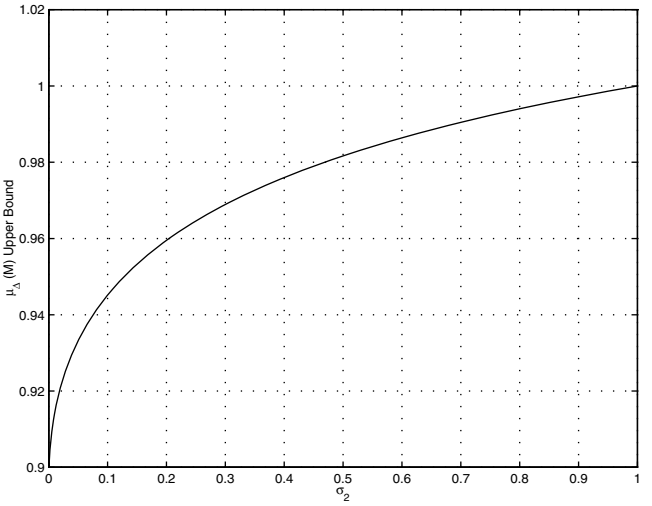


FIG. 5.1. *The upper bound on μ as a function of σ_2 for $0 \leq \sigma_2 < 1$, where $\sigma_1 = 1$ and $\bar{\phi}_1 = 0.895$ for matrix M.*

6. Summary. In this section we summarize our results. All variables are as defined in Theorem 4.1. Our main results are as follows:

- We have shown that $\mu_{\mathbf{\Delta}}(M) = 1$ if and only if $\mu_{\mathbf{\Delta}}(M_0) = 1$. It follows that determining whether there exists a gap between μ and its upper bound is equivalent to solving a μ problem of rank equal to the multiplicity of the largest singular value of M.
- We have shown that if we can obtain an upper bound on $\mu_{\mathbf{\Delta}}(M_0)$ less than 1, then we can give an upper bound on $\mu_{\mathbf{\Delta}}(M)$ which is smaller than 1.

Other secondary results are as follows:

- We used a semidefinite relaxation of the nonconvex optimization in (2.6) to derive an upper bound which admits no gap (Lemma 2.3).
- We generalized the distance to singularity results of [16] to the matrix case (Corollary 3.20).
- We solved a distance problem (Theorem 3.22).

7. Conclusions. The approach in [16] for tackling structured uncertainty was proposed around the time μ was proposed and seems to have been neglected. We have established a connection between these two approaches and we hope interest in the first approach will revive.

Hitherto, the interest in low rank μ problems has been limited to some properties concerning the gap between μ and its upper D-iteration bound, and to other theoretical considerations [20, 24, 25, 26]. We have shown that the solution of a reduced rank μ problem, which has limited direct relevance to realistic structured uncertainty descriptions, will induce an upper bound on the more relevant full rank problem. It is not clear at present whether the reduced rank μ problems constructed here have any special structure that makes them hard to compute. We hope that our work will spur more research on the reduced rank problem to help answer this question.

The approach followed here shares many characteristics with our previous work on the QIP problem [18] and potentially can be extended to a wide class of optimization problems utilizing convex relaxation techniques.

Appendix. Proof of Lemma 2.3. Let all variables be as defined in section 2. In order to fix notation, assume that

$$\mathbf{\Delta} = \{\Delta \in \mathcal{C}^{n\times n} \ : \ \Delta = \mathrm{diag}(\delta_1 I_{r_1}, \ldots, \delta_s I_{r_s}, \Delta_1, \ldots, \Delta_F), \delta_i \in \mathcal{C}, \Delta_j \in \mathcal{C}^{m_j \times m_j}\},$$

where $i \in \underline{S}$, $j \in \underline{F}$, and $n = \sum_{i=1}^{n} r_i + \sum_{i=1}^{F} m_i$. It follows that

$$\mathcal{D} = \{D = \mathrm{diag}(D_1, \ldots, D_s, d_1 I_{m_1}, \ldots, d_F I_{m_F}),\ D_i = D_i' \geq 0,\ d_j \geq 0\},$$

where $i \in \underline{S}$ and $j \in \underline{F}$. Define also the sets $\mathcal{Q} = \{\Delta \in \mathcal{B}\mathbf{\Delta} \ : \ \Delta\Delta' = I_n\}$, $\mathcal{D}_+ = \{D \in \mathcal{D} : D > 0\}$, and $\mathcal{Z} = \{D - \hat{D} : D, \hat{D} \in \mathcal{D}\}$. Clearly $\mathcal{Z}$ is a linear vector space consisting of all Hermitian matrices in $\mathcal{C}^{n\times n}$ with block diagonal structure inherited from $\mathcal{D}$, and thus $\mathcal{D} \subseteq \mathcal{Z}$. Let $M \in \mathcal{C}^{n\times n}$ have a singular value decomposition given by (2.1), in which $U_1, V_1 \in \mathcal{C}^{n\times r}$ are the parts of the left and right singular matrices of M, respectively, corresponding to the largest singular value, which is 1. Let U_1 and V_1 be partitioned conformally to the block diagonal structure of $\mathbf{\Delta}$, i.e.,

$$U_1 = \begin{bmatrix} A_1' & \cdots & A_S' & E_1' & \cdots & E_F' \end{bmatrix}', \quad V_1 = \begin{bmatrix} B_1' & \cdots & B_S' & H_1' & \cdots & H_F' \end{bmatrix}',$$

in which $A_i \in \mathcal{C}^{r_i \times r}$, $B_i \in \mathcal{C}^{r_i \times r}$, $E_j \in \mathcal{C}^{m_j \times r}$, and $H_j \in \mathcal{C}^{m_j \times r}$, for $i \in \underline{S}$ and $j \in \underline{F}$. Note that the notation used here is almost identical to that used in [21] (but see Lemma A.5).

LEMMA A.1. *Let $p, q \in \mathcal{C}^n$. Then*

1. *there exists $\Delta \in \mathcal{C}^{n\times n}$, $\|\Delta\| \leq 1$ such that $\Delta p = q$ if and only if $p'p \geq q'q$.*
2. *there exists $\delta \in \mathcal{C}$, $|\delta| \leq 1$ such that $\delta p = q$ if and only if $pp' \geq qq'$.*
3. *the following statements are equivalent:* (i) *There exists $\delta \in \mathcal{C}$, $|\delta| = 1$ such that $\delta p = q$.* (ii) *$pp' = qq'$.* (iii) *$q'Hq = p'Hp$ for every Hermitian $H \in \mathcal{C}^{n\times n}$.*
4. *there exists $Q \in \mathcal{C}^{n\times n}$, $QQ' = I_n$ such that $Qp = q$ if and only if $p'p = q'q$.*

Proof. The proof is a slight extension of parallel results in [9] and [20] (see also [22]) and is reproduced for completeness. Assume $p \neq 0$; otherwise the result is trivial.

1. If $\Delta p = q$ with $\|\Delta\| \leq 1$, then $\|p\| \geq \|q\|$ or $p'p - q'q \geq 0$. Conversely if $\|p\| \geq \|q\|$, choose the contractive rank-one matrix $\Delta = qp'/\|p\|^2$. Then $\Delta p = q$.
2. If $\delta p = q$ and $|\delta| \leq 1$, then $pp' - qq' = (1 - |\delta|^2)pp' \geq 0$. Conversely, let $pp' - qq' \geq 0$. For every x such that $p'x = 0$, we have

   $$x'(pp' - qq')x \geq 0 \Rightarrow x'qq'x = 0 \Rightarrow q'x = 0 \Rightarrow \mathcal{N}(p') \subseteq \mathcal{N}(q') \Rightarrow \mathcal{R}(q) \subseteq \mathcal{R}(p).$$

   Thus, there exists $\delta \in \mathcal{C}$ such that $q = \delta p$. Now, $0 \leq pp' - qq' = (1 - |\delta|^2)pp'$ implies that $|\delta| \leq 1$.
3. (i) $\Rightarrow$ (ii): This is immediate, since $qq' = \delta\bar{\delta}pp' = |\delta|^2 pp' = pp'$. (ii) $\Rightarrow$ (i): Using an argument similar to the one used in part 2, we conclude that $\mathcal{R}(p) = \mathcal{R}(q)$, i.e., $q = \delta p$ for some $\delta \in \mathcal{C}$, $\delta \neq 0$. Now, $qq' = |\delta|^2 pp'$ so that $|\delta| = 1$. (i) $\Rightarrow$ (iii): By direct calculation, $q'Hq = |\delta|^2 p'Hp = p'Hp$ for every $H = H'$. (iii) $\Rightarrow$ (i): Setting $H = I_n$ shows that $\|p\| = \|q\|$. Setting $H = pp'$ gives $|q'p| = \|p\|\|q\|$ and thus $q = \delta p$ for some scalar δ such that $|\delta| = 1$.
4. Similar to items 1–3 above. ∎

The next lemma characterizes the following two sets:

$$\Psi(x) = \{z \in \mathcal{C}^n : z = \Delta U_1 x, \Delta \in \mathcal{B}\mathbf{\Delta}\}$$

and

$$\Psi_o(x) = \{z \in \mathcal{C}^n : z = \Delta U_1 x, \Delta \in \mathcal{Q}\},$$

in which x is a fixed vector in $\mathcal{C}^n$.

LEMMA A.2. *Let $z \in \mathcal{C}^n$, partitioned conformally with $\mathbf{\Delta}$, i.e.,*

$$z' = \left[\begin{array}{cccccc} z_1' & \dots & z_S' & z_{S+1}' & \dots & z_{S+F}' \end{array} \right],$$

where $z_i \in \mathcal{C}^{r_i}$ for $i \in \underline{S}$ and $z_{S+i} \in \mathcal{C}^{m_i}$ for $i \in \underline{F}$. Then, for $x \in \mathcal{C}^n$,

$$\Psi(x) = \{z \in \mathcal{C}^n \ : \ z_i z_i' \le A_i x x' A_i', i \in \underline{S} \ ; \ \|z_{S+i}\| \le \|E_i x\|, i \in \underline{F}\} \tag{A.1}$$

and

$$\Psi_o(x) = \{z \in \mathcal{C}^n \ : \ z_i z_i' = A_i x x' A_i', i \in \underline{S} \ ; \ \|z_{S+i}\| = \|E_i x\|, \ i \in \underline{F}\}. \tag{A.2}$$

Further,

$$z_i' H_i z_i = x' A_i' H_i A_i x \ \text{ for every } \ z \in \Psi_o(x), \ i \in \underline{S} \text{ and } H_i = H_i' \in \mathcal{C}^{r_i \times r_i}. \tag{A.3}$$

Finally,

$$z \in \Psi_o(x) \text{ if and only if } z' Z z - x' U_1' Z U_1 x = 0 \text{ for every } Z \in \mathcal{Z}. \tag{A.4}$$

Proof. The characterization of the sets $\Psi(x)$ and $\Psi_o(x)$ follows immediately from Lemma A.1. The identity (A.3) is also a direct consequence of Lemma A.1(3). Finally, in order to show (A.4), take $Z = \text{diag}(Q_1, \dots, Q_S, q_{S+1} I_{m_1}, \dots, q_{S+F} I_{m_F})$ with $Q_i = Q_i' \in \mathcal{C}^{r_i \times r_i}$ and $q_{S+i} \in \mathcal{R}$. Then

$$z' Z z - x' U_1' Z U_1 x = \sum_{i=1}^{S} (z_i' Q_i z_i - x' A_i' Q_i A_i x) + \sum_{i=1}^{F} q_{S+i} (\|z_{S+i}\|^2 - \|E_i x\|^2) = 0, \tag{A.5}$$

using (A.2) and (A.3). Conversely, suppose that (A.5) holds for every $Z \in \mathcal{Z}$ and for each $i \in \underline{S}$ choose $Q_j = O_{m_j}$ $(j \ne i)$ and $q_{S+j} = 0 \ \forall \ j \in \underline{F}$. Then, $z' Q_i z_i = x' A_i' Q_i A_i x$ for every $Q_i = Q_i'$, which from Lemma A.1 implies that $z_i z_i' = A_i x x' A_i'$ for each $i \in \underline{S}$. Similarly, setting $Q_i = O_{m_i} \ \forall \ i \in \underline{S}$, $q_{S+j} = 0 \ \forall \ j \in \underline{F} \backslash \{i\}$, and $q_{S+i} = 1$ gives $\|z_{S+i}\| = \|E_i x\|$ and thus $z \in \Psi_o(x)$. □

Before proving Lemma 2.3 the following result is also needed.

LEMMA A.3. *Let $Z = Z'$ and define*

$$P_\alpha = \left[\begin{array}{cc} 0 & 0 \\ 0 & \Lambda \end{array} \right] - \alpha \left[\begin{array}{c} X' \\ \tilde{X}' \end{array} \right] Z \left[\begin{array}{cc} X & \tilde{X} \end{array} \right],$$

where $\Lambda = \Lambda' > 0$ and $X_o := [X \ \tilde{X}]$ is unitary. Then there exists $\alpha > 0$ such that $P_\alpha \ge 0$ if and only if $X' Z X \le 0$.

Proof. (a) If $P_\alpha \ge 0$ for some $\alpha > 0$, then the $(1,1)$ block of P_α is certainly semidefinite, i.e., $X' Z X \le 0$. (b) Assume that $X' Z X \le 0$. We will show that $X_o P_\alpha X_o' \ge 0$ for all sufficiently small $\alpha > 0$. First note that

$$X_o P_\alpha X_o' = \left[\begin{array}{cc} X & \tilde{X} \end{array} \right] \left[\begin{array}{cc} 0 & 0 \\ 0 & \Lambda \end{array} \right] \left[\begin{array}{c} X' \\ \tilde{X}' \end{array} \right] - \alpha Z = \tilde{X} \Lambda \tilde{X}' - \alpha Z.$$

Every vector $\xi \in \mathcal{C}^n$ can be written (uniquely) as $\xi = X\theta_1 + \tilde{X}\theta_2$. Now,

$$\begin{aligned} \xi' X_o P_\alpha X_o' \xi &= \xi' (\tilde{X} \Lambda \tilde{X}' - \alpha Z) \xi = (\theta_1' X' + \theta_2' \tilde{X}')(\tilde{X} \Lambda \tilde{X}' - \alpha Z)(X\theta_1 + \tilde{X}\theta_2) \\ &= \theta_2' \Lambda \theta_2 - \alpha (\theta_1' X' Z X \theta_1 + 2\text{Re}(\theta_1' X' Z \tilde{X} \theta_2) + \theta_2' \tilde{X}' Z \tilde{X} \theta_2). \end{aligned}$$

If $\theta_2 \neq 0$, this is positive for all sufficiently small α. If $\theta_2 = 0$ ($\Rightarrow \theta_1 \neq 0$ for $\xi \neq 0$), we get $X_o P_\alpha X_o' = -\alpha\theta_1' X'ZX\theta_1 \geq 0$ for every $\alpha \geq 0$. In either case, there exists $\alpha > 0$ (sufficiently small) such that $X_o P_\alpha X_o' \geq 0$, which implies that $P_\alpha \geq 0$. ∎

We can now prove the first part of Lemma 2.3.

LEMMA A.4. *With all variables as defined above,*

$$\max_{\Delta \in \mathcal{B}\mathbf{\Delta}} \|V_1'\Delta U_1\|^2 = \min_{\substack{D \in \mathcal{D} \\ D - V_1 V_1{}' \geq 0 \\ \gamma I - U_1{}' D U_1 \geq 0}} \gamma. \tag{A.6}$$

Proof. We call the maximization on the left of (A.6) the primal problem and the minimization on the right the dual problem. The primal problem can be written as

$$\max_{\substack{\Delta \in \mathcal{B}\mathbf{\Delta} \\ x'x \leq 1}} x'U_1'\Delta' V_1 V_1{}' \Delta U_1 x = \max_{\substack{x'x \leq 1 \\ z \in \Psi(x)}} z' V_1 V_1{}' z, \tag{A.7}$$

where $\Psi(x)$ is defined in (A.1). Consider the following identity:

$$\begin{aligned} z' V_1 V_1{}' z = &-\gamma(1 - x'x) - (x'U_1' D U_1 x - z' D z) - z'(D - V_1 V_1{}')z \\ &- x'(\gamma I - U_1' D U_1)x + \gamma. \end{aligned} \tag{A.8}$$

For all x, z satisfying the constraints of the primal problem in (A.7) (i.e., $x'x \leq 1$, $z \in \Psi(x)$), all D satisfying the constraints of the dual problem in (2.8) (i.e., $D \in \mathcal{D}$, $D - V_1 V_1{}' \geq 0$, and $\gamma I - U_1{}' D U_1 \geq 0$), and all $\gamma > 0$, we have that the four terms on the right of (A.8) are nonpositive, i.e.,

1. $1 - x'x \geq 0$ (primal),
2. $x'U_1' D U_1 x - z' D z \geq 0$ (primal),
3. $z'(D - V_1 V_1') z \geq 0$ (dual),
4. $x'(\gamma I - U_1' D U_1)x \geq 0$ (dual).

The nonnegativity of the second term follows from the following argument. Write

$$x'U_1' D U_1 x - z' D z = \sum_{i=1}^{S} (x' A_i' D_i A_i x - z_i' D_i z_i) + \sum_{i=1}^{F} d_i (\|E_i x\|^2 - \|z_{S+i}\|^2),$$

where $D_i = D_i' \geq 0 \ \forall\ i \in \underline{S}$ and $d_i \geq 0 \ \forall\ i \in \underline{F}$. Now, from Lemma A.2, $z \in \Psi(x)$ implies that $\|z_{S+i}\| \leq \|E_i x\| \ \forall\ i \in \underline{F}$. Further, since $z \in \Psi(x)$ we have

$$\begin{aligned} &A_i x x' A_i' - z_i z_i' \geq 0 \quad \forall\ i \in \underline{S} \\ &\quad \Rightarrow D_i^{1/2} (A_i x x' A_i' - z_i z_i') D_i^{1/2} \geq 0 \quad \forall\ D_i^{1/2} = (D_i^{1/2})' \geq 0,\ i \in \underline{S} \\ &\quad \Rightarrow D_i^{1/2} A_i x x' A_i' D_i^{1/2} \geq D_i^{1/2} z_i z_i' D_i^{1/2} \quad \forall\ D_i^{1/2} = (D_i^{1/2})' \geq 0,\ i \in \underline{S}. \end{aligned}$$

Lemma A.1(2) implies that there exist $\lambda_i \in \mathcal{C}$, $|\lambda_i| \leq 1$ such that $\lambda_i D_i^{1/2} A_i x = D_i^{1/2} z_i$, $i \in \underline{S}$, and thus $\|D_i^{1/2} A_i x\|^2 - \|D_i^{1/2} z_i\|^2 \geq 0$, so that the second term is nonnegative.

Thus, the first four terms in the global identity are all nonpositive, and hence

$$\max_{\substack{x'x \leq 1 \\ z \in \Psi(x)}} z' V_1 V_1{}' z \leq \min_{\substack{D \in \mathcal{D} \\ \gamma I - U_1{}' D U_1 \geq 0 \\ D - V_1 V_1{}' \geq 0}} \gamma.$$

It follows that the dual problem solution is an upper bound on the primal's. Next we show that they are equal.

Let γ and D be feasible for the dual problem. Then

1. $\underline{\lambda}(D - V_1V_1') = 0$, and
2. $\underline{\lambda}(\gamma I - U_1'DU_1) = 0$.

Equality in (2) is obvious: If $\gamma I - U_1'DU_1$ was not singular at optimality, then there would exist $\hat{\gamma} < \gamma$ such that $\hat{\gamma} I - U_1'DU_1 \geq 0$, contradicting the optimality of γ. To show (1), suppose for contradiction that $D - V_1V_1' > 0$. Then $(D - \mu I) - V_1V_1' \geq 0$ for some $\mu > 0$ (sufficiently small). Now $\gamma I - U_1'DU_1 \geq 0$ and so $(\gamma - \mu)I - U_1'(D - \mu I)U_1 \geq 0$, contradicting again the optimality of γ, since $\mu > 0$ can be chosen small enough so that $D - \mu I \in \mathcal{D}$. (Note that in this case the assumption $D - V_1V_1' > 0$ implies that $D > 0$.) Thus both (1) and (2) hold.

Thus $\dim \mathcal{N}(\gamma I - U_1'DU_1) \geq 1$ and $\dim \mathcal{N}(D - V_1V_1') \geq 1$, where $\mathcal{N}(\cdot)$ denotes the null space, are necessary for γ, D to be optimal for the dual problem. Let

$$D - V_1V_1' = \begin{bmatrix} X & \tilde{X} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \begin{bmatrix} X' \\ \tilde{X}' \end{bmatrix}, \tag{A.9}$$

where $X \in \mathcal{C}^{n \times m_1}$, $[X \;\; \tilde{X}]$ unitary, and $\Lambda_1 = \Lambda_1' > 0$ and let

$$\gamma I - U_1'DU_1 = \begin{bmatrix} Y & \tilde{Y} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} Y' \\ \tilde{Y}' \end{bmatrix}, \tag{A.10}$$

where $Y \in \mathcal{C}^{n \times m_2}$, $[Y \;\; \tilde{Y}]$ unitary, and $\Lambda_2 = \Lambda_2' > 0$. We prove that γ, D are optimal if and only if there does not exist $Z \in \mathcal{Z}$ such that $X'ZX \leq 0$ and $Y'U_1'ZU_1Y > 0$.

Suppose $X'ZX \leq 0$ and $Y'U_1'ZU_1Y > 0$ for some $Z \in \mathcal{Z}$. Define $D_\alpha = D - \alpha Z$ for $\alpha \geq 0$. Then it is easy to show that

$$\begin{bmatrix} X' \\ \tilde{X}' \end{bmatrix} (D_\alpha - V_1V_1') \begin{bmatrix} X & \tilde{X} \end{bmatrix} = \begin{bmatrix} -\alpha X'ZX & -\alpha X'Z\tilde{X} \\ -\alpha \tilde{X}'ZX & \Lambda_1 - \alpha \tilde{X}'Z\tilde{X} \end{bmatrix}$$

and

$$\begin{bmatrix} Y' \\ \tilde{Y}' \end{bmatrix} (\gamma I - U_1'D_\alpha U_1) \begin{bmatrix} Y & \tilde{Y} \end{bmatrix} = \begin{bmatrix} \alpha Y'U_1'ZU_1Y & \alpha Y'U_1'ZU_1\tilde{Y} \\ \alpha \tilde{Y}'U_1'ZU_1Y & \Lambda_2 + \alpha \tilde{Y}'U_1'ZU_1\tilde{Y} \end{bmatrix}.$$

It is now straightforward using Lemma A.3 and a Schur-type argument to show that there exists $\alpha > 0$ such that $D_\alpha - V_1V_1' \geq 0$ and $\gamma I - U_1'D_\alpha U_1 > 0$, and so there exists $\gamma_\alpha < \gamma$ such that $\gamma_\alpha I - U_1'D_\alpha U_1 \geq 0$, proving that γ, D are not optimal.

Conversely, suppose that γ and D are not optimal so that there exists $\gamma_o < \gamma$ and $D_o \in \mathcal{D}$ such that $D_o - V_1V_1' \geq 0$ and $\gamma_o I - U_1'D_oU_1 \geq 0$. Define $Z = D - D_o$ so that $Z \in \mathcal{Z}$. Now

$$0 = X'(D - V_1V_1')X = X'(D_o - V_1V_1' + Z)X = X'(D_o - V_1V_1')X + X'ZX,$$

and so $X'ZX \leq 0$. Also

$$\begin{aligned} 0 &= Y'(\gamma I - U_1'DU_1)Y = Y'(\gamma_o I - U_1'D_oU_1 + (\gamma - \gamma_o)I - U_1'ZU_1)Y \\ &= Y(\gamma_o I - U_1'D_oU_1)Y + (\gamma - \gamma_o)I - Y'U_1'ZU_1Y, \end{aligned}$$

and so $Y'U_1'ZU_1Y > 0$. This proves that γ, D are optimal for the dual problem if and only if there does not exist $Z \in \mathcal{Z}$ such that $X'ZX \leq 0$ and $Y'U_1'ZU_1Y > 0$.

Finally, we need to prove the absence of gap between the primal and dual problems. Let γ and D be optimal for the dual and suppose that (A.9) and (A.10) are satisfied. It follows from the identity (A.8) that there is no gap if there exist $z \in \mathcal{C}^n$ and $x \in \mathcal{C}^m$ such that

1. $x'x = 1$ (so that x is feasible and the first term on the right of (A.8) is zero);
2. $z \in \Psi_o(x)$ (so that z is feasible and the second term is zero; see Lemma A.2);
3. $z \in \mathcal{N}(D - V_1V_1')$, or equivalently, $z = X\tilde{z}$ for some $\tilde{z} \in \mathcal{C}^{m_1}$ (so that the third term is zero);
4. $x \in \mathcal{N}(\gamma I - U_1'DU_1)$, or equivalently, $x = Y\tilde{x}$ for some $\tilde{x} \in \mathcal{C}^{m_2}$ (so that the fourth term is zero).

It follows that there is no gap if there exist $\tilde{z} \in \mathcal{C}^{m_1}$ and $\tilde{x} \in \mathcal{C}^{m_2}$ such that $\tilde{x}'\tilde{x} = \tilde{z}'\tilde{z} = 1$ and $X\tilde{z} \in \Psi_o(Y\tilde{x})$, i.e., if there exist $\tilde{x}$ and $\tilde{z}$ such that $\|\tilde{x}\| = \|\tilde{z}\| = 1$ and

$$\tilde{z}'X'ZX\tilde{z} - \tilde{x}'Y'U_1'ZU_1Y\tilde{x} = 0,$$

for every $Z \in \mathcal{Z}$. We thus need to prove that such $\tilde{z}$ and $\tilde{x}$ exist if γ, D are optimal, that is, if there does not exist $Z \in \mathcal{Z}$ such that $X'ZX \le 0$ and $Y'U_1'ZU_1Y > 0$.

Assume for contradiction that $\exists Z \in \mathcal{Z}$ such that $\tilde{z}'X'ZX\tilde{z} - \tilde{x}'Y'U_1'ZU_1Y\tilde{x} \neq 0 \ \forall\ \tilde{z} \in \mathcal{C}^{m_1}$ and $\tilde{x} \in \mathcal{C}^{m_2}$ with $\|\tilde{z}\| = \|\tilde{x}\| = 1$. We need to show that this contradicts optimality of (γ, D) or, equivalently, that it implies the existence of $Z \in \mathcal{Z}$ such that $X'ZX \le 0$ and $Y'U_1ZU_1Y > 0$. Next note that since $Z \in \mathcal{Z} \Leftrightarrow -Z \in \mathcal{Z}$, the assumed condition is equivalent to the existence of $Z \in \mathcal{Z}$ such that $\tilde{z}'X'ZX\tilde{z} < \tilde{x}'Y'U_1'ZU_1Y\tilde{x}$ for every $\tilde{z} \in \mathcal{C}^{m_1}$ and $\tilde{x} \in \mathcal{C}^{m_2}$ with $\|\tilde{z}\| = \|\tilde{x}\|$. This implies, in particular, that

$$\overline{\lambda}(X'ZX) = \max_{\|\tilde{z}\|=1} \tilde{z}'X'ZX\tilde{z} < \min_{\|\tilde{x}\|=1} \tilde{x}'Y'U_1'ZU_1Y\tilde{x} = \underline{\lambda}(Y'U_1'ZU_1Y).$$

Next, set $\delta = \frac{1}{2}[\overline{\lambda}(X'ZX) + \underline{\lambda}(Y'U_1'ZU_1Y)]$. Then,

$$\tilde{z}'X'(Z - \delta I)X\tilde{z} < \tilde{x}'Y'U_1'(Z - \delta I)U_1Y\tilde{x}$$

$\forall\ \tilde{z} \in \mathcal{C}^{m_1}$ and $\tilde{x} \in \mathcal{C}^{m_2}$ with $\|\tilde{z}\| = \|\tilde{x}\|$. Now,

$$\max_{\|\tilde{z}\|=1} \tilde{z}'X'(Z - \delta I)X\tilde{z} = \overline{\lambda}(X'ZX) - \delta = \frac{1}{2}\left[\ \overline{\lambda}(X'ZX) - \underline{\lambda}(Y'Y_1'ZU_1Y)\right] < 0,$$

so that $X'(Z - \delta I)X < 0$. Similarly,

$$\min_{\|\tilde{x}\|=1} \tilde{x}'Y'U_1'(Z - \delta I)U_1Y\tilde{x} = \frac{1}{2}\left[\underline{\lambda}(Y'Y_1'ZU_1Y) - \overline{\lambda}(X'ZX)\right] > 0,$$

which implies that $Y'U_1'(Z - \delta I)U_1Y > 0$ and establishes the required contradiction since $Z - \delta I \in \mathcal{Z}$. ∎

In the remaining part of the appendix we prove the second part of Lemma 2.3. This requires two preliminary results as follows.

LEMMA A.5. *With all variables as defined above,* $\inf\{\|D^{1/2}MD^{-1/2}\| : D \in \mathcal{D}_+\} = \|M\|$ *if and only if* $\underline{\lambda}(U_1'ZU_1 - V_1'ZV_1) \le 0$ *for every* $Z \in \mathcal{Z}$.

Proof. See [21]. Note that our definition of $\mathcal{Z}$ is slightly different from the definition in [21], but this does not affect the result. ∎

LEMMA A.6. *With all variables as defined above,* $\inf\{\|D^{1/2}MD^{-1/2}\| : D \in \mathcal{D}_+\} = \|M\|$ *if and only if* $\inf\{\|D^{1/2}M_0D^{-1/2}\| : D \in \mathcal{D}_+\} = \|M\|$.

Proof. The proof is immediate from Lemma A.5 since M and M_0 share the same left and right singular vector matrices corresponding to $\|M\| = 1$ (U_1 and V_1, respectively). ∎

The second part of Lemma 2.3 can now be established as follows.

LEMMA A.7. *If the optimum in* (2.3) *is* 1*, the optimum in* (2.6) *is* 1.

Proof. First note that for every $D \in \mathcal{D}_+$,

$$M'DM - \gamma^2 D \geq 0 \Leftrightarrow D^{-1/2}M'DMD^{-1/2} - \gamma^2 I \leq 0 \Leftrightarrow \|D^{1/2}MD^{-1/2}\| \leq \gamma,$$

and hence via a continuity argument [11] the optimum in (2.3) is 1 if and only if $\inf\{\|D^{1/2}MD^{-1/2}\| : D \in \mathcal{D}_+\} = 1$. Thus, from Lemma A.6 we also have that $\inf\{\|D^{1/2}M_0D^{-1/2}\| : D \in \mathcal{D}_+\} = 1$, and hence the optimization problem $\min\{\gamma : D \in \mathcal{D}, \gamma^2 D - M_0'DM_0 \geq 0\}$ has optimum solutions $\gamma = 1$ and $D = I$. Now suppose for contradiction that $\max\{\|V_1'\Delta U_1\| : \Delta \in \mathcal{B}\mathbf{\Delta}\} < 1$, or, in view of Lemma A.4 that $\min\{\gamma : D \in \mathcal{D}, D - V_1V_1' \geq 0, \gamma I - U_1'DU_1 \geq 0\} < 1$. Write the minimum γ as $\gamma = 1-\epsilon$ for some $\epsilon > 0$ and the corresponding optimum $D = I - Z$ for some $Z \in \mathcal{Z}$. Then

$$(1-\epsilon)I - U_1'(I-Z)U_1 \geq 0 \Leftrightarrow U_1'ZU_1 \geq \epsilon I > 0.$$

Also,

$$(I-Z) - V_1V_1' \geq 0 \Rightarrow V_1'[(I-Z) - V_1V_1']V_1 \geq 0 \Rightarrow V_1'ZV_1 \leq 0.$$

Thus $U_1'ZU_1 - V_1'ZV_1 > 0$ for some $Z \in \mathcal{Z}$, which contradicts Lemma A.5. ∎

Lemma 2.3 now follows from Lemmas A.4 and A.7.

REFERENCES

[1] D. Avis and K. Fukuda, *Reverse search for enumeration*, Discrete Appl. Math., 65 (1996), pp. 21–46.

[2] V. Balakrishnan and S. Boyd, *Existence and uniqueness of optimal matrix scalings*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 29–39.

[3] G. J. Balas, J. C. Doyle, K. Glover, A. Packard, and R. Smith, *μ-Analysis and Synthesis Toolbox*, The Mathworks, Natick, MA, 1995.

[4] R. D. Braatz, P. M. Young, J. C. Doyle, and M. Morari, *Computational complexity of μ calculation*, in Proc. Amer. Control Conf. (San Francisco, CA), IEEE Press, New York, 1993, pp. 1682–1683.

[5] R. D. Braatz, P. M. Young, J. C. Doyle, and M. Morari, *Computational complexity of μ calculation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1000–1002.

[6] J. C. Doyle, J. Wall, and G. Stein, *Performance and robustness analysis for structured uncertainty*, in Proc. IEEE Conf. Dec. & Control (Orlando, FL), IEEE Press, New York, 1982, pp. 629–636.

[7] J. C. Doyle, *Structured uncertainty in control system design*, Proc. IEE-D, 129 (1982), pp. 242–250.

[8] J. C. Doyle, A. Packard, and K. Zhou, *Review of LFTs, LMIs, and μ*, in Proc. IEEE Conf. Dec. & Control (Brighton, England), IEEE Press, New York, 1991, pp. 1227–1232.

[9] G. E. Dullerud and F. Paganini, *A Course in Robust Control Theory: A Convex Approach*, Texts Appl. Math. 36, Springer, New York, 1999.

[10] M. K. H. Fan and A. L. Tits, *Characterization and efficient computation of the structured singular value*, IEEE Trans. Automat. Control, 31 (1986), pp. 734–783.

[11] M. K. H. Fan and A. L. Tits, *m-form numerical range and the computation of the structured singular value*, IEEE Trans. Automat. Control, 33 (1988), pp. 284–289.

[12] M. K. H. Fan, A. L. Tits, and J. C. Doyle, *Robustness in the presence of mixed parametric uncertainty and unmodelled dynamics*, IEEE Trans. Automat. Control, 36 (1991), pp. 25–38.

[13] S. K. Gungah, *Maximally Robust Controllers for a Class of Unstructured Uncertainty*, Ph.D. thesis, Imperial College, University of London, London, UK, 1999.

[14] S. K. Gungah, G. D. Halikias, and I. M. Jaimoukha, *Maximally robust controllers for multivariable systems*, SIAM J. Control Optim., 38 (2000), pp. 1805–1829.

[15] S. K. Gungah, U. Malik, I. M. Jaimoukha, and G. D. Halikias, *A new upper bound for the real structured singular value*, in Proc. IEEE Conf. Dec. & Control (Florida), IEEE Press, New York, 2001, pp. 247–248.

[16] N. A. Lehtomaki, D. A. Castanon, B. D. Levy, G. Stein, N. R. Sandel, and M. Athans, *Robustness and modeling error characterization*, IEEE Trans. Automat. Control, 29 (1984), pp. 212–220.

[17] U. Malik, *New Bounds for the Structured Singular Value*, Ph.D. thesis, Imperial College, University of London, London, UK, 2003.

[18] U. Malik, I. M. Jaimoukha, G. D. Halikias, and S. K. Gungah, *On the gap between the quadratic integer programming problem and its semidefinite relaxation*, Math. Program. Ser. A, 107 (2006), pp. 505–515.

[19] A. Megretski, *On the gap between structured singular values and their upper bounds*, in Proc. IEEE Conf. Dec. & Control (San Antonio, TX), IEEE Press, New York, 1993, pp. 3461–3462.

[20] G. Meinsma, Y. Shrivastava, and M. Fu, *Some properties of an upper bound for μ*, IEEE Trans. Automat. Control, 41 (1996), pp. 1326–1330.

[21] A. Packard and J. C. Doyle, *The complex structured singular value*, Automatica, 21 (1993), pp. 71–109.

[22] A. Rantzer, *On the Kalman-Yakubovich-Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.

[23] R. M. Redheffer, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269–286.

[24] R. Smith, *A study of the gap between the structured singular value and its convex upper bound for low-rank matrices*, IEEE Trans. Automat. Control, 42 (1997), pp. 1176–1179.

[25] P. M. Young, *The rank one mixed μ problem and Kharitonov-type methods*, in Proc. IEEE Conf. Dec. & Control (San Antonio, Texas), IEEE Press, New York, 1993, pp. 523–528.

[26] P. M. Young, *The rank one mixed μ problem and kharitonov-type analysis*, Automatica, 30 (1994), pp. 1899–1911.

[27] P. M. Young and J. C. Doyle, *Properties of the mixed μ problem and its bounds*, IEEE Trans. Automat. Control, 41 (1996), pp. 155–159.

[28] P. M. Young, M. P. Newlin, and J. C. Doyle, *μ analysis with real parametric uncertainty*, in Proc. IEEE Conf. Dec. & Control (Brighton, England), IEEE Press, New York, 1991, pp. 1251–1256.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1279–1296

ON A CLASS OF STOCHASTIC OPTIMAL CONTROL PROBLEMS RELATED TO BSDES WITH QUADRATIC GROWTH*

MARCO FUHRMAN†, YING HU‡, AND GIANMARIO TESSITORE§

Abstract. In this paper, we study a class of stochastic optimal control problems, where the drift term of the equation has a linear growth on the control variable, the cost functional has a quadratic growth, and the control process takes values in a closed set (not necessarily compact). This problem is related to some backward stochastic differential equations (BSDEs) with quadratic growth and unbounded terminal value. We prove that the optimal feedback control exists, and the optimal cost is given by the initial value of the solution of the related BSDE.

Key words. stochastic optimal control, backward stochastic differential equations, quadratically growing driver, unbounded final condition

AMS subject classifications. 93E20, 60H10

DOI. 10.1137/050633548

1. Introduction.

In this paper, we consider a controlled equation of the form

$$\begin{cases} dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, [dW_t + r(t, X_t, u_t)\, dt], & t \in [0, T], \\ X_0 = x. \end{cases} \tag{1.1}$$

In the equation, W is an $\mathbb{R}^d$-valued Wiener process, defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with respect to a filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfying the usual conditions; the unknown process X takes values in $\mathbb{R}^n$; x is a given element of $\mathbb{R}^n$; and u is the control process, which is assumed to be an $(\mathcal{F}_t)$-adapted process taking values in a given nonempty closed set $K \subset \mathbb{R}^m$. The control problem consists of minimizing a cost functional of the form

$$J = \mathbb{E} \int_0^T g(t, X_t, u_t)\, dt + \mathbb{E}\, \phi(X_T). \tag{1.2}$$

We suppose that r has a linear growth in u, g has quadratic growth in x and u, and ϕ has quadratic growth in x.

The main novelty of the present paper, in comparison with the existing literature, is that on the one hand we assume that neither K nor r is bounded; on the other hand we consider a degenerate control problem (since nothing is assumed on the image of σ). Moreover, we also allow ϕ to have quadratic growth.

*Received by the editors June 14, 2005; accepted for publication (in revised form) March 24, 2006; published electronically September 26, 2006.

http://www.siam.org/journals/sicon/45-4/63354.html

†Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy (marco.fuhrman@polimi.it). This author was supported by the European Community's Human Potential Programme under contracts HPRN-CT-2002-00281 (Evolution Equations) and HPRN-CT-2002-00279 (QP-Applications).

‡IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France (ying.hu@univ-rennes1.fr).

§Dipartimento di Matematica e Applicazioni, Università di Milano-Bicocca, Via Cozzi 53, 20135, Milano, Italy (gianmario.tessitore@unimib.it). This author was supported by the European Community's Human Potential Programme under contract HPRN-CT-2002-00281 (Evolution Equations).

Nonlinear backward stochastic differential equations (BSDEs) were first introduced by Pardoux and Peng [11]. A major application of BSDEs is in stochastic control; see, e.g., [4] and [12]. We also refer the reader to [9], [5], and [7]. From several points of view, the class of control problems addressed in these papers, and in the references therein, is more general than the one considered here, but assumptions implying "bounded control image" (i.e., boundedness of K or r in our notation) are a common feature of all the above papers. The special "unbounded" case corresponding to the assumptions $K = \mathbb{R}^m$ and $g = \frac{1}{2}|u|^2 + q(t,x)$ is treated in [6] by an ad hoc exponential transform. We notice that in [6] ϕ is allowed to take the value $+\infty$. Finally, the same special case (in which the Hamiltonian is exactly the square of the norm of the gradient) was treated in [8] by analytic techniques under nondegeneracy assumptions and in an infinite-dimensional framework.

The difficulty here is that the Hamiltonian corresponding to the control problem has quadratic growth in the gradient and consequently the associated BSDE has quadratic growth in the Z variable. Well-posedness for this class of BSDEs has been proved in [10] in the case of bounded terminal value. Since we allow for unbounded terminal cost, to treat such equations we have to apply the techniques recently introduced in [2]. We notice that for such BSDEs no general uniqueness results are known: we replace uniqueness with the selection of a maximal solution. Moreover, the usual application of the Girsanov technique is not allowed (since the Novikov condition is not guaranteed), and we have to develop specific arguments both to prove the fundamental relation (see section 4) and to obtain the existence of a (weak) solution to the closed loop equation (see section 5). Our main result is to prove that the optimal feedback control exists and the optimal cost is given by the value Y_0 of the maximal solution (Y, Z) of the BSDE with the quadratic growth and unbounded terminal value mentioned above. Moreover, we show that we can construct an optimal feedback in terms of the process Z. Finally, we prove that if we fix a particular optimal feedback law, then the solution of the corresponding closed loop equation is unique; see Proposition 5.4.

An alternative approach to the control problem may consist of applying the stochastic maximum principle; see, e.g., [12] for a detailed exposition. However, this would require further differentiability conditions on the coefficients of the controlled equation and would not immediately imply existence of an optimal control, since the maximum principle is usually stated as a necessary condition for optimality.

The paper is organized as follows. In the next section, we describe the control problem; in section 3, we study the related BSDE; and in section 4, we establish the fundamental relation between the optimal control problem and BSDE. The last section is devoted to the proof of the existence of optimal feedback control.

2. The control problem. We consider the optimal control problem given by a state equation of the form

$$\begin{cases} dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, [dW_t + r(t, X_t, u_t)\, dt], & t \in [0, T], \\ X_0 = x, \end{cases} \tag{2.1}$$

and given by a cost functional of the form

$$J = \mathbb{E} \int_0^T g(t, X_t, u_t)\, dt + \mathbb{E}\, \phi(X_T). \tag{2.2}$$

We work under the following assumptions.

HYPOTHESIS 2.1.

1. *The process W is a Wiener process in $\mathbb{R}^d$, defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with respect to a filtration $(\mathcal{F}_t)$ satisfying the usual conditions.*
2. *The set K is a nonempty closed subset of $\mathbb{R}^m$.*
3. *The functions $b : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : [0,T] \times \mathbb{R}^n \to \mathbb{R}^{n\times d}$, $r : [0,T] \times \mathbb{R}^n \times K \to \mathbb{R}^d$, $g : [0,T] \times \mathbb{R}^n \times K \to \mathbb{R}$, $\phi : \mathbb{R}^n \to \mathbb{R}$ are Borel measurable.*
4. *For all $t \in [0,T]$, $x \in \mathbb{R}^n$, $r(t,x,\cdot)$ and $g(t,x,\cdot)$ are continuous functions from K to $\mathbb{R}^d$ and from K to $\mathbb{R}$, respectively.*
5. *There exists a constant $C > 0$ such that for every $t \in [0,T]$, $x, x' \in \mathbb{R}^n$, $u \in K$ it holds that*

$$|b(t,x) - b(t,x')| \le C|x - x'|, \qquad |b(t,x)| \le C(1 + |x|), \tag{2.3}$$

$$|\sigma(t,x) - \sigma(t,x')| \le C|x - x'|, \tag{2.4}$$

$$|\sigma(t,x)| \le C, \tag{2.5}$$

$$|r(t,x,u) - r(t,x',u)| \le C(1 + |u|)|x - x'|, \tag{2.6}$$

$$|r(t,x,u)| \le C(1 + |u|), \tag{2.7}$$

$$0 \le g(t,x,u) \le C(1 + |x|^2 + |u|^2), \tag{2.8}$$

$$0 \le \phi(x) \le C(1 + |x|^2). \tag{2.9}$$

6. *There exist $R > 0$ and $c > 0$ such that for every $t \in [0,T]$, $x \in \mathbb{R}^n$, and every $u \in K$ satisfying $|u| \ge R$,*

$$g(t,x,u) \ge c|u|^2. \tag{2.10}$$

We will say that an $(\mathcal{F}_t)$-adapted stochastic process $\{u_t,\ t \in [0,T]\}$ with values in K is an admissible control if it satisfies

$$\mathbb{E}\int_0^T |u_t|^2 dt < \infty. \tag{2.11}$$

This square summability requirement is justified by (2.10): a control process which is not square summable would have infinite cost.

Remark 2.2. Some classes of linear quadratic control problems fall within the scope of our result. Consider the controlled system

$$\begin{cases} dX_t = A(t)X_t\, dt + b(t)\, dt + \Sigma(t)\,[dW_t + B(t)u_t\, dt], & t \in [0,T], \\ X_0 = x \end{cases} \tag{2.12}$$

and the cost functional

$$J = \frac{1}{2}\mathbb{E}\int_0^T [\langle Q(t)X_t, X_t\rangle + \langle R(t)u_t, u_t\rangle]\, dt + \frac{1}{2}\mathbb{E}\,\langle SX_T, X_T\rangle,$$

and assume that A, b, Σ, B, Q, R are bounded Borel measurable functions with values in $\mathbb{R}^{n\times n}$, $\mathbb{R}^n$, $\mathbb{R}^{n\times d}$, $\mathbb{R}^{d\times m}$, $\mathbb{R}^{n\times n}$, $\mathbb{R}^{m\times m}$, respectively, and that $S \in \mathbb{R}^{n\times n}$; also suppose that $Q(t) \geq 0$, $S \geq 0$, $R(t) \geq cI$ for some $c > 0$ and all t, in the sense of the usual matrix order. Then Hypothesis 2.1 is verified. We are not assuming $K = \mathbb{R}^m$ in general. If $K = \mathbb{R}^m$, then the usual linear quadratic theory applies and gives more general results in the sense that state- and control-dependent noise can be considered, as well as a more general drift coefficient, in (2.12).

Next we show that for every admissible control, the solution to (2.1) exists.

PROPOSITION 2.3. *Let u be an admissible control. Then there exists a unique, continuous, $(\mathcal{F}_t)$-adapted process X satisfying $\mathbb{E}\sup_{t\in[0,T]}|X_t|^2 < \infty$, and $\mathbb{P}$-a.s.,*

$$X_t = x + \int_0^t b(s, X_s)\, ds + \int_0^t \sigma(s, X_s)\, dW_s + \int_0^t \sigma(s, X_s)\, r(s, X_s, u_s)\, ds, \quad t \in [0, T].$$

Proof. The proof of Proposition 2.3 relies on an approximation procedure that will be used again in what follows. We introduce the sequence of stopping times

$$\tau_n = \inf\left\{t \in [0, T] \ : \ \int_0^t |u_s|^2 ds > n\right\},$$

with the convention that $\tau_n = T$ if the indicated set is empty. By (2.11), for $\mathbb{P}$-almost every $\omega \in \Omega$, there exists an integer $N(\omega)$ depending on ω such that

$$n \geq N(\omega) \qquad \Longrightarrow \qquad \tau_n(\omega) = T. \tag{2.13}$$

Let us fix $u^0 \in K$, and for every n, let us define

$$u_t^n = u_t\ 1_{t\leq\tau_n} + u^0\ 1_{t>\tau_n}$$

and consider the equation

$$\begin{cases} dX_t^n = b(t, X_t^n)\, dt + \sigma(t, X_t^n)\, [dW_t + r(t, X_t^n, u_t^n)\, dt], & t \in [0, T], \\ X_0^n = x. \end{cases} \tag{2.14}$$

We claim that (2.14) has a unique solution X^n in the class of $(\mathcal{F}_t)$-adapted processes X satisfying $\sup_{t\in[0,T]}\mathbb{E}\,|X_t|^2 < \infty$.

To prove the claim we use a classical argument: We first write (2.14) in the form

$$dX_t = b(t, X_t)\, dt + \widetilde{b}(t, X_t)\, dt + \sigma(t, X_t)\, dW_t,$$

where $\widetilde{b}(t, x) = \sigma(t, x, u_t^n) r(t, x, u_t^n)$ is a stochastic coefficient which satisfies, by (2.3)–(2.7),

$$|\widetilde{b}(t, x) - \widetilde{b}(t, x')| = |\sigma(t, x, u_t^n) r(t, x, u_t^n) - \sigma(t, x', u_t^n) r(t, x', u_t^n)| \leq C(1 + |u_t^n|)|x - x'|,$$

$$|\widetilde{b}(t, x)| = |\sigma(t, x, u_t^n) r(t, x, u_t^n)| \leq C(1 + |u_t^n|), \qquad t \in [0, T], \quad x, x' \in \mathbb{R}^n.$$

We define $K_t(\omega) = C(1 + |u_t^n(\omega)|)$, and we note that

$$\int_0^T |K_t|^2 dt \leq C\left(1 + \int_0^T |u_t^n|^2 dt\right) \leq C(1 + n + |u_0|^2) =: \widetilde{C}, \qquad \mathbb{P}\text{-a.s.},$$

by the definition of τ_n. Next we introduce the norm $\|X\|^2 := \sup_{t\in[0,T]} e^{-2\lambda t}\mathbb{E}\,|X_t|^2$ and prove that the mapping Γ defined by

$$\Gamma(X)_t = \int_0^t b(s,X_s)\,ds + \int_0^t \widetilde{b}(s,X_s)\,ds + \int_0^t \sigma(s,X_s)\,dW_s, \qquad t\in[0,T],$$

is a contraction with respect to this norm for sufficiently large $\lambda > 0$. For a pair of processes X, X', we have

$$\begin{aligned}
\mathbb{E}\left|\int_0^t [\widetilde{b}(s,X_s) - \widetilde{b}(s,X'_s)]\,ds\right|^2 &\le \mathbb{E}\left(\int_0^t K_s|X_s - X'_s|\,ds\right)^2 \\
&\le \mathbb{E}\left(\int_0^t |K_s|^2 ds \int_0^t |X_s - X'_s|^2\,ds\right) \\
&\le \widetilde{C}\mathbb{E}\int_0^t |X_s - X'_s|^2\,ds \\
&\le \widetilde{C}\|X - X'\|^2 \int_0^t e^{2\lambda s}\,ds,
\end{aligned}$$

and it follows that

$$\begin{aligned}
\sup_{t\in[0,T]} e^{-2\lambda t}\mathbb{E}\left|\int_0^t [\widetilde{b}(s,X_s) - \widetilde{b}(s,X'_s)]\,ds\right|^2 &\le \widetilde{C}\|X - X'\|^2 \int_0^t e^{-2\lambda(t-s)}\,ds \\
&\le \frac{\widetilde{C}}{2\lambda}\|X - X'\|^2.
\end{aligned}$$

The other verifications needed to prove the contraction property are standard. The claim is proved.

It is clear that the solution X^n of (2.14) is also continuous. Moreover, we have

$$X_t^n = X_t^{n+1} \qquad \text{for} \qquad t \le \tau_n;$$

therefore there exists a process X such that

$$X_t = X_t^n \qquad \text{for} \qquad t \le \tau_n,$$

and X is clearly the required solution. The property $\mathbb{E}\sup_{t\in[0,T]}|X_t|^2 < \infty$ is an immediate consequence of the following lemma, which concludes the proof. □

LEMMA 2.4. *Under the previous assumptions, the family of random variables*

$$\sup_{t\in[0,T]} |X_t^n|^2, \qquad n = 1,2,\dots,$$

is uniformly integrable.

Proof. We set $M_t^n = \int_0^t \sigma(s,X_s^n)\,dW_s$. By (2.14), we have

$$X_t^n = x + \int_0^t b(s,X_s^n)\,ds + \int_0^t \sigma(s,X_s^n)r(s,X_s^n,u_s^n)\,ds + M_t^n.$$

First, we claim that the family $\{\sup_{t\in[0,T]}|M_t^n|^2;\, n = 1,2,\dots\}$ is uniformly integrable; indeed it is uniformly bounded in $L^p(\Omega,\mathcal{F},\mathbb{P})$ for every $p\in[1,\infty)$, since by the Burkholder–Davis–Gundy inequalities and the boundedness of σ (see (2.5)),

$$\mathbb{E}\sup_{t\in[0,T]}|M_t^n|^{2p} \le C\mathbb{E}\left(\int_0^T |\sigma(s,X_s^n)|^2\,ds\right)^p \le C$$

for some constant independent of n. Next we note that, by (2.7),

$$\left|\int_0^t \sigma(s,X_s^n)r(s,X_s^n,u_s^n)\,ds\right|^2 \le C\int_0^t (1+|u_s^n|)^2\,ds \le C\left(1+\int_0^T |u_s|^2\,ds\right).$$

Then, setting $l_t^n = \sup_{r\in[0,t]}|X_r^n|^2$, we obtain by (2.3),

$$\begin{aligned} l_t^n &\le C\left(1+\int_0^T |u_s|^2\,ds + \sup_{t\in[0,T]}|M_t^n|^2\right) + C\int_0^t |X_s^n|^2\,ds \\ &\le C\left(1+\int_0^T |u_s|^2\,ds + \sup_{t\in[0,T]}|M_t^n|^2\right) + C\int_0^t l_s^n\,ds. \end{aligned}$$

From the Gronwall lemma we deduce

$$l_t^n \le C\left(1+\int_0^T |u_s|^2\,ds + \sup_{t\in[0,T]}|M_t^n|^2\right),$$

and the required uniform integrability follows. □

The stochastic control problem associated with (2.1)–(2.2) consists of minimizing the cost functional $J(x,u)$ among all the admissible controls.

3. The forward-backward system. We consider again the functions b,σ,g,ϕ satisfying the assumptions in Hypothesis 2.1. We define the Hamiltonian function

$$\psi(t,x,z) = \inf_{u\in K}[g(t,x,u)+z\cdot r(t,x,u)], \qquad t\in[0,T],\quad x\in\mathbb{R}^n,\quad z\in\mathbb{R}^d, \tag{3.1}$$

where $\cdot$ denotes the usual scalar product in $\mathbb{R}^d$. We collect some immediate properties of the function ψ.

LEMMA 3.1. *The map ψ is a Borel measurable function from $[0,T]\times\mathbb{R}^n\times\mathbb{R}^d$ to $\mathbb{R}$. There exists a constant $C>0$ such that*

$$-C(1+|z|^2) \le \psi(t,x,z) \le g(t,x,u)+C|z|(1+|u|) \qquad \forall u\in K. \tag{3.2}$$

Moreover, the infimum in (3.1) *is attained in a ball of radius $C(1+|x|+|z|)$, that is,*

$$\psi(t,x,z) = \min_{u\in K,|u|\le C(1+|x|+|z|)}[g(t,x,u)+z\cdot r(t,x,u)], \tag{3.3}$$

$$t\in[0,T],\quad x\in\mathbb{R}^n,\quad z\in\mathbb{R}^d,$$

and

$$\psi(t,x,z) < g(t,x,u)+z\cdot r(t,x,u) \quad \textit{if} \quad |u|>C(1+|x|+|z|). \tag{3.4}$$

Finally, for every $t\in[0,T]$ and $x\in\mathbb{R}^n$, $z\to\psi(t,x,z)$ is continuous on $\mathbb{R}^d$.

Proof. The measurability of ψ is straightforward since, by the continuity of r and g with respect to u, we have $\psi(t,x,z)=\inf_{u\in\widehat{K}}[g(t,x,u)+z\cdot r(t,x,u)]$, $t\in[0,T]$, $x\in\mathbb{R}^n$, $z\in\mathbb{R}^d$, where $\widehat{K}$ is any countable dense subset of K.

Coming now to the proof of (3.2), we notice that, since g is nonnegative and satisfies (2.10), we have $g(t,x,u) \geq c(|u|^2 - R^2)$ (c and R are the same as in (2.10)). By (2.7) and (2.8), we have

$$g(t,x,u) + z \cdot r(t,x,u) \geq c(|u|^2 - R^2) - C|z|(1+|u|), \tag{3.5}$$

and it follows that

$$\begin{aligned}\psi(t,x,z) \geq \inf_{u\in\mathbb{R}^m}[g(t,x,u) + z\cdot r(t,x,u)] &\geq \inf_{u\in\mathbb{R}^m}[c(|u|^2 - R^2) - C|z|(1+|u|)] \\ &= -C_1|z|^2 - C_2,\end{aligned}$$

by direct computation, for suitable constants C_1 and C_2. This proves the left-hand side of (3.2). The right-hand side of (3.2) is immediate, since by (2.7),

$$\psi(t,x,z) \leq g(t,x,u) + z\cdot r(t,x,u) \leq g(t,x,u) + |z|C(1+|u|).$$

We come now to the second assertion. By (3.5) we get

$$g(t,x,u) + z\cdot r(t,x,u) \geq c|u|\left(|u| - \frac{C}{c}|z|\right) - cR^2 - C|z|. \tag{3.6}$$

On the other hand, if we fix an arbitrary $u^0 \in K$, then

$$g(t,x,u^0) + z\cdot r(t,x,u^0) \leq C(1+|x|^2+|u^0|^2) + C|z|(1+|u^0|) \leq C_3(1+|x|^2+|z|). \tag{3.7}$$

Hence, there exists a constant $C > 0$ such that, if $|u| > C(1+|x|+|z|)$, then

$$g(t,x,u) + z\cdot r(t,x,u) > g(t,x,u^0) + z\cdot r(t,x,u^0),$$

and (3.3) follows from the continuity of g and r with respect to u.

Finally, the continuity of $\psi(t,x,\cdot)$ can be easily proved, taking into account (3.3). □

Next we take an arbitrarily complete probability space $(\Omega, \mathcal{F}, \mathbb{P}^\circ)$ and a Wiener process W° in $\mathbb{R}^d$ with respect to $\mathbb{P}^\circ$. We denote by $(\mathcal{F}_t^\circ)$ the associated Brownian filtration, i.e., the filtration generated by W° and augmented by the $\mathbb{P}^\circ$-null sets of $\mathcal{F}$; $(\mathcal{F}_t^\circ)$ satisfies the usual conditions.

We introduce the forward equation

$$\begin{cases} dX_t = b(t,X_t)\,dt + \sigma(t,X_t)\,dW_t^\circ, & t\in[0,T], \\ X_0 = x, \end{cases} \tag{3.8}$$

whose solution is a continuous $(\mathcal{F}_t^\circ)$-adapted process, which exists and is unique by classical results. Next we consider the associated backward equation

$$\begin{cases} dY_t = -\psi(t,X_t,Z_t)\,dt + Z_t\,dW_t^\circ, & t\in[0,T], \\ Y_T = \phi(X_T). \end{cases} \tag{3.9}$$

The solution of (3.9) exists in the sense specified by the following proposition.

PROPOSITION 3.2. *Assume that b, σ, g, ϕ satisfy Hypothesis* 2.1. *Then there exist Borel measurable functions*

$$v : [0,T]\times\mathbb{R}^n \to \mathbb{R}, \quad \zeta : [0,T]\times\mathbb{R}^n \to \mathbb{R}^d$$

with the following property: For an arbitrarily chosen complete probability space $(\Omega, \mathcal{F}, \mathbb{P}^\circ)$ *and Wiener process* W° *in* $\mathbb{R}^d$*, denoting by* X *the solution of* (3.8)*, the processes* Y, Z *defined by*

$$Y_t = v(t, X_t), \quad Z_t = \zeta(t, X_t)$$

satisfy

$$\mathbb{E}^\circ \sup_{t\in[0,T]} |Y_t|^2 < \infty, \quad \mathbb{E}^\circ \int_0^T |Z_t|^2 \, dt < \infty;$$

moreover, Y *is continuous and nonnegative, and* $\mathbb{P}^\circ$*-a.s.,*

$$Y_t + \int_t^T Z_s \, dW_s^\circ = \phi(X_T) + \int_t^T \psi(s, X_s, Z_s) \, ds, \quad t \in [0, T].$$

Finally, this solution is the maximal solution among all the solutions (Y', Z') *of* (3.9) *satisfying*

$$\mathbb{E}^\circ \left[\sup_{t\in[0,T]} |Y'_t|^2 \right] < +\infty.$$

Proof. From Lemma 3.1, there exists a constant $C > 0$ such that

$$-C(1 + |z|^2) \le \psi(t, x, z) \le g(t, x, u^0) + C(1 + |u^0|)|z|.$$

Let us first note that

$$\mathbb{E}^\circ \sup_{t\in[0,T]} |X_t|^p < \infty \ \forall p \ge 2.$$

Next, we adopt the same strategy as that in [2] to construct a maximal solution to (3.9); i.e., for each $n \ge C$, we define the globally Lipschitz continuous function,

$$\psi_n(t, x, z) = \sup\{\psi(t, x, q) - n|q - z| : q \in \mathbb{Q}^d\},$$

which is decreasing and converges to ψ; then by (Y^n, Z^n) we denote the unique solution to the BSDE with Lipschitz coefficient ψ_n,

$$\begin{cases} dY_t^n = -\psi_n(t, X_t, Z_t^n) \, dt + Z_t^n \, dW_t^\circ, & t \in [0, T], \\ Y_T^n = \phi(X_T), \end{cases} \tag{3.10}$$

and by (Y^S, Z^S) the unique solution to the BSDE,

$$\begin{cases} dY_t^S = -[g(t, X_t, u^0) + C(1 + |u^0|)|Z_t^S|] \, dt + Z_t^S \, dW_t^\circ, & t \in [0, T], \\ Y_T^S = \phi(X_T), \end{cases} \tag{3.11}$$

where C is the same as in (3.2). We notice that, since $\psi_n(t, x, 0) \ge \psi(t, x, 0) \ge 0$, then by an application of the comparison theorem (see [4]),

$$0 \le Y_t^n \le Y_t^S.$$

Then let us introduce the following stopping time: For $k \ge 1$,

$$\tau_k = \inf\{t \in [0, T] : \max(|X_t|, Y_t^S) > k\},$$

with the convention that $\tau_k = T$ if the indicated set is empty.

Then $(Y_k^n, Z_k^n) := (Y_{t\wedge\tau_k}^n, Z_t^n 1_{t\le\tau_k})$ satisfies the following BSDE:

$$Y_k^n(t) = \xi_k^n + \int_t^T 1_{s\le\tau_k}\psi_n(s, X_s, Z_k^n(s))ds - \int_t^T Z_k^n(s)dW_s^\circ,$$

where of course $\xi_k^n = Y_k^n(T) = Y_{\tau_k}^n$.

For fixed k, Y_k^n is decreasing in n and remains bounded by k. It follows from Lemma 3 in [2] (which is a slight generalization of Proposition 2.4 in [10]) that there exists a process (Y_k, Z_k) such that Y_k is a continuous process, $E\int_0^T |Z_k(s)|^2 ds < +\infty$,

$$\lim_n \sup_{t\in[0,T]} |Y_k^n(t) - Y_k(t)| = 0, \quad \lim_n E\int_0^T |Z_k^n(t) - Z_k(t)|^2 dt = 0,$$

and (Y_k, Z_k) solves the BSDE

$$Y_k(t) = \xi_k + \int_t^T 1_{s\le\tau_k}\psi(s, X_s, Z_k(s))ds - \int_t^T Z_k(s)dW_s^\circ, \tag{3.12}$$

where $\xi_k = \inf_n Y_{\tau_k}^n$.

On the other hand, $\tau_k \le \tau_{k+1}$, and, from the definition of (Y_k^n, Z_k^n), we have

$$Y_{k+1}^n(t\wedge\tau_k) = Y_k^n(t), \quad Z_{k+1}^n(t)1_{t\le\tau_k} = Z_k^n(t).$$

Sending n to infinity, we get

$$Y_{k+1}(t\wedge\tau_k) = Y_k(t), \quad Z_{k+1}(t)1_{t\le\tau_k} = Z_k(t).$$

Now we define Y and Z on $[0,T]$ by setting

$$Y_t = Y_k(t), \quad Z_t = Z_k(t) \quad \text{if} \quad t\in[0,\tau_k].$$

For $\mathbb{P}^\circ$-a.s. ω, there exists an integer $K(\omega)$ such that for $k \ge K(\omega)$, $\tau_k(\omega) = T$. Thus Y is a continuous process, $Y_T = \phi(X_T)$, and $\int_0^T |Z_t|^2 ds < \infty$, $\mathbb{P}^\circ$-a.s.

From (3.12), (Y, Z) satisfies

$$Y_{t\wedge\tau_k} = Y_{\tau_k} + \int_{t\wedge\tau_k}^{\tau_k} \psi(s, X_s, Z(s))ds - \int_{t\wedge\tau_k}^{\tau_k} Z(s)dW_s^\circ. \tag{3.13}$$

By sending k to infinity, we deduce that (Y, Z) is a solution of (3.9) and

$$\lim_n \sup_{t\in[0,T]} |Y_t^n - Y_t| = 0, \quad \lim_n \int_0^T |Z_t^n - Z_t|^2 dt = 0, \qquad \mathbb{P}^\circ\text{-a.s.}$$

Thus $|Z^n - Z|$ converges to zero in measure $d\mathbb{P}^\circ \otimes dt$, and passing, if needed, to a subsequence (that by abuse of language we still denote Z^n), we can assume that $|Z^n - Z| \to 0$, $d\mathbb{P}^\circ \otimes dt$ almost everywhere.

Now, as ψ_n is globally Lipschitz continuous, from [4] (see also [6]) there exist Borel measurable functions

$$v^n : [0,T]\times\mathbb{R}^n \to \mathbb{R}, \quad \zeta^n : [0,T]\times\mathbb{R}^n \to \mathbb{R}^d$$

such that

$$Y_t^n = v^n(t, X_t), \quad Z_t^n = \zeta^n(t, X_t).$$

It suffices to define

$$v(t,x) = \liminf_{n\to\infty} v^n(t,x), \text{ and } \zeta(t,x) = \liminf_{n\to\infty} \zeta^n(t,x)$$

(where the second lim inf is intended coordinatewise) to get

$$Y_t = v(t, X_t), \quad Z_t = \zeta(t, X_t),$$

which implies that (v, ζ) is the Borel function we look for.

Finally, $0 \le Y_t \le Y_t^S$ implies that

$$\mathbb{E} \sup_{t\in[0,T]} |Y_t|^2 < \infty,$$

and from the equation

$$|Y_t|^2 + \int_t^{\tau_k} |Z_s|^2 ds = 2\int_t^{\tau_k} Y_s \psi(s, X_s, Z_s) ds - 2\int_t^{\tau_k} Y_s Z_s dW_s^\circ,$$

taking into consideration that

$$Y_s\psi(s,X_s,Z_s) \le Y_s(g(s,X_s,u^0) + C(1+|u^0|)|Z_s|) \le Y_s^S(g(s,X_s,u^0) + C(1+|u^0|)|Z_s|),$$

we deduce, by standard arguments, that

$$\mathbb{E}^\circ \int_0^T |Z_t|^2 dt < \infty.$$

Moreover, this solution is the maximal solution among all the solutions (Y', Z') satisfying

$$\mathbb{E}^\circ \left[\sup_{t\in[0,T]} |Y_t'|^2 \right] < +\infty,$$

and it suffices to apply Proposition 5 of [2] to deduce that $Y^n \ge Y'$ and then $Y \ge Y'$. □

4. The fundamental relation. In this section we revert to the notation introduced in the first section and still assume that Hypothesis 2.1 holds.

PROPOSITION 4.1. *Let v, ζ denote the functions in the statement of Proposition 3.2. Then for every admissible control u and for the corresponding trajectory X starting at x, we have*

$$J(u) = v(0,x) + \mathbb{E}\int_0^T [-\psi(t, X_t, \zeta(t,X_t)) + \zeta(t,X_t)\cdot r(t,X_t,u_t) + g(t,X_t,u_t)]\, dt.$$

Proof. We introduce stopping times τ_n and control processes u^n as in the proof of Proposition 2.3, and we denote by X^n the solution to (2.14). Let us define

$$W_t^n = W_t + \int_0^t r(s, X_s^n, u_s^n)\, ds.$$

From the definition of τ_n and from (2.7), it follows that

(4.1)
$$\int_0^T |r(s,X_s^n,u_s^n)|^2\, ds \le C\int_0^T (1+|u_s^n|)^2\, ds \le C\int_0^{\tau_n} (1+|u_s|)^2\, ds + C \le C + Cn.$$

Therefore defining

$$\rho_n = \exp\left(\int_0^T -r(s, X_s^n, u_s^n)\, dW_s - \frac{1}{2}\int_0^T |r(s, X_s^n, u_s^n)|^2\, ds\right),$$

the Novikov condition implies that $\mathbb{E}\rho_n = 1$. Setting $d\mathbb{P}^n = \rho_n d\mathbb{P}$, by the Girsanov theorem W^n is a Wiener process under $\mathbb{P}^n$. Let us denote by $(\mathcal{F}_t^n)$ its natural augmented filtration. Since

$$\begin{cases} dX_t^n = b(t, X_t^n)\, dt + \sigma(t, X_t^n)\, dW_t^n, & t \in [0, T], \\ X_0^n = x \end{cases}$$

has a strong solution by classical results, the process X^n is also $(\mathcal{F}_t^n)$ adapted. Let us define

$$Y_t^n = v(t, X_t^n), \quad Z_t^n = \zeta(t, X_t^n).$$

Then by Proposition 3.2, we have

$$\begin{cases} dY_t^n = Z_t^n\, dW_t^n - \psi(t, X_t^n, Z_t^n)\, dt, & t \in [0, T], \\ Y_T^n = \phi(X_T^n) \end{cases} \tag{4.2}$$

and $\mathbb{E}^n \int_0^T |Z_t^n|^2\, dt < \infty$, where $\mathbb{E}^n$ denotes expectation with respect to $\mathbb{P}^n$. It follows that

$$Y_{\tau_n}^n = \phi(X_T^n) + \int_{\tau_n}^T \psi(t, X_t^n, Z_t^n)\, dt - \int_{\tau_n}^T Z_t^n\, dW_t - \int_{\tau_n}^T Z_t^n \cdot r(t, X_t^n, u_t^n)\, dt. \tag{4.3}$$

We note that for every $p \in [1, \infty)$ we have

$$\rho_n^{-p} = \exp\left(p\int_0^T r(s, X_s^n, u_s^n)\, dW_s^n - \frac{p^2}{2}\int_0^T |r(s, X_s^n, u_s^n)|^2\, ds\right)$$
$$\cdot \exp\left(\frac{p^2 - p}{2}\int_0^T |r(s, X_s^n, u_s^n)|^2\, ds\right).$$

By (4.1) the second exponential is bounded by a constant depending on n and p, while the first one has $\mathbb{P}^n$-expectation, equal to 1. So we conclude that $\mathbb{E}_n \rho_n^{-p} < \infty$. It follows that

$$\mathbb{E}\left(\int_0^T |Z_t^n|^2 dt\right)^{1/2} = \mathbb{E}^n\left(\rho_n^{-2}\int_0^T |Z_t^n|^2 dt\right)^{1/2} \le \left(\mathbb{E}^n \rho_n^{-2}\right)^{1/2}\left(\mathbb{E}^n \int_0^T |Z_t^n|^2 dt\right)^{1/2} < \infty.$$

We conclude that the stochastic integral in (4.3) has zero expectation, and we obtain

$$\mathbb{E}Y_{\tau_n}^n = \mathbb{E}\phi(X_T^n) + \mathbb{E}\int_{\tau_n}^T [\psi(t, X_t^n, Z_t^n) - Z_t^n \cdot r(t, X_t^n, u_t^n)]\, dt.$$

Since by definition, $\psi(t, x, z) - z \cdot r(t, x, u) - g(t, x, u) \le 0$, we have

$$\mathbb{E}Y_{\tau_n}^n \le \mathbb{E}\phi(X_T^n) + \mathbb{E}\int_{\tau_n}^T g(t, X_t^n, u_t^n)\, dt. \tag{4.4}$$

Now we let $n \to \infty$. By the definition of u^n and (2.8),

$$\mathbb{E}\int_{\tau_n}^{T} g(t, X_t^n, u_t^n)\, dt = \mathbb{E}\int_0^T 1_{t>\tau_n} g(t, X_t^n, u^0)\, dt$$
$$\leq C\mathbb{E}\int_0^T 1_{t>\tau_n}(1 + |X_t^n|^2 + |u^0|^2)dt \leq C\mathbb{E}\left[(T-\tau_n)\left(1 + \sup_{t\in[0,T]} |X_t^n|^2\right)\right],$$

and the right-hand side tends to 0 by Lemma 2.4 and (2.13). Next we note that, again by (2.13), for $n \geq N(\omega)$ we have $\tau_n(\omega) = T$ and

$$\phi(X_T^n) = \phi(X_{\tau_n}^n) = \phi(X_{\tau_n}) = \phi(X_T).$$

Moreover, by (2.9),

$$|\phi(X_T^n)| \leq C(1 + |X_T^n|^2) \leq C\left(1 + \sup_{t\in[0,T]} |X_t^n|^2\right),$$

and by Lemma 2.4 the right-hand side is uniformly integrable. We deduce that $\mathbb{E}\phi(X_T^n) \to \mathbb{E}\phi(X_T)$, and from (4.4) we conclude that $\limsup_{n\to\infty} \mathbb{E}Y_{\tau_n}^n \leq \mathbb{E}\phi(X_T)$. On the other hand, for $n \geq N(\omega)$ we have $\tau_n(\omega) = T$ and

$$Y_{\tau_n}^n = Y_T^n = \phi(X_T^n) = \phi(X_T).$$

Since Y^n is positive, by the Fatou lemma, $\mathbb{E}\phi(X_T) \leq \liminf_{n\to\infty} \mathbb{E}Y_{\tau_n}^n$. We have thus proved that

$$\lim_{n\to\infty} \mathbb{E}Y_{\tau_n}^n = \mathbb{E}\phi(X_T). \tag{4.5}$$

Now we return to (4.2) and write

$$Y_{\tau_n}^n = Y_0^n + \int_0^{\tau_n} -\psi(t, X_t^n, Z_t^n)\, dt + \int_0^{\tau_n} Z_t^n\, dW_t + \int_0^{\tau_n} Z_t^n \cdot r(t, X_t^n, u_t^n)\, dt.$$

Arguing as before, we conclude that the stochastic integral has zero $\mathbb{P}$-expectation. Moreover, we have $Y_0^n = v(0,x)$, and, for $t \leq \tau_n$, also have $u_t^n = u_t$, $X_t^n = X_t$, and $Z_t^n = \zeta(t, X_t)$. Thus we obtain

$$\mathbb{E}Y_{\tau_n}^n = v(0,x) + \mathbb{E}\int_0^{\tau_n} [-\psi(t, X_t, \zeta(t, X_t)) + \zeta(t, X_t) \cdot r(t, X_t, u_t)]\, dt$$

and

$$\mathbb{E}\int_0^{\tau_n} g(t, X_t, u_t)\, dt + \mathbb{E}Y_{\tau_n}^n$$
$$= v(0,x) + \mathbb{E}\int_0^{\tau_n} [-\psi(t, X_t, \zeta(t, X_t)) + \zeta(t, X_t) \cdot r(t, X_t, u_t) + g(t, X_t, u_t)]\, dt.$$

Noting that $-\psi(t,x,z) + z \cdot r(t,x,u) + g(t,x,u) \geq 0$ and recalling that $g(t,x,u) \geq 0$, by (4.5) and the monotone convergence theorem, we obtain for $n \to \infty$,

$$\mathbb{E}\int_0^T g(t, X_t, u_t)\, dt + \mathbb{E}\phi(X_T) = v(0,x)$$
$$+ \mathbb{E}\int_0^T [-\psi(t, X_t, \zeta(t, X_t)) + \zeta(t, X_t) \cdot r(t, X_t, u_t) + g(t, X_t, u_t)]\, dt,$$

which gives the required conclusion. □

COROLLARY 4.2. *For every admissible control u and any initial datum x, we have $J(u) \geq v(0,x)$, and the equality holds if and only if the following feedback law holds $\mathbb{P}$-a.s. for almost every $t \in [0,T]$:*

$$\psi(t, X_t, \zeta(t, X_t)) = \zeta(t, X_t) \cdot r(t, X_t, u_t) + g(t, X_t, u_t),$$

where X is the trajectory starting at x and corresponding to control u.

5. Existence of optimal controls: The closed loop equation. Let us consider again the functions b, σ, g, ϕ satisfying the assumptions in Hypothesis 2.1. We recall the definition of the Hamiltonian function:

$$\psi(t,x,z) = \inf_{u \in K} [g(t,x,u) + z \cdot r(t,x,u)], \qquad t \in [0,T], x \in \mathbb{R}^n, \quad z \in \mathbb{R}^d. \tag{5.1}$$

LEMMA 5.1. *There exists a Borel measurable function $\gamma : [0,T] \times \mathbb{R}^n \times \mathbb{R}^d \to K$ such that*

(5.2)
$$\psi(t,x,z) = g(t,x,\gamma(t,x,z)) + z \cdot r(t,x,\gamma(t,x,z)), \qquad t \in [0,T], \quad x \in \mathbb{R}^n, \quad z \in \mathbb{R}^d.$$

Moreover, there exists a constant $C > 0$ such that

$$|\gamma(t,x,z)| \leq C(1 + |x| + |z|). \tag{5.3}$$

Proof. Consider the function $F(t,x,z,u) = g(t,x,u) + z \cdot r(t,x,u)$, $t \in [0,T]$, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^d$. Clearly F is a Carathéodory map (that is, $F(t,x,z,\cdot)$ is continuous for all $t \in [0,T]$, $x \in \mathbb{R}^n$, and $z \in \mathbb{R}^d$, and $F(\cdot,\cdot,\cdot,u)$ is Borel measurable for all $u \in K$; see [1, p. 311]). By (3.3) we have

$$\psi(t,x,z) \in \{F(t,x,z,u) : u \in K\} \quad \forall t \in [0,T], \quad x \in \mathbb{R}^n, \quad z \in \mathbb{R}^d.$$

Thus by the Filippov theorem (see, e.g., [1, Thm. 8.2.10, p. 316]) there exists a Borel measurable map $\gamma : [0,T] \times \mathbb{R}^n \times \mathbb{R}^d \to K$ such that $F(t,x,z,\gamma(t,x,z)) = \psi(t,x,z)$; see [3] as well. The fact that $|\gamma(t,x,z)| \leq C(1+|x|+|z|)$ is an immediate consequence of (3.4). □

Next we address the problem of finding a weak solution to the so-called closed loop equation. We define

$$\underline{u}(t,x) = \gamma(t,x,\zeta(t,x)), \qquad t \in [0,T], \quad x \in \mathbb{R}^n,$$

where ζ is defined in Proposition 3.2. The closed loop equation is

$$\begin{cases} dX_t = b(t,X_t)\,dt + \sigma(t,X_t)\,[dW_t + r(t,X_t,\underline{u}(t,X_t))\,dt], \qquad t \in [0,T], \\ X_0 = x. \end{cases} \tag{5.4}$$

By a weak solution we mean a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_t)$ satisfying the usual conditions, a Wiener process W in $\mathbb{R}^d$ with respect to $\mathbb{P}$ and $(\mathcal{F}_t)$, and a continuous $(\mathcal{F}_t)$-adapted process X with values in $\mathbb{R}^n$ satisfying, $\mathbb{P}$-a.s.,

$$\int_0^T |\underline{u}(t,X_t)|^2\,dt < \infty, \tag{5.5}$$

and such that (5.4) holds. We note that by (2.7) it also follows that

$$\int_0^T |r(t,X_t,\underline{u}(t,X_t))|\,dt < \infty, \quad \mathbb{P}\text{-a.s.},$$

so that (5.4) makes sense.

PROPOSITION 5.2. *Assume that b, σ, g, ϕ satisfy Hypothesis 2.1. Then there exists a weak solution of the closed loop equation, satisfying in addition*

$$\mathbb{E}\int_0^T |\underline{u}(t, X_t)|^2 \, dt < \infty. \tag{5.6}$$

Proof. Let us take an arbitrary complete probability space $(\Omega, \mathcal{F}, \mathbb{P}^\circ)$ and a Wiener process W° in $\mathbb{R}^d$ with respect to $\mathbb{P}^\circ$. Let $(\mathcal{F}_t^\circ)$ be the associated Brownian filtration. We define the process X as the solution of the equation

$$\begin{cases} dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, dW_t^\circ, & t \in [0, T], \\ X_0 = x. \end{cases} \tag{5.7}$$

The solution is a continuous $(\mathcal{F}_t^\circ)$-adapted process, which exists and is unique by classical results. Moreover, it satisfies $\mathbb{E}^\circ \sup_{t\in[0,T]} |X_t|^p < \infty$ for every $p \in [1, \infty)$.

By Proposition 3.2, setting

$$Y_t = v(t, X_t), \quad Z_t = \zeta(t, X_t),$$

the following backward equation holds:

$$\begin{cases} dY_t = -\psi(t, X_t, Z_t)\, dt + Z_t\, dW_t^\circ, & t \in [0, T], \\ Y_T = \phi(X_T), \end{cases}$$

and we have

$$\mathbb{E}^\circ \int_0^T |Z_t|^2 \, dt < \infty. \tag{5.8}$$

By (2.7) we have $|r(t, X_t, \underline{u}(t, X_t))| \le C(1 + |\underline{u}(t, X_t)|)$, and by (5.3),

$$|\underline{u}(t, X_t)| = |\gamma(t, X_t, \zeta(t, X_t))| \le C(1 + |X_t| + |\zeta(t, X_t)|) = C(1 + |X_t| + |Z_t|). \tag{5.9}$$

Now let us define stopping times

$$\tau_n = \inf\left\{ t \in [0, T] \ : \ \int_0^t |\underline{u}(s, X_s)|^2 ds > n \right\},$$

with the convention that $\tau_n = T$ if the indicated set is empty. By (5.8) and (5.9), for $\mathbb{P}^\circ$-a.s. $\omega \in \Omega$, there exists an integer $N(\omega)$ depending on ω such that $\tau_n(\omega) = T$ for $n \ge N(\omega)$. Let us fix $u^0 \in K$, and for every n let us define

$$u_t^n = \underline{u}(t, X_t)\, 1_{t\le\tau_n} + u^0\, 1_{t>\tau_n},$$

$$M_t^n = \exp\left(\int_0^t r(s, X_s, u_s^n)\, dW_s^\circ - \frac{1}{2}\int_0^t |r(s, X_s, u_s^n)|^2 \, ds \right),$$

$$M_t = \exp\left(\int_0^t r(s, X_s, \underline{u}(s, X_s))\, dW_s^\circ - \frac{1}{2}\int_0^t |r(s, X_s, \underline{u}(s, X_s))|^2 \, ds \right),$$

$$W_t^n = W_t^\circ - \int_0^t r(s, X_s, u_s^n)\, ds,$$

$$W_t = W_t^\circ - \int_0^t r(s, X_s, \underline{u}(s, X_s))\, ds.$$

By the previous estimates, M^n, M, W^n, and W are well defined; moreover,

$$\int_0^T |r(s, X_s, u_s^n) - r(s, X_s, \underline{u}(s, X_s))|^2 \, ds \to 0, \qquad \mathbb{P}^\circ\text{-a.s.},$$

and consequently $M_T^n \to M_T$ in probability and $\sup_{t\in[0,T]} |W_t^n - W_t| \to 0$, $\mathbb{P}^\circ$-a.s. We will conclude the proof by showing that there exists a probability $\mathbb{P}$ such that W is a Wiener process with respect to $\mathbb{P}$ and $(\mathcal{F}_t^\circ)$.

The definition of τ_n and the Novikov condition implies that $\mathbb{E}^\circ M_T^n = 1$. Setting $d\mathbb{P}^n = M_T^n d\mathbb{P}^\circ$, by the Girsanov theorem W^n is a Wiener process with respect to $\mathbb{P}^n$ and $(\mathcal{F}_t^\circ)$. Writing the backward equation with respect to W^n, we obtain

$$Y_{\tau_n} = Y_0 + \int_0^{\tau_n} -\psi(t, X_t, Z_t) \, dt + \int_0^{\tau_n} Z_t \, dW_t^n + \int_0^{\tau_n} Z_t \cdot r(t, X_t, u_t^n) \, dt.$$

Arguing as in the proof of Proposition 4.1, we conclude that the stochastic integral has zero expectation with respect to $\mathbb{P}^n$. Taking into account that $u_t^n = \underline{u}(t, X_t)$ for $t \le \tau_n$, we obtain

$$\begin{aligned}
&\mathbb{E}^n Y_{\tau_n} + \mathbb{E}^n \int_0^{\tau_n} g(t, X_t, \underline{u}(t, X_t)) \, dt \\
&\quad = Y_0 + \mathbb{E}^n \int_0^{\tau_n} [-\psi(t, X_t, Z_t) + Z_t \cdot r(t, X_t, \underline{u}(t, X_t)) + g(t, X_t, \underline{u}(t, X_t))] \, dt \\
&\quad = Y_0,
\end{aligned}$$

with the last equality coming from the definition of $\underline{u}$. Recalling that Y is nonnegative, it follows that

$$\mathbb{E}^n \int_0^{\tau_n} g(t, X_t, \underline{u}(t, X_t)) \, dt \le C$$

for some constant C independent of n. By (2.10) we also deduce

$$\mathbb{E}^n \int_0^{\tau_n} |\underline{u}(t, X_t)|^2 \, dt \le C. \tag{5.10}$$

Next we prove that the family $\{M_T^n, \ n = 1, 2, \dots\}$ is uniformly integrable by showing that $\mathbb{E}^\circ[M_T^n \, 1_{\{M_T^n > c\}}] \to 0$ as $c \to \infty$, uniformly with respect to n. We have

$$\mathbb{E}^\circ[M_T^n \, 1_{\{M_T^n > c\}}] = \mathbb{E}^\circ[M_T^n \, 1_{\{M_T^n > c, \tau_n = T\}}] + \mathbb{E}^\circ[M_T^n \, 1_{\{M_T^n > c, \tau_n < T\}}]. \tag{5.11}$$

The first term in the right-hand side tends to 0 uniformly with respect to n, since

$$\mathbb{E}^\circ[M_T^n \, 1_{\{M_T^n > c, \tau_n = T\}}] = \mathbb{E}^\circ[M_T \, 1_{\{M_T > c, \tau_n = T\}}] \le \mathbb{E}^\circ[M_T \, 1_{\{M_T > c\}}] \to 0,$$

due to the fact that the equality $\mathbb{E}^\circ M_T^n = 1$ and the Fatou lemma imply that $\mathbb{E}^\circ M_T \le 1$. The second term in the right-hand side of (5.11) can be estimated as follows:

$$\begin{aligned}
\mathbb{E}^\circ[M_T^n \, 1_{\{M_T^n > c, \tau_n < T\}}] &\le \mathbb{E}^\circ[M_T^n \, 1_{\tau_n < T\}}] = \mathbb{P}^n(\tau_n < T) \\
&\le \mathbb{P}^n \left(\int_0^{\tau_n} |\underline{u}(t, X_t)|^2 dt > n \right) \le \frac{1}{n} \mathbb{E}^n \int_0^{\tau_n} |\underline{u}(t, X_t))|^2 dt \le \frac{C}{n},
\end{aligned}$$

with the last inequality coming from (5.10). The required uniform integrability follows immediately. Recalling that $M_T^n \to M_T$ in probability, we conclude that $\mathbb{E}^\circ |M_T^n - M_T| \to 0$, and in particular $\mathbb{E}^\circ M_T = 1$, and M is a $\mathbb{P}^\circ$-martingale. Thus we can define

a probability $\mathbb{P}$ by setting $d\mathbb{P} = M_T d\mathbb{P}^\circ$, and by the Girsanov theorem we conclude that W is a Wiener process with respect to $\mathbb{P}$ and $(\mathcal{F}_t^\circ)$.

It remains to prove (5.6). We define stopping times

$$\sigma_n = \inf\left\{t \in [0,T] : \int_0^t |Z_s|^2 ds > n\right\},$$

with the convention that $\sigma_n = T$ if the indicated set is empty. Writing the backward equation with respect to W, we obtain

$$Y_{\sigma_n} = Y_0 + \int_0^{\sigma_n} -\psi(t, X_t, Z_t)\, dt + \int_0^{\sigma_n} Z_t\, dW_t + \int_0^{\sigma_n} Z_t \cdot r(t, X_t, \underline{u}(t, X_t))\, dt,$$

from which we deduce that

$$\begin{aligned}
&\mathbb{E}Y_{\sigma_n} + \mathbb{E}\int_0^{\sigma_n} g(t, X_t, \underline{u}(t, X_t))\, dt \\
&\quad = Y_0 + \mathbb{E}\int_0^{\sigma_n} [-\psi(t, X_t, Z_t) + Z_t \cdot r(t, X_t, \underline{u}(t, X_t)) + g(t, X_t, \underline{u}(t, X_t))]\, dt \\
&\quad = Y_0,
\end{aligned}$$

with the last equality coming from the definition of $\underline{u}$. Recalling that Y is nonnegative, it follows that

$$\mathbb{E}\int_0^{\sigma_n} g(t, X_t, \underline{u}(t, X_t))\, dt \le C$$

for some constant C independent of n. By (2.10) and by sending n to infinity, we finally prove (5.6). □

Corollary 5.3. *By Corollary* 4.2 *it immediately follows that if X is the solution to* (5.4) *and we set $u_s^\sharp = \underline{u}(s, X_s)$, then $J(x, u^\sharp) = v(0, x)$, and consequently X is an optimal state, $u_s^\sharp$ is an optimal control, and $\underline{u}$ is an optimal feedback.*

Next we prove uniqueness in law for the closed loop equation. We remark that condition (5.5) is part of our definition of a weak solution.

Proposition 5.4. *Assume that b, σ, g, ϕ satisfy Hypothesis* 2.1. *Fix $\gamma : [0,T] \times \mathbb{R}^n \times \mathbb{R}^d \to K$ satisfying* (5.2) (*and consequently* (5.3)) *and let $\underline{u}(t,x) = \gamma(t, x, \zeta(t,x))$. Then the weak solution of the closed loop equation* (5.4) *is unique in law.*

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$, $(\mathcal{F}_t)$, W, X be a weak solution of (5.4).

Let us define

$$M_T = \exp\left(-\int_0^T r(s, X_s, \underline{u}(s, X_s))\, dW_s - \frac{1}{2}\int_0^T |r(s, X_s, \underline{u}(s, X_s))|^2\, ds\right),$$

$$W_t^\circ = W_t + \int_0^t r(s, X_s, \underline{u}(s, X_s))\, ds.$$

By (2.7) and (5.5), M_T and W° are well defined. We claim that $\mathbb{E}M_T = 1$. Assuming the claim for a moment, and setting $d\mathbb{P}^\circ = M_T d\mathbb{P}$, by the Girsanov theorem W° is a Wiener process under $\mathbb{P}^\circ$, X solves

$$\begin{cases} dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, dW_t^\circ, & t \in [0,T], \\ X_0 = x, \end{cases}$$

and

$$M_T = \exp\left(- \int_0^T r(s, X_s, \underline{u}(s, X_s))\, dW_s^\circ + \frac{1}{2} \int_0^T |r(s, X_s, \underline{u}(s, X_s))|^2\, ds \right).$$

By the Lipschitz and linear growth conditions on b and σ (see (2.3), (2.4), (2.5)) the law of (X, W°) under $\mathbb{P}^\circ$ is uniquely determined by b, σ, x. Taking into account the last displayed formula, we conclude that the law of (X, W°, M_T) under $\mathbb{P}^\circ$ is also uniquely determined, and consequently so is the law of X under $\mathbb{P}$.

To conclude the proof it remains to show that $\mathbb{E}M_T = 1$. We define stopping times

$$\tau_n = \inf \left\{ t \in [0, T] \ : \ \int_0^t |\underline{u}(s, X_s)|^2 ds > n \right\},$$

with the convention that $\tau_n = T$ if the indicated set is empty. By (5.5), for $\mathbb{P}$-almost every $\omega \in \Omega$, there exists an integer $N(\omega)$ depending on ω such that $\tau_n(\omega) = T$ for $n \geq N(\omega)$. Let us fix $u^0 \in K$, and for every n, let us define

$$u_t^n = \underline{u}(t, X_t)\, 1_{t \leq \tau_n} + u^0\, 1_{t > \tau_n},$$

$$M_T^n = \exp\left(- \int_0^T r(s, X_s, u_s^n)\, dW_s - \frac{1}{2} \int_0^T |r(s, X_s, u_s^n)|^2\, ds \right).$$

By (2.7) and the definition of τ_n, the Novikov condition shows that $\mathbb{E}M_T^n = 1$. Moreover, we have

$$\int_0^T |r(s, X_s, u_s^n) - r(s, X_s, \underline{u}(s, X_s))|^2\, ds \to 0, \qquad \mathbb{P}\text{-a.s.},$$

and consequently $M_T^n \to M_T$ in probability. In order to conclude the proof it is therefore enough to show that the family $\{M_T^n,\, n = 1, 2, \dots\}$ is uniformly integrable.

To prepare for this, let us set $d\mathbb{P}^n = M_T^n d\mathbb{P}$ and note that, by the Girsanov theorem, the process

$$W_t^n = W_t + \int_0^t r(s, X_s, u_s^n)\, ds$$

is a Wiener process under $\mathbb{P}^n$. Since X solves

$$\begin{cases} dX_t = b(t, X_t)\, dt + \sigma(t, X_t)\, dW_t^n, & t \in [0, T], \\ X_0 = x, \end{cases}$$

it follows that X is adapted to the Brownian filtration $(\mathcal{F}_t^n)$ associated to W^n, and its law under $\mathbb{P}^n$ is uniquely determined by b, σ, x. In particular, the quantities

$$C' := \mathbb{E}^n \int_0^T |X_t|^2\, dt, \qquad C'' := \mathbb{E}^n \int_0^T |\zeta(t, X_t)|^2\, dt$$

do not depend on n (here $\mathbb{E}^n$ denotes, of course, the expectation with respect to $\mathbb{P}^n$). C' is clearly finite. By Proposition 3.2, setting $Z_t = \zeta(t, X_t)$, we have

$$\mathbb{E}^n \int_0^T |\zeta(t, X_t)|^2\, dt = \mathbb{E}^n \int_0^T |Z_t|^2\, dt < \infty,$$

and it follows that C'' is also finite.

Now let us prove the uniform integrability of the family $\{M_T^n,\ n = 1, 2, \dots\}$ by showing that $\mathbb{E}[M_T^n\ 1_{\{M_T^n > c\}}] \to$ as $c \to \infty$, uniformly with respect to n. We have

$$\mathbb{E}[M_T^n\ 1_{\{M_T^n > c\}}] = \mathbb{E}[M_T^n\ 1_{\{M_T^n > c, \tau_n = T\}}] + \mathbb{E}[M_T^n\ 1_{\{M_T^n > c, \tau_n < T\}}]. \tag{5.12}$$

The first term in the right-hand side tends to 0 uniformly with respect to n, since

$$\mathbb{E}[M_T^n\ 1_{\{M_T^n > c, \tau_n = T\}}] = \mathbb{E}[M_T\ 1_{\{M_T > c, \tau_n = T\}}] \le \mathbb{E}[M_T\ 1_{\{M_T > c\}}] \to 0,$$

due to the fact that the equality $\mathbb{E}M_T^n = 1$ and the Fatou lemma imply that $\mathbb{E}M_T \le 1$. The second term in the right-hand side of (5.12) can be estimated as follows:

$$\mathbb{E}[M_T^n\ 1_{\{M_T^n > c, \tau_n < T\}}] \le \mathbb{E}[M_T^n\ 1_{\tau_n < T}] = \mathbb{P}^n(\tau_n < T)$$
$$\le \mathbb{P}^n\left(\int_0^{\tau_n} |\underline{u}(t, X_t)|^2 dt > n\right) \le \frac{1}{n}\mathbb{E}^n \int_0^{\tau_n} |\underline{u}(t, X_t)|^2 dt \le \frac{1}{n}\mathbb{E}^n \int_0^T |\underline{u}(t, X_t)|^2 dt.$$

By (5.3) we have

$$|\underline{u}(t, X_t)|^2 = |\gamma(t, X_t, \zeta(t, X_t))|^2 \le C(1 + |X_t|^2 + |\zeta(t, X_t)|^2)$$

for some constant C, and it follows that

$$\mathbb{E}[M_T^n\ 1_{\{M_T^n > c, \tau_n < T\}}] \le \frac{C}{n}\mathbb{E}^n \int_0^T (1 + |X_t|^2 + |\zeta(t, X_t)|^2)\ dt = \frac{C}{n}(T + C' + C''),$$

with C' and C'' defined as above. The required uniform integrability follows immediately, and this concludes the proof. □

Acknowledgments. The authors would like to thank two anonymous referees and the associate editor for their careful reading and helpful comments.

REFERENCES

[1] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Systems Control Found. Appl. 2, Birkhäuser Boston, Boston, MA, 1990.

[2] P. Briand and Y. Hu, *BSDE with quadratic growth and unbounded terminal value*, Probab. Theory Related Fields, to appear.

[3] C. Dellacherie and P. A. Meyer, *Probabilities and Potential*, North-Holland, Amsterdam, 1978.

[4] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.

[5] N. El Karoui and S. Hamadène, *BSDEs and risk-sensitive control, zero-sum and nonzero-sum game problems of stochastic functional differential equations*, Stochastic Process. Appl., 107 (2003), pp. 145–169.

[6] M. Fuhrman, *A class of stochastic optimal control problems in Hilbert spaces: BSDEs and optimal control laws, state constraints, conditioned processes*, Stochastic Process. Appl., 108 (2003), pp. 263–298.

[7] M. Fuhrman and G. Tessitore, *Existence of optimal stochastic controls and global solutions of forward-backward stochastic differential equations*, SIAM J. Control Optim., 43 (2004), pp. 813–830.

[8] F. Gozzi, *Global regular solutions of second order Hamilton-Jacobi equations in Hilbert spaces with locally Lipschitz nonlinearities*, J. Math. Anal. Appl., 198 (1996), pp. 399–443.

[9] S. Hamadène and J.-P. Lepeltier, *Backward equations, stochastic control and zero-sum stochastic differential games*, Stochastics Stochastics Rep., 54 (1995), pp. 221–231.

[10] M. Kobylanski, *Backward stochastic differential equations and partial differential equations with quadratic growth*, Ann. Probab., 28 (2000), pp. 558–602.

[11] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, Systems Control Lett., 14 (1990), pp. 55–61.

[12] J. Yong and X. Y. Zhou, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1297–1313

ALMOST GLOBAL STOCHASTIC STABILITY*

RAMON VAN HANDEL†

Abstract. We develop a method to prove almost global stability of stochastic differential equations in the sense that almost every initial point (with respect to the Lebesgue measure) is asymptotically attracted to the origin with unit probability. The method can be viewed as a dual to Lyapunov's second method for stochastic differential equations and extends the deterministic result of [A. Rantzer, *Syst. Control Lett.*, 42 (2001), pp. 161–168]. The result can also be used in certain cases to find stabilizing controllers for stochastic nonlinear systems using convex optimization. The main technical tool is the theory of stochastic flows of diffeomorphisms.

Key words. stochastic stability, stochastic flows, nonlinear stochastic control

AMS subject classifications. 34F05, 60H10, 93C10, 93D15, 93E15

DOI. 10.1137/040618850

1. Introduction.

Lyapunov's second method or the method of Lyapunov functions, though developed in the late 19th century, remains one of the most important tools in the study of deterministic differential equations. The power of the method lies in the fact that an important qualitative property of a differential equation, the stability of an equilibrium point, can be proved without solving the equation explicitly. The theory was generalized to stochastic differential equations in the 1960s with fundamental contributions by Has'minskiĭ [10] and Kushner [15].

Lyapunov's method also underlies many important applications in the area of nonlinear control [11]. Finding optimal controls for nonlinear systems is generally an intractable problem, but often a solution can be found which stabilizes the system. Unlike in deterministic control theory, where nonlinear control is now a major field, there are very few results on stochastic nonlinear control. It is only recently that stochastic versions of the classical stabilization results of Jurdjevic-Quinn, Artstein, and Sontag were developed by Florchinger [7, 8, 9] and backstepping designs for stochastic strict-feedback systems were developed by Deng and Krstić [5] and Deng Krstić, and Williams [6].

In this paper we will not consider stochastic stability in the sense of Has'minskiĭ; rather, we ask the following question: for a given Itô stochastic differential equation on $\mathbb{R}^n$, can we prove that for almost every initial state (with respect to the Lebesgue measure on $\mathbb{R}^n$) the solution of the equation converges to the origin almost surely as $t \to \infty$, i.e., is the origin *almost* globally stable? This notion of stability is clearly weaker than global stability in the sense of Has'minskiĭ, but is of potential interest in many cases in which global stability may not be attained.

Our main result is a Lyapunov-type theorem that can be used to prove almost global stability of stochastic differential equations, extending the deterministic result of Rantzer [23]. The theorem has several remarkable properties. It can be viewed as a "dual" to Lyapunov's second method in the following sense: whereas the Lyapunov

*Received by the editors November 13, 2004; accepted for publication (in revised form) April 3, 2006; published electronically October 3, 2006. This work was supported by the ARO under grant DAAD19-03-1-0073.

http://www.siam.org/journals/sicon/45-4/61885.html

†Departments of Physics and Control & Dynamical Systems, California Institute of Technology 266-33, Pasadena, CA 91125 (ramon@its.caltech.edu).

condition reads $\mathscr{L}V < 0$, where $\mathscr{L}$ is the characteristic operator of the stochastic differential equation and V is the Lyapunov function, the condition that guarantees almost global stability reads $\mathscr{L}^*D < 0$, where $\mathscr{L}^*$ is the formal adjoint of $\mathscr{L}$ (also known as the Fokker–Planck operator). Hence the relation between the two theorems recalls the duality between densities and expectations which is prevalent throughout the theory of stochastic processes.

A further interesting property is the following convexity property. Suppose we are given an Itô equation of the form

$$x_t = x + \int_s^t (X_0(x_\tau) + u(x_\tau)Y(x_\tau))\,d\tau + \sum_{k=1}^m \int_s^t X_k(x_\tau)\,dW_\tau^k,$$

where $X_k(0) = 0$ for $k = 0, \ldots, m$ and $u(x)$ is a state feedback control. The goal is to design $u(x)$ such that the origin is almost globally stable. It is easily verified that the set of pairs of functions $(D(x), u(x)D(x))$ which satisfy $\mathscr{L}^*D < 0$ is convex. Note that the classical Lyapunov condition $\mathscr{L}V < 0$ is not convex.

The above convexity property was used in the deterministic case by Prajna, Parrilo, and Rantzer [22] to formulate the search for almost globally stabilizing controllers as a convex optimization problem, provided that X_k, Y, D, and u are rational functions. The method applies equally to the stochastic case and thus provides a tool for computer-aided design of stochastic nonlinear controllers.

It must be emphasized that almost global stability is a global property of the flow which places very few restrictions on the local behavior near the origin. In particular, local stability is not implied.[1] A very fruitful approach to studying the local dynamical behavior of stochastic differential equations (and more general random dynamical systems) is developed by Arnold [2]. First, the flow associated to the stochastic differential equation is linearized; then Oseledec's multiplicative ergodic theorem is used to provide a suitable "time-averaged" notion of the eigenvalues of the linearized flow. To prove almost global stability we do not linearize the flow, though the proofs still rely on the flow of diffeomorphisms generated by the stochastic equation. We refer to [1] for an introduction to the dynamical approach to stochastic analysis.

This paper is organized as follows. In section 2 we fix the notation that will be used in the remainder of the paper. In section 3 we reproduce the deterministic result of Rantzer [23] with a significantly different proof that generalizes to the stochastic case. In section 4 we review the theory of stochastic flows of diffeomorphisms generated by stochastic differential equations. Section 5 is devoted to the statement and proof of our main result for the case of globally Lipschitz continuous coefficients. In section 6 the main result is extended to cases in which the global Lipschitz condition does not necessarily hold. A few examples are given in section 7. Finally, in section 8 we discuss the application to control synthesis using convex optimization.

2. Notation. Throughout this article we will consider (stochastic) differential equations in $\mathbb{R}^n$. The Lebesgue measure on $\mathbb{R}^n$ will be denoted by μ. $\mathbb{R}_+$ denotes the nonnegative real numbers and $\mathbb{Z}_+$ the nonnegative integers.

We remind the reader of the following definitions: For $0 < \alpha \leq 1$, a function $f : X \to Y$ from a normed space $(X, \|\cdot\|)$ to a normed space $(Y, \|\cdot\|_Y)$ is called

[1] The term "stability" seems a bit of a misnomer; despite that almost all points converge to the origin, a trajectory that starts close to the origin could move very far from the origin before converging to it. We have used the term that has been used in the deterministic literature, e.g., [18].

globally Hölder continuous of order α if there exists a positive constant C such that

$$\|f(x) - f(y)\|_Y \le C\|x - y\|^\alpha \qquad \forall x, y \in X. \tag{2.1}$$

f is locally Hölder continuous of order α if it satisfies the condition (2.1) on every bounded subset of X. f is called globally (locally) Lipschitz continuous if it is globally (locally) Hölder continuous of order 1. f is called a $C^{k,\alpha}$ function if it is k times continuously differentiable and the kth derivatives are locally Hölder continuous of order α for some $k \in \mathbb{Z}_+$ and $0 < \alpha \le 1$.

3. The deterministic case. In this section we give a new proof of Rantzer's theorem [23], which is a deterministic counterpart of our main result. Our proof demonstrates the main features of the proof of the stochastic result in the simpler deterministic case.

The following lemma is similar to Lemma A.1 in [23], and we omit the proof.

LEMMA 3.1. *Let $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ be globally Lipschitz continuous, let $S \subset \mathbb{R}^n$ be an invariant set of $\dot{x}(t) = f(x(t))$, and let $Z \subset S$ be μ-measurable. Let $D \in C^1(S, \mathbb{R})$ be integrable on Z. Then*

$$\int_{\phi_t^{-1}(Z)} D(x)\,dx = \int_Z D(x)\,dx - \int_0^t \int_{\phi_\tau^{-1}(Z)} [\nabla \cdot (fD)](x)\,dx\,d\tau, \tag{3.1}$$

where $\phi_t : \mathbb{R}^n \to \mathbb{R}^n$ is the flow of f.

THEOREM 3.2. *Let $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ be globally Lipschitz continuous and let $f(0) = 0$. Suppose there exists $D \in C^1(\mathbb{R}^n \backslash \{0\}, \mathbb{R}_+)$ such that D is integrable on $\{x \in \mathbb{R}^n : |x| > 1\}$ and $[\nabla \cdot (fD)](x) > 0$ for μ-almost all x. Then for μ-almost all initial states $x(0)$ the solution of $\dot{x}(t) = f(x(t))$ tends to the origin as $t \to \infty$.*

Proof. Let $S = \mathbb{R}^n \backslash \{0\}$, $\varepsilon > 0$ and $Z = \{x \in \mathbb{R}^n : |x| > \varepsilon\}$. Note that $\phi_t(x)$ is a diffeomorphism for every $t \in \mathbb{R}$; hence $\phi_t(x)$ is one-to-one, and as $\phi_t(0) = 0$, $t \in \mathbb{R}$, is a solution of $\dot{x}(t) = f(x(t))$ there can be no $x \in S$ such that $\phi_t(x) = 0$ for some $t \in \mathbb{R}$. We have thus verified the invariance of S under the flow $\phi_t(x)$. We now invoke Lemma 3.1. As $D(x)$ is nonnegative, expression (3.1) is also nonnegative. Furthermore, (3.1) is finite because D is integrable on Z, and is nonincreasing due to $[\nabla \cdot (fD)](x) \ge 0$. By monotone convergence the limit as $t \to \infty$ exists and is finite. Hence

$$\int_0^\infty \mathbb{D}(\phi_\tau^{-1}(Z))\,d\tau < \infty, \qquad \mathbb{D}(A) = \int_A [\nabla \cdot (fD)](x)\,dx.$$

Note that the assumption $[\nabla \cdot (fD)](x) \ge 0$ implies that $\mathbb{D}$ is a measure on S. The measure space $(S, \mathbb{D})$ is σ-finite as $\mathbb{D}(\{x \in S : \frac{1}{k} < |x| < k\}) < \infty$ for all $k > 1$ and $\bigcup_{n=2}^\infty \{x \in S : \frac{1}{k} < |x| < k\} = S$.

We now fix some $m \in \mathbb{N}$ and divide the halfline into bins $S_k^m = [(k-1)2^{-m}, k2^{-m}]$, $k \in \mathbb{N}$. From each bin we choose a time $t_k^m \in S_k^m$ such that

$$\mathbb{D}(\phi_{t_k^m}^{-1}(Z)) \le \inf_{t \in S_k^m} \mathbb{D}(\phi_t^{-1}(Z)) + 2^{-k}.$$

For fixed m, we denote this discrete grid by $T_m = \{t_k^m : k \in \mathbb{N}\}$. We now have

$$2^{-m} \sum_{k=1}^\infty \mathbb{D}(\phi_{t_k^m}^{-1}(Z)) \le 2^{-m} + \int_0^\infty \mathbb{D}(\phi_\tau^{-1}(Z))\,d\tau < \infty.$$

As $\mathbb{D}$ is σ-finite we can now apply the Borel–Cantelli lemma, which gives

$$\mathbb{D}\left(\limsup_{k\to\infty}\phi_{t_k^m}^{-1}(Z)\right)=\mu\left(\limsup_{k\to\infty}\phi_{t_k^m}^{-1}(Z)\right)=0,$$

where the first equality follows as $[\nabla\cdot(fD)](x)>0$ μ-a.e. implies $\mu\ll\mathbb{D}$. Consequently

$$\mu\left(\bigcup_{m=1}^{\infty}\limsup_{t\in T_m}\phi_t^{-1}(Z)\right)\le\sum_{m=1}^{\infty}\mu\left(\limsup_{t\in T_m}\phi_t^{-1}(Z)\right)=0.$$

We have thus shown that the set of initial states x for which there are, for some m, infinitely many times $t\in T_m$ such that $\phi_t(x)\in Z$ has Lebesgue measure zero.

We now claim that if $\limsup_{t\to\infty}|\phi_t(x)|>\varepsilon$, then we can choose m so that there are infinitely many times t in T_m such that $\phi_t(x)\in Z$. The statement is trivial if also $\liminf_{t\to\infty}|\phi_t(x)|>\varepsilon$; let us thus assume that $\liminf_{t\to\infty}|\phi_t(x)|\le\varepsilon$. We will need the following result. Due to the global Lipschitz condition and $f(0)=0$, we have

$$|\phi_t(x)|\le|\phi_s(x)|+\int_s^t|f(\phi_\sigma(x))|\,d\sigma\le|\phi_s(x)|+C\int_s^t|\phi_\sigma(x)|\,d\sigma \tag{3.2}$$

for some constant $C>0$. Thus Gronwall's lemma gives $|\phi_t(x)|\le|\phi_s(x)|\,e^{C(t-s)}$. Now note that $\liminf_{t\to\infty}|\phi_t(x)|\le\varepsilon<\limsup_{t\to\infty}|\phi_t(x)|$ implies that there exist $\varepsilon''>\varepsilon'>\varepsilon$ such that (i) there are infinitely many upcrossings of the curve $|\phi_t(x)|$ through ε', and (ii) $|\phi_t(x)|$ crosses ε'' infinitely often. Denote by t'' a time such that $|\phi_{t''}(x)|=\varepsilon''$ and by t' the latest time previous to t'' that $|\phi_{t'}(x)|=\varepsilon'$. Then clearly $t''-t'\ge\frac{1}{C}\log\frac{\varepsilon''}{\varepsilon'}$. As this happens infinitely often, we conclude that $\phi_t(x)$ infinitely often spends a time in excess of $\frac{1}{C}\log\frac{\varepsilon''}{\varepsilon'}$ in Z. But then clearly m can be chosen large enough so that every such interval includes at least one of the $t_k^m\in T_m$.

We have now shown that for μ-almost all $x\in\mathbb{R}^n$ we have $\limsup_{t\to\infty}|\phi_t(x)|\le\varepsilon$, i.e., for μ-almost all $x\in\mathbb{R}^n$ $\exists t_e>0$ such that $|\phi_t(x)|\le\varepsilon$ for $t\ge t_e$. But as this holds for any $\varepsilon>0$ the trajectories must converge to the origin. □

4. Stochastic flows. The purpose of this section is to review, without proofs, some results of the theory of stochastic flows of diffeomorphisms generated by stochastic differential equations. A detailed exposition on the subject can be found in [13, 14], and shorter treatments are in [1, 2, 3, 12].

Throughout this article $(\Omega,\mathcal{F},\mathbb{P})$ denotes the canonical Wiener space of the m-dimensional Brownian motion W_t with two-sided time $\mathbb{R}$. We also introduce the two-parameter filtration $\mathcal{F}_s^t=\sigma\{W_u^k-W_v^k:s\le v\le u\le t,\ 1\le k\le m\}$. The extension to two-sided time is important in that it allows us to treat the Wiener process as a dynamical system [1, 2, 3].

THEOREM 4.1. *There exists a one-parameter group $\{\theta_t:t\in\mathbb{R}\}$ of measure-preserving transformations of $(\Omega,\mathcal{F},\mathbb{P})$ such that $W_t(\theta_s\omega)=W_{t+s}(\omega)-W_s(\omega)$ for all $\omega\in\Omega$ and $s,t\in\mathbb{R}$.*

We will consider Itô stochastic differential equations of the form

$$x_t=x+\int_s^t X_0(x_\tau)\,d\tau+\sum_{k=1}^m\int_s^t X_k(x_\tau)\,dW_\tau^k \tag{4.1}$$

with the following assumptions:

1. $x\in\mathbb{R}^n$.

2. $X_k : \mathbb{R}^n \to \mathbb{R}^n$, $k = 0, \dots, m$, are globally Lipschitz continuous.

The global Lipschitz condition guarantees many nice properties of the solutions; we will assume it for the time being, and later relax this requirement somewhat in section 6.

Denote by $\xi_{s,t}(x, \omega)$ (or simply $\xi_{s,t}(x)$) the solution of (4.1) at time $t \geq s$ given the initial condition $x_s = x$. It is well known that in the case of globally Lipschitz continuous coefficients there exists a unique, nonexploding solution $\xi_{s,t}(x)$ which is an $\mathcal{F}_s^t$-semimartingale and is in L^p for any $p \geq 1$ (see, e.g., [13]).

THEOREM 4.2 (see [13, 2]). *Suppose X_k, $k = 0, \dots, m$, are globally Lipschitz continuous and let $s < t$. Then we have the following properties:*

1. *$\xi_{s,s}(x, \omega) = x$ for all s and ω.*
2. *For any u we have $\xi_{s,t}(\cdot, \theta_u \omega) = \xi_{s+u,t+u}(\cdot, \omega)$.*
3. *For almost all ω we have $\xi_{s,t}(\cdot, \omega) = \xi_{r,t}(\xi_{s,r}(\cdot, \omega), \omega)$ for all $s < r < t$.*
4. *$\xi_{s,t}(x)$ is $\mathbb{P}$-a.s. continuous in (s, t, x).*
5. *For almost all ω the map $\xi_{s,t}(\cdot, \omega) : \mathbb{R}^n \to \mathbb{R}^n$ is a homeomorphism for all $s < t$.*

The following result establishes that, under additional smoothness conditions, $\xi_{s,t}(x)$ is in fact a *stochastic flow of diffeomorphisms*.

THEOREM 4.3 (see [13]). *Suppose X_k, $k = 0, \dots, m$, are globally Lipschitz continuous and that they are $C^{p,\alpha}$ functions for some $p \geq 1$ and $0 < \alpha < 1$. Then for almost all ω the map $\xi_{s,t}(\cdot, \omega) : \mathbb{R}^n \to \mathbb{R}^n$ is a C^p diffeomorphism for any $s \leq t$, and*

$$\frac{\partial \xi_{s,t}(x)^i}{\partial x^j} = \delta_j^i + \sum_{\beta=1}^{n} \int_s^t \frac{\partial X_0^i}{\partial x^\beta}(\xi_{s,\tau}(x)) \frac{\partial \xi_{s,\tau}(x)^\beta}{\partial x^j} \, d\tau + \sum_{k=1}^{m} \sum_{\beta=1}^{n} \int_s^t \frac{\partial X_k^i}{\partial x^\beta}(\xi_{s,\tau}(x)) \frac{\partial \xi_{s,\tau}(x)^\beta}{\partial x^j} \, dW_\tau^k. \tag{4.2}$$

It will be convenient for our purposes to work with the inverse flow $\xi_{s,t}^{-1}(x)$, considered as a backward stochastic process in the time variable s (with t fixed). This will not give rise to ordinary Itô integrals as s behaves like a time-reversed variable, and hence the adaptedness of the process runs backward in time. The Itô backward integral is defined as [13]

$$\int_s^t f_\sigma \overleftarrow{dW}_\sigma \equiv \lim \text{ in prob } \sum_{k=0}^{n-1} f_{t_{k+1}} (W_{t_{k+1}} - W_{t_k}),$$

where f_s is a backward predictable process with $\int_s^t |f_u|^2 \, du < \infty$ almost surely, and the formal construction of the integral from simple functions proceeds along the usual lines. The backward integral has similar properties to the forward integral; in particular, it is a backward $\mathcal{F}_s^t$-local martingale (for fixed t) and satisfies an Itô formula (see, e.g., [4, p. 124]), which is proved in the same way as its forward counterpart: given $\xi_s = \xi_t + \int_s^t a_\sigma \, d\sigma + \sum_k \int_s^t (b_\sigma)_k \overleftarrow{dW}_\sigma^k$ with backward predictable processes a_s, $(b_s)_k$ such that $\int_s^t a_\sigma \, d\sigma < \infty$ almost surely, $\int_s^t |(b_\sigma)_k|^2 \, d\sigma < \infty$ almost surely, then for any C^2 function $F : \mathbb{R}^n \to \mathbb{R}$

$$F(\xi_s) = F(\xi_t) + \frac{1}{2} \sum_{k=1}^{m} \sum_{i,j=1}^{n} \int_s^t (b_\sigma)_k^i (b_\sigma)_k^j \frac{\partial^2 F}{\partial x^i \partial x^j}(\xi_\sigma) \, d\sigma + \sum_{i=1}^{n} \int_s^t a_\sigma^i \frac{\partial F}{\partial x^i}(\xi_\sigma) \, d\sigma + \sum_{k=1}^{m} \sum_{i=1}^{n} \int_s^t (b_\sigma)_k^i \frac{\partial F}{\partial x^i}(\xi_\sigma) \overleftarrow{dW}_\sigma^k. \tag{4.3}$$

We can now formulate the following result.

THEOREM 4.4 (see [13]). *Suppose X_k, $k = 0, \ldots, m$, are globally Lipschitz continuous and that they are $C^{p,\alpha}$ functions for some $p \geq 2$ and $0 < \alpha < 1$. Then*

$$\xi_{s,t}^{-1}(x) = x - \int_s^t \tilde{X}_0(\xi_{\sigma,t}^{-1}(x))\, d\sigma - \sum_{k=1}^m \int_s^t X_k(\xi_{\sigma,t}^{-1}(x))\, \overleftarrow{dW}_\sigma,$$

where we have defined

$$\tilde{X}_0(x) = X_0(x) - \sum_{k=1}^m \sum_{\beta=1}^n X_k^\beta(x) \frac{\partial}{\partial x^\beta} X_k(x).$$

This expression can be manipulated much in the same way as its forward counterpart. In particular, under the conditions of Theorem 4.4 and using (4.3), we obtain for any C^2 function $F : \mathbb{R}^n \to \mathbb{R}$ the backward Itô formula

$$\begin{aligned} F(\xi_{s,t}^{-1}(x)) = F(x) &+ \frac{1}{2} \sum_{k,i,j} \int_s^t X_k^i(\xi_{\sigma,t}^{-1}(x)) X_k^j(\xi_{\sigma,t}^{-1}(x)) \frac{\partial^2 F}{\partial x^i \partial x^j}(\xi_{\sigma,t}^{-1}(x))\, d\sigma \\ &- \sum_i \int_s^t \tilde{X}_0^i(\xi_{\sigma,t}^{-1}(x)) \frac{\partial F}{\partial x^i}(\xi_{\sigma,t}^{-1}(x))\, d\sigma - \sum_{k,i} \int_s^t X_k^i(\xi_{\sigma,t}^{-1}(x)) \frac{\partial F}{\partial x^i}(\xi_{\sigma,t}^{-1}(x))\, \overleftarrow{dW}_\sigma^k. \end{aligned} \tag{4.4}$$

Similarly we can differentiate the inverse flow, giving

$$\begin{aligned} \frac{\partial \xi_{s,t}^{-1}(x)^i}{\partial x^j} = \delta_j^i &- \sum_{\beta=1}^n \int_s^t \frac{\partial \tilde{X}_0^i}{\partial x^\beta}(\xi_{\sigma,t}^{-1}(x)) \frac{\partial \xi_{\sigma,t}^{-1}(x)^\beta}{\partial x^j}\, d\sigma \\ &- \sum_{k=1}^m \sum_{\beta=1}^n \int_s^t \frac{\partial X_k^i}{\partial x^\beta}(\xi_{\sigma,t}^{-1}(x)) \frac{\partial \xi_{\sigma,t}^{-1}(x)^\beta}{\partial x^j}\, \overleftarrow{dW}_\sigma^k. \end{aligned} \tag{4.5}$$

This expression is obtained, in the same way as its forward counterpart (4.2), by letting $y \to 0$ in the backward expression corresponding to [13, p. 219, eqn. (4)].

5. The main result. We consider an Itô equation of the form (4.1). We write

$$\mathscr{L}^* f(x) = \frac{1}{2} \sum_{k=1}^m \sum_{i,j=1}^n \frac{\partial^2}{\partial x^i \partial x^j}(X_k^i(x) X_k^j(x) f(x)) - \sum_{i=1}^n \frac{\partial}{\partial x^i}(X_0^i(x) f(x)).$$

The following is our main result.

THEOREM 5.1. *Let X_k, $k = 0, \ldots, m$, be globally Lipschitz continuous and $C^{2,\alpha}$ functions for some $\alpha > 0$, and let $X_k(0) = 0$. Suppose there exists $D \in C^2(\mathbb{R}^n \backslash \{0\}, \mathbb{R}_+)$ such that D is integrable on $\{x \in \mathbb{R}^n : |x| > 1\}$ and $\mathscr{L}^* D(x) < 0$ for μ-almost all x. Then for every initial time s and μ-a.e. initial state x the flow $\xi_{s,t}(x)$ tends to the origin as $t \to \infty$ $\mathbb{P}$-a.s.*

Before we prove the theorem, let us prove a stochastic version of Lemma 3.1.

LEMMA 5.2. *Suppose X_k, $k = 0, \ldots, m$, are globally Lipschitz continuous and $C^{2,\alpha}$ functions for some $\alpha > 0$. Let $S_\ell \subset S_{\ell+1} \subset \mathbb{R}^n$ be an increasing sequence of open sets such that $\tau_\ell = \sup\{s < t : \xi_{s,t}^{-1}(x) \notin S_\ell\} \to -\infty$ as $\ell \to \infty$ $\mathbb{P}$-a.s. for every $x \in S = \bigcup_\ell S_\ell$. Suppose there is a $D \in C^2(S, \mathbb{R}_+)$ that is integrable on a*

measurable set $Z \subset S$, that obeys $\mathscr{L}^* D \le 0$ on S, and such that for each ℓ there is a $D_\ell \in C^2(\mathbb{R}^n, \mathbb{R}_+)$ that coincides with D on S_ℓ. Then

$$0 \le \int_Z D(x)\,dx + \int_s^t \mathbb{E} \int_{\xi_{\sigma,t}^{-1}(Z)} \mathscr{L}^* D(x)\,dx\,d\sigma$$

for all $s \le t$, and in particular the limit as $s \to -\infty$ of this expression is well defined.

Proof. Denote by $\mathbf{J}_{s,t}(x)$ the matrix with elements $\mathbf{J}_{s,t}(x)^i_j = \partial \xi^{-1}_{s,t}(x)^i / \partial x^j$, i.e.,

$$\mathbf{J}_{s,t}(x)^i_j = \delta^i_j - \sum_\alpha \int_s^t \frac{\partial \tilde{X}^i_0}{\partial x^\alpha}(\xi^{-1}_{\sigma,t}(x))\mathbf{J}_{\sigma,t}(x)^\alpha_j\,d\sigma - \sum_{k,\alpha} \int_s^t \frac{\partial X^i_k}{\partial x^\alpha}(\xi^{-1}_{\sigma,t}(x))\mathbf{J}_{\sigma,t}(x)^\alpha_j \circ\!\!\leftarrow dW^k_\sigma$$

by (4.5). Denote by $|\mathbf{J}_{s,t}(x)|$ its determinant, i.e.,

$$|\mathbf{J}_{s,t}(x)| = \sum_{j_1 \cdots j_n = 1}^n \varepsilon_{j_1,\ldots,j_n} \mathbf{J}_{s,t}(x)^1_{j_1} \mathbf{J}_{s,t}(x)^2_{j_2} \cdots \mathbf{J}_{s,t}(x)^n_{j_n},$$

where $\varepsilon_{j_1,\ldots,j_n}$ is the antisymmetric tensor. Using Itô's rule and straightforward calculations we obtain

$$\begin{aligned}|\,\mathbf{J}_{s,t}(x)| = 1 &- \sum_i \int_s^t \frac{\partial \tilde{X}^i_0}{\partial x^i}(\xi^{-1}_{\sigma,t}(x))|\mathbf{J}_{\sigma,t}(x)|\,d\sigma - \sum_{k,i} \int_s^t \frac{\partial X^i_k}{\partial x^i}(\xi^{-1}_{\sigma,t}(x))|\mathbf{J}_{\sigma,t}(x)| \circ\!\!\leftarrow dW^k_\sigma \\ &+ \frac{1}{2}\sum_{k,i,j} \int_s^t \left[\frac{\partial X^i_k}{\partial x^i}(\xi^{-1}_{\sigma,t}(x))\frac{\partial X^j_k}{\partial x^j}(\xi^{-1}_{\sigma,t}(x)) - \frac{\partial X^i_k}{\partial x^j}(\xi^{-1}_{\sigma,t}(x))\frac{\partial X^j_k}{\partial x^i}(\xi^{-1}_{\sigma,t}(x))\right] |\mathbf{J}_{\sigma,t}(x)|\,d\sigma.\end{aligned}$$

Note that as $\xi^{-1}_{s,t}(\cdot)$ is a diffeomorphism almost surely, its Jacobian $\mathbf{J}_{s,t}(\cdot)$ must almost surely be an invertible matrix; but as $|\mathbf{J}_{s,t}(x)|$ has almost surely continuous sample paths and $|\mathbf{J}_{t,t}(x)| = 1$, this implies that almost surely $|\mathbf{J}_{s,t}(x)| > 0$ for all $s < t$. Using (4.4) with $F = D_\ell$ and Itô's rule we obtain

$$\begin{aligned}0 \le D_\ell(\xi^{-1}_{s,t}(x))|\mathbf{J}_{s,t}(x)| = D_\ell(x) &+ \int_s^t (\mathscr{L}^* D_\ell)(\xi^{-1}_{\sigma,t}(x))|\mathbf{J}_{\sigma,t}(x)|\,d\sigma \\ &- \sum_{k=1}^m \sum_{i=1}^n \int_s^t \frac{\partial X^i_k D_\ell}{\partial x^i}(\xi^{-1}_{\sigma,t}(x))|\mathbf{J}_{\sigma,t}(x)| \circ\!\!\leftarrow dW^k_\sigma.\end{aligned}$$

Now note that as D_ℓ coincides with D on S_ℓ, we can identify $(\mathscr{L}^* D_\ell)(\xi^{-1}_{s\vee\tau_\ell,t}(x)) = (\mathscr{L}^* D)(\xi^{-1}_{s\vee\tau_\ell,t}(x))$ for every ℓ. Moreover, as the last term in the expression above is a backward local martingale, there exists a sequence of stopping times $\tau'_p \searrow -\infty$ such that the stochastic integral stopped at τ'_p is a martingale. Replacing s by $s \vee \tau_\ell \vee \tau'_p$ in the expression above and taking the expectation gives

$$0 \le D(x) + \mathbb{E} \int_{s\vee\tau_\ell\vee\tau'_p}^t (\mathscr{L}^* D)(\xi^{-1}_{\sigma,t}(x))|\mathbf{J}_{\sigma,t}(x)|\,d\sigma.$$

We can now let $\ell, p \to \infty$ by monotone convergence. Integrating both sides gives

$$0 \le \int_Z D(x)\,dx + \int_s^t \mathbb{E} \int_Z (\mathscr{L}^* D)(\xi^{-1}_{\sigma,t}(x))|\mathbf{J}_{\sigma,t}(x)|\,dx\,d\sigma,$$

where we have used Tonelli's theorem to change the order of integration. The result follows after a change of coordinates. □

The proof of Theorem 5.1 is similar to the proof of Theorem 3.2. The stochastic version of the argument following (3.2), however, is a little more subtle, as we do not have a pathwise upper bound on the rate of growth of sample paths. On the other hand, we can establish such a bound *in probability* which, together with the strong Markov property, is sufficient for our purposes; a similar argument was used in [16] to the same effect. For this purpose we give the following lemma, various versions of which appear in the literature (the result below is adapted from [6]).

LEMMA 5.3. *Let X_k, $k = 0, \ldots, m$, be locally Lipschitz continuous and let $\lambda > 0$. Then*

$$\mathbb{P}\left[\sup_{0\le\delta\le\Delta} |\xi_{s,s+\delta}(x) - x| \ge \lambda\right] \le K_1\Delta + K_2\Delta^2,$$

where $K_1, K_2 < \infty$ are constants that depend only on λ and $|x|$.

Proof. Let $\mathbf{W}_t$ be the m-vector with elements W_t^k and let $\mathbf{X}(\cdot)$ be the $n\times m$-matrix with entries $X_k^i(\cdot)$, $k = 1, \ldots, m$. For $r > 0$, define $B_r(x') = \{x \in \mathbb{R}^n : |x - x'| < r\}$, $B_r = B_r(0)$, and

$$\rho_0(r) = \sup_{|y|<r} |X_0(y)|, \qquad \rho_1(r) = \sup_{|y|<r} \|\mathbf{X}(y)\| = \sup_{|y|<r} \mathrm{tr}[\mathbf{X}(y)^T\mathbf{X}(y)]^{1/2}.$$

Let τ_r be the first exit time of $\xi_{s,t}(x)$ from B_r. In [6, p. 1240] it was established that

$$\mathbb{E}\left[\sup_{0\le\delta\le\Delta} |\xi_{s,(s+\delta)\wedge\tau_r}(x) - x|^2\right] \le 2\rho_0(r)^2\Delta^2 + 8\rho_1(r)^2\Delta.$$

Hence we have by Markov's inequality

$$\mathbb{P}\left[\sup_{0\le\delta\le\Delta} |\xi_{s,(s+\delta)\wedge\tau_r}(x) - x| \ge \lambda\right] \le \lambda^{-2}(2\rho_0(r)^2\Delta^2 + 8\rho_1(r)^2\Delta).$$

Now note that $B_\lambda(x)$ is strictly included in $B_{|x|+2\lambda}$, so that the first exit time from $B_\lambda(x)$ is no later than $\tau_{|x|+2\lambda}$. But then the events

$$\left\{\omega : \sup_{0\le\delta\le\Delta} |\xi_{s,(s+\delta)\wedge\tau_{|x|+2\lambda}}(x) - x| \ge \lambda\right\}, \quad \left\{\omega : \sup_{0\le\delta\le\Delta} |\xi_{s,s+\delta}(x) - x| \ge \lambda\right\}$$

are equivalent; after all, the events are equivalent on $\tau_{|x|+2\lambda} > s + \Delta$ by construction, whereas if $\tau_{|x|+2\lambda} \le s + \Delta$, both events must be true as $|\xi_{s,\tau_{|x|+2\lambda}}(x) - x| \ge \lambda$. Hence

$$\mathbb{P}\left[\sup_{0\le\delta\le\Delta} |\xi_{s,s+\delta}(x) - x| \ge \lambda\right] \le \lambda^{-2}(2\rho_0(|x| + 2\lambda)^2\Delta^2 + 8\rho_1(|x| + 2\lambda)^2\Delta),$$

where we have set $r = |x| + 2\lambda$. This completes the proof. □

We now turn to the proof of the main theorem.

Proof of Theorem 5.1. Let $\varepsilon > 0$ and $Z = \{x \in \mathbb{R}^n : |x| > \varepsilon\}$. We begin by applying Lemma 5.2. To this end, define $S_\ell = \{x \in \mathbb{R}^n : |x| > \ell^{-1}\}$, so $S = \bigcup_\ell S_\ell = \mathbb{R}^n\backslash\{0\}$. Clearly D is integrable on Z and there exists a $C^2(\mathbb{R}^n, \mathbb{R}_+)$-approximation D_ℓ of D for each ℓ. It remains to check that $\tau_\ell \to -\infty$. Suppose that this is not the case; then given $x \in S$ there must be a positive probability that $\xi_{s,t}^{-1}(x) = 0$ for some

$-\infty < s < t$. But $\xi_{s,t}^{-1}(0) = 0$ for all s and almost surely $\xi_{s,t}^{-1}(x)$ is one-to-one for all $s < t$, so this cannot happen. Hence all the conditions of Lemma 5.2 are satisfied, and we have

$$0 \le \int_Z D(x)\,dx + \int_s^t \mathbb{E} \int_{\xi_{\sigma,t}^{-1}(Z)} \mathscr{L}^* D(x)\,dx\,d\sigma. \tag{5.1}$$

Now note that (5.1) is nonincreasing with decreasing s due to $\mathscr{L}^* D \le 0$ and is finite because D is integrable on Z. By monotone convergence the limit as $s \to -\infty$ exists and is finite. Hence

$$\int_{-\infty}^t \mathbb{D}(\xi_{\sigma,t}^{-1}(Z))\,d\sigma < \infty, \qquad \mathbb{D}(A) = -\int_A \mathscr{L}^* D(x)\,(\mathbb{P}(d\omega) \times \mu(dx)),$$

where we have used Tonelli's theorem to convert the iterated integral to a single integral with respect to the product measure, and we slightly abuse our notation by writing $\mathbb{D}(\xi_{\sigma,t}^{-1}(Z)) = \mathbb{D}(\{(\omega, x) \in \Omega \times S : \xi_{\sigma,t}(x, \omega) \in Z\})$. Note that $\mathscr{L}^* D \le 0$ implies that $\mathbb{D}$ is a measure on $\Omega \times S$, and $\mathbb{D}$ is σ-finite as $\mathbb{D}(\Omega \times \{x \in S : \frac{1}{k} < |x| < k\}) < \infty$ for all $k > 1$ and $\bigcup_{n=2}^{\infty}(\Omega \times \{x \in S : \frac{1}{k} < |x| < k\}) = \Omega \times S$.

We now fix some $m \in \mathbb{N}$ and divide the halfline into bins $S_k^m = [(k-1)2^{-m}, k2^{-m}]$, $k \in \mathbb{N}$. From each bin we choose a time $t_k^m \in S_k^m$ such that

$$\mathbb{D}(\xi_{t-t_k^m,t}^{-1}(Z)) \le \inf_{s \in S_k^m} \mathbb{D}(\xi_{t-s,t}^{-1}(Z)) + 2^{-k}.$$

For fixed m, we denote this discrete grid by $T_m = \{t_k^m : k \in \mathbb{N}\}$. We now have

$$2^{-m} \sum_{k=1}^{\infty} \mathbb{D}(\xi_{t-t_k^m,t}^{-1}(Z)) \le 2^{-m} + \int_{-\infty}^t \mathbb{D}(\xi_{\sigma,t}^{-1}(Z))\,d\sigma < \infty.$$

Using the fact that the transformation θ_t of Theorem 4.1 is $\mathbb{P}$-preserving to shift the times t_k^m to the forward variable, we obtain

$$\sum_{k=1}^{\infty} \mathbb{D}(\xi_{s,s+t_k^m}^{-1}(Z)) = \sum_{k=1}^{\infty} \mathbb{D}(\xi_{t-t_k^m,t}^{-1}(Z)) < \infty.$$

As $\mathbb{D}$ is σ-finite we can now apply the Borel–Cantelli lemma, which gives

$$\mathbb{D}\left(\limsup_{k\to\infty} \xi_{s,s+t_k^m}^{-1}(Z)\right) = (\mathbb{P} \times \mu)\left(\limsup_{k\to\infty} \xi_{s,s+t_k^m}^{-1}(Z)\right) = 0,$$

where the first equality follows as $\mathscr{L}^* D(x) < 0$ μ-a.e. implies $\mathbb{P} \times \mu \ll \mathbb{D}$. Consequently

$$(\mathbb{P} \times \mu)\left(\bigcup_{m=1}^{\infty} \limsup_{t\in T_m} \xi_{s,s+t}^{-1}(Z)\right) \le \sum_{m=1}^{\infty} (\mathbb{P} \times \mu)\left(\limsup_{t\in T_m} \xi_{s,s+t}^{-1}(Z)\right) = 0.$$

We have thus shown that for all initial states x, except in a set $N \subset \mathbb{R}^n$ of Lebesgue measure zero, there is $\mathbb{P}$-a.s. for any m only a finite number of times t in the discrete grid T_m such that $\xi_{s,s+t}(x) \in Z$.

Let us fix an $x \notin N$. We now claim that the fact that $\mathbb{P}$-a.s. for any m there is only a finite number of times $t \in T_m$ such that $\xi_{s,s+t}(x) \in Z$ implies that $\mathbb{P}$-a.s. $\limsup_{t\to\infty} |\xi_{s,t}(x)| \le \varepsilon$. To see this, suppose $\mathbb{P}[\limsup_{t\to\infty} |\xi_{s,t}(x)| > \varepsilon] = \delta > 0$.

By monotone convergence $\mathbb{E}[\chi_{\limsup|\xi_{s,t}(x)|>\varepsilon'}] \nearrow \delta$ as $\varepsilon' \searrow \varepsilon$; hence there exists an $\varepsilon' > \varepsilon$ such that $\mathbb{P}[\limsup_{t\to\infty}|\xi_{s,t}(x)| > \varepsilon'] > 0$. We have already shown, however, that almost surely $|\xi_{s,t}(x)| \le \varepsilon$ for infinitely many times $t_n \nearrow \infty$. Hence

$$\mathbb{P}\left[\limsup_{t\to\infty}|\xi_{s,t}(x)| > \varepsilon'\right] > 0 \quad\Longrightarrow\quad \mathbb{P}[|\xi_{s,t}(x)| \text{ crosses } \varepsilon \text{ and } \varepsilon' \text{ infinitely often}] > 0.$$

Once we disprove the latter statement, the claim is proved by contradiction.

To this end, introduce the following sequence of predictable stopping times. Let $\sigma_0 = \inf\{t > s : |\xi_{s,t}(x)| \le \varepsilon\}$, $\tau_0 = \inf\{t > \sigma_0 : |\xi_{s,t}(x)| \ge \varepsilon'\}$, and for any $n > 0$ we set $\sigma_n = \inf\{t > \tau_{n-1} : |\xi_{s,t}(x)| \le \varepsilon\}$, $\tau_n = \inf\{t > \sigma_n : |\xi_{s,t}(x)| \ge \varepsilon'\}$. Define

$$\Omega_n(\Delta) = \{\omega \in \Omega : \tau_n < \infty,\ |\xi_{s,\tau_n+\delta}(x)| > \varepsilon\ \forall\, 0 \le \delta \le \Delta\}.$$

For any $\Delta > 0$, the set of $\omega \in \Omega$ such that $\omega \in \Omega_n(\Delta)$ for infinitely many n must be of $\mathbb{P}$-measure zero; after all, we can choose m sufficiently large so that every time interval of length Δ contains at least one point in T_m, and for points $t \in T_m$ we have $|\xi_{s,t}(x)| > \varepsilon$ only finitely often $\mathbb{P}$-a.s. Thus $\sum_n \chi_{\Omega_n(\Delta)} < \infty$ $\mathbb{P}$-a.s. To proceed, we use the following argument (see [17, pp. 398–399]). Introduce the discrete filtration $\mathcal{B}_k = \mathcal{F}_s^{\tau_{k+1}}$ and define $Z_k = X_k - Y_k$ with

$$X_k = \sum_{n=1}^{k} \chi_{\Omega_n(\Delta)}, \qquad Y_k = \sum_{n=1}^{k} \mathbb{E}[\chi_{\Omega_n(\Delta)}|\mathcal{B}_{n-1}].$$

As $\Omega_k(\Delta) \in \mathcal{B}_n$ for all $k \le n$, Z_k is a $\mathcal{B}_k$-martingale. Now define for $a > 0$ the stopping time $\kappa(a) = \inf\{n : Z_n > a\}$. As $|Z_k - Z_{k-1}| \le 1$ almost surely, the stopped process $Z'_k = Z_{k\wedge\kappa(a)}$ is a martingale that is bounded from above, and by the martingale convergence theorem Z'_k converges almost surely as $k \to \infty$ to a finite random variable Z'_∞. But as Z'_k and Z_k coincide on $\{\omega : \sup_n Z_n < a\}$ and $a > 0$ was chosen arbitrarily, we conclude that $Z_k \to Z_\infty < \infty$ on $\{\omega : \sup_n Z_n < \infty\}$ (modulo a null set). Note, however, that X_n and Y_n are both positive increasing processes and we have already established that $\sup_n X_n < \infty$ $\mathbb{P}$-a.s., so $\sup_n Z_n < \infty$ $\mathbb{P}$-a.s. But this implies that Z_k, and hence also Y_k, converges to a finite value $\mathbb{P}$-a.s. Thus we have established

$$\sum_{n=1}^{\infty} \mathbb{E}[\chi_{\Omega_n(\Delta)}|\mathcal{F}_s^{\tau_n}] < \infty \quad \mathbb{P}\text{-a.s.} \quad \text{for any } \Delta > 0.$$

Note that by the continuity of the sample paths, $|\xi_{s,\tau_n}(x)| = \varepsilon'$ on $\tau_n < \infty$. By Lemma 5.3, we can choose $\overline{\Delta} > 0$ sufficiently small such that

$$P(y) = \mathbb{P}\left[\sup_{0\le\delta\le\overline{\Delta}} |\xi_{s,s+\delta}(y) - y| < \frac{\varepsilon' - \varepsilon}{2}\right] \ge \frac{1}{2}$$

for all $|y| = \varepsilon'$. Using the strong Markov property, we can write

$$\infty > \sum_{n=1}^{\infty} \mathbb{E}[\chi_{\Omega_n(\overline{\Delta})}|\mathcal{F}_s^{\tau_n}] \ge \sum_{n=1}^{\infty} P(\xi_{s,\tau_n}(x))\,\chi_{\tau_n<\infty} \ge \frac{1}{2}\sum_{n=1}^{\infty} \chi_{\tau_n<\infty} \quad \mathbb{P}\text{-a.s.}$$

But this implies that $\tau_n < \infty$ finitely often $\mathbb{P}$-a.s., contradicting the assertion that $\mathbb{P}[|\xi_{s,t}(x)| \text{ crosses } \varepsilon \text{ and } \varepsilon' \text{ infinitely often}] > 0$. This is the desired result.

We have now shown that for μ-almost all $x \in \mathbb{R}^n$, $\mathbb{P}$-a.s. $\limsup_{t\to\infty} |\xi_{s,t}(x)| \le \varepsilon$, i.e., for μ-almost all $x \in \mathbb{R}^n$ $\mathbb{P}$-a.s., $\exists t_e > s$ such that $|\xi_{s,t}(x)| \le \varepsilon$ for $t \ge t_e$. But as this holds for any $\varepsilon > 0$ the flow must converge to the origin. □

The proof of Theorem 5.1 is readily extended to prove other assertions, such as the following instability theorem.

THEOREM 5.4. *Let X_k, $k = 0, \dots, m$, be globally Lipschitz continuous and $C^{2,\alpha}$ functions for some $\alpha > 0$. Suppose there exists a $D \in C^2(\mathbb{R}^n, \mathbb{R}_+)$ such that $\mathscr{L}^* D(x) < 0$ for μ-almost all x. Then for every initial time s and μ-a.e. initial state x the flow escapes to infinity, i.e., $|\xi_{s,t}(x)| \to \infty$ as $t \to \infty$ $\mathbb{P}$-a.s.*

Proof. Let $\varepsilon > 0$ and $Z' = \{x \in \mathbb{R}^n : |x| < \varepsilon\}$. Again we begin by applying Lemma 5.2. We can simply choose $S_\ell = S = \mathbb{R}^n$ for all ℓ; by nonexplosion $\tau_\ell = -\infty$ and the remaining conditions are evident. Hence

$$0 \le \int_{Z'} D(x)\, dx + \int_s^t \mathbb{E} \int_{\xi_{\sigma,t}^{-1}(Z')} \mathscr{L}^* D(x)\, dx\, d\sigma.$$

Proceeding in exactly the same way as in the proof of Theorem 5.1 we can now show that for μ-almost all $x \in \mathbb{R}^n$, $\mathbb{P}$-a.s. $\liminf_{t\to\infty} |\xi_{s,t}(x)| \ge \varepsilon$, i.e., for μ-almost all $x \in \mathbb{R}^n$ $\mathbb{P}$-a.s., $\exists t_e > s$ such that $|\xi_{s,t}(x)| \ge \varepsilon$ for $t \ge t_e$. But as this holds for any $\varepsilon > 0$ the flow must escape to infinity. □

Remark. At first sight the statements of Theorems 5.1 and 5.4 may seem contradictory, but this is not the case. The essential difference between the theorems is the region in $\mathbb{R}^n$ on which D is integrable. Roughly speaking, the idea behind the proofs of Theorems 5.1 and 5.4 is to show that if $\mathscr{L}^* D < 0$ μ-a.e., then the solution of the Itô equation can spend only a finite amount of time in any region on which D is integrable. Hence in Theorem 5.1 the solution will be attracted to the origin, whereas in Theorem 5.4 the solution is attracted to infinity.

If we try to satisfy the conditions of Theorems 5.1 and 5.4 simultaneously, we will run into problems. Suppose we have a nonnegative $D \in C^2(\mathbb{R}^n)$, as in Theorem 5.4, which is integrable as in Theorem 5.1. Then D is a normalizable density function, i.e., we could normalize D and interpret it as the density of the Itô equation at some point in time. But then $\mathscr{L}^* D < 0$ would imply that the associated Fokker–Planck equation does not preserve normalization of the density. Evidently Theorem 5.4 can only be satisfied if D is not integrable, whereas Theorem 5.1 requires D to have a singularity at the origin. See section 7 for examples.

6. Further results. In this section we extend the main result to cases in which the global Lipschitz condition is not necessarily satisfied. We first show that the result of Theorem 5.1 still holds if we can convert the coefficients of (4.1) to be globally Lipschitz continuous through a suitably chosen time transformation. In particular, this allows us to treat the case that X_k, $k = 0, \dots, m$, and their first derivatives are polynomially bounded, provided that some additional integrability conditions on D are satisfied. We also extend the main result to the case in which the flow is restricted to an invariant subset of $\mathbb{R}^n$ with compact closure.

THEOREM 6.1. *Let $X_k : \mathbb{R}^n \to \mathbb{R}^n$, $k = 0, \dots, m$, be measurable and let $X_k(0) = 0$. Suppose there is a strictly positive measurable map $c : \mathbb{R}^n \to (0, \infty)$ such that $c(x)$ and $c(x)^{-1}$ are locally bounded, and such that $c(x)X_0(x)$ and $\sqrt{c(x)}\, X_k(x)$, $k = 1, \dots, m$, are globally Lipschitz continuous and $C^{2,\alpha}$ functions for some $\alpha > 0$. Suppose there exists $D : \mathbb{R}^n \backslash \{0\} \to \mathbb{R}_+$ such that $D(x)/c(x)$ is C^2 and is integrable on $\{x \in \mathbb{R}^n : |x| > 1\}$, and $\mathscr{L}^* D(x) < 0$ for μ-almost all x. Then for every initial*

time s and μ-a.e. initial state x the solution x_t of (4.1) tends to the origin as $t \to \infty$ $\mathbb{P}$-a.s.

Proof. Consider the Itô equation

$$y_t = y_s + \int_s^t c(y_\tau) X_0(y_\tau)\, d\tau + \sum_{k=1}^m \int_s^t \sqrt{c(y_\tau)}\, X_k(y_\tau)\, dW_\tau^k. \tag{6.1}$$

We will write $Y_0(y) = c(y)X_0(y)$, $Y_k(y) = \sqrt{c(y)}\, X_k(y)$ $(k = 1, \ldots, m)$, and $\tilde{D}(y) = D(y)/c(y)$. Note that by construction $\tilde{\mathscr{L}}^* \tilde{D}(y) = \mathscr{L}^* D(y)$, where $\tilde{\mathscr{L}}^*$ is the adjoint generator of (6.1). By our assumptions we can apply Theorem 5.1 to the Itô equation (6.1). Thus for all $y_s \in \mathbb{R}^n$, except in a set N with $\mu(N) = 0$, $y_t \to 0$ as $t \to \infty$ $\mathbb{P}$-a.s.

Now choose any $y_s \notin N$ and define

$$\beta_t = \int_s^t c(y_\tau)\, d\tau, \qquad\qquad \alpha_t = \inf\{s : \beta_s > t\}.$$

Note that α_τ is an $\mathcal{F}_s^t$-stopping time for each τ. We claim that $\beta_t \to \infty$ as $t \to \infty$ almost surely; indeed y_t almost surely spends an infinite amount of time in an arbitrarily small neighborhood of the origin, and as $c(x)^{-1}$ is locally bounded $c(x) \geq \delta > 0$ in any such neighborhood. Moreover, $\beta_t < \infty$ almost surely for any t as $c(x)$ is locally bounded, and hence $\alpha_t \to \infty$ as $t \to \infty$ almost surely. From [24, section V.26, p. 175], it follows that the time rescaled solution y_{α_t} is equivalent in law to the solution x_t of (4.1). But as almost all paths of the process y_t go to zero asymptotically and as $\alpha_t \to \infty$ almost surely, the result follows. □

Corollary 6.2. *Suppose that all the conditions of Theorem 5.1 are satisfied except the global Lipschitz condition. Suppose that additionally X_k, $k = 1, \ldots, m$, satisfy $|X_k(x)| \leq C_k(1 + |x|^{p+1})$, $|\partial X_k(x)/\partial x^i| \leq C_k'(1 + |x|^p)$, and $|X_0(x)| \leq C_0(1 + |x|^{2p+1})$, $|\partial X_0(x)/\partial x^i| \leq C_0'(1+|x|^{2p})$ for some $p \geq 1$ and positive constants $C_k, C_k' < \infty$. If $(1+|x|^{2p})D(x)$ is integrable on $\{x \in \mathbb{R}^n : |x| > 1\}$, then Theorem 5.1 still holds.*

Proof. Let $c(x) = (1 + |x|^{2p})^{-1}$, and note that $c(x)$ is smooth, strictly positive and that $c(x)$ and $c(x)^{-1}$ are locally bounded. Let $Y_0(x) = c(x)X_0(x)$ and $Y_k(x) = \sqrt{c(x)}\, X_k(x)$, $k = 1, \ldots, m$; as $\sqrt{c(x)}$ is smooth and X_k, $k = 0, \ldots, m$, are $C^{2,\alpha}$, the coefficients Y_k are also $C^{2,\alpha}$. To show that Y_k are globally Lipschitz continuous it suffices to show that their first derivatives are bounded. Let us write $y_{k,i}(x) = |\partial Y_k/\partial x^i|(x)$. First consider the case $k = 0$. Then

$$y_{0,i}(x) \leq \frac{|\partial X_0(x)/\partial x^i|}{1 + |x|^{2p}} + \frac{2p|x|^{2p-2}|x^i|}{(1 + |x|^{2p})^2}\, |X_0(x)| \leq C_0' + C_0 \frac{2p|x|^{2p-2}|x^i|}{(1 + |x|^{2p})^2}(1 + |x|^{2p+1}),$$

which is bounded. Similarly, for $k \geq 1$

$$y_{k,i}(x) \leq C_k' \frac{1 + |x|^p}{\sqrt{1 + |x|^{2p}}} + C_k \frac{p|x|^{2p-2}|x^i|}{(1 + |x|^{2p})^{3/2}}(1 + |x|^{p+1})$$

is bounded. Finally, $D(x)/c(x) = (1 + |x|^{2p})D(x)$ is C^2, as $D(x)$ is C^2 and $c(x)^{-1}$ is smooth on $\mathbb{R}^n \backslash \{0\}$. Hence, provided $D(x)/c(x)$ is integrable on $\{x \in \mathbb{R}^n : |x| > 1\}$, we can apply Theorem 6.1. □

We now turn our attention to stochastic differential equations which evolve on an invariant set. The following notion of invariance is sufficient for our purposes.

Definition 6.3. *A set K is called backward invariant with respect to the flow $\xi_{s,t}$ if $\xi_{s,t}^{-1}(K) \subset K$ almost surely for all $s < t$.*

We can now formulate the following result.

THEOREM 6.4. *Suppose the Itô equation* (4.1) *evolves on a backward invariant open set K with compact closure $\overline{K}$. Let X_k, $k = 0, \ldots, m$, be $C^{2,\alpha}$ for some $\alpha > 0$ and let $X_k(0) = 0 \in \overline{K}$. Suppose there exists $D \in C^2(\overline{K}\backslash\{0\}, \mathbb{R}_+)$ such that $\mathscr{L}^* D(x) < 0$ for μ-almost all $x \in K$. Then for every initial time s and μ-a.e. initial state $x \in K$ the flow $\xi_{s,t}(x)$ tends to the origin as $t \to \infty$ $\mathbb{P}$-a.s.*

Proof. We will assume without loss of generality that X_k, $k = 0, \ldots, m$, are globally Lipschitz continuous. Indeed, we can smoothly modify X_k outside K to have compact support without changing the properties of the flow in K, and as X_k are already locally Lipschitz continuous their modifications will be globally Lipschitz continuous.

Let $\varepsilon > 0$ and $Z = \{x \in K : |x| > \varepsilon\}$. D is integrable on Z, as D is bounded on Z and Z has compact closure. Let S_ℓ be an increasing sequence of open sets whose closure is strictly contained in $S = K\backslash\{0\}$, such that $\bigcup_\ell S_\ell = S$. Then there exists a $C^2(\mathbb{R}^n, \mathbb{R}_+)$-approximation D_ℓ of D for each ℓ, obtained by smoothly modifying D outside S_ℓ so that its support is contained in $\overline{K}$. That $\tau_\ell \to -\infty$ follows from backward invariance and from the one-to-one property of the flow. Hence all the conditions of Lemma 5.2 are satisfied, and we have

$$0 \leq \int_Z D(x)\, dx + \int_s^t \mathbb{E} \int_{\xi_{\sigma,t}^{-1}(Z)} \mathscr{L}^* D(x)\, dx\, d\sigma.$$

The remainder of the proof proceeds along the same lines as Theorem 5.1. ∎

7. Examples.

Example 1. Consider the Itô equation

$$\begin{aligned} dx_t &= (x_t^2 - 2x_t - z_t^2)\, dt + x_t\, dW_t, \\ dz_t &= 2z_t(x_t - 1)\, dt + z_t\, dW_t. \end{aligned} \tag{7.1}$$

Note that the line $z = 0$ is invariant under the flow of (7.1), where the solution $(x_t, 0)$ for an initial state $(x_0, 0)$ is given by

$$dx_t = (x_t^2 - 2x_t)\, dt + x_t\, dW_t.$$

This equation has an explicit solution (see also [2] for a detailed analysis of the dynamical behavior of this system):

$$x_t = \frac{x_0 e^{-2t} e^{W_t - \frac{1}{2}t}}{1 - x_0 \int_0^t e^{-2s} e^{W_s - \frac{1}{2}s} ds}.$$

Clearly $x_t(\omega)$, $\omega \in \Omega$, explodes in finite time if

$$x_0 > \left(\int_0^\infty e^{-2s} e^{W_s - \frac{1}{2}s} ds \right)^{-1} < \infty.$$

Hence the system (7.1) is certainly not globally stable.

Nonetheless, almost all points $(x_0, z_0) \in \mathbb{R}^2$ are attracted to the origin. To show this, apply Corollary 6.2 with

$$D(x, z) = \frac{1}{(x^2 + z^2)^2}, \qquad \mathscr{L}^* D(x, z) = -\frac{3}{(x^2 + z^2)^2} < 0.$$

Hence for almost every $(x_0, z_0) \in \mathbb{R}^2$, almost surely $(x_t, z_t) \to (0,0)$ as $t \to \infty$.

Example 2. Consider the Itô equation

$$\begin{aligned} dx_t &= 12\,(2z_t - 1)x_t z_t\,dt - \tfrac{1}{2}x_t\,dt + (1-2z_t)x_t\,dW_t, \\ dy_t &= -\tfrac{1}{2}y_t\,dt + (1-2z_t)y_t\,dW_t, \\ dz_t &= -12 z_t x_t^2\,dt + 2(1-z_t)z_t\,dW_t. \end{aligned} \tag{7.2}$$

Let $R_t = 2z_t - 2z_t^2 - x_t^2 - y_t^2$. By Itô's rule we have

$$dR_t = -4\,(1-z_t)z_t R_t\,dt + 2\,(1-2z_t)R_t\,dW_t.$$

Evidently the ellipse $\{(x,y,z) \in \mathbb{R}^3 : 2z - 2z^2 - x^2 - y^2 = 0\}$ is invariant under (7.2). Local uniqueness of the solution implies that the interior of the ellipse is also invariant. Hence $K = \{(x,y,z) \in \mathbb{R}^3 : 2z - 2z^2 - x^2 - y^2 > 0\}$ is a (backward) invariant set for the system (7.2). Consider

$$D(x,y,z) = \frac{1}{z^2}, \qquad \mathscr{L}^* D(x,y,z) = -\frac{12x^2}{z^2} < 0 \quad \mu\text{-a.e.}$$

Hence by Theorem 6.4 for almost every $(x_0, y_0, z_0) \in K$, almost surely $(x_t, y_t, z_t) \to (0,0,0)$ as $t \to \infty$. Note that $(0,0,0)$ is certainly not globally stable: it is easily verified that any point with $x_0 = 0$ and $z_0 \neq 0$ is not attracted to $(0,0,0)$ almost surely, as in this case z_t has a constant nonzero mean.

Example 3. We consider again the system (7.2), but now we are interested in the behavior of points in the invariant set $K' = \{(x,y,z) \in \mathbb{R}^3 : 2z - 2z^2 - x^2 = 0,\ y = 0\}$. As K' is not an open set, we cannot apply Theorem 6.4 to study this case.

Define the transformation $(x,z) \mapsto p = x/z$. Note that p is the stereographic projection of $(x,y,z) \in K'$ which maps $(0,0,0) \mapsto \infty$. As the fixed point $(0,0,0)$ cannot be reached in finite time, we expect that the stereographic projection gives a well-defined dynamical system on $\mathbb{R}$. Using Itô's rule and $2z - 2z^2 - x^2 = 0$ we obtain the Itô equation

$$dp_t = \left(\frac{3}{2} + \frac{20}{2+p^2}\right) p_t\,dt - p_t\,dW_t.$$

Note that this expression satisfies a global Lipschitz condition. Now consider

$$D(p) = \sqrt{2+p^2}, \qquad \mathscr{L}^* D(p) = -\frac{42}{(2+p^2)^{3/2}} < 0.$$

Hence by Theorem 5.4 for almost every $p_0 \in \mathbb{R}$, almost surely $p_t \to \infty$ as $t \to \infty$. This implies that the point $(x,y,z) = (0,0,0)$ is almost globally stable in K'.

8. Application to control synthesis. Consider an Itô equation of the form

$$x_t = x + \int_s^t (X_0(x_\tau) + u_\tau \tilde{X}_0(x_\tau))\,d\tau + \sum_{k=1}^m \int_s^t X_k(x_\tau)\,dW_\tau^k, \tag{8.1}$$

where $\tilde{X}_0$ and X_k, $k = 0, \ldots, m$, are $C^{2,\alpha}$ for some $\alpha > 0$, $X_k(0) = 0$ and u_t is a scalar control input. We consider instantaneous state feedback of the form $u_t = u(x_t)$ where $u(x)$ is $C^{2,\beta}$ for some $\beta > 0$ and $u(0) = 0$. Then by Theorem 5.1 or 6.4 or by

Corollary 6.2, $x_t \to 0$ as $t \to \infty$ almost surely for almost every x_0 if there exists a $D(x)$, with additional properties required by the appropriate theorem, such that

$$\begin{aligned}\mathscr{L}^* D(x) = \frac{1}{2}\sum_{k=1}^{m}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x^i \partial x^j}(X_k^i(x)X_k^j(x)D(x)) \\ -\sum_{i=1}^{n}\frac{\partial}{\partial x^i}(X_0^i(x)D(x)+\tilde{X}_0^i(x)u(x)D(x)) < 0 \qquad \mu\text{-a.e.}\end{aligned} \tag{8.2}$$

Note that this expression is affine in $D(x)$ and $u(x)D(x)$ and that the set of functions $(D(x), u(x)D(x))$ which satisfy (8.2) is convex. This fact has been used in the deterministic case by Prajna, Parrilo, and Rantzer [22] to search for "almost stabilizing" controllers for systems with polynomial vector fields using convex optimization. As the stochastic case enjoys the same convexity properties as the deterministic Theorem 3.2, this approach can also be applied to find stabilizing controllers for stochastic nonlinear systems. Note that that convex optimization cannot be used to search for globally stabilizing controllers using LaSalle's theorem [22] as LaSalle's convergence criterion [16, 6] is *not* convex.

The purpose of this section is to briefly outline the method of [22] for the synthesis of stabilizing controllers using convex optimization. We will also discuss a simple example.

Suppose that $\tilde{X}_0$ and X_k, $k = 0, \ldots, m$, are polynomial functions (the case of rational functions can be treated in a similar way). Consider $D(x)$ and $u(x)$ parametrized in the following way:

$$D(x) = \frac{a(x)}{b(x)^\gamma}, \qquad u(x) = \frac{c(x)}{a(x)}. \tag{8.3}$$

Here $b(x)$ is a nonnegative polynomial which vanishes only at the origin, $a(x)$ is a polynomial that is nonnegative in a neighborhood of the origin and is such that $u(x)$ is $C^{2,\beta}$, $c(x)$ is a polynomial that vanishes at the origin, and $\gamma > 0$ is a constant. The orders of the polynomials and γ can be chosen in such a way that $D(x)$ satisfies the integrability and growth requirements of Corollary 6.2. For fixed $b(x)$ and γ consider the expression

$$-b(x)^{\gamma+2}\mathscr{L}^* D(x) > 0 \qquad \mu\text{-a.e.} \tag{8.4}$$

with $D(x)$ and $u(x)$ given by (8.3) and $\mathscr{L}^*$ given by (8.2). Then (8.4) is a polynomial inequality that is linear in the polynomial coefficients of $a(x)$ and $c(x)$. Our goal is to formulate the search for these coefficients as a convex optimization problem.

Verifying whether (8.4) is satisfied comes down to testing nonnegativity of a polynomial (a nonnegative polynomial can only vanish on a finite set of points, and hence is positive μ-a.e.). This problem is known to be NP-hard in general; however, a powerful convex relaxation was suggested by Parrilo [20]. Instead of testing (8.4) directly we may ask whether the polynomial can be written as a *sum of squares*, i.e., whether $-b(x)^{\gamma+2}\mathscr{L}^* D(x) = \sum_i p_i(x)^2$ for a set of polynomials $p_i(x)$. The power of this relaxation comes from the fact that every sum of squares polynomial up to a specified order can be represented by a positive semidefinite matrix; hence the search for a sum of squares representation can be performed using semidefinite programming. As (8.4) is convex in $a(x)$ and $c(x)$ the following is a convex optimization problem:

Find polynomials $a(x)$, $c(x)$ such that $-b(x)^{\gamma+2}\mathscr{L}^* D(x)$ is a sum of squares.

This type of problem, known as a sum of squares program, can be solved in a highly efficient manner using the software SOSTOOLS [21]. We refer to [20, 21, 22] for further details on the computational technique.

Remark. Note that $a(0)$ and $c(0)$ depend only on the value of the constant coefficient of the polynomials $a(x)$ and $c(x)$. Thus $c(0) = 0$ can easily be enforced by fixing the constant coefficient of $c(x)$. To make sure $a(x)$ is nonnegative near the origin and $u(x)$ does not blow up, we can, for example, require $a(x)$ to be of the form $\lambda + p(x)$ with $\lambda > 0$ and $p(x)$ to be a sum of squares that vanishes at the origin.

Note that if the Itô equation (8.1) evolves on an invariant open set K with compact closure, then the sum of squares relaxation is overly restrictive. A related relaxation that guarantees only polynomial nonnegativity on K for the case that K is a semialgebraic set is considered in, e.g., [19].

Example. The following example is similar to an example in [22]. Consider the Itô equation

$$\begin{aligned} dx_t &= (2x_t^3 + x_t^2 y_t - 6x_t y_t^2 + 5y_t^3)\,dt + (x_t^2 + y_t^2)\,dW_t, \\ dy_t &= u_t\,dt - (x_t^2 + y_t^2)\,dW_t. \end{aligned}$$

We choose $b(x, y) = x^2 + y^2$ and $\gamma = 2.5$. Using SOSTOOLS we find a solution with controller of order 3 and a constant $a(x)$. Note that these choices satisfy the integrability requirements of Corollary 6.2. We obtain a stabilizing controller

$$u(x) = -2.7\,x^3 + 4.6\,x^2 y - 6.7\,xy^2 - 3.4\,y^3,$$

where

$$\begin{aligned} -(x^2 + y^2)^{4.5}\mathscr{L}^*(x^2 + y^2)^{-2.5} &= 0.35\,y^6 - 0.0015\,xy^5 \\ &\quad + 0.6\,x^2 y^4 + 0.0026\,x^3 y^3 + 0.33\,x^4 y^2 + 0.004\,x^5 y + 0.13\,x^6 \end{aligned}$$

is a sum of squares polynomial.

Remark. A drawback of this method is that $b(x)$ and γ must be fixed at the outset. We have found that the method is very sensitive to the choice of $b(x)$ and γ even in the deterministic case; often an unfortunate choice will cause the search to be infeasible. Moreover it is not clear, even if there exists for polynomial $\tilde{X}_0$, X_k a rational u which almost globally stabilizes the system, that a rational D can always be found that satisfies $\mathscr{L}^* D < 0$. Nonetheless the method can be successful in cases where other methods fail, and as such could be a useful addition to the stochastic nonlinear control engineer's toolbox.

Acknowledgments. This work was performed in Hideo Mabuchi's group, and the author gratefully acknowledges his support. The author would like to thank Anders Rantzer, Luc Bouten, Stephen Prajna, Paige Randall, and especially Houman Owhadi for insightful discussions and comments. The author is particularly thankful to an anonymous referee for pointing out a gap in the proofs and for his careful reading of the manuscript, which has significantly improved the presentation.

REFERENCES

[1] L. Arnold, *The unfolding of dynamics in stochastic analysis*, Mat. Apl. Comput., 16 (1997), pp. 3–25.

[2] L. Arnold, *Random Dynamical Systems*, Springer-Verlag, Berlin, 1998.

[3] P. H. Baxendale, *Properties of stochastic flows of diffeomorphisms*, in Diffusion Processes and Related Problems in Analysis, Volume II, M. A. Pinsky and V. Wihstutz, eds., Progr. Probab. 27, Birkhäuser Boston, Boston, MA, 1992, pp. 3–35.
[4] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
[5] H. Deng and M. Krstić, *Stochastic nonlinear stabilization—part* I: *A backstepping design*, Syst. Control Lett., 32 (1997), pp. 143–150.
[6] H. Deng, M. Krstić, and R. J. Williams, *Stabilization of stochastic nonlinear systems driven by noise of unknown covariance*, IEEE Trans. Automat. Control, 46 (2001), pp. 1237–1253.
[7] P. Florchinger, *A universal formula for the stabilization of control stochastic differential equations*, Stochastic Anal. Appl., 11 (1993), pp. 155–162.
[8] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.
[9] P. Florchinger, *A stochastic Jurdjevic–Quinn theorem*, SIAM J. Control Optim., 41 (2002), pp. 83–88.
[10] R. Z. Has'minskiĭ, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Germantown, MD, 1980.
[11] A. Isidori, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, Berlin, 1995.
[12] H. Kunita, *On the decomposition of solutions of stochastic differential equations*, in Stochastic Integrals, D. Williams, ed., Lecture Notes in Math. 851, Springer-Verlag, Berlin, 1981, pp. 213–255.
[13] H. Kunita, *Stochastic differential equations and stochastic flows of diffeomorphisms*, in École d'Été de Probabilités de Saint-Flour XII, P. L. Hennequin, ed., Lecture Notes in Math. 1097, Springer-Verlag, Berlin, 1984, pp. 143–303.
[14] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1990.
[15] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[16] H. J. Kushner, *Stochastic stability*, in Stability of Stochastic Dynamical Systems, R. F. Curtain, ed., Lecture Notes in Math. 294, Springer-Verlag, Berlin, 1972, pp. 97–124.
[17] M. Loève, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, NJ, 1963.
[18] P. Monzón, *On necessary conditions for almost global stability*, IEEE Trans. Automat. Control, 48 (2003), pp. 631–634.
[19] A. Papachristodoulou and S. Prajna, *On the construction of Lyapunov functions using the sum of squares decomposition*, in Proceedings of the 41st IEEE Conference on Decision and Control, 2002, pp. 3482–3487.
[20] P. A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program., 96 (2003), pp. 293–320.
[21] S. Prajna, A. Papachristodoulou, P. Seiler, and P. A. Parrilo, *SOSTOOLS: Sum of Squares Optimization Toolbox for MATLAB. User's Guide, Version* 2.00. Available from http://www.cds.caltech.edu/sostools, 2004.
[22] S. Prajna, P. A. Parrilo, and A. Rantzer, *Nonlinear control synthesis by convex optimization*, IEEE Trans. Automat. Control, 49 (2004), pp. 310–314.
[23] A. Rantzer, *A dual to Lyapunov's stability theorem*, Syst. Control Lett., 42 (2001), pp. 161–168.
[24] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales, Volume* 2: *Itô calculus*, 2nd ed., Cambridge University Press, Cambridge, UK, 2000.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1314–1328

FORWARD STABILIZED SPLINE INTERPOLATION*

DAVID L. RUSSELL†

Abstract. Forward interpolation, as discussed here, refers to the problem of interpolating data points revealed, one at a time, in a sequential manner. As such, forward spline interpolation is an inherently unstable process. In this paper we study forward interpolation using a class of "generalized splines": piecewise polynomials of odd degree $2m-1$, continuous to order $m-1$, where $m \geq 2$ is an integer. We show that the problem can be interpreted as a state estimation problem for a discrete linear system and show how successful, stable procedures can be obtained by applying optimal state estimation techniques of "LQG" type long familiar in the theory of such systems. A number of examples are presented.

Key words. spline interpolation, linear recursion equation, linear difference equation, linear discrete control, linear discrete state estimation, discrete system stabilization

AMS subject classifications. 41A05, 41A15, 65D07, 93B07

DOI. 10.1137/040621727

1. Forward splines in a systems context.

A *standard spline function* of degree n is a piecewise nth degree polynomial $w(t)$ satisfying the differential equation

$$\frac{d^n w}{dt^n} = f(t),$$

with $f(t)$ piecewise constant. The discontinuities of $f(t)$ occur at an increasing sequence of "knots" $\{t_k \,|\, k = 0, 1, 2, 3, \ldots\}$, with $f(t) \equiv f_{k-1}$, $t \in [t_{k-1}, t_k)$. Setting $h_k = t_k - t_{k-1}$ and defining W_k to be the n-dimensional vector whose jth component is $w^{(j)}(t_k)$, $j = 0, 1, 2, \ldots, n-1$, integration shows that W_k satisfies

$$W_k = \mathbf{E}(h_k)\, W_{k-1} + \mathbf{B}(h_k)\, f_{k-1}, \quad k = 1, 2, 3, \ldots, \tag{1.1}$$

where, for any real h,

$$\mathbf{E}(h) = \begin{pmatrix} 1 & h & \frac{h^2}{2} & \cdots & \frac{h^{n-2}}{(n-2)!} & \frac{h^{n-1}}{(n-1)!} \\ 0 & 1 & h & \cdots & \frac{h^{n-3}}{(n-3)!} & \frac{h^{n-2}}{(n-2)!} \\ 0 & 0 & 1 & \cdots & \frac{h^{n-4}}{(n-4)!} & \frac{h^{n-3}}{(n-3)!} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}, \quad \mathbf{B}(h) = \begin{pmatrix} \frac{h^n}{n!} \\ \frac{h^{n-1}}{(n-1)!} \\ \frac{h^{n-2}}{(n-2)!} \\ \vdots \\ h \end{pmatrix}. \tag{1.2}$$

In the standard spline interpolation problem, data values z_k, $k = 0, 1, 2, \ldots, K$, are given in advance, and one wishes to find a spline function $w(t)$ of specified degree, such that $w(t_k) = z_k$ for these values of k. The subject possesses a vast and familiar literature; cf. [3], [11], [6], which we do not add to here. Rather, we are concerned with what we will call *forward spline interpolation*, the name being suggested by comparison with Newton's forward interpolation formula. The data point z_0 is given

*Received by the editors December 29, 2004; accepted for publication (in revised form) April 5, 2006; published electronically October 3, 2006.
http://www.siam.org/journals/sicon/45-4/62172.html
†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060 (russell@math.vt.edu).

initially, and thereafter the z_k, $k = 1, 2, 3, \ldots$, are supplied *one at a time* in the obvious order. The values z_ℓ, $\ell > k$, are unknown, except possibly for some statistical information, at time t_{k-1}. At that point we assume a spline function $w(t)$ has been constructed on $[t_0, t_{k-1})$ interpolating $z_0, z_1, \ldots, z_{k-1}$; z_k then being revealed, it is desired to extend $w(t)$ to the larger interval $[t_0, t_k)$, not changing its definition on $[t_0, t_{k-1})$, with $w(t_k) = z_k$, the overall, repeated process being stable. It turns out this stability does not obtain automatically; it needs to be designed into the process using system theoretic tools.

Splines were studied as discrete linear systems by deBoor and Schoenberg [4] early in the history of the subject but in a "two point boundary value" setting not applicable to the forward interpolation problem. More recently some connections between linear control theory and spline approximation/interpolation have been made in papers by Agwu and Martin [1], Zhang, Tomlinson, and Martin [12], Jackson and Crouch [5], and by Renardy and the present author in [10]. These all deal with spline interpolation in its standard setting and, except for the last cited, in the ordinary differential equation context. The present article studies the forward interpolation problem in the context of *discrete* linear control systems and is developed in terms of *system stabilization* and *state estimation* for such systems.

A picturesque illustration can be given in terms of a ship following a series of lighted buoys through a foggy channel. Visibility is such that, just as each buoy is reached, the next becomes visible. As each buoy is reached a course is plotted out for the next buoy, using a spline function $w(t)$, which can be followed with appropriate motion of the rudder over the time interval $[t_{k-1}, t_k)$. At $t = t_{k-1}$ the position and heading of the ship, assumed to vary continuously with t, correspond to $w_{k-1,0} \equiv w(t_{k-1})$ and $w_{k-1,1} \equiv w'(t_{k-1})$. If $w(t)$ is a piecewise quadratic polynomial, the second derivative $w''(t) \equiv f_{k-1}$ is constant over $[t_{k-1}, t_k)$, corresponding to a constant rudder setting. Assuming we already have $w(t_{k-1}) = z_{k-1}$ and the value $w'(t_{k-1})$, these together with the new requirement $w(t_k) = z_k$ determine a quadratic on $[t_{k-1}, t_k)$. Assuming the intervals $[t_{k-1}, t_k)$ to have constant length h we have

$$w_{k,0} = z_{k-1} + w_{k-1,1}\, h + f_{k-1}\, \frac{h^2}{2}, \qquad w_{k,1} = w_{k-1,1} + f_{k-1}\, h.$$

Taking $w_{k,0} = z_k$, solving for f_{k-1} in the first equation, and substituting the result into the second equation we have

$$w_{k,1} = w_{k-1,1} + \frac{2}{h}\,(z_k - z_{k-1} - w_{k-1,1}h) = -w_{k-1,1} + 2\,\frac{z_k - z_{k-1}}{h}. \tag{1.3}$$

The scalar linear recursion system (1.3) is not asymptotically stable; its single eigenvalue is $\lambda = -1$, and it is not difficult to see that this approach does not yield very satisfactory results. In Figure 1, corresponding to

$$h = 1/3, \quad t_0 = 0, \quad z_k = \begin{cases} 0, & k = 0, \\ t_{k-1}, & k = 1, 2, 3, \\ 1, & k \geq 4, \end{cases} \tag{1.4}$$

the interpolation data z_k are indicated by small circles and the spline interpolant $w(t)$ by a solid curve.

The extension of this "naive" forward spline interpolation process to higher degree standard splines can be described as follows. Taking the first equation of (1.1) for

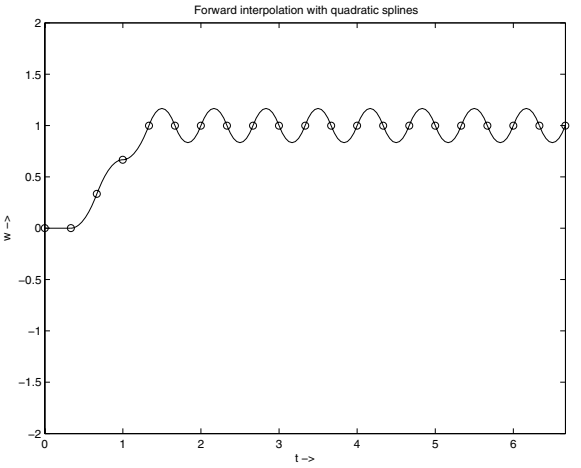


FIG. 1. *Forward interpolation with quadratic splines.*

general $n \geq 2$ we have

$$w_{k,0} = w_{k-1,0} + \sum_{j=1}^{n-1} w_{k-1,j} \frac{h^j}{j!} + f_{k-1} \frac{h^n}{n!}.$$

Setting $w_{k,0} = z_k$, $w_{k-1,0} = z_{k-1}$ and solving for f_{k-1} we have

$$f_{k-1} = \frac{n!}{h^n} \left[(z_k - z_{k-1}) - \mathbf{H}(h) \hat{W}_{k-1} \right], \tag{1.5}$$

where $\mathbf{H}(h) = \left(h \quad \frac{h^2}{2} \quad \cdots \quad \frac{h^{n-1}}{(n-1)!} \right)$ and $\hat{W}$ is a vector of dimension $n-1$ consisting of the $n-1$ highest indexed components of W in (1.1).

THEOREM 1.1. *The naive forward spline procedure just described fails to be asymptotically stable for every $n \geq 2$ and is strongly unstable for $n \geq 3$.*

Proof. Let $\hat{\mathbf{E}}$ denote the $(n-1) \times (n-1)$ matrix obtained by deleting the first column and first row of $\mathbf{E}(h)$ in (1.2), and let $\hat{\mathbf{B}}(h)$ be the $(n-1) \times 1$ matrix $\big((n-1)$-dimensional column vector$\big)$ consisting of the $n-1$ highest indexed components of $\mathbf{B}(h)$ there. Substituting (1.5) into the last $n-1$ equations of (1.1) we obtain the $(n-1)$-dimensional system

$$\hat{W}_k = \left(\hat{\mathbf{E}}(h) - \frac{n!}{h^n} \hat{\mathbf{B}}(h) \mathbf{H}(h) \right) \hat{W}_{k-1} - \frac{n!}{h^n} \hat{\mathbf{B}}(h) \left(z_k - z_{k-1} \right), \quad k = 1, 2, 3, \ldots. \tag{1.6}$$

The matrix $\hat{\mathbf{A}}(h) = \big(\hat{\mathbf{E}}(h) - \frac{n!}{h^n} \hat{\mathbf{B}}(h) \mathbf{H}(h) \big)$ has been studied in the literature; the reader is referred to [4] for the earlier background and to the more recent [10] for a direct, self-contained treatment. From these sources it is known that $\hat{\mathbf{A}}(h)$ has distinct negative (real) eigenvalues occurring in pairs λ, λ^{-1}, except for the single eigenvalue -1 when n is even. (For example, when $n = 3$ the 2×2 matrix $\hat{\mathbf{A}}(h)$ has eigenvalues $-2 \pm \sqrt{3}$.) Clearly we do not have asymptotic stability in these cases. For $n \geq 3$ there must be a real eigenvalue $\lambda < -1$, and we have instability in the strong sense. The proof is complete. ∎

Thus, if we are to pursue our objective further, we must identify a forward spline process which is either inherently stable or which can be stabilized by some means. We will see that stabilization can be accomplished with the use of a class of *generalized splines* (see, e.g., [11, sect. 8.3]); specifically we use piecewise polynomials of odd degree $n = 2m-1$, $m \geq 2$, with $w^{(j)}(t)$, $j = 0, 1, 2, \ldots, m-1$, continuous across the knots. The problem is interpreted in the context of a discrete linear control system similar to (1.1) to which we apply "LQG" stabilization techniques standard for such systems.

We conclude our introductory section with a discussion of the most elementary version of stabilized forward spline interpolation for the case $m = 2, n = 2m-1 = 3$. The generalized splines used here are piecewise polynomials $w(t)$ of degree 3, with $w(t)$ and $w'(t)$ continuous across the equally spaced ($h_k \equiv h > 0$) knots t_k, where they have respective values $w_{k,0}$, $w_{k,1}$. The second and third derivatives have, in general, discontinuities at the t_k; the limits from the right there are denoted by f_k, g_k, respectively. The equations replacing (1.1) then take the form

$$\begin{pmatrix} w_{k,0} \\ w_{k,1} \end{pmatrix} = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix} \begin{pmatrix} w_{k-1,0} \\ w_{k-1,1} \end{pmatrix} + \begin{pmatrix} h^2/2 & h^3/6 \\ h & h^2/2 \end{pmatrix} \begin{pmatrix} f_{k-1} \\ g_{k-1} \end{pmatrix}, \quad k = 1, 2, 3, \ldots. \tag{1.7}$$

We suppose that a data sequence z_j, $j = 0, 1, 2, \ldots$, is made available one term at a time. At time t_{k-1} a generalized spline $w(t)$, piecewise a polynomial of degree 3 with $w(t)$ and $w'(t)$ continuous, has been constructed on $[t_0, t_{k-1})$, interpolating the data z_j, $j = 0, 1, 2, \ldots, k-1$. We wish to develop a stable process whereby $w(t)$ may be extended to $[t_{k-1}, t_k)$, interpolating the new value z_k and maintaining continuity of $w(t)$ and $w'(t)$ at $t = t_{k-1}$.

For real α to be determined, we set

$$g_k = \frac{\alpha f_k}{h}, \quad k = 0, 1, 2, \ldots. \tag{1.8}$$

The first equation of (1.7) then becomes

$$w_{k,0} = w_{k-1,0} + h\, w_{k-1,1} + h^2 \left(\frac{1}{2} + \frac{\alpha}{6}\right) f_{k-1}, \quad k = 1, 2, 3, \ldots. \tag{1.9}$$

It should be noted that (1.8) relates $w''(t_{k-1})$ and $w'''(t_{k-1})$ but does not imply a fixed ratio between $w''(t)$ and $w'''(t)$ over the interval $[t_{k-1}, t_k)$; indeed, we have $w''(t) = f_k + \frac{\alpha f_k}{h}(t - t_{k-1}) = f_k + w'''(t_{k-1})(t - t_{k-1})$ on that interval. Since $w_{k-1,0} = z_{k-1}$ and we are to have $w_{k,0} = z_k$, solving (1.9) for f_{k-1} yields

$$f_{k-1} = \frac{1}{h^2\,(1/2 + \alpha/6)} \left(z_k - z_{k-1} - h\, w_{k-1,1}\right). \tag{1.10}$$

The second equation of (1.7) then takes the "closed loop" form

$$w_{k,1} = \kappa\, w_{k-1,1} + c \left(\frac{z_k - z_{k-1}}{h}\right) \tag{1.11}$$

with

$$c = \frac{1 + \alpha/2}{1/2 + \alpha/6} = \frac{6 + 3\,\alpha}{3 + \alpha}, \qquad \kappa = 1 - c; \text{ reciprocally, } \alpha = -\frac{3\kappa + 3}{\kappa + 2}.$$

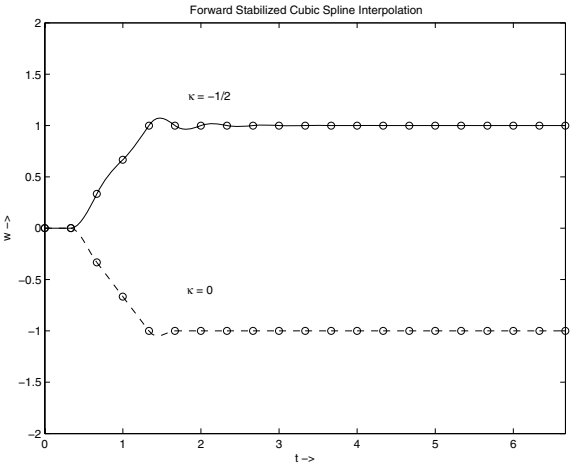


FIG. 2. *Forward stabilized cubic spline interpolation.*

The quadratic spline case corresponds to $\alpha = 0$, $\kappa = -1$, for which the system (1.11) fails to be asymptotically stable. We exclude the case $\alpha = -3$, for which (1.10) is invalid; we cannot achieve $\kappa = -2$; it does not correspond to any valid choice of α.

In discrete linear systems theory a closed loop system wherein all eigenvalues are reduced to zero is called a "deadbeat" system. For most applications deadbeat systems are avoided, but we will see that they play a central role for stabilized forward splines. Here the deadbeat case is $\kappa = 0$, corresponding to $\alpha = -3/2$, $c = 1$, giving $w_{k,1} = \frac{z_k - z_{k-1}}{h}$; $w'(t_k)$ is set equal to the difference quotient $\frac{z_k - z_{k-1}}{h}$. Other cases of interest are $\kappa = 1/2$ and $\kappa = -1/2$, for which (1.11) is exponentially asymptotically stable, corresponding to $\alpha = -9/5$ and $\alpha = -1$, respectively. In all of these cases $w(t)$ is given on $[t_{k-1}, t_k)$, "synthesizing" f_{k-1} via (1.10), by

$$w(t) = w_{k-1,0} + w_{k-1,1}(t - t_{k-1}) + f_{k-1}\left(\frac{(t - t_{k-1})^2}{2} + \alpha\frac{(t - t_{k-1})^3}{6h}\right).$$

We take the case $\kappa = -1/2$, $\alpha = -1$ and apply the method to the data (1.4). The result is shown in the upper part of Figure 2. We display in the lower part of the figure results for the deadbeat case $\kappa = 0$, using the negatives of the z_k in (1.4).

2. Stabilized forward splines of odd degree; error analysis. Let $m \geq 2$ be an integer, let $h_k \equiv h > 0$ constant, and let the "knots" t_k and data z_k be defined as in the preceding sections. Our purpose is to construct a generalized spline function $w(t)$, piecewise of degree $2m-1$ with $w^{(j)}(t)$, $j = 0, 1, 2, \ldots, m-1$, continuous across the knots, sequentially interpolating the data as described earlier. We again define $w_{k,j} = w^{(j)}(t_k)$, $j = 0, 1, 2, \ldots, m-1$. For $j = 0$ we have

$$w_{k,0} = w_{k-1,0} + \sum_{j=1}^{m-1} w_{k-1,j}\frac{h^k}{k!} + f_{k-1}\frac{h^m}{m!} + \sum_{j=1}^{m-1} g_{k-1,j}\frac{h^{j+m}}{(j+m)!}, \tag{2.1}$$

with the "controls" f_{k-1} and $g_{k-1,j}$, $j = 1, 2, \ldots, m-1$, representing the values of $w^{(m)}(t_{k-1}+)$, $w^{(j+m)}(t_k+)$, $j = 1, 2, \ldots, m-1$, respectively. Assuming $w(t)$ already

interpolates the data $z_0, z_1, \ldots, z_{k-1}$ at the knots $t_0, t_1, \ldots, t_{k-1}$, we set

$$g_{k-1,j} \;=\; \frac{a_j\, f_{k-1}}{h^j}, \quad j \;=\; 1, 2, \ldots, m-1, \tag{2.2}$$

for coefficients a_j independent of k but yet to be determined. We substitute (2.2) into (2.1), reminding ourselves that $w_{k-1,0} \;=\; z_{k-1}$ and that we are to have $w_{k,0} \;=\; z_k$. Solving for f_{k-1}, with $a_0 \;=\; 1$,

$$f_{k-1} \;=\; \frac{z_k - z_{k-1} - \sum_{\ell=1}^{m-1} w_{k-1,\ell}\, h^\ell/\ell!}{h^m \left(\sum_{\ell=0}^{m-1} a_\ell/(\ell+m)!\right)}. \tag{2.3}$$

The equations for the derivatives $w^{(j)}(t_\ell) \;=\; w_{\ell,j}$, $j \;=\; 1, 2, \ldots, m-1$, are

$$w_{k,j} = w_{k-1,j} + \sum_{\ell=j+1}^{m-1} w_{k-1,\ell} \frac{h^{\ell-j}}{(\ell-j)!} + f_{k-1} \frac{h^{m-j}}{(m-j)!} + \sum_{\ell=1}^{m-1} g_{k-1,\ell} \frac{h^{\ell+m-j}}{(\ell+m-j)!}. \tag{2.4}$$

Substituting (2.3) into (2.4) we obtain the closed loop equations, for $j \;=\; 1, 2, \ldots, m-1$,

$$w_{k,j} \;=\; w_{k-1,j} \;+\; \sum_{\ell=j+1}^{m-1} w_{k-1,\ell}\, \frac{h^{\ell-j}}{(\ell-j)!} \;+\; \frac{c_j}{h^j} \left(z_k - z_{k-1} \;-\; \sum_{\ell=1}^{m-1} w_{k-1,\ell}\, \frac{h^\ell}{\ell!} \right), \tag{2.5}$$

where

$$c_j \;=\; \frac{\sum_{\ell=0}^{m-1} a_\ell/(\ell+m-j)!}{\sum_{\ell=0}^{m-1} a_\ell/(\ell+m)!}, \quad j \;=\; 1, 2, \ldots, m-1. \tag{2.6}$$

The equations (2.5) combine to form the closed loop system

$$\tilde{W}_k \;=\; \left(\tilde{\mathbf{E}}(h) \;-\; \mathbf{C}(h)\mathbf{H}(h) \right) \tilde{W}_{k-1} \;+\; \mathbf{C}(h)\, (z_k - z_{k-1})\,, \tag{2.7}$$

where $\tilde{\mathbf{E}}(h)$ is the $(m-1)$-dimensional counterpart of $\mathbf{E}(h)$ shown in (1.2), $\tilde{W}$ is the $(m-1)$-dimensional vector consisting of components 1 through $m-1$ of the vector W there, $\mathbf{C}(h)$ is the $(m-1)$-dimensional vector with components c_j/h^j, $j \;=\; 1, 2, \ldots, m-1$, appearing in (2.5), and

$$\mathbf{H}(h) \;=\; \left(h \quad \tfrac{h^2}{2} \quad \cdots \quad \tfrac{h^{m-1}}{(m-1)!} \right). \tag{2.8}$$

From this point on, for notational convenience, we will revert to the use of $\mathbf{E}(h)$ and W, abandoning the $\tilde{}$ modifier. Writing $\zeta_{k-1} \;=\; z_k \;-\; z_{k-1}$ we have

$$W_k = \mathbf{A}(h)\, W_{k-1} + \mathbf{C}(h)\zeta_{k-1}, \quad k \;=\; 1, 2, 3, \ldots; \;\; \mathbf{A}(h) = \mathbf{E}(h) - \mathbf{C}(h)\mathbf{H}(h). \tag{2.9}$$

The row vector $\mathbf{H}(h)$ is fixed as soon as the step length h is chosen. The "feedforward" matrix $\mathbf{C}(h)$ with components c_j/h^j, as shown in (2.5), depends on the choice of the a_j in (2.2), as made clear by (2.6). The a_j, and hence $\mathbf{C}(h)$, are to be chosen so that the system (2.9) is stable and exhibits "desirable" performance characteristics. The stability requirement corresponds to our being able to choose $\mathbf{C}(h)$ so that $\mathbf{A}(h)$ is a discrete stability matrix; i.e., all of its eigenvalues have modulus less than 1. To be

sure we can do this we need to know that the pair $\big(\mathbf{H}(h), \mathbf{E}(h)\big)$ is *observable* ([2], p. 222).

THEOREM 2.1. *The pair* $\big(\mathbf{H}(h), \mathbf{E}(h)\big)$ *is observable.*

Proof. Let P_k, $k = 0, 1, 2, \ldots$, be a vector solution of the homogeneous system $P_k = \mathbf{E}(h)\, P_{k-1}$. Then there is a polynomial $p(t)$ of degree $\leq m-2$ such that $p^{(j)}(t_k) = P_{k,j+1}$, $j = 1, 2, \ldots, m-2$, $k = 0, 1, 2, \ldots$. Let $p^{(-1)}(t)$ be the antiderivative of $p(t)$ such that $p(t_0) = 0$. For $k = 1, 2, 3, \ldots$ we have $p^{(-1)}(t_k) = p^{(-1)}(t_{k-1}) + \mathbf{H}(h)P_{k-1} = \mathbf{H}(h)\mathbf{E}(h)^{k-1}\, P_0$. If the theorem were false, there would be a nonzero vector P_0 such that $\mathbf{H}(h)\mathbf{E}(h)^{k-1}\, P_0 = 0$ for all $k \geq 1$. But then $p^{(-1)}(t_k) = p^{(-1)}(t_0) = 0$, $k = 1, 2, 3, \ldots$, which implies $p^{(-1)}(t) \equiv 0$, and then $P_k \equiv 0$, $k = 0, 1, 2, \ldots$, which is a contradiction. The theorem follows.

With Theorem 2.1 in hand we can conclude, from standard theorems in the control and estimation theory of discrete linear systems (cf. [2], [8], [9]) that $\mathbf{C}(h)$, and hence the c_j appearing in (2.5) can be chosen so that $\mathbf{A}(h) = \mathbf{E}(h) - \mathbf{C}(h)\mathbf{H}(h)$ is a discrete stability matrix; in fact the coefficients c_j can be chosen to achieve any desired eigenvalues for that matrix, including the "deadbeat observer" choice for which all of the eigenvalues of that matrix are zero, and thus $\mathbf{A}(h)$ is nilpotent. □

To begin the error analysis we suppose the data values z_k, $k = 0, 1, 2, \ldots$, are obtained by a sampling process $z_k = z(t_k)$ from an $(m-1)$ times continuously differentiable function $z(t)$. In the context of this paper we would not know a formula for $z(t)$, but we might assume some information about its properties would be available. The accuracy of the generated spline $w(t)$ as an approximation to $z(t)$ for t between the knots t_k depends on how accurately the derivatives of $z(t)$ are matched at the knots by the derivatives of $w(t)$ and on the remainder term $r(t, t_{k-1})$ in the Newton formula

$$z(t) = z(t_{k-1}) + \sum_{\ell=1}^{m-1} \frac{z^{(\ell)}(t_{k-1})}{\ell!}(t - t_{k-1})^{\ell} + r(t, t_{k-1}). \tag{2.10}$$

Repeated differentiation of (2.10) yields, for $j = 1, 2, \ldots, m-1$, the formulae

$$z^{(j)}(t) = z^{(j)}(t_{k-1}) + \sum_{\ell=1}^{m-1-\ell} \frac{z^{(j+\ell)}(t_{k-1})}{\ell!}(t - t_{k-1})^{\ell} + r_j(t, t_{k-1}), \tag{2.11}$$

where $r_j(t, t_{k-1}) = \frac{\partial^j r}{\partial t^j}(t, t_{k-1})$. Taking Z_k, $R_k(h)$ to be the vectors in E^{m-1} with components $Z_{k,j} = z^{(j)}(t_k)$, $R_{k,j}(h) = r_j(h, t_k)$, we obtain the *discrete observed linear system* consisting of the vector counterpart of (2.11),

$$Z_k = \mathbf{E}(h)Z_{k-1} + R_{k-1}(h), \quad k = 1, 2, 3, \ldots, \tag{2.12}$$

and the *linear observation relationship*, or "output,"

$$\zeta_{k-1} = z_k - z_{k-1} = \mathbf{H}(h)Z_{k-1} + r_{k-1}(h) \tag{2.13}$$

which generates the data differences $\zeta_{k-1} = z_k - z_{k-1}$. In (2.12) and (2.13), $\mathbf{E}(h)$ and $\mathbf{H}(h)$ are the $(m-1)\times(m-1)$ matrices discussed following (2.8) and $r_k(h) = r(h, t_k)$. In this context (2.9) can be viewed in a new light; it is a *state estimator system* for the "plant" (2.12) via the output (2.13).

To make use of this insight we introduce the *error vectors*

$$X_k = W_k - Z_k, \quad k = 0, 1, 2, \ldots, \tag{2.14}$$

and see immediately, substracting (2.12) from (2.9), that

$$X_k = \Big(\mathbf{E}(h) - \mathbf{C}(h)\mathbf{H}(h)\Big)X_{k-1} + \mathbf{C}(h)r_{k-1}(h) - R_{k-1}(h)$$

$$\equiv \mathbf{A}(h)X_{k-1} + Q_{k-1}(h), \quad k = 1, 2, 3, \ldots, \tag{2.15}$$

with $\mathbf{A}(h)$ as in (2.9). This *error system* allows us to assess the asymptotic relationship between the spline approximation $w(t)$ and the data source function $z(t)$.

The eigenvalues, but not the eigenvectors, of $\mathbf{A}(h)$ are independent of $h \neq 0$. This follows immediately from

$$\mathbf{J}(h)\mathbf{A}(h)\mathbf{J}(h)^{-1} = \mathbf{A}(1) \equiv \mathbf{A}, \qquad \mathbf{J}(h) = \operatorname{diag}(1 \quad h \quad \cdots \quad h^{m-2}).$$

If we further suppose $z(t)$ to be m times continuously differentiable, we can write

$$r_k(h) = \tilde{r}_k(h)\, h^m, \quad R_{k,\,j}(h) = \tilde{R}_{k,\,j}(h)\, h^{m-j}, \quad j = 1, 2, \ldots, m-1,$$

where $\tilde{r}_k(h), \tilde{R}_{k,\,j}(h)$ are uniformly bounded for t_k in a fixed, compact interval. Then (2.15) transforms into

$$\tilde{X}_k = \mathbf{A}\tilde{X}_{k-1} + \tilde{Q}_{k-1}(h)h^{m-1}, \qquad \tilde{X}_k = \mathbf{J}(h)\, X_k, \tag{2.16}$$

where the $\tilde{Q}_k(h)$ are similarly uniformly bounded. Then we have the following theorem.

THEOREM 2.2. *Let $z(t)$ be m times continuously differentiable for t in a compact interval $\mathcal{I}$, and let all knots t_k under consideration lie in $\mathcal{I}$. Let $\mathbf{C}(1)$ be chosen so that all eigenvalues λ of $\mathbf{A} = \mathbf{E}(1) - \mathbf{C}(1)\mathbf{H}(1)$ satisfy $|\lambda| < 1$. Let $z_k = z(t_k)$, $k = 0, 1, 2, \ldots$, and let $W_0 = 0$. Then there exist constants γ_0, γ_1, μ, $0 \leq \gamma_0 < 1$, $\gamma_1 \geq 0$, $\mu \geq 0$, independent of k and $h \neq 0$, such that*

$$\left\|\tilde{X}_k\right\| \leq \mu\, \gamma_0^k \left\|\tilde{X}_0\right\| + \gamma_1\, h^{m-1}, \quad k = 1, 2, 3, \ldots. \tag{2.17}$$

Proof. From (2.16) we have, for $k = 1, 2, 3, \ldots$,

$$\tilde{X}_k = \mathbf{A}^k \tilde{X}_0 + h^{m-1} \sum_{\ell=0}^{k-1} \mathbf{A}^{k-1-\ell} \tilde{Q}_{k-1}(h) \equiv \tilde{X}_{0,k} + \tilde{X}_{1,k}. \tag{2.18}$$

Assuming $\mathbf{C}(1)$ is chosen as indicated, we take γ_0 to be the largest modulus of any eigenvalue of $\mathbf{A}$. Then there is a positive number μ such that $\left\|\mathbf{A}^\ell\right\| \leq \mu\, \gamma_0^\ell$ for all nonnegative integers ℓ. Thus

$$\left\|\tilde{X}_k\right\| \leq \mu\, \gamma_0^k \left\|\tilde{X}_0\right\| + \mu\, h^{m-1} \sum_{\ell=0}^{k-1} \gamma_0^{k-1-\ell} \left\|\tilde{Q}_{k-1}(h)\right\|.$$

Under our assumptions there is an upper bound $\tilde{Q}$ for the norms $||\tilde{Q}_{k-1}(h)||$, independent of k and h. Then

$$\left\|\tilde{X}_k\right\| \leq \mu\, \gamma_0^k \left\|\tilde{X}_0\right\| + \frac{\mu\, h^{m-1}\tilde{Q}}{1-\gamma_0} = \mu\, \gamma_0^k \left\|\tilde{X}_0\right\| + \gamma_1 h^{m-1}, \tag{2.19}$$

and the proof is complete. □

From (2.16) and (2.18) we have

$$X_k = \mathbf{J}(h)^{-1}\tilde{X}_{0,k} + \mathbf{J}(h)^{-1}\tilde{X}_{1,k} \equiv X_{0,k} + X_{1,k}.$$

These are the *transient* and *steady state* components of the error vector X_k. For the components $X_{1,k,j}$ of the steady state error we clearly have

$$|X_{1,k,j}| \le \gamma_1 h^{m-1}h^{1-j} = \gamma_1 h^{m-j}, \quad j = 1, 2, \ldots, m-1.$$

For the components $X_{0,k,j}$ of the transient error $X_{0,k}$, on the other hand, we have

$$|X_{0,k,j}| \le h^{1-j}\mu\gamma_0^k \left\|\mathbf{J}(h)X_0\right\|.$$

To pursue a different matter, let us suppose $\mathbf{C}(1)$ is chosen so that $\mathbf{A}$ has only zero eigenvalues and hence is nilpotent. Then (2.9) serves as a deadbeat state estimator for (2.12). If the data z_k correspond to values $z(t_k)$, with $z(t)$ an arbitrary polynomial of degree $\le m-1$, then the terms R_{k-1}, r_{k-1} in (2.12) and (2.15) are all zero and the latter becomes

$$X_k = \mathbf{A}(h)X_{k-1}, \quad k = 1, 2, 3, \ldots,$$

implying that $X_k = 0$, $k \ge m-1$. Then using $\mathbf{A}(h)^{m-1} = \mathbf{O}$ together with (2.9) we have

$$Z_{m-1} = W_{m-1} = \sum_{k=1}^{m-1} \mathbf{A}(h)^{m-1-k}\mathbf{C}(h)\left(z_k - z_{k-1}\right), \tag{2.20}$$

showing the components of W_{m-1}, i.e., the derivatives $w^{(j)}(t_{m-1})$, $j = 1, 2, \ldots, m-1$, of $w(t)$ at $t = t_{m-1}$, to be linear combinations of z_k, $k = 0, 1, 2, \ldots, m-1$. Since that is true for all polynomials $z(t)$ of degree $\le m-1$, and since $W_{m-1} = Z_{m-1}$, we conclude that the component equations of (2.20) must agree with the backward finite difference expressions for the derivatives $z^{(j)}(t_{m-1})$ in terms of z_k, $k = 0, 1, 2, \ldots, m-1$ (see, e.g., [6]). The simplest case, for stabilized forward cubic spline interpolation, has already been studied in section 1, where we saw that the deadbeat case corresponds to setting $w'(t_k) = \frac{z_k - z_{k-1}}{h}$, $k \ge 1$. Significantly, in his text [3] de Boor suggests the possible use of such expressions as alternate boundary conditions for spline interpolation (as compared, for example, with the "natural" boundary conditions described there).

3. Stochastic optimality. In the previous sections we have established stabilizability of the forward interpolation system via choice of the feedforward vector $\mathbf{C}(h)$, equivalently $\mathbf{C}(1)$, and we have obtained some general error bounds. The next question naturally arising concerns which choice of $\mathbf{C}(h)$ is "best." This clearly depends on the criteria used and on our notion of how the data values z_k are generated, in other words on the formulation of a *model process* upon whose operation the z_k are assumed to constitute an *observation*. We have started on this in the preceding section, interpreting our process as a plant/state estimator system. Now we need to go further and introduce some statistical concepts into that system. The following development is based on several sources, including [2], [8], and [9].

The first step is to replace (2.12) by

$$Z_k = \mathbf{E}(h)\, Z_{k-1} + N_{k-1}, \tag{3.1}$$

wherein the vectors N_k are now assumed to constitute discrete white noise; i.e., N_k and N_j are uncorrelated for $k \neq j$, with zero mean and constant covariance matrix $\mathbf{N} \equiv \mathcal{E}\,(N_k\, N_k^*)\,,\ \mathcal{E}$ denoting the expected value. These noise terms will be referred to as the *process noise*. Then we replace (2.13) by

$$\zeta_{k-1} \,=\, z(t_k) - z(t_{k-1}) \,=\, \mathbf{H}(h)\, Z_{k-1} \,+\, n_{k-1}, \tag{3.2}$$

where $\mathbf{H}(h)$ is given in (2.8) and the sequence $\{n_{k-1}\}$ is assumed to be scalar white noise with zero mean and covariance $\mathbf{n}$; this is the *observation noise*.

In most applications the remainders R_k, r_k of the preceding section are not white noise, but it will often be advantageous to treat them as if they were. This parallels situations in systems theory wherein unmodelled states, nonlinearities, disturbances, etc., are often lumped together and treated as white noise in design procedures. Our model process might be described as a *stochastically interrupted polynomial* of degree $m-1$. On each interval $[t_{k-1}, t_k)$, $q_k(t) \,\equiv\, z(t_{k-1}) \,+\, \mathbf{H}(t - t_{k-1})Z(t_{k-1})$ is a polynomial of degree $\leq m-1$. If the N_k and n_k were identically zero, we would have $q_k(t) \equiv q(t)$, where $q(t)$ is a fixed polynomial of degree $\leq m-1$ for all t. If there is an integer μ, $1 \leq \mu \leq m-1$, such that $N_{k,j} \equiv 0$ when $j \leq \mu$, we say that $q(t)$ has *smoothness level* μ, since derivatives of order $\leq \mu$ are continuous across the knots.

From (2.9), (3.1), and (3.2) we have the combined plant/state estimator system

$$W_k \,=\, \mathbf{E}(h)\, W_{k-1} \,+\, \mathbf{C}(h)\Big(\zeta_k \,-\, \mathbf{H}(h)W_{k-1}\Big),$$

$$\zeta_k \,=\, \mathbf{H}(h)\, Z_{k-1} \,+\, n_{k-1}, \qquad Z_k \,=\, \mathbf{E}(h)\, Z_{k-1} \,+\, N_{k-1}, \tag{3.3}$$

in which the model system (3.1) plays the role of the plant and the spline generator system (2.9) plays the role of state estimator. Defining the *state estimation error* X_k as in (2.14) and subtracting the last equation of (3.3) from the first we have

$$X_k \,=\, \Big(\mathbf{E}(h) \,-\, \mathbf{C}(h)\mathbf{H}(h)\Big)X_{k-1} \,+\, \mathbf{C}(h)n_{k-1} - N_{k-1}. \tag{3.4}$$

Normally $W_0 \,=\, 0$ because we have no information on derivatives at $t \,=\, t_0$. This gives $X_0 \,=\, -\,Z_0$. Let $\mathbf{X}_k$ be the covariance matrix of X_k. Then (see, e.g., [2])

$$\mathbf{X}_k \,=\, \mathbf{A}(h)\mathbf{X}_{k-1}\mathbf{A}(h)^* \,+\, \mathbf{N} \,+\, \mathbf{n}\,\mathbf{C}(h)\mathbf{C}(h)^*, \quad k \,=\, 1, 2, \ldots, \quad \mathbf{X}_0 \,=\, \mathbf{Z}_0, \tag{3.5}$$

wherein $\mathbf{A}(h)$ is the "closed loop matrix," $\mathbf{A}(h) \,=\, \mathbf{E}(h) \,-\, \mathbf{C}(h)\mathbf{H}(h)$, and $\mathbf{Z}_0$ is the covariance matrix for the (generally unknown) initial "plant state" Z_0. If the feedforward vector $\mathbf{C}(h)$ is chosen so that $\mathbf{A}(h)$ is a discrete stability matrix, then

$$\lim_{k \,\to\, \infty} \mathbf{X}_k \,\equiv\, \mathbf{X} \tag{3.6}$$

exists and represents the ultimate, "steady state" covariance of the error vectors X_k.

The goal of optimal state estimation in this context is to find $\hat{\mathbf{C}}(h)$ such that $\mathbf{A}(h)$ is a discrete stability matrix and the corresponding limiting error covariance $\hat{\mathbf{X}}$ is minimal; i.e., $\hat{\mathbf{X}} \,\leq\, \mathbf{X}$ for $\mathbf{X}$ associated with any other stabilizing feedforward matrix $\mathbf{C}(h)$. This problem has a long history, and the solution is well known; cf., e.g., [2], [7]. The optimal $\hat{\mathbf{C}}(h)$ is given by

$$\hat{\mathbf{C}}(h) \,=\, -\,\mathbf{E}(h)\hat{\mathbf{X}}\mathbf{H}(h)^*\Big(\mathbf{H}(h)\hat{\mathbf{X}}\mathbf{H}(h)^* + \mathbf{n}\Big)^{-1}.$$

The matrix $\hat{\mathbf{X}}$ is given by (3.6) with $\hat{\mathbf{X}}_k$ replacing $\mathbf{X}_k$ and $\hat{\mathbf{C}}(h)$ replacing $\mathbf{C}(h)$ in (3.5). The circularity of this description is escaped by algebraic manipulations which eliminate $\hat{\mathbf{C}}(h)$ from the corresponding version of (3.5); the $\hat{\mathbf{X}}_k$ satisfy the quadratic matrix recursion equation

$$\hat{\mathbf{X}}_k \;=\; \mathbf{E}(h)\hat{\mathbf{X}}_{k-1}\mathbf{E}(h)^* \;+\; \mathbf{N}$$

$$-\,\mathbf{E}(h)\hat{\mathbf{X}}_{k-1}\mathbf{H}(h)^*\Big(\mathbf{H}(h)\hat{\mathbf{X}}_{k-1}\mathbf{H}(h)^* + \mathbf{n}\Big)^{-1}\mathbf{H}(h)\hat{\mathbf{X}}_{k-1}\mathbf{E}(h)^*. \tag{3.7}$$

As is well known and documented in the indicated references,

$$\lim_{k\,\to\,\infty}\,\hat{\mathbf{X}}_k \;\equiv\; \hat{\mathbf{X}}, \tag{3.8}$$

which, in turn, satisfies the steady state quadratic matrix equation

$$\hat{\mathbf{X}} \;-\; \mathbf{E}(h)\hat{\mathbf{X}}\mathbf{E}(h)^* \;-\; \mathbf{N}$$

$$+\,\mathbf{E}(h)\hat{\mathbf{X}}\mathbf{H}(h)^*\Big(\mathbf{H}(h)\hat{\mathbf{X}}\mathbf{H}(h)^* + \mathbf{n}\Big)^{-1}\mathbf{H}(h)\hat{\mathbf{X}}\mathbf{E}(h)^* \;=\; 0. \tag{3.9}$$

The recursion formula (3.7) can be used with $\hat{\mathbf{X}}_0 \;=\; \mathbf{N}$ to produce a sequence $\hat{\mathbf{X}}_k$ which converges quadratically as $k \;\to\; \infty$ to the unique positive symmetric solution $\hat{\mathbf{X}}$ of (3.9). This is frequently used as a numerical procedure to compute the latter matrix; care has to be taken to keep the computed $\hat{\mathbf{X}}_k$ symmetric at each step.

4. An example: $m \;=\; 3,\ n \;=\; 5$. After the case $n \;=\; 3$ discussed in section 1 the next relevant case is $n \;=\; 5,\ m \;=\; 3$. For $k \;=\; 1, 2, 3, \ldots$ we have

$$\begin{pmatrix} w_{k,0} \\ w_{k,1} \\ w_{k,2} \end{pmatrix} \;=\; \begin{pmatrix} 1 & h & \frac{h^2}{2} \\ 0 & 1 & h \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_{k-1,0} \\ w_{k-1,1} \\ w_{k-1,2} \end{pmatrix} + \begin{pmatrix} \frac{h^3}{6} & \frac{h^4}{24} & \frac{h^5}{120} \\ \frac{h^2}{2} & \frac{h^3}{6} & \frac{h^4}{24} \\ h & \frac{h^2}{2} & \frac{h^3}{6} \end{pmatrix} \begin{pmatrix} f_{k-1} \\ g_{k-1,1} \\ g_{k-1,2} \end{pmatrix}. \tag{4.1}$$

The specialization of (2.2) here is

$$g_{k-1,1} \;=\; \frac{a_1 f_{k-1}}{h},\quad g_{k-1,2} \;=\; \frac{a_2 f_{k-1}}{h^2},\quad k \;=\; 1, 2, 3, \ldots. \tag{4.2}$$

We substitute (4.2) into (4.1), recall that $w_{k,0} \;=\; z_k$, the given data, and use the resulting first equation to obtain the feedback relation

$$f_k \;=\; \frac{z_k - z_{k-1} - w_{k-1,1}h - w_{k-1,2}h^2/2}{h^3\,(1/6 + a_1/24 + a_2/120)}. \tag{4.3}$$

Then we substitute (4.3) into the second and third equations of (4.1) to obtain

$$\begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix} \;=\; \begin{pmatrix} 1 - c_1 & (1 - c_1/2)\,h \\ -c_2/h & 1 - c_2/2 \end{pmatrix} \begin{pmatrix} w_{k-1,1} \\ w_{k-1,2} \end{pmatrix} + \begin{pmatrix} c_1/h \\ c_2/h^2 \end{pmatrix} (z_k - z_{k-1}), \tag{4.4}$$

where

$$c_1 \;=\; \frac{1/2 + a_1/6 + a_2/24}{1/6 + a_1/24 + a_2/120}, \qquad c_2 \;=\; \frac{1 + a_1/2 + a_2/6}{1/6 + a_1/24 + a_2/120}. \tag{4.5}$$

The equations (4.5) can be solved for a_1 and a_2 to give

$$a_1 = -8\left(\frac{7\,c_1 - c_2 - 15}{8\,c_1 - c_2 - 20}\right), \qquad a_2 = 20\left(\frac{6\,c_1 - c_2 - 12}{8\,c_1 - c_2 - 20}\right), \tag{4.6}$$

provided c_1 and c_2 are such that $8\,c_1 - c_2 - 20 \neq 0$.

PROPOSITION 4.1. *The restriction* $8\,c_1 - c_2 - 20 \neq 0$ *excludes no pairs* c_1, c_2 *for which the matrix*

$$\tilde{\mathbf{A}}(h) = \begin{pmatrix} 1 - c_1 & (1 - c_1/2)\,h \\ -c_2/h & 1 - c_2/2 \end{pmatrix} \tag{4.7}$$

in (4.4) *is a discrete stability matrix.*

Remark. We saw in section 1 that a similar result applies for the cubic case $m = 2$, $n = 3$. Whether this remains valid for general m, $n = 2m - 1$, remains open.

Proof. The matrix (4.7) has characteristic polynomial with roots

$$\lambda = \frac{1}{2}\left(2 - \left(c_1 + \frac{c_2}{2}\right) \pm \sqrt{\left(c_1 + \frac{c_2}{2}\right)^2 - 4c_2}\right). \tag{4.8}$$

Solving $8c_1 - c_2 - 20 = 0$ for c_2 and substituting into (4.8) we obtain

$$\lambda = \frac{1}{2}\left(12 - 5c_1 \pm \sqrt{25\,c_1^2 - 132\,c_1 + 180}\right).$$

The discriminant of the quadratic under the square root sign is $(132)^2 - 18000 = -576$ so that quadratic is positive for every choice of c_1. For any choice of c_1 the absolute value of the difference between the two roots is $\sqrt{25\,c_1^2 - 132\,c_1 + 180}$, which has a minimum value $12/5 = 2.4$. Since this minimum exceeds 2, no choice of c_1, hence no choice of c_1, c_2, satisfying $8c_1 - c_2 - 20 = 0$ yields both roots with modulus less than 1. We conclude that the exceptional cases described by $8c_1 - c_2 - 20 = 0$ do not limit the choice of stabilizing pairs c_1, c_2; the proof is complete. ∎

Once the closed loop system (4.4) has been obtained the next step is to obtain c_1, c_2 stabilizing the corresponding homogeneous system. Confirming Theorem 2.1, the pair consisting of $\mathbf{H} = (\,1 \quad 1/2\,)$ and $\mathbf{E} = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ is readily seen to be *observable*, i.e., $\operatorname{rank}\left(\begin{smallmatrix} \mathbf{H} \\ \mathbf{HE} \end{smallmatrix}\right) = 2$, so the system is stabilizable [2]; by proper choice of c_1 and c_2 the eigenvalues of $\tilde{\mathbf{A}}$ can be placed at will. From the form (4.8) of the roots, deadbeat stabilization is achieved with $c_1 = 3/2$, $c_2 = 1$, which in turn, from (4.6), corresponds to $a_1 = -44/9 = -4.8889$, $a_2 = 80/9 = 8.8889$. Here system (4.4) becomes

$$\begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix} = \begin{pmatrix} -1/2 & h/4 \\ -1/h & 1/2 \end{pmatrix} \begin{pmatrix} w_{k-1,1} \\ w_{k-1,2} \end{pmatrix} + \begin{pmatrix} 3/2h \\ 1/h^2 \end{pmatrix} (z_k - z_{k-1})\,. \tag{4.9}$$

Substituting (4.9), with k replaced by $k - 1$, back into (4.9) one obtains

$$\begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix} = \begin{pmatrix} \frac{3}{2h}\,(z_k - z_{k-1}) - \frac{1}{2h}\,(z_{k-1} - z_{k-2}) \\ \frac{1}{h^2}\,(z_k - 2\,z_{k-1} + z_{k-2}) \end{pmatrix}.$$

Thus, with the indicated values of c_1 and c_2, after two steps the procedure sets the derivative, $w_{k,1}$, and the second derivative, $w_{k,2}$, equal to backward divided differences known [6] to be exact expressions for $w'(t_k)$ and $w''(t_k)$ when $w(t)$ is of degree 2.

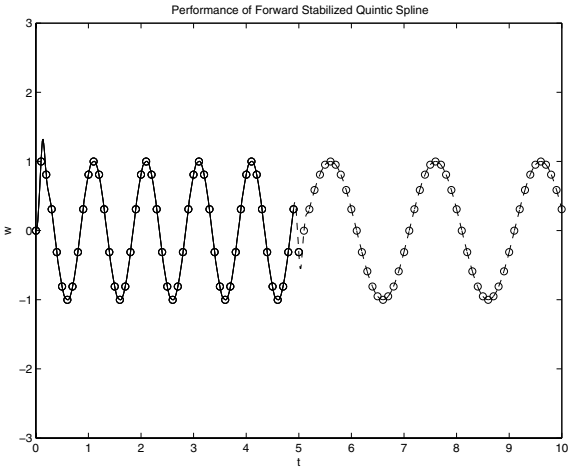


FIG. 3. *Performance of forward stabilized quintic spline.*

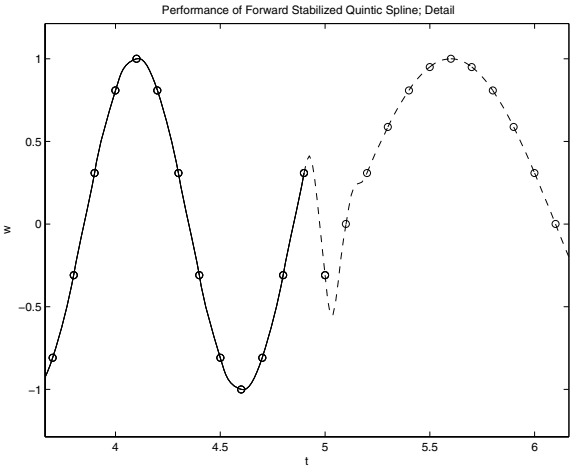


FIG. 4. *Performance of forward stabilized quintic spline; detail.*

Figures 3 and 4 show results when the deadbeat forward quintic spline procedure is carried out with data taken from a function $z(t)$ which combines two different trigonometric functions. The overall interval has length 10 and $h = .2$. Initially $z_0 = z(0) = 0$; thereafter, while $t < 5$, we have $z(t) = \cos(2\pi t)$; a discontinuity occurs at $t = 5$; we have $z(t) = \sin(\pi(t - 5.1))$ for $t \geq 5$. Figure 4 shows the situation near $t = 5$ in greater detail. The discontinuity is treated by the procedure in a controlled manner over the interval $[t_{24}, t_{27}] = [4.8, 5.2]$, and convergence to splines closely approximating $\sin(\pi(t - 5.1))$ occurs rapidly (in two steps, essentially) thereafter.

In the present quintic case a straightforward computation with (3.9) shows that the deadbeat case, wherein the eigenvalues of the closed loop matrix $\tilde{\mathbf{A}}$ are both equal to zero, corresponds to $\mathbf{N} = \operatorname{diag}(1 \quad 4)$, $\mathbf{n} = 0$; i.e., zero observation noise is assumed. This result was obtained by a trial and error process after initially guessing

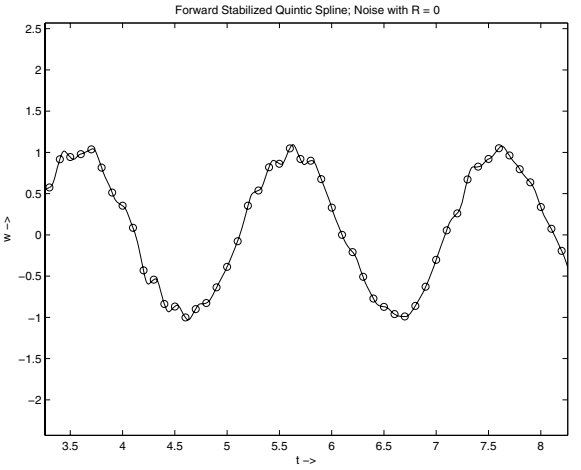


FIG. 5. *Forward stabilized quintic spline; noise with R=0.*

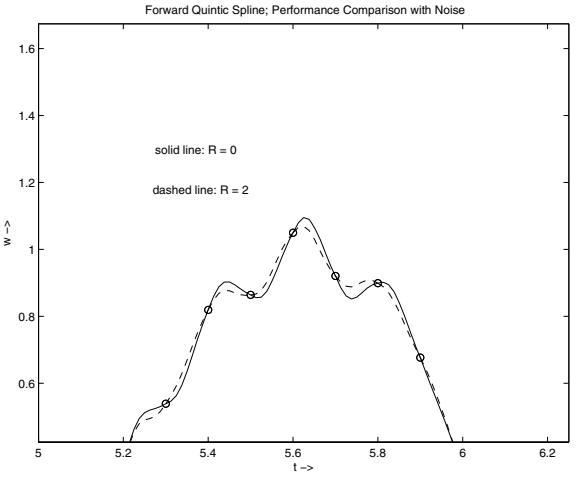


FIG. 6. *Forward quintic spline; performance comparison with noise.*

that the deadbeat estimator would correspond to the zero observation noise situation. A similarly unsystematic treatment of the seventh degree case $m = 4$ (not detailed here) shows the deadbeat state estimator to be achieved in that case with $\mathbf{n} = 0$ with process noise covariance matrix $\mathbf{N} = \text{diag}\,(1 \quad 6 \quad 18)$. Whether for general m the deadbeat estimator always corresponds to diagonal $\mathbf{N}$ with integer diagonal weights combined with $\mathbf{n} = 0$, and what the general form of those diagonal weights might be, are questions for which we have no answer at the present time.

In Figure 5 we show results for this case with data obtained from a sinusoidal function with knot values perturbed by random noise in the range $[-.25, .25]$ supplied by MATLAB. In Figure 6 we look more closely at the central "peak" of Figure 5. The solid curve corresponds to the deadbeat case $\mathbf{n} = 0$ of Figure 6, while the dashed curve corresponds to the same $\mathbf{N}$ but with $\mathbf{n} = 2$, which amounts to assuming observation noise present with that covariance. We observe that the assumption that

noise is present in the observation appears to lead to interpolation with less overshoot between the knots.

REFERENCES

[1] N. N. Agwu and C. F. Martin, *Optimal control of dynamic systems: Application to spline approximations*, Appl. Math. Comput., 97 (1998), pp. 99–138.

[2] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice–Hall, Englewood Cliffs, NJ, 1979.

[3] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, 1978.

[4] C. de Boor and I. J. Schoenberg, *Cardinal interpolation and spline functions;* VIII: *The Budan-Fourier theorem for splines and applications*, in Spline Functions (Proc. Internat. Sympos., Karlsruhe, 1975), Lecture Notes in Math. 501, Springer-Verlag, Berlin, 1976, pp. 1–79.

[5] J. W. Jackson and P. E. Crouch, *Dynamic interpolation and application to flight control*, J. Guidance Control Dynam., 14 (1991), pp. 814–822.

[6] D. Kincaid and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, 3rd ed., Brooks/Cole, Pacific Grove, CA, 2002.

[7] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[8] V. Kucera, *Discrete Linear Control: The Polynomial Equation Approach*, John Wiley and Sons, Chichester, UK, 1979.

[9] M. S. Mahmoudi and M. G. Singh, *Discrete Systems, Analysis, Control, and Optimization*, Springer-Verlag, Berlin, New York, 1984.

[10] M. Renardy and D. L. Russell, *Spline interpolation in a discrete control systems context*, Linear Algebra Appl., 406 (2005), pp. 77–98.

[11] L. L. Schumaker, *Spline Functions: Basic Theory*, John Wiley and Sons, New York, 1981; reprinted, Krieger, Malabar, FL, 1993.

[12] Z. Zhang, J. Tomlinson, and C. F. Martin, *Splines and linear control theory*, Acta. Appl. Math., 49 (1997), pp. 1–34.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1329–1358

OPTIMAL DISTURBANCE ATTENUATION FOR DISCRETE-TIME SWITCHED AND MARKOVIAN JUMP LINEAR SYSTEMS*

JI-WOONG LEE† AND GEIR E. DULLERUD‡

Abstract. An exact condition for uniform stabilization and disturbance attenuation for switched linear systems is given in the discrete-time domain via the union of an increasing family of linear matrix inequality conditions. Associated with each Markovian jump linear system is a switched linear system, so we obtain a necessary and sufficient condition for almost sure uniform stabilization and disturbance attenuation for Markovian jump linear systems as well. The results lead to semidefinite programming–based controller synthesis techniques, from which optimal finite-path-dependent linear dynamic output feedback controllers arise naturally. In particular, under the notion of path-by-path optimal disturbance attenuation, finite-path-dependent controllers can outperform the usual mode-dependent ones.

Key words. discrete linear inclusions, dynamic output feedback, H^∞ control, linear matrix inequalities, linear time-varying systems, semidefinite programming

AMS subject classifications. 93B36, 93B12, 90C22, 93C55

DOI. 10.1137/050627538

1. Introduction.

The switched linear system is defined as a family of linear time-varying systems whose parameters vary within a single finite set, and it serves as an abstraction of hybrid systems, where continuous system dynamics and discrete events coexist and depend on each other in complex ways [31]. On the other hand, the Markovian jump linear system is a linear system whose parameters jump according to the state transitions of a finite-state Markov chain, and it typically arises in the context of networked control systems, where the feedback loop is subject to random delays [32]. The parameters of these systems are indexed, where the indices are called the modes of the systems. This paper focuses on the discrete-time domain and considers the problems of uniform disturbance attenuation for switched linear systems and almost sure uniform disturbance attenuation for Markovian jump linear systems, where the uniformity refers to the stability and ℓ^2-gain of the sequences of modes, called switching sequences, that are admissible. We develop semidefinite programming formulations [41] for the solutions to these problems, and their generalizations, without any assumption on the parameters or admissible switching sequences.

The most fundamental problem of discrete-time switched linear systems is determining the asymptotic stability of the discrete linear inclusion (i.e., the family of linear time-varying systems over all infinite sequences of parameters in a given finite set) [12, 21]. This problem basically involves checking the stability of all possible (uncountably many) infinite matrix products due to the fact that the finiteness conjecture (i.e., the conjecture that it suffices to check the stability of periodic products) is generally false [8]. The problem is hence considered undecidable [7]; more precisely,

*Received by the editors March 24, 2005; accepted for publication (in revised form) April 4, 2006; published electronically October 3, 2006. This work was supported by AFOSR MURI grant F49620-02-1-0325 and by AFOSR URI grant F49620-01-1-0365.

http://www.siam.org/journals/sicon/45-4/62753.html

†Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 (jiwoong@ufl.edu).

‡Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 (dullerud@uiuc.edu).

it is considered semidecidable (i.e., there exists an algorithm that is guaranteed to correctly determine, after a finite amount of computation, the stability of switched linear systems that are indeed stable) [6]. Recently, the authors gave a restatement of the result of [6] and introduced an increasing family of linear matrix inequality conditions whose union characterizes the uniform stability and stabilizability of discrete-time switched linear systems [30]. This result generalizes the multiple Lyapunov function approaches that have provided the most useful tools for stability analysis [9].

This paper complements the previous work on uniform stabilization described above and extends it to disturbance attenuation. The problem of disturbance attenuation reduces to that of $\mathcal{H}^\infty$ control of linear time-invariant systems [16, 26] if the given set of system parameters is a singleton. On the other hand, if the family of admissible switching sequences is a singleton, the problem reduces to that of ℓ^2-induced norm minimization for linear time-varying systems [36, 4], where the set of possible parameter values is finite. Therefore, our work draws on the results in the $\mathcal{H}^\infty$ control and ℓ^2-gain minimization literature. In particular, from the linear operator inequality approach for analyzing linear time-varying systems [17], we deduce that the Riccati difference inequality associated with any stable and contractive linear time-varying system admits a uniformly stabilizing solution that has a finite memory of past parameters. This finite-memory property leads to our analysis results. On the other hand, our controller synthesis results are based on a straightforward extension of the linear matrix inequality approach originally developed for linear time-invariant systems [33, 20].

We relax the standard restriction to mode-dependent controllers (i.e., controllers that perfectly observe the present mode of the system but do not recall past modes) and consider finite-path-dependent controllers (i.e., controllers that not only perfectly observe the present mode but also have a finite memory of past modes). This relaxation, along with the aforementioned finite-memory property, enables us to derive a complete characterization of the existence of finite-path-dependent linear dynamic output feedback controllers that stabilize the switched linear system and achieve a uniform disturbance attenuation level. This characterization is given in terms of the "union" of an increasing family of systems of linear matrix inequality conditions. Although our result inherits the semidecidable nature of the underlying problem, this limitation is not likely to pose difficulties in practice, as examples show that it usually suffices to check the feasibility of the first few systems of linear matrix inequalities from an increasing family. Moreover, the result is amenable to the standard linear matrix inequality–based controller synthesis technique, from which admissible finite-path-dependent controller syntheses arise naturally.

The notion of finite-path-dependent controllers not only serves as a relaxation to achieve an exact condition for disturbance attenuation but also is required for optimality under the notion of path-by-path disturbance attenuation. Roughly speaking, the path-by-path performance is defined as a finite family of disturbance attenuation levels over all admissible switching paths of a given length. This notion of performance is a natural extension of that of uniform disturbance attenuation and can be used to improve upon the performance of a controller synthesis that achieves a given uniform disturbance attenuation level. As long as the path-by-path optimality is concerned, examples show that finite-path-dependent controllers can outperform mode-dependent controllers.

The result described above carries over to the almost sure uniform disturbance attenuation for Markovian jump linear systems because the uniformity requirement for stability and performance enables one to consider each Markovian jump linear

system as a switched linear system, where the underlying Markov chain of the former defines the switching path constraint of the latter. An immediate consequence is that, as long as almost sure uniform disturbance attenuation is concerned, the performance of Markovian jump linear systems is intrinsically robust against variations that preserve the directed graph of the underlying Markov chain (i.e., against sparsity pattern–preserving variations from the given transition probability matrix and initial distribution). On the other hand, under the notion of path-by-path disturbance attenuation, one can reduce the conservatism associated with this robustness property while still maintaining almost sure uniform stability; two Markov chains with the same directed graph can result in widely different path-by-path performances, depending on the actual values of initial and transition probabilities.

A key contribution of this paper to the control of switched linear systems and Markovian jump linear systems is that exact "control-oriented" conditions for (almost sure) stabilization and disturbance attenuation are provided. These conditions are control oriented in the sense that they lead to semidefinite programming–based techniques, which render optimal controllers very efficiently. There has been little work done on the control of disturbance attenuation performances of switched linear systems, other than that some partial analysis results in the continuous-time domain exist [44, 23]. On the other hand, the usual approach in the literature to disturbance attenuation for Markovian jump linear systems has been based on the notion of stochastic stability in both continuous time [28, 34, 13] and discrete time [11, 19, 10, 37]. These results, however, though either partly satisfactory or exact, are not well-suited for efficient optimal controller synthesis. For instance, in the case of independent and identically distributed switching, an exact linear matrix inequality–based synthesis condition for discrete-time jump systems has been obtained in [38]; for general Markov switching, this condition is only sufficient. Moreover, even though stochastically stable Markovian jump linear systems are almost surely stable [29], the usual approach does not guarantee almost sure disturbance attenuation.

The paper is organized as follows. Section 2 analyzes the performance of general linear time-varying systems. Then, based on this analysis, main results on uniform disturbance attenuation for switched linear systems and Markovian jump linear systems are derived in sections 3 and 4, respectively. Section 5 introduces the notion of path-by-path disturbance attenuation and provides some examples including optimal disturbance attenuation for a planar robot, called the Pendubot, subject to random delays in the feedback loop. Finally, concluding remarks are given in section 6.

Notation. If $\mathbf{X} \in \mathbb{R}^{m\times n}$, the range (or image) of $\mathbf{X}$ is denoted by $\operatorname{Im}\mathbf{X}$, the null space (or kernel) of $\mathbf{X}$ by $\operatorname{Ker}\mathbf{X}$, and the rank of $\mathbf{X}$ by $\operatorname{rank}\mathbf{X}$; denoted by $\mathbf{N}(\mathbf{X})$ is any particular full-rank matrix such that $\operatorname{Im}\mathbf{N}(\mathbf{X}) = \operatorname{Ker}\mathbf{X}$. The indicator matrix $\mathbf{1}(\mathbf{X}) = (\mu_{ij}) \in \{0,1\}^{m\times n}$ of an $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{m\times n}$ is such that $\mu_{ij} = 1$ if $x_{ij} \neq 0$ and $\mu_{ij} = 0$ if $x_{ij} = 0$. The matrices whose entries are all 0 (resp., all 1) are denoted by $\mathbf{0}$ (resp., $\mathbf{1}$) whenever m and n are understood. If $\mathbf{X}$, $\mathbf{Y} \in \mathbb{R}^{n\times n}$ are symmetric and $\mathbf{X} - \mathbf{Y}$ is positive definite (resp., nonnegative definite), we write $\mathbf{X} > \mathbf{Y}$ (resp., $\mathbf{X} \geq \mathbf{Y}$). The identity matrix is denoted by $\mathbf{I}$ with n understood.

For $x \in \mathbb{R}^n$, denoted by $\|x\|$ is the Euclidean vector norm $\|x\| = \sqrt{x^{\mathrm{T}}x}$ of x. If $\mathbf{X} \in \mathbb{R}^{m\times n}$, the Euclidean vector norm induces the spectral norm $\|\mathbf{X}\|$ of $\mathbf{X}$ given by $\|\mathbf{X}\| = \sup\{\sqrt{\lambda}\colon \lambda \text{ is an eigenvalue of } \mathbf{X}^{\mathrm{T}}\mathbf{X}\}$. Given a symmetric positive definite matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$, $\|\cdot\|_{\mathbf{X}}$ is the Hilbert norm of $x \in \mathbb{R}^n$ defined by $\|x\|_{\mathbf{X}} = \sqrt{x^{\mathrm{T}}\mathbf{X}x}$. If $\boldsymbol{x} = (x(0), x(1), \dots)$ is a sequence in $\mathbb{R}^n$, then we write $\boldsymbol{x} \in \ell^2(\mathbb{R}^n)$ whenever the ℓ^2 norm of $\boldsymbol{x}$, defined by $\|\boldsymbol{x}\| = \sqrt{\sum_{s=0}^{\infty}\|x(s)\|^2}$, is finite.

2. Analysis of linear time-varying systems. In this section, we analyze the disturbance attenuation performance of the discrete-time linear time-varying system and the asymptotic property of the associated Riccati difference equation. We choose a notation that is compatible with the latter sections of this paper. Let a subset $\boldsymbol{\mathcal{G}}$ of $\mathbb{R}^{n\times n}\times\mathbb{R}^{n\times m}\times\mathbb{R}^{l\times n}\times\mathbb{R}^{l\times m}$ and a sequence $\boldsymbol{\theta}=(\theta(0),\theta(1),\dots)$ in $\{0,1,\dots\}$ be as follows:

$$\boldsymbol{\mathcal{G}}=\{(\mathbf{A}_0,\mathbf{B}_0,\mathbf{C}_0,\mathbf{D}_0),(\mathbf{A}_1,\mathbf{B}_1,\mathbf{C}_1,\mathbf{D}_1),\dots\},\quad \boldsymbol{\theta}=(0,1,\dots). \tag{2.1}$$

With (2.1), we have $(\mathbf{A}_{\theta(t)},\mathbf{B}_{\theta(t)},\mathbf{C}_{\theta(t)},\mathbf{D}_{\theta(t)})=(\mathbf{A}_t,\mathbf{B}_t,\mathbf{C}_t,\mathbf{D}_t)$, and the pair $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$ is identified with the discrete-time linear time-varying system

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

where $\boldsymbol{\mathcal{G}}$ defines an indexed family of parameter quadruples, and the sequence $\boldsymbol{\theta}$ chooses one quadruple among $\boldsymbol{\mathcal{G}}$ for each $t\ge 0$. Given the initial state $x(0)$ and disturbance sequence $\boldsymbol{w}=(w(0),w(1),\dots)$, (2.2) determines the state sequence $\boldsymbol{x}=(x(0),x(1),\dots)$ and output sequence $\boldsymbol{z}=(z(0),z(1),\dots)$. If $\mathbf{B}_i$, $\mathbf{C}_i$, $\mathbf{D}_i$ are all zero matrices, we write $(\boldsymbol{\mathcal{A}},\boldsymbol{\theta})$ for $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$, where

$$\boldsymbol{\mathcal{A}}=\{\mathbf{A}_0,\mathbf{A}_1,\dots\}. \tag{2.3}$$

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

$$\|x(t)\|\le c\,\lambda^{t-t_0}\,\|x(t_0)\| \tag{2.4}$$

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

DEFINITION 2.2. *The system $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$ is said to be* uniformly (strictly) contractive *if there exists a $\gamma\in(0,1)$ such that, whenever $x(t_0)=0$,*

$$\sum\nolimits_{s=t_0}^{t}\|z(s)\|^2\le\gamma^2\sum\nolimits_{s=t_0}^{t}\|w(s)\|^2 \tag{2.5}$$

for $t\ge t_0\ge 0$ and for $\boldsymbol{w}\in\ell^2(\mathbb{R}^m)$.

Remark 1. It is clear that the uniform contractiveness is equivalent to the condition that the ℓ^2-induced gain from $\boldsymbol{w}$ to $\boldsymbol{z}$ be less than one; that is, for some $\gamma\in(0,1)$, $\|\boldsymbol{z}\|\le\gamma\|\boldsymbol{w}\|$ whenever $x(0)=0$ and $\boldsymbol{w}\in\ell^2(\mathbb{R}^m)$. The infimum of all $\gamma>0$ that satisfy (2.5) is called the ℓ^2-induced norm of the system $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$.

LEMMA 2.3. *Let $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\theta}$ be as in* (2.1)*; let $\boldsymbol{\mathcal{G}}$ be bounded. The following are equivalent:*

(a) *The system $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$ is uniformly exponentially stable and uniformly strictly contractive.*

(b) *There exist α_1, $\beta_1>0$, and $\mathbf{X}_t\in\mathbb{R}^{n\times n}$, $t=0,1,\dots$, such that, for all t,*

$$\alpha_1\mathbf{I}\le\mathbf{X}_t\le\beta_1\mathbf{I}; \tag{2.6a}$$

$$\begin{bmatrix}\mathbf{A}_t&\mathbf{B}_t\\ \mathbf{C}_t&\mathbf{D}_t\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{X}_{t+1}&\mathbf{0}\\ \mathbf{0}&\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A}_t&\mathbf{B}_t\\ \mathbf{C}_t&\mathbf{D}_t\end{bmatrix}-\begin{bmatrix}\mathbf{X}_t&\mathbf{0}\\ \mathbf{0}&\mathbf{I}\end{bmatrix}\le-\alpha_1\mathbf{I}. \tag{2.6b}$$

(c) *There exist α_2, $\beta_2>0$, and $\mathbf{Y}_t\in\mathbb{R}^{n\times n}$, $t=0,1,\dots$, such that, for all t,*

$$\alpha_2\mathbf{I}\le\mathbf{Y}_t\le\beta_2\mathbf{I}; \tag{2.7a}$$

$$\begin{bmatrix}\mathbf{A}_t&\mathbf{B}_t\\ \mathbf{C}_t&\mathbf{D}_t\end{bmatrix}\begin{bmatrix}\mathbf{Y}_t&\mathbf{0}\\ \mathbf{0}&\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A}_t&\mathbf{B}_t\\ \mathbf{C}_t&\mathbf{D}_t\end{bmatrix}^{\mathrm{T}}-\begin{bmatrix}\mathbf{Y}_{t+1}&\mathbf{0}\\ \mathbf{0}&\mathbf{I}\end{bmatrix}\le-\alpha_2\mathbf{I}. \tag{2.7b}$$

Moreover, if either (b) *or* (c) *holds, one may take*

$$\mathbf{X}_t = \mathbf{Y}_t^{-1} \tag{2.8}$$

for all t.

Proof. The equivalence of (a) and (b) is shown in [17, Thm. 11] via an operator theoretic approach. A simple Schur complement argument, together with the matrix inversion formula, shows that (b) and (c) are equivalent via relation (2.8). ∎

PROPOSITION 2.4. *Let* $\boldsymbol{\mathcal{A}}$ *be as in* (2.3) *and bounded; let* $\boldsymbol{\theta}$ *be as in* (2.1). *Then the system* $(\boldsymbol{\mathcal{A}}, \boldsymbol{\theta})$ *is uniformly exponentially stable if and only if there exist* α_1, $\beta_1 > 0$ *and* $\mathbf{X}_t \in \mathbb{R}^{n\times n}$ *(resp.,* α_2, $\beta_2 > 0$ *and* $\mathbf{Y}_t \in \mathbb{R}^{n\times n}$*),* $t = 0, 1, \dots$, *such that*

$$\alpha_1\mathbf{I} \le \mathbf{X}_t \le \beta_1\mathbf{I}; \quad \mathbf{A}_t^{\mathrm{T}}\mathbf{X}_{t+1}\mathbf{A}_t - \mathbf{X}_t \le -\alpha_1\mathbf{I}$$
$$\textit{(resp., } \alpha_2\mathbf{I} \le \mathbf{Y}_t \le \beta_2\mathbf{I}; \quad \mathbf{A}_t\mathbf{Y}_t\mathbf{A}_t^{\mathrm{T}} - \mathbf{Y}_{t+1} \le -\alpha_2\mathbf{I} \textit{)}$$

for all $t \ge 0$.

Proof. Set the matrices $\mathbf{B}_t$, $\mathbf{C}_t$, $\mathbf{D}_t$, $t \ge 0$, to zero in Lemma 2.3. ∎

Lemma 2.3 is a time-varying version of the classical Kalman–Yacubovitch–Popov (KYP) lemma (see, e.g., [35, 18]). Inequality (2.6b) is called the (extended) KYP inequality, and (2.7b) is its dual form. The solutions of these inequalities are obtained by solving the associated Riccati difference equations. Let $\mathbb{S}$ be the set of all symmetric matrices in $\mathbb{R}^{n\times n}$. For $i = 0, 1, \dots$, let $\mathbb{X}_i$ be the set of symmetric matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$ such that

$$\mathcal{W}_i(\mathbf{X}) = \mathbf{I} - \mathbf{D}_i^{\mathrm{T}}\mathbf{D}_i - \mathbf{B}_i^{\mathrm{T}}\mathbf{X}\mathbf{B}_i$$

is invertible. Define $\mathcal{S}_i \colon \mathbb{X}_i \to \mathbb{S}$, $i = 0, 1, \dots$, by

$$\mathcal{S}_i(\mathbf{X}) = \mathbf{A}_i^{\mathrm{T}}\mathbf{X}\mathbf{A}_i + \mathbf{C}_i^{\mathrm{T}}\mathbf{C}_i + (\mathbf{A}_i^{\mathrm{T}}\mathbf{X}\mathbf{B}_i + \mathbf{C}_i^{\mathrm{T}}\mathbf{D}_i)\mathcal{W}_i(\mathbf{X})^{-1}(\mathbf{B}_i^{\mathrm{T}}\mathbf{X}\mathbf{A}_i + \mathbf{D}_i^{\mathrm{T}}\mathbf{C}_i)$$

for $\mathbf{X} \in \mathbb{X}_i$. Similarly, for $i = 0, 1, \dots$, let $\mathbb{Y}_i$ be the set of symmetric matrices $\mathbf{Y} \in \mathbb{R}^{n\times n}$ such that

$$\mathcal{V}_i(\mathbf{Y}) = \mathbf{I} - \mathbf{D}_i\mathbf{D}_i^{\mathrm{T}} - \mathbf{C}_i\mathbf{Y}\mathbf{C}_i^{\mathrm{T}}$$

is invertible, and define $\mathcal{R}_i \colon \mathbb{Y}_i \to \mathbb{S}$, $i = 0, 1, \dots$, by

$$\mathcal{R}_i(\mathbf{Y}) = \mathbf{A}_i\mathbf{Y}\mathbf{A}_i^{\mathrm{T}} + \mathbf{B}_i\mathbf{B}_i^{\mathrm{T}} + (\mathbf{A}_i\mathbf{Y}\mathbf{C}_i^{\mathrm{T}} + \mathbf{B}_i\mathbf{D}_i^{\mathrm{T}})\mathcal{V}_i(\mathbf{Y})^{-1}(\mathbf{C}_i\mathbf{Y}\mathbf{A}_i^{\mathrm{T}} + \mathbf{D}_i\mathbf{B}_i^{\mathrm{T}})$$

for $\mathbf{Y} \in \mathbb{Y}_i$.

LEMMA 2.5. *Let* $\boldsymbol{\mathcal{G}}$ *and* $\boldsymbol{\theta}$ *be as in* (2.1)*; let* $\boldsymbol{\mathcal{G}}$ *be bounded. The following are equivalent:*

(a) *The system* $(\boldsymbol{\mathcal{G}}, \boldsymbol{\theta})$ *is uniformly exponentially stable and uniformly strictly contractive.*

(b) *There exist* ε_1, δ_1, $\eta_1 > 0$ *such that for all* $T = 0, 1, \dots$, *and* $\varepsilon \in [0, \varepsilon_1]$, *the equation*

$$\mathbf{X}_t^{(\varepsilon,T)} = \mathcal{S}_t\big(\mathbf{X}_{t+1}^{(\varepsilon,T)}\big) + \varepsilon\mathbf{I}, \tag{2.9a}$$

with the terminal condition $\mathbf{X}_{T+1}^{(\varepsilon,T)} = \varepsilon\mathbf{I}$, *satisfies for* $t = 0, \dots, T$ *that*

$$\mathcal{W}_t\big(\mathbf{X}_{t+1}^{(\varepsilon,T)}\big) \ge \eta_1\mathbf{I}; \quad \varepsilon\mathbf{I} \le \mathbf{X}_t^{(\varepsilon,T)} \le \mathbf{X}_t^{(\varepsilon,T+1)} \le \delta_1\mathbf{I}. \tag{2.9b}$$

(c) *There exist* ε_2, δ_2, $\eta_2 > 0$ *such that for all* $t_0 = 0, 1, \dots,$ *and* $\varepsilon \in [0, \varepsilon_2]$, *the equation*

$$\mathbf{Y}_{t+1}^{(\varepsilon,t_0)} = \mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) + \varepsilon\mathbf{I}, \tag{2.10a}$$

with the initial condition $\mathbf{Y}_{t_0}^{(\varepsilon,t_0)} = \varepsilon\mathbf{I}$, *satisfies for* $t = t_0, t_0+1, \dots$ *that*

$$\mathcal{V}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) \geq \eta_2\mathbf{I}; \quad \varepsilon\mathbf{I} \leq \mathbf{Y}_{t+1}^{(\varepsilon,t_0+1)} \leq \mathbf{Y}_{t+1}^{(\varepsilon,t_0)} \leq \delta_2\mathbf{I}. \tag{2.10b}$$

Moreover, if (b) *holds, then one may take* $\mathbf{X}_t = \lim_{T\to\infty} \mathbf{X}_t^{(\varepsilon,T)}$ *in* (2.6)*; if* (c) *holds, then one may take* $\mathbf{Y}_t = \mathbf{Y}_t^{(\varepsilon,0)}$ *in* (2.7).

Equation (2.9a) is the (generalized) Riccati difference equation associated with the KYP inequality, and (2.10a) is its dual form. The latter evolves forward in time and so provides an explicit recursive expression for computing the solution to the KYP inequalities (2.6)–(2.7). Similar, but less explicit, results from operator theoretic points of view exist [22, 27]. The proof of Lemma 2.5 is based on the following standard lemma. For $\mathbf{Y} \in \mathbb{Y}_i$, let

$$\mathcal{A}_i(\mathbf{Y}) = \mathbf{A}_i + (\mathbf{A}_i\mathbf{Y}\mathbf{C}_i^{\mathrm{T}} + \mathbf{B}_i\mathbf{D}_i^{\mathrm{T}})\mathcal{V}_i(\mathbf{Y})^{-1}\mathbf{C}_i. \tag{2.11}$$

LEMMA 2.6. *Let* $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)} \in \mathbb{Y}_i$*; let* $\boldsymbol{\Delta}^{(12)} = \mathbf{Y}^{(1)} - \mathbf{Y}^{(2)}$. *Then*

$$\begin{aligned}\mathcal{R}_i\big(\mathbf{Y}^{(1)}\big) - \mathcal{R}_i\big(\mathbf{Y}^{(2)}\big) &= \mathcal{A}_i\big(\mathbf{Y}^{(2)}\big)\boldsymbol{\Delta}^{(12)}\mathcal{A}_i\big(\mathbf{Y}^{(2)}\big)^{\mathrm{T}} \\ &\quad + \mathcal{A}_i\big(\mathbf{Y}^{(2)}\big)\boldsymbol{\Delta}^{(12)}\mathbf{C}_i^{\mathrm{T}}\mathcal{V}_i\big(\mathbf{Y}^{(1)}\big)^{-1}\mathbf{C}_i\boldsymbol{\Delta}^{(12)}\mathcal{A}_i\big(\mathbf{Y}^{(2)}\big)^{\mathrm{T}} \qquad (2.12a)\\ &= \mathcal{A}_i\big(\mathbf{Y}^{(1)}\big)\boldsymbol{\Delta}^{(12)}\mathcal{A}_i\big(\mathbf{Y}^{(2)}\big)^{\mathrm{T}}. \qquad (2.12b)\end{aligned}$$

Proof. It follows from [22, Lem. 11, p. 77] that, for $\mathbf{Y} \in \mathbb{Y}_i$, we may write

$$\mathcal{R}_i(\mathbf{Y}) = \overline{\mathbf{A}}_i\mathbf{Y}\overline{\mathbf{A}}_i^{\mathrm{T}} + \overline{\mathbf{Q}}_i + \overline{\mathbf{A}}_i\mathbf{Y}\mathbf{C}_i^{\mathrm{T}}\mathcal{V}_i(\mathbf{Y})^{-1}\mathbf{C}_i\mathbf{Y}\overline{\mathbf{A}}_i^{\mathrm{T}} \tag{2.13}$$

with

$$\begin{aligned}\overline{\mathbf{A}}_i &= \mathbf{A}_i + \mathbf{B}_i\mathbf{D}_i^{\mathrm{T}}(\mathbf{I} - \mathbf{D}_i\mathbf{D}_i^{\mathrm{T}})^{-1}\mathbf{C}_i,\\ \overline{\mathbf{Q}}_i &= \mathbf{B}_i\mathbf{B}_i^{\mathrm{T}} + \mathbf{B}_i\mathbf{D}_i^{\mathrm{T}}(\mathbf{I} - \mathbf{D}_i\mathbf{D}_i^{\mathrm{T}})^{-1}\mathbf{D}_i\mathbf{B}_i^{\mathrm{T}}.\end{aligned}$$

Then [15, Lem. 3.1] leads to (2.12a). Since

$$\mathcal{V}_i\big(\mathbf{Y}^{(1)}\big)^{-1} - \mathcal{V}_i\big(\mathbf{Y}^{(2)}\big)^{-1} - \mathcal{V}_i\big(\mathbf{Y}^{(2)}\big)^{-1}\mathbf{C}_i\boldsymbol{\Delta}^{(12)}\mathbf{C}_i^{\mathrm{T}}\mathcal{V}_i\big(\mathbf{Y}^{(1)}\big)^{-1} = \mathbf{0},$$

it is easily seen that

$$\begin{aligned}\mathcal{A}_i\big(\mathbf{Y}^{(2)}\big) &= \mathcal{A}_i\big(\mathbf{Y}^{(1)}\big)\big(\mathbf{I} - \boldsymbol{\Delta}^{(12)}\mathbf{C}_i^{\mathrm{T}}\mathcal{V}_i\big(\mathbf{Y}^{(2)}\big)^{-1}\mathbf{C}_i\big),\\ \big(\mathbf{I} - \boldsymbol{\Delta}^{(12)}\mathbf{C}_i^{\mathrm{T}}\mathcal{V}_i\big(\mathbf{Y}^{(2)}\big)^{-1}\mathbf{C}_i\big)\big(\mathbf{I} + \boldsymbol{\Delta}^{(12)}\mathbf{C}_i^{\mathrm{T}}\mathcal{V}_i\big(\mathbf{Y}^{(1)}\big)^{-1}\mathbf{C}_i\big) &= \mathbf{I}.\end{aligned}$$

These equalities and (2.12a) yield (2.12b). ∎

Lemma 2.6 is useful in proving asymptotic properties of the Riccati difference equation. In particular, an immediate consequence of (2.12a) is that $\mathcal{R}_i\big(\mathbf{Y}^{(1)}\big) \geq \mathcal{R}_i\big(\mathbf{Y}^{(2)}\big)$ whenever $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)} \in \mathbb{Y}_i$, and $\mathbf{Y}^{(1)} \geq \mathbf{Y}^{(2)}$.

Proof of Lemma 2.5. To first show that (c) is equivalent to (a), suppose that (c) holds. Let $\varepsilon = \varepsilon_2$. Since $\boldsymbol{\mathcal{G}}$ is bounded, (2.10) implies that

$$\begin{aligned}\mathbf{A}_t\mathbf{Y}_t^{(\varepsilon,t_0)}\mathbf{A}_t^{\mathrm{T}} + \mathbf{B}_t\mathbf{B}_t^{\mathrm{T}} - \mathbf{Y}_{t+1}^{(\varepsilon,t_0)} + \alpha\mathbf{I} \\ + \big(\mathbf{A}_t\mathbf{Y}_t^{(\varepsilon,t_0)}\mathbf{C}_t^{\mathrm{T}} + \mathbf{B}_t\mathbf{D}_t^{\mathrm{T}}\big)\mathcal{V}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)} - \alpha\mathbf{I}\big)^{-1}\big(\mathbf{C}_t\mathbf{Y}_t^{(\varepsilon,t_0)}\mathbf{A}_t^{\mathrm{T}} + \mathbf{D}_t\mathbf{B}_t^{\mathrm{T}}\big) \le \mathbf{0}\end{aligned}$$

for some small $\alpha > 0$. Then the Schur complement formula yields

$$\begin{bmatrix}\mathbf{A}_t & \mathbf{B}_t\\ \mathbf{C}_t & \mathbf{D}_t\end{bmatrix}\begin{bmatrix}\mathbf{Y}_t^{(\varepsilon,t_0)} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A}_t & \mathbf{B}_t\\ \mathbf{C}_t & \mathbf{D}_t\end{bmatrix}^{\mathrm{T}} - \begin{bmatrix}\mathbf{Y}_{t+1}^{(\varepsilon,t_0)} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} \le -\alpha\mathbf{I}.$$

Putting $\alpha_2 = \min\{\alpha,\varepsilon\}$, $\beta_2 = \delta_2$, and $\mathbf{Y}_t = \mathbf{Y}_t^{(\varepsilon,0)}$, we obtain condition (c) of Lemma 2.3, which in turn leads to the uniform stability and contractiveness of $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$.

Conversely, suppose that $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$ is uniformly stable and contractive so that condition (c) of Lemma 2.3 holds. Taking the Schur complement of $\mathbf{C}_t\mathbf{Y}_t\mathbf{C}_t^{\mathrm{T}} + \mathbf{D}_t\mathbf{D}_t^{\mathrm{T}} - \mathbf{I}$ from (2.7b), we obtain

$$\mathcal{V}_t(\mathbf{Y}_t) \ge \alpha_2\mathbf{I}; \quad \mathcal{R}_t(\mathbf{Y}_t) + \alpha_2\mathbf{I} \le \mathbf{Y}_{t+1}$$

for all t. Choose an arbitrary $\varepsilon \in [0,\alpha_2]$. For all $t_0 \ge 0$ and all $t \ge t_0$, define $\mathbf{Y}_{t_0}^{(\varepsilon,t_0)} = \varepsilon\mathbf{I}$ and $\mathbf{Y}_{t+1}^{(\varepsilon,t_0)} = \mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) + \varepsilon\mathbf{I}$. Since $\mathbf{Y}_t \ge \alpha_2\mathbf{I}$ for all t, we have $\mathbf{Y}_{t_0} \ge \mathbf{Y}_{t_0}^{(\varepsilon,t_0)}$; by induction, together with (2.12a), we obtain

$$\mathcal{V}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) \ge \alpha_2\mathbf{I}; \quad \varepsilon\mathbf{I} \le \mathbf{Y}_t^{(\varepsilon,t_0)} \le \mathbf{Y}_t \le \beta_2\mathbf{I}$$

for $t \ge t_0 \ge 0$. Moreover, $\mathbf{Y}_t^{(\varepsilon,t_0)} \ge \varepsilon\mathbf{I}$ for all t, and so we have $\mathbf{Y}_{t_0+1}^{(\varepsilon,t_0+1)} \le \mathbf{Y}_{t_0+1}^{(\varepsilon,t_0)}$ from (2.12a); by induction, we obtain that

$$\mathbf{Y}_{t+1}^{(\varepsilon,t_0+1)} \le \mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) + \varepsilon\mathbf{I} = \mathbf{Y}_{t+1}^{(\varepsilon,t_0)}$$

for $t \ge t_0 \ge 0$. Putting $\varepsilon_2 = \alpha_2$, $\delta_2 = \beta_2$, and $\eta_2 = \alpha_2$, we obtain condition (c). This concludes the proof of the equivalence of (a) and (c). The proof that (a) and (b) are equivalent is analogous, and so is omitted. ∎

THEOREM 2.7. *Let $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\theta}$ be as in* (2.1)*; let $\boldsymbol{\mathcal{G}}$ be bounded. Suppose that the system $(\boldsymbol{\mathcal{G}},\boldsymbol{\theta})$ is uniformly exponentially stable and uniformly strictly contractive so that condition* (c) *of Lemma* 2.5 *holds. For $\varepsilon \in (0,\varepsilon_2)$ and $t_0 \ge 0$, let $\mathbf{Y}_i^{(\varepsilon,t_0)}$ and $\mathcal{A}_i(\cdot)$, $i = t_0, t_0+1, \dots$, be as in* (2.10a) *and* (2.11)*, respectively, where $\mathbf{Y}_{t_0}^{(\varepsilon,t_0)} = \varepsilon\mathbf{I}$. Then the following hold:*

(a) *For each $\varepsilon \in (0,\varepsilon_2)$ and $t_0 \ge 0$, define*

$$\boldsymbol{\mathcal{A}}^{(\varepsilon,t_0)} = \big\{\mathcal{A}_i\big(\mathbf{Y}_i^{(\varepsilon,t_0)}\big)\colon i = t_0, t_0+1, \dots\big\}, \quad \boldsymbol{\theta}^{(t_0)} = (t_0, t_0+1, \dots).$$

Then each system $\big(\boldsymbol{\mathcal{A}}^{(\varepsilon,t_0)},\boldsymbol{\theta}^{(t_0)}\big)$ is uniformly exponentially stable. Moreover, for each $\varepsilon \in (0,\varepsilon_2)$, there exist $c_\varepsilon \ge 1$ and $\lambda_\varepsilon \in (0,1)$ such that

$$\big\|\mathcal{A}_{t-1}\big(\mathbf{Y}_{t-1}^{(\varepsilon,t_0)}\big)\cdots\mathcal{A}_{t_0}\big(\mathbf{Y}_{t_0}^{(\varepsilon,t_0)}\big)\big\| \le c_\varepsilon\,\lambda_\varepsilon^{t-t_0} \tag{2.14}$$

for $t > t_0 \ge 0$.

(b) *For each $\varepsilon \in (0, \varepsilon_2)$, there exist a nonnegative integer M and α_2, $\beta_2 > 0$ such that* (2.7) *is satisfied with*

$$\mathbf{Y}_t = \begin{cases} \mathbf{Y}_t^{(\varepsilon,0)} & \text{for } t < M; \\ \mathbf{Y}_t^{(\varepsilon,t-M)} & \text{for } t \geq M \end{cases} \tag{2.15}$$

for $t \geq 0$.

Proof. Fix ε, $\varepsilon' \in (0, \varepsilon_2)$ such that $\varepsilon < \varepsilon'$. For $t \geq t_0 \geq 0$, let $\mathbf{Y}_t^{(t_0)} = \mathbf{Y}_t^{(\varepsilon',t_0)} - \mathbf{Y}_t^{(\varepsilon,t_0)}$. Then (2.12a) yields that

$$\mathbf{Y}_{t+1}^{(t_0)} \geq \mathcal{A}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big)\mathbf{Y}_t^{(t_0)}\mathcal{A}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big)^{\mathrm{T}} + (\varepsilon' - \varepsilon)\mathbf{I}$$

for all $t \geq t_0$. Since $\varepsilon' - \varepsilon > 0$ and $(\varepsilon' - \varepsilon)\mathbf{I} \leq \mathbf{Y}_t^{(t_0)} \leq (\delta_2 - \epsilon)\mathbf{I}$ for $t \geq t_0 \geq 0$, we have that, with $\mathbf{X}_t^{(t_0)} = \big(\mathbf{Y}_t^{(t_0)}\big)^{-1}$, there exist α_1, $\beta_1 > 0$ such that

$$\alpha_1\mathbf{I} \leq \mathbf{X}_t^{(t_0)} \leq \beta_1\mathbf{I}; \quad \mathcal{A}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big)^{\mathrm{T}}\mathbf{X}_{t+1}^{(t_0)}\mathcal{A}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) - \mathbf{X}_t^{(t_0)} \leq -\alpha_1\mathbf{I}$$

for all t. The uniform stability of the systems $\big(\boldsymbol{\mathcal{A}}^{(\varepsilon,t_0)}, \boldsymbol{\theta}^{(t_0)}\big)$ follows from Proposition 2.4. Due to [30, Lem. 4], inequality (2.14) holds for $t > t_0 \geq 0$ with $c_\varepsilon = \sqrt{\beta_1/\alpha_1}$ and $\lambda_\varepsilon = \sqrt{1 - \alpha_1/\beta_1}$. Thus assertion (a) holds true.

To prove (b), pick a nonnegative integer M such that $c_\varepsilon^2\lambda_\varepsilon^{2M} < \varepsilon/(\delta_2 - \varepsilon)$. Then, using (2.12b), we deduce that there exists an $\varepsilon'' \in (0, \varepsilon)$ such that

$$\begin{aligned} &\mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,t-M)}\big) - \mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,t-M+1)}\big) \\ &\qquad = \mathcal{A}_t\big(\mathbf{Y}_t^{(\varepsilon,t-M)}\big)\cdots\mathcal{A}_{t-M+1}\big(\mathbf{Y}_{t-M+1}^{(\varepsilon,t-M)}\big)\big(\mathcal{R}_{t-M+1}(\varepsilon\mathbf{I}) - \varepsilon\mathbf{I}\big) \\ &\qquad\quad \times \mathcal{A}_{t-M+1}\big(\mathbf{Y}_{t-M+1}^{(\varepsilon,t-M+1)}\big)^{\mathrm{T}}\cdots\mathcal{A}_t\big(\mathbf{Y}_t^{(\varepsilon,t-M+1)}\big)^{\mathrm{T}} \\ &\qquad \leq (\delta_2 - \varepsilon)c_\varepsilon^2\lambda_\varepsilon^{2M}\mathbf{I} \\ &\qquad \leq \varepsilon''\mathbf{I} \end{aligned}$$

for $t \geq M$; or equivalently,

$$\mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,t-M)}\big) - \mathbf{Y}_{t+1}^{(\varepsilon,t-M+1)} \leq -(\varepsilon - \varepsilon'')\mathbf{I}$$

for $t \geq M$. On the other hand,

$$\mathcal{R}_t\big(\mathbf{Y}_t^{(\varepsilon,0)}\big) - \mathbf{Y}_{t+1}^{(\varepsilon,0)} = -\varepsilon\mathbf{I} \leq -(\varepsilon - \varepsilon'')\mathbf{I}.$$

Therefore, $\mathbf{Y}_t$ defined by (2.15) satisfies

$$\mathcal{R}_t(\mathbf{Y}_t) - \mathbf{Y}_{t+1} \leq -(\varepsilon - \varepsilon'')\mathbf{I}$$

for all $t \geq 0$. Using the Schur complement formula, we have that there exists an $\alpha > 0$ such that

$$\begin{bmatrix} \mathbf{A}_t & \mathbf{B}_t \\ \mathbf{C}_t & \mathbf{D}_t \end{bmatrix}\begin{bmatrix} \mathbf{Y}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{A}_t & \mathbf{B}_t \\ \mathbf{C}_t & \mathbf{D}_t \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} \mathbf{Y}_{t+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \leq -\alpha\mathbf{I}$$

holds for $t \geq 0$. Putting $\alpha_2 = \min\{\alpha, \epsilon\}$ and $\beta_2 = \delta_2$, we see that $\mathbf{Y}_t$ satisfies (2.7) for all $t \geq 0$, and hence assertion (b) holds. ∎

Remark 2. Using (2.13), we may write

$$\mathbf{Y}_{t+1}^{(\varepsilon,t_0)} = \overline{\mathbf{A}}_t \mathbf{Y}_t^{(\varepsilon,t_0)} \overline{\mathbf{A}}_t^{\mathrm{T}} + (\overline{\mathbf{Q}}_t + \varepsilon \mathbf{I}) + \overline{\mathbf{A}}_t \mathbf{Y}_t^{(\varepsilon,t_0)} \mathbf{C}_t^{\mathrm{T}} \mathcal{V}_t\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big)^{-1} \mathbf{C}_t \mathbf{Y}_t^{(\varepsilon,t_0)} \overline{\mathbf{A}}_t^{\mathrm{T}}$$

for $\varepsilon \in (0, \varepsilon_2)$ and $t \geq t_0 \geq 0$. Let

$$\boldsymbol{\mathcal{G}}' = \{(\overline{\mathbf{A}}_0, \overline{\mathbf{Q}}_0 + \varepsilon \mathbf{I}, \mathbf{C}_0, \mathbf{0}), (\overline{\mathbf{A}}_1, \overline{\mathbf{Q}}_1 + \varepsilon \mathbf{I}, \mathbf{C}_1, \mathbf{0}), \dots\}.$$

If $(\boldsymbol{\mathcal{G}}, \boldsymbol{\theta})$ is uniformly stable, it is easy to verify that $(\boldsymbol{\mathcal{G}}', \boldsymbol{\theta})$ is uniformly detectable and uniformly stabilizable. Then, by [1, 14], the system $\big(\boldsymbol{\mathcal{A}}^{(\varepsilon,t_0)}, \boldsymbol{\theta}^{(t_0)}\big)$ is uniformly stable for each $t_0 \geq 0$; it follows from (2.12b) that, for each $t_0 \geq 0$, $\mathbf{Y}_t^{(\varepsilon,t_0)}$ converges to the unique "moving equilibrium" (i.e., the maximal solution) of the Riccati difference equation (2.10a) as $t - t_0 \to \infty$ [14]. However, part (a) of Theorem 2.7 says that this convergence is uniform in (t, t_0), as the uniform stability of $\big(\boldsymbol{\mathcal{A}}^{(\varepsilon,t_0)}, \boldsymbol{\theta}^{(t_0)}\big)$ is again uniform in t_0.

It is known that, if a given linear time-varying system is uniformly stable, the corresponding Lyapunov inequality admits a solution that has a finite memory of past parameters (see, e.g., [30]). Part (b) of Theorem 2.7 is an extension of this and says that the KYP inequality (or equivalently, the Riccati inequality) associated with a uniformly stable and contractive linear time-varying system has a solution that has a finite memory of past parameters.

3. Control of switched linear systems. The switched linear system is a family of linear time-varying systems whose parameters vary within a single finite set. Fix a positive integer N and define

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

where $\mathbf{A}_i \in \mathbb{R}^{n \times n}$, $\mathbf{B}_i \in \mathbb{R}^{n \times m}$, $\mathbf{C}_i \in \mathbb{R}^{l \times n}$, $\mathbf{D}_i \in \mathbb{R}^{l \times m}$ for $i = 1, \dots, N$. Let $\boldsymbol{\Omega}$ be the set of all infinite sequences in $\{1, \dots, N\}$; each member of $\boldsymbol{\Omega}$ is called a *switching sequence*. If $\boldsymbol{\Theta}$ is a nonempty subset of $\boldsymbol{\Omega}$, then the pair $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ defines the *switched linear system*, where $\boldsymbol{\Theta}$ is the set of *admissible* switching sequences: for initial states $x(0)$, disturbance sequences $\boldsymbol{w}$, switching sequences $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $t \geq 0$, the system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ has the state-space representation

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

If $\theta(t) = i$, then the system is said to be in *mode* i at time t, and its parameters at time t are given by the quadruple $(\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i, \mathbf{D}_i)$. When the set $\boldsymbol{\Theta}$ is equal to the entire set $\boldsymbol{\Omega}$, the pair $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Omega})$ defines the *discrete linear inclusion*, which is the switched linear system without a switching path constraint; on the other hand, if $\boldsymbol{\Theta} = \{\boldsymbol{\theta}\}$ is a singleton, then the pair $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ is nothing but the linear time-varying system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\theta})$. We require that the stability and contractiveness of the system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ be uniform over all switching sequences in $\boldsymbol{\Theta}$.

DEFINITION 3.1. *The system* $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ *is said to be* uniformly (exponentially) stable *if there exist* $c \geq 1$ *and* $\lambda \in (0, 1)$ *such that, whenever* $\boldsymbol{w} = 0$, *inequality* (2.4) *holds for* $t \geq t_0 \geq 0$, *for* $x(t_0) \in \mathbb{R}^n$, *and for* $\theta \in \boldsymbol{\Theta}$.

DEFINITION 3.2. *The system* $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ *is said to be* uniformly (strictly) contractive *if there exists a* $\gamma \in (0, 1)$ *such that, whenever* $x(t_0) = 0$, *inequality* (2.5) *holds for* $t \geq t_0 \geq 0$, *for* $\boldsymbol{w} \in \ell^2(\mathbb{R}^m)$, *and for* $\theta \in \boldsymbol{\Theta}$.

For the sake of convenience, we introduce a dummy mode 0 and think of each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as a two-sided sequence $(\dots, \theta(-1), \theta(0), \theta(1), \dots)$ by putting $\theta(t) = 0$ for $t < 0$. Fix a nonnegative integer L. Any finite sequence in $\{0, \dots, N\}$ will be called a *(finite) switching path*; in particular, elements of the set $\{0, \dots, N\}^{L+1}$ are switching paths of length L and are called *L-paths*. For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t \geq 0$, let

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

An L-path $(i_0, \dots, i_L)$ is said to *occur in* $\boldsymbol{\Theta}$ if $\theta_L(t) = (i_0, \dots, i_L)$ for some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and some $t \geq 0$. Each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ generates an *L-path switching sequence* $\boldsymbol{\theta}_L$ defined by

$$\boldsymbol{\theta}_L = (\theta_L(0), \theta_L(1), \dots).$$

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

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

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

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

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

$$\begin{aligned}(3.3)\quad &\big(i_0^{\theta(0)}, \dots, i_{L-1}^{\theta(0)}, \theta(0)\big),\ \big(i_1^{\theta(0)}, \dots, i_{L-1}^{\theta(0)}, \theta(0), \theta(1)\big),\\ &\qquad\qquad \dots,\ \big(i_{L-1}^{\theta(0)}, \theta(0), \dots, \theta(L-1)\big) \in \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})\end{aligned}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If $L = 0$, then let $\boldsymbol{\mathcal{M}}_0(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{L}}_0(\boldsymbol{\Theta})$. The sets $\boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})$, $L = 0, 1, \dots$, are unique and so are well defined. Let

$$\boldsymbol{\mathcal{M}}_L^-(\boldsymbol{\Theta}) = \{\boldsymbol{i}_- \colon \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})\}, \quad \boldsymbol{\mathcal{L}}_L^-(\boldsymbol{\Theta}) = \{\boldsymbol{i}_- \colon \boldsymbol{i} \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})\}.$$

In general, we have

$$\begin{gathered}\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}) \cap \{1, \dots, N\}^{L+1} \subset \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}) \subsetneq \{0, \dots, N\}^{L+1},\\ \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{M}}_{L+1}^-(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{L}}_{L+1}^-(\boldsymbol{\Theta}).\end{gathered}$$

Example 1. Let $\boldsymbol{\Theta}$ consist of three periodic sequences $(1, 1, \dots)$, $(1, 2, 1, 2, \dots)$, and $(2, 2, \dots)$. To simplify notation, write $i_0 \cdots i_L$ for $(i_0, \dots, i_L)$. Then we have

$$\begin{gathered}\boldsymbol{\mathcal{L}}_0(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{M}}_0(\boldsymbol{\Theta}) = \{1, 2\}, \quad \boldsymbol{\mathcal{L}}_1(\boldsymbol{\Theta}) = \{01, 02, 11, 12, 21, 22\},\\ \boldsymbol{\mathcal{L}}_2(\boldsymbol{\Theta}) = \{001, 002, 011, 012, 022, 111, 121, 212, 222\},\end{gathered}$$

and so on. For $L = 1$, we can replace the one-path $(0, 1)$ in $\boldsymbol{\mathcal{L}}_1(\boldsymbol{\Theta})$ with $(1, 1)$ or $(2, 1)$; similarly, the one-path $(0, 2)$ can be replaced with either $(1, 2)$ or $(2, 2)$. Hence, we have

$$\boldsymbol{\mathcal{M}}_1(\boldsymbol{\Theta}) = \{11, 12, 21, 22\}.$$

However, for $L = 2$, if we choose any $(i_0, i_1) \neq (0, 0)$, then we have that at least one of the two-paths $(i_0, i_1, 1)$, $(i_1, 1, 1)$, $(i_1, 1, 2)$ does not belong to $\boldsymbol{\mathcal{L}}_2(\boldsymbol{\Theta})$ and so $(0, 0, 1)$

cannot be replaced with any of $(0,1,1)$, $(1,1,1)$, $(1,2,1) \in \boldsymbol{\mathcal{L}}_2(\boldsymbol{\Theta})$; on the other hand, both the two-paths $(0,0,2)$ and $(0,2,2)$ can be replaced with $(2,2,2) \in \boldsymbol{\mathcal{L}}_2(\boldsymbol{\Theta})$. Hence,

$$\boldsymbol{\mathcal{M}}_2(\boldsymbol{\Theta}) = \{001, 011, 012, 111, 121, 212, 222\}.$$

This $\boldsymbol{\Theta}$ will be used later in Examples 4 and 7.

Example 2. For each positive integer k, let $\boldsymbol{\Theta}^{(k)}$ consist of a single sequence

$$\boldsymbol{\theta}^{(k)} = (\underbrace{1,\dots,1}_{k},\underbrace{2,\dots,2}_{k},\underbrace{1,\dots,1}_{k},\underbrace{2,\dots,2}_{k},\dots),$$

where k 1's and k 2's alternate in $\boldsymbol{\theta}^{(k)}$. Then

$$\boldsymbol{\mathcal{L}}_0\big(\boldsymbol{\Theta}^{(k)}\big) = \boldsymbol{\mathcal{M}}_0\big(\boldsymbol{\Theta}^{(k)}\big) = \{1,2\} \quad \text{for all } k;$$

$$\boldsymbol{\mathcal{L}}_1\big(\boldsymbol{\Theta}^{(k)}\big) = \begin{cases} \{01,12,21\}, & k=1; \\ \{01,11,12,21,22\}, & k \ge 2; \end{cases} \quad \boldsymbol{\mathcal{M}}_1\big(\boldsymbol{\Theta}^{(k)}\big) = \begin{cases} \{12,21\}, & k=1; \\ \{11,12,21,22\}, & k \ge 2; \end{cases}$$

$$\boldsymbol{\mathcal{L}}_2\big(\boldsymbol{\Theta}^{(k)}\big) = \begin{cases} \{001,012,121,212\}, & k=1; \\ \{001,011,112,122,211,221\}, & k=2; \\ \{001,011,111,112,122,211,221,222\}, & k \ge 3; \end{cases}$$

$$\boldsymbol{\mathcal{M}}_2\big(\boldsymbol{\Theta}^{(k)}\big) = \begin{cases} \{121,212\}, & k=1; \\ \{112,122,211,221\}, & k=2; \\ \{111,112,122,211,221,222\} & k \ge 3, \end{cases}$$

and so on. These sets will be used in Example 3. The two-path switching sequence generated by $\boldsymbol{\theta}^{(2)}$, for instance, is

$$\boldsymbol{\theta}_2^{(2)} = (001, 011, 112, 122, 221, 211, 112, 122, 221, 211, \dots).$$

THEOREM 3.3. *Let $\boldsymbol{\mathcal{G}}$ be as in* (3.1)*; let $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ be nonempty. The system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ is uniformly exponentially stable and uniformly strictly contractive if and only if there exist a nonnegative integer M and an indexed family $\{\mathbf{X}_{\boldsymbol{j}} : \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})\}$ of symmetric positive definite matrices $\mathbf{X}_{\boldsymbol{j}} \in \mathbb{R}^{n\times n}$ such that*

$$\begin{bmatrix} \mathbf{A}_{i_M} & \mathbf{B}_{i_M} \\ \mathbf{C}_{i_M} & \mathbf{D}_{i_M} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{X}_{\boldsymbol{i}_+} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{i_M} & \mathbf{B}_{i_M} \\ \mathbf{C}_{i_M} & \mathbf{D}_{i_M} \end{bmatrix} - \begin{bmatrix} \mathbf{X}_{\boldsymbol{i}_-} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} < \mathbf{0} \tag{3.4}$$

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

Proof. To show sufficiency, fix an M and suppose (3.4) holds for $\boldsymbol{i} = (i_0, \dots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$. Assume $M > 0$ without loss of generality. By the definition of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, one can choose $\big(i_0^j, \dots, i_{M-1}^j\big) \in \{0,\dots,N\}^M$, $j \in \{\theta(0) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, such that (3.3), with L replaced by M, holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Put

$$\mathbf{X}_t = \begin{cases} \mathbf{X}_{\left(i_0^{\theta(0)},\dots,i_{M-1}^{\theta(0)}\right)} & \text{for } t = 0; \\ \mathbf{X}_{\left(i_t^{\theta(0)},\dots,i_{M-1}^{\theta(0)},\theta(0),\dots,\theta(t-1)\right)} & \text{for } 0 < t < M; \\ \mathbf{X}_{(\theta(t-M),\dots,\theta(t-1))} & \text{for } t \ge M. \end{cases}$$

Then, since $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ is finite, one can choose α, $\beta > 0$ independently of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that, in particular,

$$\alpha\mathbf{I} \le \mathbf{X}_t \le \beta\mathbf{I}; \quad \mathbf{A}_{\theta(t)}^{\mathrm{T}} \mathbf{X}_{t+1} \mathbf{A}_{\theta(t)} - \mathbf{X}_t < -\alpha\mathbf{I}$$

for $t \geq 0$; hence it follows from Proposition 2.4 and the first half of the proof of Theorem 2.7 that the system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ is uniformly stable. On the other hand, the finiteness of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ implies that there exists an $\eta \in (0,1)$, which is independent of $\boldsymbol{\theta}$, such that

$$\begin{bmatrix} x(t) \\ w(t) \end{bmatrix}^{\mathrm{T}} \left(\begin{bmatrix} \mathbf{A}_{\theta(t)} & \mathbf{B}_{\theta(t)} \\ \mathbf{C}_{\theta(t)} & \mathbf{D}_{\theta(t)} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{X}_{t+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{\theta(t)} & \mathbf{B}_{\theta(t)} \\ \mathbf{C}_{\theta(t)} & \mathbf{D}_{\theta(t)} \end{bmatrix} - \begin{bmatrix} \mathbf{X}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} + \eta \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right) \begin{bmatrix} x(t) \\ w(t) \end{bmatrix} \leq 0$$

for all $x(t) \in \mathbb{R}^n$ and $w(t) \in \mathbb{R}^m$ so that

$$\|z(t)\|^2 + \|x(t+1)\|^2_{\mathbf{X}_{t+1}} - \|x(t)\|^2_{\mathbf{X}_t} \leq (1-\eta)\|w(t)\|^2$$

for $t \geq 0$. If $t \geq t_0 \geq 0$ and $x(t_0) = 0$, then this inequality, as well as the fact that $\mathbf{X}_{\boldsymbol{i}} > \mathbf{0}$ for all $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}^-_M(\boldsymbol{\Theta})$, leads to

$$\sum\nolimits_{s=t_0}^{t} \|z(s)\|^2 \leq (1-\eta) \sum\nolimits_{s=t_0}^{t} \|w(s)\|^2.$$

Putting $\gamma = \sqrt{1-\eta}$ yields (2.5) for $t \geq t_0 \geq 0$. Since γ is independent of $\boldsymbol{\theta}$, this inequality holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, too. Hence the system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ is uniformly contractive as well as uniformly stable. This proves sufficiency.

To show necessity, suppose that $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ is uniformly stable and contractive. Consider the augmented disturbance signal $\tilde{w}(t) = [w(t)^{\mathrm{T}}\ v(t)^{\mathrm{T}}]^{\mathrm{T}}$ with $v(t) \in \mathbb{R}^n$, $t \geq 0$, and the perturbed system $\big(\boldsymbol{\mathcal{G}}^{(\varepsilon)}, \boldsymbol{\Theta}\big)$, where

$$\boldsymbol{\mathcal{G}}^{(\varepsilon)} = \big\{\big(\mathbf{A}_1, \mathbf{B}_1^{(\varepsilon)}, \mathbf{C}_1, \mathbf{D}_1^{(\varepsilon)}\big), \ldots, \big(\mathbf{A}_N, \mathbf{B}_N^{(\varepsilon)}, \mathbf{C}_N, \mathbf{D}_N^{(\varepsilon)}\big)\big\},$$
$$\mathbf{B}_i^{(\varepsilon)} = [\mathbf{B}_i\ \sqrt{\varepsilon}\mathbf{I}] \in \mathbb{R}^{n \times (m+n)}, \quad \mathbf{D}_i^{(\varepsilon)} = [\mathbf{D}_i\ \mathbf{0}] \in \mathbb{R}^{l \times (m+n)}.$$

Then there exists a sufficiently small $\varepsilon_2 > 0$, dependent on γ in (2.5) but independent of $\boldsymbol{\Theta}$, such that $(\boldsymbol{\mathcal{G}}^{(\varepsilon)}, \boldsymbol{\Theta})$ is uniformly stable and contractive for all $\varepsilon \in (0, \varepsilon_2)$. If we fix an $\varepsilon \in (0, \varepsilon_2)$, then by Lemma 2.5, there exist δ_2, $\eta_2 > 0$ such that the (dual) Riccati equation $\mathbf{Y}_{t+1}^{(\varepsilon,t_0)} = \mathcal{R}_{\theta(t)}\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) + \varepsilon\mathbf{I}$, with the initial condition $\mathbf{Y}_{t_0}^{(\varepsilon,t_0)} = \varepsilon\mathbf{I}$, satisfies

$$\mathcal{V}_{\theta(t)}\big(\mathbf{Y}_t^{(\varepsilon,t_0)}\big) \geq \eta_2\mathbf{I}; \quad \varepsilon\mathbf{I} \leq \mathbf{Y}_{t+1}^{(\varepsilon,t_0)} \leq \delta_2\mathbf{I}$$

for $t \geq t_0 \geq 0$ and for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Part (b) of Theorem 2.7 and its proof then imply that there exists a nonnegative integer M, which depends only on ε, δ_2, η_2, and $\boldsymbol{\mathcal{G}}$, such that for some α_2, $\beta_2 > 0$, $\mathbf{Y}_t$ given by (2.15), with $\mathbf{A}_t = \mathbf{A}_{\theta(t)}$, $\mathbf{B}_t = \mathbf{B}_{\theta(t)}$, $\mathbf{C}_t = \mathbf{C}_{\theta(t)}$, and $\mathbf{D}_t = \mathbf{D}_{\theta(t)}$, satisfies (2.7) for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t \geq 0$. Then Lemmas 2.3 and 2.5 imply that the symmetric positive definite matrices $\mathbf{X}_t$, $t \geq 0$, satisfying (2.6) can be taken to be of the form

$$\mathbf{X}_t = f(\theta(t-M), \ldots, \theta(t-1))$$

for some function $f\colon \{0, \ldots, N\}^M \to \mathbb{S}$, where $\theta(s) = 0$ for $s < 0$. Putting

$$\mathbf{X}_{(\theta(t-M),\ldots,\theta(t-1))} = f(\theta(t-M), \ldots, \theta(t-1))$$

leads to (3.4) for $\boldsymbol{i} = (i_0, \dots, i_M) \in \boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$. Since $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$, and since the path length M is independent of $\boldsymbol{\theta}$, we obtain the desired result. ∎

Remark 3. In Theorem 3.3 the number of linear matrix inequalities (3.4) to solve simultaneously, given M, is equal to the cardinality of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ and bounded above by $\sum_{k=1}^{M+1} N^k$; the number of matrix variables to solve for is equal to the cardinality of $\boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$. In particular, if either $\boldsymbol{\Theta} = \boldsymbol{\Omega}$ or $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for some $\boldsymbol{\theta}$ which is recurrent with respect to $\{1, \dots, N\}$, so that every finite switching path in $\{1, \dots, N\}$ occurs infinitely many times in $\boldsymbol{\theta}$, then the cardinality of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ is precisely N^{M+1}.

Remark 4. If $N = 1$, then we have $\boldsymbol{\Theta} = \{(1, 1, \dots)\}$, and the set $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ is a singleton for each M, and so Theorem 3.3 reduces to the classical KYP lemma.

If inequalities (3.4), with $\mathbf{X}_{\boldsymbol{j}} > \mathbf{0}$ for $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\theta})$, are feasible for some M, then it is also feasible when M is replaced with any integer greater than M. Hence Theorem 3.3 characterizes the performance of switched linear systems via the countably infinite union of an increasing family of systems of linear matrix inequalities. For uniform stability and contractiveness, not only is each member of this family sufficient, but also the union of the family is necessary.

The condition in Theorem 3.3 simplifies if we focus on the uniform stability only. For nonnegative integers L and $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$, let $\boldsymbol{\mathcal{N}}_L(\boldsymbol{\Theta})$ be the largest subset of $\boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})$ satisfying the following: For each $(i_0, \dots, i_L) \in \boldsymbol{\mathcal{N}}_L(\boldsymbol{\Theta})$, there exist an integer $M > L$ and a switching path $(i_0, \dots, i_M)$ such that $(i_{M-L}, \dots, i_M) = (i_0, \dots, i_L)$ and $(i_t, \dots, i_{t+L}) \in \boldsymbol{\mathcal{N}}_L(\boldsymbol{\Theta})$ for $0 \le t \le M - L$. Then we have

$$\boldsymbol{\mathcal{N}}_L(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta}) \cap \{1, \dots, N\}^{M+1}.$$

For example, if

$$\boldsymbol{\Theta} = \{(1, 2, 2, 2, 2, \dots), (1, 1, 2, 2, 2, \dots), (1, 1, 1, 2, 2, \dots), \dots\},$$

then $\boldsymbol{\mathcal{N}}_0(\boldsymbol{\Theta}) = \{1, 2\}$, $\boldsymbol{\mathcal{N}}_1(\boldsymbol{\Theta}) = \{11, 22\}$, $\boldsymbol{\mathcal{N}}_2(\boldsymbol{\Theta}) = \{111, 222\}$, and so on.

COROLLARY 3.4. *Let $\boldsymbol{\mathcal{G}}$ be as in* (3.1)*; let $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ be nonempty. The system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ is uniformly exponentially stable if and only if there exist a nonnegative integer M and matrices $\mathbf{X}_{\boldsymbol{j}} > \mathbf{0}$ such that*

$$\mathbf{A}_{i_M}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{i}_+} \mathbf{A}_{i_M} - \mathbf{X}_{\boldsymbol{i}_-} < \mathbf{0} \tag{3.5}$$

for all $\boldsymbol{i} = (i_0, \dots, i_M) \in \boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta})$.

Proof. Set the matrices $\mathbf{B}_i$, $\mathbf{C}_i$, $\mathbf{D}_i$, $i = 1, \dots, N$, to zero in (3.4) to obtain (3.5). Since $\boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, it suffices to show sufficiency. If (3.5) holds for all $\boldsymbol{i} = (i_0, \dots, i_M) \in \boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta})$, perform the following algorithm:

0. Set $\widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta})$.
1. If $\widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, then stop; otherwise, choose an $\boldsymbol{i} = (i_0, \dots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta}) \setminus \widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta})$ such that $(i_1, \dots, i_M, i_{M+1}) \in \widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta})$ for some $i_{M+1} \in \{1, \dots, N\}$. (By definition of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, such an M-path $\boldsymbol{i}$ exists.)
2. If $\boldsymbol{i}_- \notin \{\hat{\boldsymbol{\imath}}_- : \hat{\boldsymbol{\imath}} \in \widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta})\}$, then choose an $\alpha > 0$ such that $\mathbf{A}_{i_M}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{i}_+} \mathbf{A}_{i_M} - \alpha \mathbf{I} < \mathbf{0}$, put $\mathbf{X}_{\boldsymbol{i}_-} = \alpha \mathbf{I}$, and go to step 4.
3. Choose an $\alpha > 0$ such that $\mathbf{A}_{i_M}^{\mathrm{T}} \mathbf{X}_{\boldsymbol{i}_+} \mathbf{A}_{i_M} - \alpha \mathbf{X}_{\boldsymbol{i}_-} < \mathbf{0}$, and substitute $\mathbf{X}_{\boldsymbol{i}_-}$ with $\alpha \mathbf{X}_{\boldsymbol{i}_-}$. (By definition of $\boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta})$, $\boldsymbol{i}_-$ is not equal to $\boldsymbol{i}_+$.) Whenever there exists an integer $L > 0$ and a switching path $(i_{-L}, \dots, i_{-1})$ such that $(i_{t-L}, \dots, i_{t+M-L})_- \in \{\hat{\boldsymbol{\imath}}_- : \hat{\boldsymbol{\imath}} \in \widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta})\}$ for all $0 \le t \le L$, then substitute $\mathbf{X}_{(i_{-L}, \dots, i_{M-L})_-}$ with $\alpha \mathbf{X}_{(i_{-L}, \dots, i_{M-L})_-}$, too. (Again, by definition of $\boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta})$, such a path $(i_{-L}, \dots, i_{M-L})_-$ cannot be equal to $\boldsymbol{i}_+$.)

4. Substitute $\widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta})$ with $\widetilde{\boldsymbol{\mathcal{M}}}_M(\boldsymbol{\Theta}) \cup \{\boldsymbol{i}\}$ and go to step 1.

Since the cardinality of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ is finite, we will have reconstructed an entire family $\{\mathbf{X}_{\boldsymbol{j}} : \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})\}$ of matrices $\mathbf{X}_{\boldsymbol{j}} > \mathbf{0}$ at the termination of this algorithm so that (3.5) holds for all $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$. □

Remark 5. If $\boldsymbol{\Theta}$ is defined via a strongly connected directed graph, then Corollary 3.4 reverts to the stability analysis result in [30].

Now, consider the set

$$\boldsymbol{\mathcal{T}} = \{(\mathbf{A}_i, \mathbf{B}_{1,i}, \mathbf{B}_{2,i}, \mathbf{C}_{1,i}, \mathbf{C}_{2,i}, \mathbf{D}_{11,i}, \mathbf{D}_{12,i}, \mathbf{D}_{21,i}) : i = 1, \ldots, N\} \tag{3.6}$$

with $\mathbf{A}_i \in \mathbb{R}^{n\times n}$, $\mathbf{B}_{1,i} \in \mathbb{R}^{n\times m_1}$, $\mathbf{B}_{2,i} \in \mathbb{R}^{n\times m_2}$, $\mathbf{C}_{1,i} \in \mathbb{R}^{l_1\times n}$, $\mathbf{C}_{2,i} \in \mathbb{R}^{l_2\times n}$, $\mathbf{D}_{11,i} \in \mathbb{R}^{l_1\times m_1}$, $\mathbf{D}_{12,i} \in \mathbb{R}^{l_1\times m_2}$, $\mathbf{D}_{21,i} \in \mathbb{R}^{l_2\times m_1}$ for $i = 1, \ldots, N$. If $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ and is nonempty, then the pair $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ defines the *controlled switched linear system* represented by

$$\begin{aligned} x(t+1) &= \mathbf{A}_{\theta(t)}x(t) + \mathbf{B}_{1,\theta(t)}w(t) + \mathbf{B}_{2,\theta(t)}u(t), \\ z(t) &= \mathbf{C}_{1,\theta(t)}x(t) + \mathbf{D}_{11,\theta(t)}w(t) + \mathbf{D}_{12,\theta(t)}u(t), \\ y(t) &= \mathbf{C}_{2,\theta(t)}x(t) + \mathbf{D}_{21,\theta(t)}w(t). \end{aligned} \tag{3.7}$$

Given the initial state $x(0)$, disturbance sequence $\boldsymbol{w} = (w(t))$, control sequence $\boldsymbol{u} = (u(t))$, and switching sequence $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, this system of equations defines the evolution of the state $x(t)$, controlled output $z(t)$, and measured output $y(t)$ for $t \geq 0$. Based on the analysis result given by Theorem 3.3, we will be deriving a necessary and sufficient condition for controller synthesis.

We make the standard assumption that *the mode $\theta(t)$ is perfectly observed at each time instant t*; however, relaxing the standard restriction to mode-dependent controllers (i.e., controllers that do not recall past modes), we consider all controllers that have a finite memory of past modes as well as a perfect observation of the current mode. Fix a nonnegative integer L. Let

$$\boldsymbol{\Theta}_L = \{\boldsymbol{\theta}_L : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

be the set of L-path switching sequences generated by $\boldsymbol{\Theta}$; let

$$\boldsymbol{\mathcal{K}} = \{(\mathbf{A}_{K,\boldsymbol{i}}, \mathbf{B}_{K,\boldsymbol{i}}, \mathbf{C}_{K,\boldsymbol{i}}, \mathbf{D}_{K,\boldsymbol{i}}) : \boldsymbol{i} \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})\} \tag{3.8}$$

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

$$\begin{aligned} x_K(t+1) &= \mathbf{A}_{K,\theta_L(t)}x_K(t) + \mathbf{B}_{K,\theta_L(t)}y(t), \\ u(t) &= \mathbf{C}_{K,\theta_L(t)}x_K(t) + \mathbf{D}_{K,\theta_L(t)}y(t) \end{aligned} \tag{3.9}$$

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

$$\mathbf{K}_{\boldsymbol{i}} = \begin{bmatrix} \mathbf{A}_{K,\boldsymbol{i}} & \mathbf{B}_{K,\boldsymbol{i}} \\ \mathbf{C}_{K,\boldsymbol{i}} & \mathbf{D}_{K,\boldsymbol{i}} \end{bmatrix} \in \mathbb{R}^{(n_K+m_2)\times(n_K+l_2)}, \quad \boldsymbol{i} \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}).$$

Given a finite-path-dependent controller $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}_L)$, where L is the path length, let

$$\begin{aligned}\widetilde{\mathbf{A}}_{\boldsymbol{i}} &= \widehat{\mathbf{A}}_{i_L} + \widehat{\mathbf{B}}_{2,i_L}\mathbf{K}_{\boldsymbol{i}}\widehat{\mathbf{C}}_{2,i_L}, \quad \widetilde{\mathbf{B}}_{\boldsymbol{i}} = \widehat{\mathbf{B}}_{1,i_L} + \widehat{\mathbf{B}}_{2,i_L}\mathbf{K}_{\boldsymbol{i}}\widehat{\mathbf{D}}_{21,i_L},\\ \widetilde{\mathbf{C}}_{\boldsymbol{i}} &= \widehat{\mathbf{C}}_{1,i_L} + \widehat{\mathbf{D}}_{12,i_L}\mathbf{K}_{\boldsymbol{i}}\widehat{\mathbf{C}}_{2,i_L}, \widetilde{\mathbf{D}}_{\boldsymbol{i}} = \mathbf{D}_{11,i_L} + \widehat{\mathbf{D}}_{12,i_L}\mathbf{K}_{\boldsymbol{i}}\widehat{\mathbf{D}}_{21,i_L}\end{aligned} \tag{3.10}$$

for $\boldsymbol{i} = (i_0, \dots, i_L) \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$, with

$$\widehat{\mathbf{A}}_i = \begin{bmatrix}\mathbf{A}_i & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{(n+n_K)\times(n+n_K)},$$

$$\widehat{\mathbf{B}}_{1,i} = \begin{bmatrix}\mathbf{B}_{1,i}\\ \mathbf{0}\end{bmatrix} \in \mathbb{R}^{(n+n_K)\times m_1}, \qquad \widehat{\mathbf{B}}_{2,i} = \begin{bmatrix}\mathbf{0} & \mathbf{B}_{2,i}\\ \mathbf{I} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{(n+n_K)\times(n_K+m_2)},$$

$$\widehat{\mathbf{C}}_{1,i} = \begin{bmatrix}\mathbf{C}_{1,i} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{l_1\times(n+n_K)}, \quad \widehat{\mathbf{C}}_{2,i} = \begin{bmatrix}\mathbf{0} & \mathbf{I}\\ \mathbf{C}_{2,i} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{(n_K+l_2)\times(n+n_K)},$$

$$\widehat{\mathbf{D}}_{12,i} = \begin{bmatrix}\mathbf{0} & \mathbf{D}_{12,i}\end{bmatrix} \in \mathbb{R}^{l_1\times(n_K+m_2)}, \widehat{\mathbf{D}}_{21,i} = \begin{bmatrix}\mathbf{0}\\ \mathbf{D}_{21,i}\end{bmatrix} \in \mathbb{R}^{(n_K+l_2)\times m_1}$$

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

$$\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}} = \big\{\big(\widetilde{\mathbf{A}}_{\boldsymbol{i}}, \widetilde{\mathbf{B}}_{\boldsymbol{i}}, \widetilde{\mathbf{C}}_{\boldsymbol{i}}, \widetilde{\mathbf{D}}_{\boldsymbol{i}}\big) : \boldsymbol{i} \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})\big\}.$$

If we define the closed-loop state by

$$\tilde{x}(t) = \begin{bmatrix} x(t)^{\mathrm{T}} & x_K(t)^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{n+n_K},$$

then the *closed-loop system* $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}}, \boldsymbol{\Theta}_L)$ has the representation

$$\begin{aligned}\tilde{x}(t+1) &= \widetilde{\mathbf{A}}_{\theta_L(t)}\tilde{x}(t) + \widetilde{\mathbf{B}}_{\theta_L(t)}w(t),\\ z(t) &= \widetilde{\mathbf{C}}_{\theta_L(t)}\tilde{x}(t) + \widetilde{\mathbf{D}}_{\theta_L(t)}w(t)\end{aligned} \tag{3.11}$$

for each L-path switching sequence $\boldsymbol{\theta}_L \in \boldsymbol{\Theta}_L$.

Let N_L be the cardinality of $\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$. If we label the elements of $\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$ from 1 to N_L, then each L-path switching sequence $\boldsymbol{\theta}_L = (\theta_L(0), \theta_L(1), \dots) \in \boldsymbol{\Theta}_L$ can be considered a *closed-loop switching sequence* in $\{1, \dots, N_L\}$; letting $\theta_L(t) = 0$ for $t < 0$, the *closed-loop switching path* $(\theta_L(t-M), \dots, \theta_L(t))$ can be identified with the switching path $(\theta(t-L-M), \dots, \theta(t))$ for each triple (t, L, M) of nonnegative integers. This leads to the following identities for all integers $L > 0$ and $M \geq 0$:

$$\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta}_L) = \boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta}_L) = \boldsymbol{\mathcal{L}}_{M+L}(\boldsymbol{\Theta}). \tag{3.12}$$

Hence, even if $L > 0$, the closed-loop system $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}}, \boldsymbol{\Theta}_L)$ is a switched linear system, where the *closed-loop modes* are the L-paths in $\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$, and the *closed-loop M-paths* are the $(M+L)$-paths in $\boldsymbol{\mathcal{L}}_{M+L}(\boldsymbol{\Theta})$ for each nonnegative integer M.

LEMMA 3.5. *Let $\boldsymbol{\mathcal{T}}$ be as in* (3.6)*; let $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ be nonempty. Suppose that $\boldsymbol{\mathcal{K}}$ is finite-path-dependent as in* (3.8) *with some nonnegative integer L. Then the closed-loop system $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}}, \boldsymbol{\Theta}_L)$ is uniformly exponentially stable and uniformly strictly contractive if and only if there exist an integer $M \geq L$ and an indexed family $\{\mathbf{X}_{\boldsymbol{j}} : \boldsymbol{j} \in \boldsymbol{\mathcal{L}}_M^-(\boldsymbol{\Theta})\}$ of symmetric positive definite matrices $\mathbf{X}_{\boldsymbol{j}} \in \mathbb{R}^{(n+n_K)\times(n+n_K)}$ such that*

$$\mathbf{H}_{(i_0,\dots,i_M)} + \mathbf{G}_{i_M}^{\mathrm{T}}\mathbf{K}_{(i_{M-L},\dots,i_M)}^{\mathrm{T}}\mathbf{F}_{i_M} + \mathbf{F}_{i_M}^{\mathrm{T}}\mathbf{K}_{(i_{M-L},\dots,i_M)}\mathbf{G}_{i_M} < \mathbf{0} \tag{3.13}$$

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

$$\mathbf{H}_{\boldsymbol{i}}=\begin{bmatrix}-\mathbf{X}_{\boldsymbol{i}_+}^{-1} & \widehat{\mathbf{A}}_{i_M} & \widehat{\mathbf{B}}_{1,i_M} & \mathbf{0}\\ \widehat{\mathbf{A}}_{i_M}^{\mathrm{T}} & -\mathbf{X}_{\boldsymbol{i}_-} & \mathbf{0} & \widehat{\mathbf{C}}_{1,i_M}^{\mathrm{T}}\\ \widehat{\mathbf{B}}_{1,i_M}^{\mathrm{T}} & \mathbf{0} & -\mathbf{I} & \mathbf{D}_{11,i_M}^{\mathrm{T}}\\ \mathbf{0} & \widehat{\mathbf{C}}_{1,i_M} & \mathbf{D}_{11,i_M} & -\mathbf{I}\end{bmatrix}\in\mathbb{R}^{(2n+2n_K+m_1+l_1)\times(2n+2n_K+m_1+l_1)}$$

for $\boldsymbol{i}=(i_0,\dots,i_M)\in\boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$ and where

$$\begin{bmatrix}\mathbf{F}_i\\ \mathbf{G}_i\end{bmatrix}=\begin{bmatrix}\widehat{\mathbf{B}}_{2,i}^{\mathrm{T}} & \mathbf{0} & \mathbf{0} & \widehat{\mathbf{D}}_{12,i}^{\mathrm{T}}\\ \mathbf{0} & \widehat{\mathbf{C}}_{2,i} & \widehat{\mathbf{D}}_{21,i} & \mathbf{0}\end{bmatrix}\in\mathbb{R}^{(2n_K+m_2+l_2)\times(2n+2n_K+m_1+l_1)}$$

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

Proof. Let us first assume that $L>0$. It follows from Theorem 3.3 and (3.12) that $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}},\boldsymbol{\Theta}_L)$ is uniformly stable and contractive if and only if there exists a nonnegative integer M satisfying the following: either $M=0$ and there is a single $\mathbf{X}_0>\mathbf{0}$ such that

$$\begin{bmatrix}\widetilde{\mathbf{A}}_{\boldsymbol{i}} & \widetilde{\mathbf{B}}_{\boldsymbol{i}}\\ \widetilde{\mathbf{C}}_{\boldsymbol{i}} & \widetilde{\mathbf{D}}_{\boldsymbol{i}}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{X}_0 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{A}}_{\boldsymbol{i}} & \widetilde{\mathbf{B}}_{\boldsymbol{i}}\\ \widetilde{\mathbf{C}}_{\boldsymbol{i}} & \widetilde{\mathbf{D}}_{\boldsymbol{i}}\end{bmatrix}-\begin{bmatrix}\mathbf{X}_0 & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}<\mathbf{0}\tag{3.14a}$$

for $\boldsymbol{i}=(i_0,\dots,i_L)\in\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$, or $M>0$ and there are $\mathbf{X}_{(j_1,\dots,j_{M+L})}>\mathbf{0}$ such that

$$\begin{aligned}&\begin{bmatrix}\widetilde{\mathbf{A}}_{(i_M,\dots,i_{M+L})} & \widetilde{\mathbf{B}}_{(i_M,\dots,i_{M+L})}\\ \widetilde{\mathbf{C}}_{(i_M,\dots,i_{M+L})} & \widetilde{\mathbf{D}}_{(i_M,\dots,i_{M+L})}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{X}_{\boldsymbol{i}_+} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\\ &\qquad\times\begin{bmatrix}\widetilde{\mathbf{A}}_{(i_M,\dots,i_{M+L})} & \widetilde{\mathbf{B}}_{(i_M,\dots,i_{M+L})}\\ \widetilde{\mathbf{C}}_{(i_M,\dots,i_{M+L})} & \widetilde{\mathbf{D}}_{(i_M,\dots,i_{M+L})}\end{bmatrix}-\begin{bmatrix}\mathbf{X}_{\boldsymbol{i}_-} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}<\mathbf{0}\end{aligned}\tag{3.14b}$$

for $\boldsymbol{i}=(i_0,\dots,i_{M+L})\in\boldsymbol{\mathcal{L}}_{M+L}(\boldsymbol{\Theta})$. If $M=0$, then we may put $\mathbf{X}_{(j_1,\dots,j_L)}=\mathbf{X}_0$ for all L-paths $(j_1,\dots,j_L)$ and write (3.14a) as

$$\begin{bmatrix}\widetilde{\mathbf{A}}_{\boldsymbol{i}} & \widetilde{\mathbf{B}}_{\boldsymbol{i}}\\ \widetilde{\mathbf{C}}_{\boldsymbol{i}} & \widetilde{\mathbf{D}}_{\boldsymbol{i}}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{X}_{\boldsymbol{i}_+} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{A}}_{\boldsymbol{i}} & \widetilde{\mathbf{B}}_{\boldsymbol{i}}\\ \widetilde{\mathbf{C}}_{\boldsymbol{i}} & \widetilde{\mathbf{D}}_{\boldsymbol{i}}\end{bmatrix}-\begin{bmatrix}\mathbf{X}_{\boldsymbol{i}_-} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}<\mathbf{0};\tag{3.14c}$$

if (3.14c) holds for $\boldsymbol{i}\in\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$, with some $\mathbf{X}_{(j_1,\dots,j_L)}>\mathbf{0}$, then Lemma 2.3 and the finiteness of $\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta})$ imply that $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}},\boldsymbol{\Theta}_L)$ is uniformly stable and contractive. Now, as in [20, p. 431], rewrite (3.14b) and (3.14c) (the former for $M>0$ and the latter for $M=0$) as inequalities of the form (3.13) using the decompositions (3.10) along with the Schur complement formula and an appropriate congruence transformation. Replacing M with $M-L$ then yields the desired result.

If $L=0$, then $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta}_L)=\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, which is not equal to $\boldsymbol{\mathcal{L}}_{M+L}(\boldsymbol{\Theta})=\boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$ in general. However, the proof of Theorem 3.3 shows that the existence of $\mathbf{X}_{\boldsymbol{j}}>\mathbf{0}$, $\boldsymbol{j}\in\boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$, such that (3.4) holds for all $\boldsymbol{i}=(i_0,\dots,i_M)\in\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, suffices for the existence of $\mathbf{X}_{\boldsymbol{j}}>\mathbf{0}$, $\boldsymbol{j}\in\boldsymbol{\mathcal{L}}_M^-(\boldsymbol{\Theta})$, such that (3.4) holds for all $\boldsymbol{i}=(i_0,\dots,i_M)\in\boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$. Therefore, the proof of the result for $L=0$ is identical to the case of $L>0$. ∎

Note that Lemma 3.5 is stated in terms of the closed-loop $(M-L)$-paths in $\boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$, $M\geq L$, rather than the closed-loop M-paths in $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta}_L)$, because the former are easier to deal with. Inequality (3.13) is amenable to the standard linear matrix

inequality embedding technique, originally developed for linear time-invariant systems [33, 20]. Finite-path-dependent controllers arise naturally from this technique.

DEFINITION 3.6. *The controller* $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}_L)$ *is said to be an* admissible (L-path-dependent) synthesis (of order n_K) *for the system* $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ *if the closed-loop system* $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}}, \boldsymbol{\Theta}_L)$ *is uniformly exponentially stable and uniformly strictly contractive.*

THEOREM 3.7. *Let* $\boldsymbol{\mathcal{T}}$ *be as in* (3.6)*; let* $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ *be nonempty. Suppose that* $n_K \geq n$*. There exists an admissible finite-path-dependent synthesis of order* n_K *for the system* $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ *if and only if there exist a nonnegative integer* M *and an indexed family* $\{(\mathbf{R}_{\boldsymbol{j}}, \mathbf{S}_{\boldsymbol{j}}) \colon \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})\}$ *of pairs of symmetric positive definite matrices* $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}} \in \mathbb{R}^{n \times n}$ *such that*

(3.15a)
$$\mathbf{N}_{F,i_M}^{\mathrm{T}} \left(\begin{bmatrix} \mathbf{A}_{i_M} & \mathbf{B}_{1,i_M} \\ \mathbf{C}_{1,i_M} & \mathbf{D}_{11,i_M} \end{bmatrix} \begin{bmatrix} \mathbf{R}_{\boldsymbol{i}_-} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{i_M} & \mathbf{B}_{1,i_M} \\ \mathbf{C}_{1,i_M} & \mathbf{D}_{11,i_M} \end{bmatrix}^{\mathrm{T}} - \begin{bmatrix} \mathbf{R}_{\boldsymbol{i}_+} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right) \mathbf{N}_{F,i_M} < \mathbf{0},$$

(3.15b)
$$\mathbf{N}_{G,i_M}^{\mathrm{T}} \left(\begin{bmatrix} \mathbf{A}_{i_M} & \mathbf{B}_{1,i_M} \\ \mathbf{C}_{1,i_M} & \mathbf{D}_{11,i_M} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{S}_{\boldsymbol{i}_+} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{i_M} & \mathbf{B}_{1,i_M} \\ \mathbf{C}_{1,i_M} & \mathbf{D}_{11,i_M} \end{bmatrix} - \begin{bmatrix} \mathbf{S}_{\boldsymbol{i}_-} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right) \mathbf{N}_{G,i_M} < \mathbf{0},$$

(3.15c)
$$\begin{bmatrix} \mathbf{R}_{\boldsymbol{i}_-} & \mathbf{I} \\ \mathbf{I} & \mathbf{S}_{\boldsymbol{i}_-} \end{bmatrix} \geq \mathbf{0}$$

for all M*-paths* $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$*, where*

$$\mathbf{N}_{F,i} = \mathbf{N}\big(\begin{bmatrix} \mathbf{B}_{2,i}^{\mathrm{T}} & \mathbf{D}_{12,i}^{\mathrm{T}} \end{bmatrix} \big), \quad \mathbf{N}_{G,i} = \mathbf{N}\big(\begin{bmatrix} \mathbf{C}_{2,i} & \mathbf{D}_{21,i} \end{bmatrix} \big)$$

for $i \in \{1, \ldots, N\}$*. Moreover, if* (3.15) *holds for all* $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$*, then there exist a nonnegative integer* $L \leq M$ *and matrices* $\mathbf{K}_{(j_0,\ldots,j_L)} \in \mathbb{R}^{(n_K+m_2)\times(n_K+l_2)}$ *such that* (3.13) *holds for all* $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ *with*

(3.16)
$$\mathbf{X}_{\boldsymbol{j}} = \begin{bmatrix} \mathbf{S}_{\boldsymbol{j}} & \mathbf{U}_{\boldsymbol{j}} \mathbf{V}_{\boldsymbol{j}}^{\frac{1}{2}} \\ \mathbf{V}_{\boldsymbol{j}}^{\frac{1}{2}} \mathbf{U}_{\boldsymbol{j}}^{\mathrm{T}} & \mathbf{V}_{\boldsymbol{j}} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_{\boldsymbol{j}} & -\mathbf{R}_{\boldsymbol{j}} \mathbf{U}_{\boldsymbol{j}} \mathbf{V}_{\boldsymbol{j}}^{-\frac{1}{2}} \\ -\mathbf{V}_{\boldsymbol{j}}^{-\frac{1}{2}} \mathbf{U}_{\boldsymbol{j}}^{\mathrm{T}} \mathbf{R}_{\boldsymbol{j}} & \mathbf{V}_{\boldsymbol{j}}^{-\frac{1}{2}} (\mathbf{I} + \mathbf{U}_{\boldsymbol{j}}^{\mathrm{T}} \mathbf{R}_{\boldsymbol{j}} \mathbf{U}_{\boldsymbol{j}}) \mathbf{V}_{\boldsymbol{j}}^{-\frac{1}{2}} \end{bmatrix}^{-1} > \mathbf{0}$$

for $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$*, where* $\mathbf{U}_{\boldsymbol{j}} \in \mathbb{R}^{n \times n_K}$, $\mathbf{V}_{\boldsymbol{j}} \in \mathbb{R}^{n_K \times n_K}$ *are any matrices such that* $\mathbf{U}_{\boldsymbol{j}} \mathbf{U}_{\boldsymbol{j}}^{\mathrm{T}} = \mathbf{S}_{\boldsymbol{j}} - \mathbf{R}_{\boldsymbol{j}}^{-1}$ *and* $\mathbf{V}_{\boldsymbol{j}} > \mathbf{0}$*; in particular, one may take* $L = M$*.*

Proof. It follows from [20, Lem. 3.1] that, with $L = M$, inequality (3.13) in $\mathbf{K}_{(i_{M-L},\ldots,i_M)} = \mathbf{K}_{(i_0,\ldots,i_M)}$ is feasible if and only if

$$\mathbf{N}(\mathbf{F}_{i_M})^{\mathrm{T}} \mathbf{H}_{(i_0,\ldots,i_M)} \mathbf{N}(\mathbf{F}_{i_M}) < \mathbf{0}; \quad \mathbf{N}(\mathbf{G}_{i_M})^{\mathrm{T}} \mathbf{H}_{(i_0,\ldots,i_M)} \mathbf{N}(\mathbf{G}_{i_M}) < \mathbf{0}.$$

These inequalities are equivalent to (3.15) due to the Schur complement arguments in [20, sect. 5] together with the matrix inverse completion result in [33, Lem. 6.2]. The proof of the latter shows that $\mathbf{X}_{\boldsymbol{j}}$, $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$, can be reconstructed from $\mathbf{R}_{\boldsymbol{j}}$ and $\mathbf{S}_{\boldsymbol{j}}$ through (3.16). Now, since $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta}) \subset \boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$, the existence of matrices $\mathbf{K}_{(i_0,\ldots,i_M)}$, such that (3.13) holds for $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$, implies the feasibility of (3.15) for $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$.

To show the converse, suppose that there are a nonnegative integer M and matrices $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}} > \mathbf{0}$, $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$, such that (3.15) holds for $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$. Assume $M > 0$ without loss of generality. Reconstruct $\mathbf{X}_{\boldsymbol{j}} > \mathbf{0}$, $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$, through

(3.16). Then there exist matrices $\mathbf{K}_{(i_0,\ldots,i_M)}$, $(i_0,\ldots,i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, such that (3.13), with $L = M$, holds for $(i_0,\ldots,i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$. By the definition of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, for each $j \in \boldsymbol{\mathcal{L}}_0(\boldsymbol{\Theta})$, one can choose a switching path $\left(i_0^j,\ldots,i_{M-1}^j\right) \in \{0,\ldots,N\}^M$ such that (3.3), with $L = M$, holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. If we put

$$\mathbf{X}_{\theta_M(t)} = \begin{cases} \mathbf{X}_{\left(i_0^{\theta(0)},\ldots,i_{M-1}^{\theta(0)}\right)}, & t = 0; \\ \mathbf{X}_{\left(i_t^{\theta(0)},\ldots,i_{M-1}^{\theta(0)},\theta(0),\ldots,\theta(t)\right)}, & 0 < t < M; \end{cases}$$

$$\mathbf{K}_{\theta_M(t)} = \begin{cases} \mathbf{K}_{\left(i_0^{\theta(0)},\ldots,i_{M-1}^{\theta(0)}\right)}, & t = 0; \\ \mathbf{K}_{\left(i_t^{\theta(0)},\ldots,i_{M-1}^{\theta(0)},\theta(0),\ldots,\theta(t)\right)}, & 0 < t < M, \end{cases}$$

for all $\boldsymbol{\theta}_M \in \boldsymbol{\Theta}_M$ and $t \geq 0$ such that $\theta_M(t) \notin \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, then we recover (3.13) for all $(i_0,\ldots,i_M) \in \boldsymbol{\mathcal{L}}_M(\boldsymbol{\Theta})$. The result then follows from Lemma 3.5. ∎

Remark 6. Given a nonnegative integer M, the number of systems of linear matrix inequalities (3.15) to solve simultaneously is equal to the cardinality of $\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, and the feasibility of inequalities (3.15) is sufficient for the existence of an admissible L-path-dependent synthesis for some $L \leq M$. However, given a nonnegative integer L, the feasibility of (3.15) for some $M \leq L$ is sufficient but not necessary for the existence of an admissible L-path-dependent synthesis. In fact, even if the existence of an admissible L-path-dependent controller guarantees that inequalities (3.15) are feasible for some finite M, there is no upper bound on such an M. See Example 3.

Remark 7. In the case of reduced order controllers with $n_K < n$, the matrices $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}}$, $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$, must satisfy

$$\operatorname{rank} \begin{bmatrix} \mathbf{R}_{\boldsymbol{i}_-} & \mathbf{I} \\ \mathbf{I} & \mathbf{S}_{\boldsymbol{i}_-} \end{bmatrix} \leq n + n_K,$$

in addition to (3.15), for all $\boldsymbol{i} = (i_0,\ldots,i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$.

COROLLARY 3.8. *Let $\boldsymbol{\mathcal{T}}$ be as in* (3.6)*; let $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ be nonempty. Suppose that $n_K \geq n$. There exists a finite-path-dependent linear output feedback controller of order n_K that uniformly stabilizes the system $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ if and only if there exist a nonnegative integer M and matrices $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}} > \mathbf{0}$ such that*

$$\mathbf{N}(\mathbf{B}_{i_M}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{A}_{i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{A}_{i_M}^{\mathrm{T}} - \mathbf{R}_{\boldsymbol{i}_+})\mathbf{N}(\mathbf{B}_{i_M}^{\mathrm{T}}) < \mathbf{0}, \tag{3.17a}$$

$$\mathbf{N}(\mathbf{C}_{i_M})^{\mathrm{T}}(\mathbf{A}_{i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{A}_{i_M} - \mathbf{S}_{\boldsymbol{i}_-})\mathbf{N}(\mathbf{C}_{i_M}) < \mathbf{0}, \tag{3.17b}$$

$$\begin{bmatrix} \mathbf{R}_{\boldsymbol{i}_-} & \mathbf{I} \\ \mathbf{I} & \mathbf{S}_{\boldsymbol{i}_-} \end{bmatrix} \geq \mathbf{0} \tag{3.17c}$$

for all $\boldsymbol{i} = (i_0,\ldots,i_M) \in \boldsymbol{\mathcal{N}}_M(\boldsymbol{\Theta})$.

Proof. In (3.15), set the matrices $\mathbf{B}_{1,i}$, $\mathbf{C}_{1,i}$, $\mathbf{D}_{11,i}$, $\mathbf{D}_{12,i}$, and $\mathbf{D}_{21,i}$ to zero for all $i = 1, \ldots, N$, and obtain (3.17). The rest of the proof proceeds similarly to that of Corollary 3.4. ∎

Remark 8. Corollary 3.8 reverts to the stabilization result in [30] if a strongly connected directed graph defines $\boldsymbol{\Theta}$.

Suppose that a set of matrices $\mathbf{K}_{\boldsymbol{i}}$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})$, is obtained by solving (3.13) via Theorem 3.7 for some nonnegative integer $L \leq M$. If $L = 0$, then it follows from $\boldsymbol{\mathcal{M}}_0(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{L}}_0(\boldsymbol{\Theta})$ that we have all the matrices $\mathbf{K}_i$, $i \in \boldsymbol{\mathcal{L}}_0(\boldsymbol{\Theta})$, that define an admissible zero-path-dependent controller synthesis. If $L > 0$, on the other hand,

then choose switching paths $\big(i_0^j,\dots,i_{L-1}^j\big)\in\{0,\dots,N\}^L$, $j\in\boldsymbol{\mathcal{L}}_0(\boldsymbol{\Theta})$, such that (3.3) holds for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, and put

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

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

Example 3. Let $\boldsymbol{\theta}^{(k)}$ be as in Example 2 for positive integers k. Let

$$\boldsymbol{\mathcal{T}}=\{(1,-1,1,-1,1,1,-1,-1),(1,0,0,0,0,0,0,0)\}$$

and $\boldsymbol{\Theta}^{(k)}=\big\{\boldsymbol{\theta}^{(k)}\big\}$ so that the controlled system $\big(\boldsymbol{\mathcal{T}},\boldsymbol{\Theta}^{(k)}\big)$ has the representation

$$\begin{aligned}x(t+1)&=x(t)-w(t)+u(t),\\ z(t)&=-x(t)+w(t)-u(t),\\ y(t)&=x(t)-w(t)\end{aligned}$$

in mode 1 (i.e., when $\theta^{(k)}(t)=1$), and

$$x(t+1)=x(t),\quad z(t)=0,\quad y(t)=0$$

in mode 2 (i.e., when $\theta^{(k)}(t)=2$). Let M be a nonnegative integer. If $M<k$, then, because mode 2 is not stabilizable and because $\boldsymbol{\mathcal{M}}_M\big(\boldsymbol{\Theta}^{(k)}\big)$ contains the switching path $(2,\dots,2)$ that consists of mode 2 only, inequality (3.15) cannot be satisfied for all $\boldsymbol{i}=(i_0,\dots,i_M)\in\boldsymbol{\mathcal{M}}_M\big(\boldsymbol{\Theta}^{(k)}\big)$. On the other hand, the cardinality of $\boldsymbol{\mathcal{M}}_M\big(\boldsymbol{\Theta}^{(k)}\big)$ is equal to that of $\boldsymbol{\mathcal{M}}_k\big(\boldsymbol{\Theta}^{(k)}\big)$ for all $M>k$. Hence, to design an admissible synthesis for $\big(\boldsymbol{\mathcal{T}},\boldsymbol{\Theta}^{(k)}\big)$, it suffices to consider the single path length $M=k$ in (3.15). It is readily verified that there indeed is an admissible finite-path-dependent controller for $\big(\boldsymbol{\mathcal{T}},\boldsymbol{\Theta}^{(k)}\big)$ for each k. This shows that there does not exist a general upper bound on the path length M in Theorem 3.7.

In this particular example, it is easy to find solutions $\mathbf{R}_{\boldsymbol{j}},\mathbf{S}_{\boldsymbol{j}}>\mathbf{0}$, $\boldsymbol{j}\in\boldsymbol{\mathcal{M}}_M^-\big(\boldsymbol{\Theta}^{(k)}\big)$, to (3.15) with $M=k$, such that $L=0$ (which leads to mode-dependent controllers) suffices for (3.13) to be feasible. This indicates that the existence of an admissible L-path controller synthesis does not necessarily lead to the feasibility of (3.15) with $M=L$.

DEFINITION 3.9. *Let $\gamma>0$. The system $(\boldsymbol{\mathcal{G}},\boldsymbol{\Theta})$ is said to* satisfy uniform disturbance attenuation level γ *if there exists a $\tilde{\gamma}\in(0,\gamma)$ such that, whenever $x(t_0)=0$,*

$$\sum\nolimits_{s=t_0}^{t}\|z(s)\|^2\le\tilde{\gamma}^2\sum\nolimits_{s=t_0}^{t}\|w(s)\|^2$$

for $t\ge t_0\ge0$, for $\boldsymbol{w}\in\ell^2(\mathbb{R}^m)$, and for $\boldsymbol{\theta}\in\boldsymbol{\Theta}$.

DEFINITION 3.10. *Let $\gamma>0$. The controller $(\boldsymbol{\mathcal{K}},\boldsymbol{\Theta}_L)$ is said to be a* γ-admissible (L-path-dependent) synthesis (of order n_K), *or to* achieve uniform disturbance attenuation level γ, *for the system $(\boldsymbol{\mathcal{T}},\boldsymbol{\Theta})$ if the closed-loop system $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}},\boldsymbol{\Theta}_L)$ is uniformly exponentially stable and satisfies uniform disturbance attenuation level γ.*

Given a $\gamma>0$, let

$$\boldsymbol{\mathcal{G}}^{(\gamma)}=\big\{(\mathbf{A}_i,\gamma^{-1/2}\mathbf{B}_i,\gamma^{-1/2}\mathbf{C}_i,\gamma^{-1}\mathbf{D}_i)\colon i=1,\dots,N\big\}.$$

Then $(\boldsymbol{\mathcal{G}},\boldsymbol{\Theta})$ satisfies uniform disturbance attenuation level γ if and only if $\big(\boldsymbol{\mathcal{G}}^{(\gamma)},\boldsymbol{\Theta}\big)$ is uniformly strictly contractive. Using this fact and Theorem 3.7, and applying the Schur complement formula to (3.15a) and (3.15b), we obtain the following.

COROLLARY 3.11. *Let $\boldsymbol{\mathcal{T}}$ be as in* (3.6)*, and let $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ be nonempty. Let $\gamma > 0$. Suppose that $n_K \geq n$. There exists a γ-admissible finite-path-dependent synthesis of order n_K for the system $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ if and only if there exist a nonnegative integer M and an indexed family $\{(\mathbf{R}_{\boldsymbol{j}}, \mathbf{S}_{\boldsymbol{j}}) \colon \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})\}$ of pairs of symmetric positive definite matrices $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}} \in \mathbb{R}^{n\times n}$ such that*

$$\begin{bmatrix} \mathbf{N}_{F,i_M} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{A}_{i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{A}_{i_M}^{\mathrm{T}} - \mathbf{R}_{\boldsymbol{i}_+} & \mathbf{A}_{i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{C}_{1,i_M}^{\mathrm{T}} & \mathbf{B}_{1,i_M} \\ \mathbf{C}_{1,i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{A}_{i_M}^{\mathrm{T}} & \mathbf{C}_{1,i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{C}_{1,i_M}^{\mathrm{T}} - \gamma\mathbf{I} & \mathbf{D}_{11,i_M} \\ \mathbf{B}_{1,i_M}^{\mathrm{T}} & \mathbf{D}_{11,i_M}^{\mathrm{T}} & -\gamma\mathbf{I} \end{bmatrix} \times \begin{bmatrix} \mathbf{N}_{F,i_M} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} < \mathbf{0}, \tag{3.18a}$$

$$\begin{bmatrix} \mathbf{N}_{G,i_M} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{A}_{i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{A}_{i_M} - \mathbf{S}_{\boldsymbol{i}_-} & \mathbf{A}_{i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{B}_{1,i_M} & \mathbf{C}_{1,i_M}^{\mathrm{T}} \\ \mathbf{B}_{1,i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{A}_{i_M} & \mathbf{B}_{1,i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{B}_{1,i_M} - \gamma\mathbf{I} & \mathbf{D}_{11,i_M}^{\mathrm{T}} \\ \mathbf{C}_{1,i_M} & \mathbf{D}_{11,i_M} & -\gamma\mathbf{I} \end{bmatrix} \times \begin{bmatrix} \mathbf{N}_{G,i_M} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} < \mathbf{0}, \tag{3.18b}$$

$$\begin{bmatrix} \mathbf{R}_{\boldsymbol{i}_-} & \mathbf{I} \\ \mathbf{I} & \mathbf{S}_{\boldsymbol{i}_-} \end{bmatrix} \geq \mathbf{0} \tag{3.18c}$$

for all $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$. Moreover, if (3.18) *holds for all $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, then there exist a nonnegative integer $L \leq M$ and matrices $\mathbf{K}_{(i_{M-L},\ldots,i_M)} \in \mathbb{R}^{(n_K+m_2)\times(n_K+l_2)}$ such that* (3.13) *holds for $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ with*

$$\mathbf{H}_{(i_0,\ldots,i_M)} = \begin{bmatrix} -\mathbf{X}_{(i_0,\ldots,i_M)_+}^{-1} & \widehat{\mathbf{A}}_{i_M} & \widehat{\mathbf{B}}_{1,i_M} & \mathbf{0} \\ \widehat{\mathbf{A}}_{i_M}^{\mathrm{T}} & -\mathbf{X}_{(i_0,\ldots,i_M)_-} & \mathbf{0} & \widehat{\mathbf{C}}_{1,i_M}^{\mathrm{T}} \\ \widehat{\mathbf{B}}_{1,i_M}^{\mathrm{T}} & \mathbf{0} & -\gamma\mathbf{I} & \mathbf{D}_{11,i_M}^{\mathrm{T}} \\ \mathbf{0} & \widehat{\mathbf{C}}_{1,i_M} & \mathbf{D}_{11,i_M} & -\gamma\mathbf{I} \end{bmatrix}, \tag{3.19}$$

where the matrices $\mathbf{X}_{\boldsymbol{j}}$ are reconstructed via (3.16) *for $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$; in particular, one may take $L = M$.*

Example 4. Let

$$\boldsymbol{\mathcal{T}} = \{(0.3, 0, 0, 1, 1, 0, 1, 0), (3, 0.5, 1, 1, 1, 0, 1, 0)\};$$

let $\boldsymbol{\Theta}$ be as in Example 1. Then the controlled system $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ has the representation

$$x(t+1) = 0.3x(t), \quad z(t) = x(t) + u(t), \quad y(t) = x(t)$$

in mode 1, and

$$x(t+1) = 3x(t) + 0.5w(t) + u(t), \quad z(t) = x(t) + u(t), \quad y(t) = x(t)$$

in mode 2. With $M = 0$, the system of linear matrix inequalities (3.18) is feasible for any $\gamma > 1$; it is easy to see that, if $\gamma = 1$, then inequalities (3.18) are not feasible for any $M \geq 0$. Solving the semidefinite program of minimizing γ subject to (3.18), and applying Corollary 3.11, we obtain a mode-dependent controller $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}_0)$ with

$$\mathbf{K}_1 = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{K}_2 = \begin{bmatrix} 0 & 0 \\ 0 & -3 \end{bmatrix}.$$

The resulting controller $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}_0)$ has the representation

$$u(t) = \begin{cases} -y(t) & \text{if } \theta(t) = 1, \\ -3y(t) & \text{if } \theta(t) = 2. \end{cases}$$

This controller is *optimal* in the sense that it achieves any uniform disturbance attenuation level greater than one, and that no finite-path-dependent linear dynamic output feedback controller achieves the uniform disturbance attenuation level equal to one. (The optimal controller is static in this case because $\mathbf{C}_{2,1} = \mathbf{C}_{2,2} = 1$ and $\mathbf{D}_{21,1} = \mathbf{D}_{21,2} = 0$ lead to perfect observation of the state.) The controller, however, is not optimal under the notion of path-by-path disturbance attenuation—see Example 7.

4. Control of Markovian jump linear systems. In Markovian jump linear systems, the switching sequence is modeled as a finite-state homogeneous Markov chain. Let $\boldsymbol{\mathcal{G}}$ be as in (3.1). Let $p = (p_i) \in \mathbb{R}^{1\times N}$ be a row vector whose entries are nonnegative and sum to one; let $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{N\times N}$ be a (row) stochastic matrix so that each row of $\mathbf{P}$ has nonnegative entries that sum to one. Then the discrete-time *Markovian jump linear system*, defined by the triple $(\boldsymbol{\mathcal{G}}, \mathbf{P}, p)$, has the representation (3.2). Here, the switching sequence $\boldsymbol{\theta}$ is a realization of the Markov chain defined by the pair $(\mathbf{P}, p)$, where $\mathbf{P}$ is the transition probability matrix and p the initial distribution. The state $\theta(t)$ of the chain $(\mathbf{P}, p)$ at time t defines the mode of $(\boldsymbol{\mathcal{G}}, \mathbf{P}, p)$ at time t; the distribution of the mode at time t is given by $p\mathbf{P}^t$. As in the previous section, let $\boldsymbol{\Omega}$ be the space of all infinite sequences in $\{1, \dots, N\}$. Let P be the unique consistent probability measure [39] on $\boldsymbol{\Omega}$ such that

$$P\{\, \theta(t+1) = j \mid \theta(t) = i \,\} = p_{ij}, \quad P\{\, \theta(0) = i \,\} = p_i$$

for all i, j, and t.

DEFINITION 4.1. *The system* $(\boldsymbol{\mathcal{G}}, \mathbf{P}, p)$ *is said to be* almost surely uniformly (exponentially) stable *if there exists a set* $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ *with* $P(\boldsymbol{\Theta}) = 1$ *such that the system* $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ *is uniformly exponentially stable.*

DEFINITION 4.2. *The system* $(\boldsymbol{\mathcal{G}}, \mathbf{P}, p)$ *is said to be* almost surely uniformly (strictly) contractive *if there exists a set* $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ *with* $P(\boldsymbol{\Theta}) = 1$ *such that the system* $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ *is uniformly strictly contractive.*

A switching sequence $\boldsymbol{\theta}$ in $\{1, \dots, N\}$ is said to be *admissible* with respect to $(\mathbf{P}, p)$ if $p_{\theta(0)} > 0$ and $p_{\theta(t)\theta(t+1)} > 0$ for $t \geq 0$. If we define

$$\boldsymbol{\Theta}(\mathbf{P}, p) = \{\boldsymbol{\theta} \in \boldsymbol{\Omega} \colon \boldsymbol{\theta} \text{ is admissible with respect to } (\mathbf{P}, p)\}$$

and let

$$\boldsymbol{\mathcal{M}}_L(\mathbf{P}, p) = \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta}(\mathbf{P}, p)), \quad \boldsymbol{\mathcal{M}}_L^-(\mathbf{P}, p) = \boldsymbol{\mathcal{M}}_L^-(\boldsymbol{\Theta}(\mathbf{P}, p))$$

for nonnegative integers L, then we have $P(\Theta(\mathbf{P}, p)) = 1$; on the other hand, whenever $(i_0, \dots, i_L) \in \boldsymbol{\mathcal{M}}_L(\mathbf{P}, p)$, we have that $P\{\boldsymbol{\theta} \in \boldsymbol{\Omega} \colon (i_0, \dots, i_L) \in \boldsymbol{\mathcal{L}}_L(\{\boldsymbol{\theta}\})\} > 0$ so that $\boldsymbol{\mathcal{M}}_L(\mathbf{P}, p) \subset \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})$ whenever $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ and $P(\boldsymbol{\Theta}) = 1$.

Example 5. Let $N = 3$, and let $(\mathbf{P}, p)$ be a Markov chain with

$$\mathbf{1}(\mathbf{P}) = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

If $\mathbf{1}(p) = [1\ 0\ 0]$, then

$$\boldsymbol{\mathcal{L}}_0(\mathbf{P},p) = \boldsymbol{\mathcal{M}}_0(\mathbf{P},p) = \{1,2,3\};$$
$$\boldsymbol{\mathcal{L}}_1(\mathbf{P},p) = \{01,11,12,13,23,33\}, \quad \boldsymbol{\mathcal{M}}_1(\mathbf{P},p) = \{11,12,13,23,33\},$$

and so on. However, if $\mathbf{1}(p) = [0\ 1\ 1]$, then

$$\boldsymbol{\mathcal{L}}_0(\mathbf{P},p) = \boldsymbol{\mathcal{M}}_0(\mathbf{P},p) = \{2,3\};$$
$$\boldsymbol{\mathcal{L}}_1(\mathbf{P},p) = \{02,03,23,33\}, \quad \boldsymbol{\mathcal{M}}_1(\mathbf{P},p) = \{02,23,33\},$$

and so on.

THEOREM 4.3. *Let $\boldsymbol{\mathcal{G}}$ be as in* (3.1)*; let* $(\mathbf{P},p)$ *be a Markov chain. The system* $(\boldsymbol{\mathcal{G}},\mathbf{P},p)$ *is almost surely uniformly exponentially stable and almost surely uniformly strictly contractive if and only if there exist a nonnegative integer M and an indexed family* $\{\mathbf{X}_{\boldsymbol{j}} : \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\mathbf{P},p)\}$ *of symmetric positive definite matrices* $\mathbf{X}_{\boldsymbol{j}} \in \mathbb{R}^{n\times n}$ *such that* (3.4) *holds for all M-paths* $\boldsymbol{i} = (i_0,\dots,i_M) \in \boldsymbol{\mathcal{M}}_M(\mathbf{P},p)$.

Proof. The result is immediate from Theorem 3.3: sufficiency follows from $P(\Theta(\mathbf{P},p)) = 1$, and necessity from the fact that $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ and $P(\boldsymbol{\Theta}) = 1$ implies $\boldsymbol{\mathcal{M}}_L(\mathbf{P},p) \subset \boldsymbol{\mathcal{M}}_L(\boldsymbol{\Theta})$. □

Remark 9. Define

$$n_i(0) = 1, \quad n_i(L+1) = \sum\nolimits_{\{j\,:\,p_{ij}>0\}} n_j(L)$$

for $i \in \{1,\dots,N\}$ and for $L = 0, 1, \dots$. Given a Markov chain $(\mathbf{P},p)$, let

$$S(p) = \{j : p_j > 0\}, \quad T(\mathbf{P},p) = \big\{j : p_i p_{ij}^{(k)} > 0 \text{ for some } (i,k)\big\},$$

where $\mathbf{P}^k = \big(p_{ij}^{(k)}\big)$ is the k-step transition probability matrix [24]. Then, in Theorem 4.3, the number of linear matrix inequalities (3.4) to solve simultaneously, with a fixed M, is equal to the cardinality of $\boldsymbol{\mathcal{M}}_M(\mathbf{P},p)$, which is precisely given by

$$\sum\nolimits_{j\in S(p)\setminus T(\mathbf{P},p)} \sum\nolimits_{k=0}^{M-1} n_j(k) + \sum\nolimits_{j\in S(p)\cup T(\mathbf{P},p)} n_j(M).$$

In particular, if $\mathbf{P}$ is irreducible (i.e., if the directed graph of $\mathbf{P}$ is strongly connected—see, e.g., [25]), then the cardinality of $\boldsymbol{\mathcal{M}}_M(\mathbf{P},p)$ is equal to $\sum_{j=1}^N n_j(M)$.

Remark 10. Theorem 4.3 implies that the Markovian jump linear system $(\boldsymbol{\mathcal{G}},\mathbf{P},p)$ is almost surely uniformly stable and contractive if and only if the switched linear system $(\boldsymbol{\mathcal{G}},\boldsymbol{\Theta}(\mathbf{P},p))$ is uniformly stable and contractive. Therefore, Markovian jump linear systems can be treated as if they are switched linear systems. Moreover, since the set $\boldsymbol{\Theta}(\mathbf{P},p)$ depends only on the sparsity patterns of $\mathbf{P}$ and p, the almost sure uniform stability and contractiveness of $(\boldsymbol{\mathcal{G}},\mathbf{P},p)$ is robust against sparsity pattern–preserving deviations from $\mathbf{P}$ and p.

Remark 11. When only the almost sure uniform stability is considered, it is immediate from Corollary 3.4 that it suffices to consider the "irreducible parts" of the Markov chain $(\mathbf{P},p)$; see [30].

Let $\boldsymbol{\mathcal{T}}$ be as in (3.6), and let $(\mathbf{P},p)$ be a Markov chain. Then the triple $(\boldsymbol{\mathcal{T}},\mathbf{P},p)$ defines the *controlled Markovian jump linear system* described by (3.7), where $\boldsymbol{\theta}$ is a realization of $(\mathbf{P},p)$. As in the previous section, we make the standard assumption that *the state $\theta(t)$ of the chain* $(\mathbf{P},p)$ *is perfectly observed at each time instant t*; we consider all finite-path-dependent controllers.

Fix a nonnegative integer L. Let

$$\boldsymbol{\mathcal{K}} = \{(\mathbf{A}_{K,\boldsymbol{i}}, \mathbf{B}_{K,\boldsymbol{i}}, \mathbf{C}_{K,\boldsymbol{i}}, \mathbf{D}_{K,\boldsymbol{i}}) \colon \boldsymbol{i} \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}(\mathbf{P}, p))\}.$$

Then the pair $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}(\mathbf{P}, p)_L)$ defines an L-path-dependent controller, whose representation is given by (3.9). Label the L-paths in $\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}(\mathbf{P}, p))$ in dictionary order from 1 to N_L, where N_L is the cardinality of $\boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}(\mathbf{P}, p))$. Define $\mathbf{P}^{(L)} = (q_{ij}) \in \mathbb{R}^{N_L \times N_L}$ as follows: Whenever $(i_0, \ldots, i_L)$ and $(j_0, \ldots, j_L)$ are L-paths labeled i and j, respectively, set $q_{ij} = p_{i_L j_L}$ if $(i_0, \ldots, i_L)_+ = (j_0, \ldots, j_L)_-$; otherwise, set $q_{ij} = 0$. Also, define a row vector $p^{(L)} = (q_i) \in \mathbb{R}^{N_L}$ as follows: Whenever $(i_0, \ldots, i_L)$ is an L-path labeled i, set $q_i = p_{i_L}$ if $(i_0, \ldots, i_L)_- = (0, \ldots, 0)$; otherwise, set $q_i = 0$. Then the pair $(\mathbf{P}^{(L)}, p^{(L)})$ defines the *L-path Markov chain* generated by $(\mathbf{P}, p)$. Consequently, the closed-loop system, given by the triple $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}}, \mathbf{P}^{(L)}, p^{(L)})$ with

$$\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}} = \left\{ \left(\widetilde{\mathbf{A}}_{\boldsymbol{i}}, \widetilde{\mathbf{B}}_{\boldsymbol{i}}, \widetilde{\mathbf{C}}_{\boldsymbol{i}}, \widetilde{\mathbf{D}}_{\boldsymbol{i}} \right) \colon \boldsymbol{i} \in \boldsymbol{\mathcal{L}}_L(\boldsymbol{\Theta}(\mathbf{P}, p)) \right\},$$

is a Markovian jump linear system whose representation is of the form (3.11) for each realization $\boldsymbol{\theta}_L$ of $(\mathbf{P}^{(L)}, p^{(L)})$.

DEFINITION 4.4. *Let $\gamma > 0$. The system $(\boldsymbol{\mathcal{G}}, \mathbf{P}, p)$ is said to* satisfy almost sure uniform disturbance attenuation level γ *if there exists a set $\boldsymbol{\Theta} \subset \boldsymbol{\Omega}$ with $P(\boldsymbol{\Theta}) = 1$ such that the system $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ satisfies uniform disturbance attenuation level γ.*

DEFINITION 4.5. *Let $\gamma > 0$. The controller $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}(\mathbf{P}, p)_L)$ is said to be a* γ-admissible (L-path-dependent) synthesis (of order n_K), *or to* achieve almost sure uniform disturbance attenuation level γ, *for the system $(\boldsymbol{\mathcal{T}}, \mathbf{P}, p)$ if the closed-loop system $(\boldsymbol{\mathcal{T}}_{\boldsymbol{\mathcal{K}}}, \mathbf{P}^{(L)}, p^{(L)})$ satisfies almost sure uniform disturbance attenuation level γ.*

THEOREM 4.6. *Let $\boldsymbol{\mathcal{T}}$ be as in* (3.6)*; let $(\mathbf{P}, p)$ be a Markov chain. Let $\gamma > 0$. Suppose that $n_K \geq n$. There exists a γ-admissible finite-path-dependent synthesis of order n_K for the system $(\boldsymbol{\mathcal{T}}, \mathbf{P}, p)$ if and only if there exist a nonnegative integer M and an indexed family $\{(\mathbf{R}_{\boldsymbol{j}}, \mathbf{S}_{\boldsymbol{j}}) \colon \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\mathbf{P}, p)\}$ of pairs of symmetric positive definite matrices $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}} \in \mathbb{R}^{n \times n}$ such that* (3.18) *holds for all M-paths $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\mathbf{P}, p)$. Moreover, if* (3.18) *holds for all $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\mathbf{P}, p)$, then there exist a nonnegative integer $L \leq M$ and matrices $\mathbf{K}_{(i_{M-L}, \ldots, i_M)} \in \mathbb{R}^{(n_K + m_2) \times (n_K + l_2)}$ such that* (3.13)*, with* (3.19)*, holds for $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\mathbf{P}, p)$, where the matrices $\mathbf{X}_{\boldsymbol{j}}$ are given by* (3.16) *for $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\mathbf{P}, p)$; in particular, one may take $L = M$.*

Proof. The result immediately follows from Corollary 3.11. ∎

Remark 12. The existence of finite-path-dependent controller syntheses achieving an almost sure uniform disturbance attenuation level for the Markovian jump linear system $(\boldsymbol{\mathcal{T}}, \mathbf{P}, p)$ is robust against sparsity pattern–preserving deviations from $\mathbf{P}$ and p. The conservatism associated with this robustness property can be reduced via path-by-path disturbance attenuation—see the examples in the next section.

Example 6. Let the Markov chain $(\mathbf{P}, p)$ have

$$\mathbf{1}(\mathbf{P}) = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{1}(p) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}. \tag{4.1}$$

Then we have

$$\boldsymbol{\mathcal{M}}_0(\mathbf{P}, p) = \{1, 2, 3\}, \quad \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p) = \{12, 13, 23, 31\},$$
$$\boldsymbol{\mathcal{M}}_2(\mathbf{P}, p) = \{123, 131, 231, 312, 313\},$$

and so on. Let

$$\boldsymbol{\mathcal{T}} = \{(0.5,0,0,0,0,0,0,0),(3,0,0,0,0,0,0,0),(2,1,1,1,1,1,1,1)\}$$

so that the controlled Markovian jump linear system $(\boldsymbol{\mathcal{T}}, \mathbf{P}, p)$ has the representation

$$\begin{aligned} x(t+1) &= 0.5x(t), \quad z(t) = 0, \quad y(t) = 0; \\ x(t+1) &= 3x(t), \quad z(t) = 0, \quad y(t) = 0; \end{aligned}$$

or

$$\begin{aligned} x(t+1) &= 2x(t) + w(t) + u(t), \\ z(t) &= x(t) + w(t) + u(t), \\ y(t) &= x(t) + w(t), \end{aligned}$$

depending on whether the mode at time t is 1, 2, or 3, respectively. With $M = 0$, the system of linear matrix inequalities (3.18), over $i_0 \in \boldsymbol{\mathcal{M}}_0(\mathbf{P}, p)$, is not feasible for any $\gamma > 0$. However, with $M = 1$, the semidefinite program of minimizing γ subject to (3.18) over $(i_0, i_1) \in \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)$ leads to $\gamma = 0.834$; the same is true for all $M > 1$. Setting $L = M = 1$ in (3.13) with (3.19), we obtain

$$\mathbf{K}_{(1,2)} = \mathbf{K}_{(3,1)} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{K}_{(1,3)} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{K}_{(2,3)} = \begin{bmatrix} 0 & 0.0000 \\ 0 & -1.5555 \end{bmatrix}. \tag{4.2}$$

The resulting one-path-dependent controller is optimal (up to the third digit below the decimal point) in the sense that no finite-path-dependent linear dynamic output feedback controller achieves the disturbance attenuation level 0.833. This controller is applied as follows: For $t = 0$, use $\mathbf{K}_{(3,1)}$ if $\theta(0) = 1$, use $\mathbf{K}_{(1,2)}$ if $\theta(0) = 2$, and use either $\mathbf{K}_{(1,3)}$ or $\mathbf{K}_{(2,3)}$ if $\theta(0) = 3$; for $t > 0$, use $\mathbf{K}_{(\theta(t-1),\theta(t))}$.

It turns out that setting $L = 0$ and $M = 1$ in (3.13) with (3.19) also works and results in a mode-dependent optimal controller with

$$\mathbf{K}_1 = \mathbf{K}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{K}_3 = \begin{bmatrix} 0 & 0.0000 \\ 0 & -1.5555 \end{bmatrix}. \tag{4.3}$$

This mode-dependent controller, however, is not optimal under a path-by-path disturbance attenuation criterion—see Example 8.

5. Path-by-path disturbance attenuation. In this section, we formulate a refined disturbance attenuation problem for switched linear systems. The result holds for Markovian jump linear systems as well. We introduce the notion of path-by-path disturbance attenuation that improves upon the uniform disturbance attenuation performance presented in previous sections. It turns out that, under the notion of path-by-path disturbance attenuation, finite-path-dependent controllers can outperform mode-dependent ones—see Examples 7 and 8. We shall use the same notation as in section 3.

DEFINITION 5.1. *If there exist a nonnegative integer M, a positive integer n_K, an indexed family $\Gamma = \{\gamma_{\boldsymbol{i}} : \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})\}$ of positive numbers $\gamma_{\boldsymbol{i}}$, and an indexed family $\{(\mathbf{R}_{\boldsymbol{j}}, \mathbf{S}_{\boldsymbol{j}}) : \boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})\}$ of pairs of symmetric positive definite matrices $\mathbf{R}_{\boldsymbol{j}}$, $\mathbf{S}_{\boldsymbol{j}} \in \mathbb{R}^{n \times n}$ such that*

$$\text{(5.1a)}\quad \begin{bmatrix}\mathbf{N}_{F,i_M} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^{\mathrm{T}} \begin{bmatrix}\mathbf{A}_{i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{A}_{i_M}^{\mathrm{T}} - \mathbf{R}_{\boldsymbol{i}_+} & \mathbf{A}_{i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{C}_{1,i_M}^{\mathrm{T}} & \mathbf{B}_{1,i_M}\\ \mathbf{C}_{1,i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{A}_{i_M}^{\mathrm{T}} & \mathbf{C}_{1,i_M}\mathbf{R}_{\boldsymbol{i}_-}\mathbf{C}_{1,i_M}^{\mathrm{T}} - \gamma_{\boldsymbol{i}}\mathbf{I} & \mathbf{D}_{11,i_M}\\ \mathbf{B}_{1,i_M}^{\mathrm{T}} & \mathbf{D}_{11,i_M}^{\mathrm{T}} & -\gamma_{\boldsymbol{i}}\mathbf{I}\end{bmatrix} \times \begin{bmatrix}\mathbf{N}_{F,i_M} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} < \mathbf{0},$$

$$\text{(5.1b)}\quad \begin{bmatrix}\mathbf{N}_{G,i_M} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}^{\mathrm{T}} \begin{bmatrix}\mathbf{A}_{i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{A}_{i_M} - \mathbf{S}_{\boldsymbol{i}_-} & \mathbf{A}_{i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{B}_{1,i_M} & \mathbf{C}_{1,i_M}^{\mathrm{T}}\\ \mathbf{B}_{1,i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{A}_{i_M} & \mathbf{B}_{1,i_M}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{i}_+}\mathbf{B}_{1,i_M} - \gamma_{\boldsymbol{i}}\mathbf{I} & \mathbf{D}_{11,i_M}^{\mathrm{T}}\\ \mathbf{C}_{1,i_M} & \mathbf{D}_{11,i_M} & -\gamma_{\boldsymbol{i}}\mathbf{I}\end{bmatrix} \times \begin{bmatrix}\mathbf{N}_{G,i_M} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix} < \mathbf{0},$$

$$\text{(5.1c)}\quad \begin{bmatrix}\mathbf{R}_{\boldsymbol{i}_-} & \mathbf{I}\\ \mathbf{I} & \mathbf{S}_{\boldsymbol{i}_-}\end{bmatrix} \geq \mathbf{0} \quad \left(\textit{and, if } n_K < n, \quad \operatorname{rank}\begin{bmatrix}\mathbf{R}_{\boldsymbol{i}_-} & \mathbf{I}\\ \mathbf{I} & \mathbf{S}_{\boldsymbol{i}_-}\end{bmatrix} \leq n + n_K\right)$$

for all M-paths $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, then an L-path-dependent controller $(\boldsymbol{\mathcal{K}}, \boldsymbol{\Theta}_L)$ (of order n_K) is said to achieve path-by-path disturbance attenuation level Γ *for the system $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ whenever it is constructed in the following manner: Choose a nonnegative integer $L \leq M$, symmetric positive definite matrices $\mathbf{X}_{(i_0,\ldots,i_M)_-} \in \mathbb{R}^{(n+n_K)\times(n+n_K)}$, and matrices $\mathbf{K}_{(i_{M-L},\ldots,i_M)} \in \mathbb{R}^{(n_K+m_2)\times(n_K+l_2)}$ such that* (3.13) *holds for all M-paths $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ with*

$$\text{(5.2)}\quad \mathbf{H}_{(i_0,\ldots,i_M)} = \begin{bmatrix}-\mathbf{X}^{-1}_{(i_0,\ldots,i_M)_+} & \widehat{\mathbf{A}}_{i_M} & \widehat{\mathbf{B}}_{1,i_M} & \mathbf{0}\\ \widehat{\mathbf{A}}_{i_M}^{\mathrm{T}} & -\mathbf{X}_{(i_0,\ldots,i_M)_-} & \mathbf{0} & \widehat{\mathbf{C}}_{1,i_M}^{\mathrm{T}}\\ \widehat{\mathbf{B}}_{1,i_M}^{\mathrm{T}} & \mathbf{0} & -\gamma_{(i_0,\ldots,i_M)}\mathbf{I} & \mathbf{D}_{11,i_M}^{\mathrm{T}}\\ \mathbf{0} & \widehat{\mathbf{C}}_{1,i_M} & \mathbf{D}_{11,i_M} & -\gamma_{(i_0,\ldots,i_M)}\mathbf{I}\end{bmatrix}.$$

The only difference between the synthesis conditions for the uniform disturbance attenuation and path-by-path disturbance attenuation is that, while a single variable γ is used in (3.18) and (3.19) for the former, multiple variables $\gamma_{(i_0,\ldots,i_M)}$ are used in (5.1) and (5.2) for the latter. Suppose that there exists a finite-path-dependent controller synthesis of order $n_K = n$ for $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ that achieves a uniform disturbance attenuation level $\gamma_0 > 0$, and that the system of inequalities (3.18) is feasible with $M = M_0$. Then one can improve the closed-loop performance further by obtaining a *path-by-path optimal* controller synthesis through semidefinite programming as follows.

ALGORITHM 5.2.

Step 0. *Set $n_K = n$. Choose $\gamma \geq \gamma_0$, $M \geq M_0$, and $\lambda_{\boldsymbol{i}} > 0$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$.*

Step 1. *Minimize $\sum_{\boldsymbol{i}\in\boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})} \lambda_{\boldsymbol{i}}\gamma_{\boldsymbol{i}}$ subject to* (5.1)*, with the additional constraint $\gamma_{\boldsymbol{i}} \leq \gamma$, for $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$.*

Step 2. *Reconstruct $\mathbf{X}_{\boldsymbol{j}}$ from $(\mathbf{R}_{\boldsymbol{j}}, \mathbf{S}_{\boldsymbol{j}})$ via* (3.16) *for all $\boldsymbol{j} \in \boldsymbol{\mathcal{M}}_M^-(\boldsymbol{\Theta})$.*

Step 3. *Solve* (3.13) *with* (5.2) *for all $(i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, and obtain a nonnegative integer $L \leq M$ and matrices $\mathbf{K}_{(i_{M-L},\ldots,i_M)}$.*

The path-by-path optimal controller $(\mathbf{K}, \boldsymbol{\Theta}_L)$ resulting from Algorithm 5.2 shall be said to be *M-path Pareto optimal* because of the following property: If $(\mathbf{K}, \boldsymbol{\Theta}_L)$ achieves path-by-path disturbance attenuation level $\{\gamma_{\boldsymbol{i}} \colon \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})\}$, then no M-path-dependent controller can achieve a disturbance attenuation level $\{\tilde{\gamma}_{\boldsymbol{i}} \colon \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})\}$ such that $\tilde{\gamma}_{\boldsymbol{i}} \leq \gamma_{\boldsymbol{i}}$ for all $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$ and such that $\tilde{\gamma}_{\boldsymbol{i}} < \gamma_{\boldsymbol{i}}$ for some

$\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$. In general, different sets of weights $\lambda_{\boldsymbol{i}}$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, result in different interpretations of optimality—see Examples 7 and 8—and hence different Pareto optimal path-by-path disturbance attenuation levels—see Example 9.

Example 7. Let us revisit the controlled switched linear system $(\boldsymbol{\mathcal{T}}, \boldsymbol{\Theta})$ considered in Example 4. In this example, we set $\gamma = 10^3$ and $\lambda_{\boldsymbol{i}} = 1$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_M(\boldsymbol{\Theta})$, for Algorithm 5.2. Running Algorithm 5.2 with $M = 0$, it turns out that the mode-dependent controller obtained in Example 4 achieves any path-by-path disturbance attenuation level $\{\gamma_i\}$ satisfying $\gamma_1 > 0$ and $\gamma_2 > 1$. In particular, this mode-dependent controller is γ_1-admissible for $(\boldsymbol{\mathcal{T}}, (1, 1, \dots))$ and γ_2-admissible for $(\boldsymbol{\mathcal{T}}, \{(1, 2, 1, 2, \dots), (2, 2, \dots)\})$ whenever $\gamma_1 > 0$ and $\gamma_2 > 1$. Running Algorithm 5.2 with $M = 1$ leads to $\gamma_{(1,1)} = \gamma_{(2,1)} = 0$ and $\gamma_{(1,2)} = 0.795$, but $\gamma_{(2,2)} = 1.16$, which is greater than one.

However, running Algorithm 5.2 with $M = 2$ leads to a one-path-dependent controller $(L = 1)$ with

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

This controller is two-path Pareto optimal and achieves any path-by-path disturbance attenuation level $\{\gamma_{\boldsymbol{i}} : \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_2(\boldsymbol{\Theta})\}$ such that $\gamma_{\boldsymbol{i}} > 1$ for $\boldsymbol{i} = (2, 2, 2)$, and $\gamma_{\boldsymbol{i}} > 0$ otherwise. In particular, this one-path-dependent controller is γ_1-admissible for $(\boldsymbol{\mathcal{T}}, \{(1, 1, \dots), (1, 2, 1, 2, \dots)\})$ and γ_2-admissible for $(\boldsymbol{\mathcal{T}}, (2, 2, \dots))$ whenever $\gamma_1 > 0$ and $\gamma_2 > 1$. Clearly, this is an improvement over the path-by-path performance of the mode-dependent controller obtained in Example 4 with $M = 0$. In fact, it is not difficult to see that, in this example, no mode-dependent controller can achieve a path-by-path disturbance attenuation level $\{\gamma_{\boldsymbol{i}} : \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_2(\boldsymbol{\Theta})\}$ such that, say, $\gamma_{\boldsymbol{i}} = 0.01$ for $\boldsymbol{i} \neq (2, 2, 2)$.

Example 8. This example revisits the controlled Markovian jump linear system $(\boldsymbol{\mathcal{T}}, \mathbf{P}, p)$ considered in Example 6. Since $\mathbf{1}(\mathbf{P}^5) = \mathbf{1}$, the Markov chain $(\mathbf{P}, p)$ is irreducible and aperiodic [25]. Let $\pi = [\pi_1 \ \pi_2 \ \pi_3]$ be the unique steady-state distribution of the chain. For each nonnegative integer M, let

$$\lambda_{(i_0,\dots,i_M)} = \pi_{i_0} p_{i_0 i_1} \cdots p_{i_{M-1} i_M}, \quad (i_0, \dots, i_M) \in \boldsymbol{\mathcal{M}}_M(\mathbf{P}, p). \tag{5.3}$$

Then Algorithm 5.2 minimizes the "average steady-state disturbance attenuation level over M-paths," where each M-path $(i_0, \dots, i_M)$ is interpreted as the mode i_M preceded by the path $(i_0, \dots, i_{M-1})$. With $\gamma = \infty$ and $M = 1$, we obtain the one-path-dependent controller $(L = 1)$ given by (4.2). This controller achieves the path-by-path disturbance attenuation level $\{\gamma_{\boldsymbol{i}} : \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)\}$ where, e.g., $\gamma_{(1,2)} = \gamma_{(1,3)} = \gamma_{(3,1)} = 0.001$ and $\gamma_{(2,3)} = 0.834$; the weighted sum of these disturbance attenuation levels $\gamma_{\boldsymbol{i}}$ is $0.001 + 0.833\pi_2$. On the other hand, the mode-dependent controller given by (4.3) achieves the path-by-path disturbance attenuation level $\{\tilde{\gamma}_{\boldsymbol{i}} : \boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)\}$ where, e.g., $\tilde{\gamma}_{(1,2)} = \tilde{\gamma}_{(3,1)} = 0.001$ and $\tilde{\gamma}_{(1,3)} = \tilde{\gamma}_{(2,3)} = 0.834$. Clearly, the one-path-dependent controller performs better than the mode-dependent controller in terms of average steady-state performance. In fact, it is not difficult to see that, for some chain $(\mathbf{P}, p)$ with (4.1), there does not exist a mode-dependent controller that achieves the average steady-state performance level $0.001 + 0.833\pi_2$.

Example 9. We are to balance the Pendubot [40], a two-link planar robot with revolute joints and actuation at the shoulder, subject to random but bounded delays in the feedback loop from the relative angular position sensor to the actuator. We

use the following linearized model (of order 4) borrowed from [43]:

$$\begin{aligned} x(t+1) &= \mathbf{A}x(t) + \mathbf{B}_1 w(t) + \mathbf{B}_2 u(t), \\ z(t) &= \mathbf{C}_1 x(t) + \mathbf{D}_{11} w(t) + \mathbf{D}_{12} u(t), \\ y(t) &= \mathbf{C}_2 x(t) + \mathbf{D}_{21} w(t); \end{aligned}$$

$$\mathbf{A} = \begin{bmatrix} 0.9992 & 0.0050 & 0.0003 & 0.0000 \\ -0.3369 & 0.9992 & 0.1242 & 0.0003 \\ 0.0008 & 0.0000 & 1.0007 & 0.0050 \\ 0.3263 & 0.0008 & 0.2786 & 1.0007 \end{bmatrix}, \quad \mathbf{B}_1 = \begin{bmatrix} -0.0001 \\ -0.0232 \\ 0.0012 \\ 0.4742 \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} 0.0006 \\ 0.2243 \\ -0.0001 \\ -0.0232 \end{bmatrix},$$

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{D}_{11} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{D}_{12} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{C}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{D}_{21} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The random delays in the feedback loop are modeled as a Markov chain such that the amount of delay at time t is given by $\theta(t) - 1$ if $\theta(t)$ is the state of the chain at time t. If the maximum possible amount of delay is $N - 1$, then the augmented state

$$\hat{x}(t) = \begin{bmatrix} x(t)^{\mathrm{T}} & y(t-1)^{\mathrm{T}} & \cdots & y(t-N+1)^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$$

and delayed measurement $\hat{y}(t)$ yield the Markovian jump linear system (of order 14)

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

where

$$\mathbf{A}_i = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_{1,i} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{D}_{21} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_{2,i} = \begin{bmatrix} \mathbf{B}_2 \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix},$$

$$\mathbf{C}_{1,i} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \quad \mathbf{D}_{11,i} = \mathbf{D}_{11}, \quad \mathbf{D}_{12,i} = \mathbf{D}_{12}$$

for $i = 1, \ldots, N$, and

$$\begin{bmatrix} \mathbf{C}_{2,1} \\ \mathbf{C}_{2,2} \\ \vdots \\ \mathbf{C}_{2,N} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} \end{bmatrix}, \quad \begin{bmatrix} \mathbf{D}_{21,1} \\ \mathbf{D}_{21,2} \\ \vdots \\ \mathbf{D}_{21,N} \end{bmatrix} = \begin{bmatrix} \mathbf{D}_{21} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

The measurements are time-stamped, so the controller perfectly observes $\theta(t)$ at each time t. Let $N = 6$. Let the transition probability matrix $\mathbf{P}$ of the Markov chain be

$$\mathbf{P} = \begin{bmatrix} 0.3 & 0.7 & 0 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.6 & 0 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.4 & 0 & 0 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 & 0 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.3 & 0.3 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.3 & 0.3 \end{bmatrix},$$

where the fact that the controller can always use the most recent measurement is accounted for [42]; the initial distribution p is assumed to be equal to the unique steady-state distribution

$$\pi = \begin{bmatrix} 0.2142 & 0.2999 & 0.2699 & 0.1440 & 0.0504 & 0.0216 \end{bmatrix}.$$

The semidefinite program minimizing the uniform disturbance attenuation level γ subject to (3.18) over $\boldsymbol{i} = (i_0, \ldots, i_M) \in \boldsymbol{\mathcal{M}}_M(\mathbf{P}, p)$ yields approximately $\gamma = 11.1$ for $M = 0$ and $\gamma = 10.6$ for $M \geq 1$. Now running Algorithm 5.2 with $\gamma = 100$, $M = 1$, and $\lambda_{\boldsymbol{i}}$ as in (5.3) yields $\gamma_{\boldsymbol{i}}$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)$, such that $\sum_{\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P},p)} \lambda_{\boldsymbol{i}} \gamma_{\boldsymbol{i}} = 9.96$, where, if $\boldsymbol{\Lambda} = (\lambda_{(i,j)})$ and $\boldsymbol{\Gamma} = (\gamma_{(i,j)})$, then

$$\boldsymbol{\Lambda} = \begin{bmatrix} 0.0643 & 0.1499 & 0 & 0 & 0 & 0 \\ 0.0600 & 0.0600 & 0.1799 & 0 & 0 & 0 \\ 0.0540 & 0.0540 & 0.0540 & 0.1080 & 0 & 0 \\ 0.0228 & 0.0228 & 0.0228 & 0.0228 & 0.0228 & 0 \\ 0.0050 & 0.0050 & 0.0050 & 0.0050 & 0.0151 & 0.0151 \\ 0.0022 & 0.0022 & 0.0022 & 0.0022 & 0.0065 & 0.0065 \end{bmatrix},$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} 9.52 & 9.56 & 0 & 0 & 0 & 0 \\ 9.48 & 9.52 & 9.39 & 0 & 0 & 0 \\ 9.61 & 9.65 & 9.53 & 9.24 & 0 & 0 \\ 10.1 & 9.99 & 9.87 & 9.57 & 10.1 & 0 \\ 17.7 & 17.7 & 17.6 & 17.1 & 10.4 & 10.9 \\ 27.9 & 28.0 & 27.7 & 27.2 & 18.9 & 11.5 \end{bmatrix}.$$

On the other hand, running Algorithm 5.2 with $\gamma = 100$, $M = 1$, and $\lambda_{\boldsymbol{i}} = 1$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)$, yields the disturbance attenuation levels $\tilde{\gamma}_{\boldsymbol{i}}$, $\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)$, such that $\frac{1}{N_1} \sum_{\boldsymbol{i} \in \boldsymbol{\mathcal{M}}_1(\mathbf{P},p)} \tilde{\gamma}_{\boldsymbol{i}} = 10.3$, where N_1 is the cardinality of $\boldsymbol{\mathcal{M}}_1(\mathbf{P}, p)$ and equals 26. If $\widetilde{\boldsymbol{\Gamma}} = (\tilde{\gamma}_{(i,j)})$, then

$$\widetilde{\boldsymbol{\Gamma}} = \begin{bmatrix} 13.3 & 16.9 & 0 & 0 & 0 & 0 \\ 9.90 & 12.9 & 17.7 & 0 & 0 & 0 \\ 7.05 & 8.86 & 12.4 & 16.9 & 0 & 0 \\ 5.74 & 6.75 & 8.84 & 12.0 & 16.0 & 0 \\ 5.11 & 5.80 & 7.05 & 9.09 & 11.9 & 13.5 \\ 5.32 & 5.68 & 6.75 & 8.61 & 11.2 & 12.1 \end{bmatrix}.$$

The two sets of disturbance attenuation levels given by $\boldsymbol{\Gamma}$ and $\widetilde{\boldsymbol{\Gamma}}$ are very different from each other, yet they are both one-path Pareto optimal. □

6. Conclusion. This paper dealt with switched linear systems and Markovian jump linear systems in the discrete-time domain and developed complete conditions for (almost sure) uniform disturbance attenuation and (almost sure) path-by-path disturbance attenuation. These conditions naturally give rise to finite-path-dependent controllers and admit semidefinite programming algorithms for optimal dynamic output feedback controller synthesis. Limitations of these algorithms include that, in the worst case, the computational complexity grows exponentially in the number M of past modes that the optimal controller recalls, and that nonexistence of an admissible controller synthesis is not guaranteed to be correctly determined after a finite amount of computation. These limitations are due to the problem's nature and considered

unavoidable; nevertheless, they will not pose difficulties in most cases because M is usually very small.

There is at least one conceptually important question unanswered in this paper. This question is whether mode-dependent controllers perform as well as finite-path-dependent controllers as long as uniform disturbance attenuation is concerned; a more specific question is whether, given a Markovian jump linear system, there exists a stabilizing mode-dependent controller whenever a finite-path-dependent controller can stabilize the system. Under the notion of path-by-path disturbance attenuation, however, we showed that finite-path-dependent controllers can outperform mode-dependent ones.

The optimal disturbance attenuation example in Example 9 showed that our results provide a new contribution to the study of networked control systems. The proposed controller synthesis techniques are expected to be applicable to other time-delay systems. Related to switched and Markovian jump linear systems are linear parameter-varying systems, where the common design approaches for gain scheduling [5, 2, 3] are similar to those for mode-dependent controllers. A possible future research direction is to investigate if the approach of finite-path-dependent controller synthesis applies to the control of linear parameter-varying systems.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.

[2] P. Apkarian and P. Gahinet, *A convex characterization of gain-scheduled H_∞ controllers*, IEEE Trans. Automat. Control, 40 (1995), pp. 853–864.

[3] P. Apkarian, P. Gahinet, and G. Becker, *Self-scheduled H_∞ control of linear parameter-varying systems: A design example*, Automatica J. IFAC, 31 (1995), pp. 1251–1261.

[4] T. Başar, *A dynamic games approach to controller design: Disturbance rejection in discrete-time*, IEEE Trans. Automat. Control, 36 (1991), pp. 936–952.

[5] G. Becker and A. Packard, *Robust performance of linear parametrically varying systems using parametrically-dependent linear feedback*, Systems Control Lett., 23 (1994), pp. 205–215.

[6] P.-A. Bliman and G. Ferrari-Trecate, *Stability analysis of discrete-time switched systems through Lyapunov functions with nonminimal state*, in Proceedings of the IFAC Conference on the Analysis and Design of Hybrid Systems, Elsevier, Oxford, UK, 2003, pp. 325–330.

[7] V. D. Blondel and A. Megretski, eds., *Unsolved Problems in Mathematical Systems and Control Theory*, Princeton University Press, Princeton, NJ, 2004.

[8] V. D. Blondel, J. Theys, and A. A. Vladimirov, *An elementary counterexample to the finiteness conjecture*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 963–970.

[9] M. S. Branicky, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.

[10] Y.-Y. Cao and J. Lam, *Stochastic stabilizability and H_∞ control for discrete-time jump linear systems with time delay*, J. Franklin Inst., 336 (1999), pp. 1263–1281.

[11] O. L. V. Costa and M. D. Fragoso, *Stability results on discrete-time linear systems with Markovian jumping parameters*, J. Math. Anal. Appl., 179 (1993), pp. 154–178.

[12] I. Daubechies and J. C. Lagarias, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[13] D. P. de Farias, J. C. Geromel, J. B. R. do Val, and O. L. V. Costa, *Output feedback control of Markov jump linear systems in continuous-time*, IEEE Trans. Automat. Control, 45 (2000), pp. 944–949.

[14] G. de Nicolao, *On the time-varying Riccati difference equation of optimal filtering*, SIAM J. Control Optim., 30 (1992), pp. 1251–1269.

[15] C. E. de Souza, *On stabilizing properties of solutions of the Riccati difference equation*, IEEE Trans. Automat. Control, 34 (1989), pp. 1313–1316.

[16] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[17] G. E. Dullerud and S. Lall, *A new approach for analysis and synthesis of time-varying systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 1486–1497.
[18] G. E. Dullerud and F. Paganini, *A Course in Robust Control Theory: A Convex Approach*, Springer-Verlag, New York, 2000.
[19] M. D. Fragoso, J. B. Ribeiro do Val, and D. L. Pinto, Jr., *Jump linear H^∞ control: The discrete-time case*, Control Theory Adv. Tech., 10 (1995), pp. 1459–1474.
[20] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.
[21] L. Gurvits, *Stability of discrete linear inclusion*, Linear Algebra Appl., 231 (1995), pp. 47–85.
[22] A. Halanay and V. Ionescu, *Time-Varying Discrete Linear Systems*, Oper. Theory Adv. Appl. 68, Birkhäuser, Basel, Switzerland, 1994.
[23] J. Hespanha, *Root-mean-square gains of switched linear systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 2040–2045.
[24] P. G. Hoel, S. C. Port, and C. J. Stone, *Introduction to Stochastic Processes*, Waveland Press, Long Grove, IL, 1987.
[25] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
[26] P. A. Iglesias and K. Glover, *State-space approach to discrete-time H_∞ control*, Internat. J. Control, 54 (1991), pp. 1031–1073.
[27] P. A. Iglesias and M. A. Peters, *On the induced norms of discrete-time and hybrid time-varying systems*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 811–833.
[28] Y. Ji and H. J. Chizeck, *Controllability, stability, and continuous-time Markovian jump linear quadratic control*, IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[29] Y. Ji, H. J. Chizeck, X. Feng, and K. A. Loparo, *Stability and control of discrete-time jump linear systems*, Control Theory Adv. Tech., 7 (1991), pp. 247–270.
[30] J.-W. Lee and G. E. Dullerud, *Uniform stabilization of discrete-time switched and Markovian jump linear systems*, Automatica J. IFAC, 42 (2006), pp. 205–218.
[31] D. Liberzon, *Switching in Systems and Control*, Birkhäuser Boston, Boston, MA, 2003.
[32] M. Mariton, *Jump Linear Systems in Automatic Control*, Marcel Dekker, New York, 1990.
[33] A. Packard, *Gain scheduling via linear fractional transformations*, Systems Control Lett., 22 (1994), pp. 79–92.
[34] Z. Pan and T. Başar, *H^∞-control of Markovian jump systems and solutions to associated piecewise-deterministic differential games*, in New Trends in Dynamic Games and Applications, G. J. Olsder, ed., Birkhäuser Boston, Boston, MA, 1995, pp. 61–94.
[35] A. Rantzer, *On the Kalman–Yakubovich–Popov lemma*, Systems Control Lett., 28 (1996), pp. 7–10.
[36] R. Ravi, K. M. Nagpal, and P. P. Khargonekar, *H^∞ control of linear time-varying systems: A state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.
[37] P. Seiler and R. Sengupta, *A bounded real lemma for jump systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1651–1654.
[38] P. Seiler and R. Sengupta, *An H_∞ approach to networked control*, IEEE Trans. Automat. Control, 50 (2005), pp. 356–364.
[39] A. N. Shiryayev, *Probability*, Springer-Verlag, New York, 1996.
[40] M. W. Spong and D. J. Block, *The Pendubot: A mechatronic system for control research and education*, in Proceedings of the 34th IEEE Conference on Decision and Control, Vol. 1, IEEE, Piscataway, NJ, 1995, pp. 555–556.
[41] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[42] L. Xiao, A. Hassibi, and J. P. How, *Control with random communication delays via a discrete-time jump system approach*, in Proceedings of the American Control Conference, Vol. 3, IEEE, Los Alamitos, CA, 2000, pp. 2199–2204.
[43] B. Yoo, *Real-Time Control of Dynamic Systems Over Communication Networks*, Master's thesis, University of Illinois at Urbana-Champaign, Urbana, IL, 2003.
[44] G. Zhai, B. Hu, K. Yasuda, and A. N. Michel, *Disturbance attenuation properties of time-controlled switched systems*, J. Franklin Inst., 338 (2001), pp. 765–779.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1359–1368

AN APPROXIMATION ALGORITHM FOR THE DISCRETE TEAM DECISION PROBLEM*

RANDY COGILL† AND SANJAY LALL‡

Abstract. In this paper we study a discrete version of the classical team decision problem. It has been shown previously that the general discrete team decision problem is NP-hard. Here we present an efficient approximation algorithm for this problem. For the maximization version of this problem with nonnegative rewards, this algorithm computes decision rules which are guaranteed to be within a fixed bound of optimal.

Key words. team theory, decentralized control, approximation algorithms

AMS subject classifications. 93A14, 68Q25, 68W25

DOI. 10.1137/050628374

1. Introduction.

The problem of decentralized control arises when several decision makers, each with limited information, must make simultaneous cooperative decisions. The fact that multiple decisions are made on the basis of incomplete and differing observations of the overall system state is what sets decentralized control problems apart from conventional control problems. Such problems arise naturally in applications involving sensing and communication networks. An example is detection by multiple sensors in a sensor network [7]. The job of the sensors is to take measurements of the environment and transmit a minimal amount of information to a fusion center which estimates the state of the environment based on the information received from the sensors. Each sensor is faced with a decentralized decision making problem: based on its limited measurement of the environment, it must decide what information it should send to maximize the probability that the fusion center makes a correct estimate.

Marschak was probably the first to study the problem of decentralized decision making. In his work [5], he introduced *team theory* as a framework for studying decision making problems in organizations. The class of problems considered by team theory, called team decision problems, are analogous in some ways to static games. In these problems, the state of the system is chosen randomly according to some specified probability, and each decision maker partially observes the state. Based on their observations, each decision maker chooses an action. The goal is to choose decision rules which maximize the expected value of a reward which is jointly a function of the system state and all actions.

The later work of Radner [6] presented optimality conditions for a class of team decision problems with reward functions which are concave and differentiable in the decision variables. The main result of that work is that, subject to some technical conditions, *person-by-person optimal* policies are globally optimal for such problems.

*Received by the editors April 1, 2005; accepted for publication (in revised form) April 3, 2006; published electronically October 6, 2006.

http://www.siam.org/journals/sicon/45-4/62837.html

†Department of Electrical Engineering, Stanford University, Stanford, CA 94305 (rcogill@stanford.edu). This author was partially supported by a Stanford Graduate Fellowship.

‡Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 (lall@stanford.edu). This author was partially supported by the Stanford URI Architectures for Secure and Robust Distributed Infrastructures, AFOSR DoD award 49620-01-1-0365.

A set of decision rules is person-by-person optimal if no improvement can be obtained by changing the decision rule for one decision maker while leaving the decision rules for others fixed. A person-by-person optimal policy can be computed by sequentially optimizing the decision rule for each decision maker while leaving the remaining decision rules fixed.

The problems we are interested in have state and action sets which are finite, so the results regarding global optimality of person-by-person optimal policies do not apply. In fact, the problems that we consider may have many person-by-person optimal policies, and there may exist a person-by-person optimal policy which is quite poor compared to the globally optimal policy. Although problems with finite state and action spaces were presented in the original work by Marschak [5], no positive results regarding computation of optimal policies for these problems followed. Tsitsiklis showed in [8] that the general team decision problem with finite state and action spaces is NP-hard, explaining the lack of positive results for this problem. Another recent paper which considers the discrete team decision problem is [9]. In that paper, the authors consider problems with finite state and action spaces and reward functions which are *multimodular* in the decision variables. Multimodularity of functions on discrete spaces is analogous to concavity of functions on continuous spaces. The authors attempt to use multimodularity to extend the results of Radner to discrete problems. They show that necessary and sufficient optimality conditions are not assured by multimodularity, but this property still can be exploited when searching for optimal policies.

Since it is highly unlikely that an algorithm exists which can efficiently compute optimal solutions to the general team decision problem, this leads us to ask if an efficient algorithm exists which can compute acceptable suboptimal solutions. Many problems, although NP-hard, admit efficient *approximation algorithms* which produce solutions guaranteed to be within a fixed bound of optimal [1]. In this paper we present an approximation algorithm for the team decision problem. Specifically, the contributions of this paper are the following:

- For the general team problem with two decision makers, we present an algorithm that runs in $\mathcal{O}(|Y_1||Y_2||U_1||U_2|)$ time. When the objective is to maximize a nonnegative reward function, this algorithm computes a decentralized policy guaranteed to be within a factor of $1/\min\{|U_1|, |U_2|\}$ of optimal. Here, $|Y_1|, |Y_2|$ are the number of states and $|U_1|, |U_2|$ are the number of actions associated with decision makers 1 and 2. Even the special case with $\min\{|U_1|, |U_2|\} = 2$ is still NP-hard, and in this case we obtain a $1/2$ approximation factor.
- We consider the class of team decision problems with two decision makers and multimodular reward functions, as in [9]. We show that this special case is still NP-hard, although the presence of multimodularity leads to a significantly tighter approximation ratio of $1/2 - 1/(2\max\{|U_1|, |U_2|\})$ for our algorithm.

An outline of this paper is as follows. In section 2, we introduce the general discrete team decision problem and provide a proof that this problem is NP-hard. In section 3, we present an approximation algorithm for the general team decision problem and prove an approximation ratio for this algorithm. In section 4, we consider a special class of team decision problems with multimodular cost functions. In this section, we show that this class of problems is still NP-hard, but the approximation ratio for our algorithm is significantly tighter on these problems.

2. Problem formulation and complexity. In this section we will present the TEAM DECISION PROBLEM (TDP), the problem which is the subject of this paper. Before presenting the formal definition, we will briefly describe this problem in words. The problem is essentially a feedback control problem, where the goal is to choose actions in response to observations with the goal of maximizing some reward. The system is in a state described by the variables (y_1, y_2). After observing the system state, we would like to choose a pair of actions (u_1, u_2) to make the reward $c(y_1, y_2, u_1, u_2)$ as large as possible. However, the actions must be chosen in a decentralized manner. That is, action u_1 is chosen based only on an observation of y_1 and action u_2 is chosen based only on an observation of y_2. Our goal is to choose decision rules $\gamma_1 : Y_1 \to U_1$ and $\gamma_2 : Y_2 \to U_2$ to maximize some measure of total reward. A formal definition of the problem is the following.

TEAM DECISION PROBLEM. *Given finite sets Y_1, Y_2, U_1, U_2 and a nonnegative reward function $c : Y_1 \times Y_2 \times U_1 \times U_2 \to \mathbb{Q}_+$, find decision rules $\gamma_i : Y_i \to U_i$, $i = 1, 2$ which maximize the reward*

$$J(\gamma_1, \gamma_2) = \sum_{y_1, y_2} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2)).$$

A few words are in order regarding the formulation given above. First of all, this formulation only considers the case in which two decision makers are choosing actions. In general, we have problems where n decision makers are independently choosing actions $u_1, \ldots, u_n$. It turns out, as we will see shortly, that this problem is computationally intractable even in the two decision maker case. This is the simplest special case which exhibits the inherent computational complexity of this problem, and this case is the focus of this paper. It is worth noting that the approach taken in this paper can be extended to problems with more than two decision makers, although the resulting approximation guarantees degrade exponentially as the number of decision makers increases.

In [8], this problem is posed as maximization of the *expected* reward with respect to some probability mass function $p(y_1, y_2)$. While this formulation may relate more naturally to the applications where this problem is of interest, the formulation in terms of expected reward and the formulation in terms of total reward are essentially equivalent. That is, given any instance of the problem of maximizing expected reward, we can easily modify the reward function to obtain an equivalent instance of maximizing total reward. Conversely, any instance of the problem of maximizing total reward is equivalent to an instance of maximizing expected reward where $p(y_1, y_2)$ is uniform. Therefore, we consider the problem of maximizing total reward simply to reduce the amount of required notation.

Also, it is interesting to note that the corresponding centralized problem is trivial. By the centralized problem, we mean the problem of choosing a policy $\gamma : Y_1 \times Y_2 \to U_1 \times U_2$, to maximize the total reward. This problem is solved by simply choosing the (u_1, u_2) which maximizes $c(y_1, y_2, u_1, u_2)$ for each (y_1, y_2). Although the centralized problem is easy, it was shown in [8] that TDP is NP-hard, even when $|U_1| = |U_2| = 2$.

Here we will present a simple new proof of NP-hardness to keep our treatment self contained. We will do this by also showing that the special case of TDP with $|U_1| = |U_2| = 2$, which we refer to as TDP-2, is NP-hard. Our proof of NP-hardness of TDP-2 involves reducing the problem MAXIMUM CUT [1] to TDP-2. The problem MAXIMUM CUT is the following. As input, we are given an undirected graph $G = (V, E)$. The goal is to partition the set of vertices V into two sets S and $\overline{S}$ so that

the number of edges crossing from vertices in S to vertices in $\overline{S}$ is maximized. Given that MAXIMUM CUT is NP-hard, we have the following theorem.

THEOREM 2.1. *The problem* TDP-2 *is NP-hard.*

Proof. Consider an arbitrary instance of MAXIMUM CUT specified by an undirected graph $G = (V, E)$. We construct a corresponding instance of TDP-2 as follows. Let $Y_1 = Y_2 = V$. Define c_1 as

$$c_1(y_1, y_2, u_1, u_2) = \begin{cases} |V|^2 & \text{if } y_1 = y_2 \text{ and } u_1 = u_2, \\ 0 & \text{otherwise.} \end{cases}$$

Define c_2 as

$$c_2(y_1, y_2, u_1, u_2) = \begin{cases} 1 & \text{if } (y_1, y_2) \in E \text{ and } u_1 \neq u_2, \\ 0 & \text{otherwise.} \end{cases}$$

The goal is to show that the instance of MAXIMUM CUT is solved by finding an optimal policy for the instance of TDP-2 with reward

$$c(y_1, y_2, u_1, u_2) = c_1(y_1, y_2, u_1, u_2) + c_2(y_1, y_2, u_1, u_2).$$

The optimal reward satisfies $|V|^3 < J(\gamma_1^*, \gamma_2^*)$. The lower bound is achieved by choosing any policy satisfying $\gamma_1 = \gamma_2$. Such policies maximize the component of the reward associated with c_1. As long as G contains one edge, we can always achieve reward strictly greater than $|V|^3$. Moreover, if $\gamma_1(i) \neq \gamma_2(i)$ for some i, then $J(\gamma_1, \gamma_2) \leq |V|^3$. This is because, in this case, the component of the reward associated with c_1 is at most $(|V| - 1)|V|^2$ and the component of the reward associated with c_2 is at most $|V|^2$. Hence $\gamma_1^* = \gamma_2^*$.

Given any decision rules γ_1 and γ_2 satisfying $\gamma_1 = \gamma_2$, we obtain a cut for the instance of MAXIMUM CUT by letting

$$v_i \in \begin{cases} S & \text{if } \gamma_1(v_i) = 1, \\ \overline{S} & \text{if } \gamma_1(v_i) = 0 \end{cases}$$

for all $v_i \in V$. The capacity of this cut is exactly $\frac{1}{2}J(\gamma_1, \gamma_2) - \frac{1}{2}|V|^3$. Hence, any algorithm which computes optimal policies for arbitrary instances of TDP-2 in polynomial time can be easily modified to compute optimal cuts for arbitrary instances of MAXIMUM CUT in polynomial time. □

3. An approximation algorithm. Given that TDP is NP-hard, it is highly unlikely that an efficient algorithm for computing optimal policies exists. However, for many NP-hard problems there exist efficient *approximation algorithms* which produce suboptimal solutions which achieve a reward guaranteed to be within some constant factor of optimal. In this section we will give an approximation algorithm for TDP which produces a suboptimal policy in $\mathcal{O}(|Y_1||Y_2||U_1||U_2|)$ operations.

Before presenting our approximation algorithm, we would like to first consider algorithms that produce *person-by-person* optimal policies [6]. A person-by-person optimal policy is a pair of decision rules γ_1, γ_2 for which no improvement can be obtained by modifying γ_1 while leaving γ_2 fixed or by modifying γ_2 while leaving γ_1 fixed. A person-by-person optimal solution can be considered analogous in some ways to a Nash equilibrium. Person-by-person optimal solutions can be computed by choosing initial decision rule γ_1 and γ_2 and alternately maximizing over one decision

TABLE 3.1

For this reward function there is a person-by-person optimal policy with reward $J(\gamma_1, \gamma_2) = 8$. The optimal policy has reward $J(\gamma_1^*, \gamma_2^*) = 400$.

	$y_1 = 0$ $u_1 = 0$	$y_1 = 0$ $u_1 = 1$	$y_1 = 1$ $u_1 = 0$	$y_1 = 1$ $u_1 = 1$
$y_2 = 0$ $u_2 = 0$	2	1	1	2
$y_2 = 0$ $u_2 = 1$	1	100	100	1
$y_2 = 1$ $u_2 = 0$	1	100	100	1
$y_2 = 1$ $u_2 = 1$	2	1	1	2

rule while leaving the other decision rule fixed until no more improvement is obtained. For discrete problems, this process will terminate in a finite number of iterations.

For problems with rewards which are continuous, differentiable, and concave in the variables u_1 and u_2, it was shown that person-by-person optimality implies global optimality [6]. This is probably the best-known result for team decision problems. It is natural, then, to believe that person-by-person optimal policies may be acceptable solutions to discrete team decision problems. The purpose for considering person-by-person optimal solutions here is to show that, in fact, person-by-person optimal solutions may achieve a reward which is arbitrarily far from the globally optimal reward in discrete problems.

Consider the problem instance with the reward function given in Table 3.1. For this instance, consider the policy $\gamma_1(0) = 0$, $\gamma_1(1) = 1$, $\gamma_2(0) = 0$, $\gamma_2(1) = 1$. This policy is person-by-person optimal and achieves a reward of $J(\gamma_1, \gamma_2) = 8$. However, the optimal policy achieves a reward of $J(\gamma_1^*, \gamma_2^*) = 400$. Here, a person-by-person optimal policy is suboptimal by a factor of 50. Of course, we can modify the example to make the reward achieved by the person-by-person optimal as far from the optimal reward as we like. Therefore, algorithms which start with an arbitrary policy and seek person-by-person optimal solutions cannot produce polices which are within a guaranteed bound of optimal.

We will now present our approximation algorithm for TDP. Roughly speaking, this algorithm first constructs a "marginalized" reward function

$$c_1(y_1, u_1) = \sum_{u_2, y_2} c(y_1, y_2, u_1, u_2),$$

then computes a decision rule $\gamma_1 : Y_1 \to U_1$ which is optimal for this marginalized reward. Then, given γ_1, we compute the decision rule $\gamma_2 : Y_2 \to U_2$ which is optimal with respect to γ_1. Unlike an arbitrary person-by-person optimal policy, this policy is guaranteed to be within a constant factor of optimal, where the factor depends on $|U_1|$ and $|U_2|$. To obtain the tightest suboptimality guarantee, we may first compute γ_2 and then γ_1.

ALGORITHM 3.1. *Assume, without loss of generality, that $|U_1| \geq |U_2|$.*

1. *Let $\gamma_1(y_1) \in argmax_{u_1}\{\sum_{u_2}\sum_{y_2} c(y_1, y_2, u_1, u_2)\}$ for all $y_1 \in Y_1$.*
2. *Let $\gamma_2(y_2) \in argmax_{u_2}\{\sum_{y_1} c(y_1, y_2, \gamma_1(y_1), u_2)\}$ for all $y_2 \in Y_2$.*

If $|U_2| > |U_1|$, then we just compute γ_2 first, followed by γ_1. The approximation guarantee for the previous algorithm is given in the following theorem.

THEOREM 3.2. *For* TDP*, Algorithm* 3.1 *produces a policy* γ_1, γ_2 *with value satisfying*

$$J(\gamma_1, \gamma_2) \geq \frac{1}{\min\{|U_1|, |U_2|\}} J(\gamma_1^*, \gamma_2^*).$$

Proof. Without any loss of generality, here we will assume that $|U_1| \geq |U_2|$. Since $\gamma_1(y_1) \in \operatorname{argmax}_{u_1} \sum_{u_2} \sum_{y_2} c(y_1, y_2, u_1, u_2)$, we have

$$\begin{aligned}\sum_{y_2} \sum_{u_2} c(y_1, y_2, \gamma_1(y_1), u_2) &\geq \sum_{y_2} \sum_{u_2} c(y_1, y_2, \gamma_1^*(y_1), u_2) \\ &\geq \sum_{y_2} c(y_1, y_2, \gamma_1^*(y_1), \gamma_2^*(y_2))\end{aligned}$$

for each $y_1 \in Y_1$. Also, $\gamma_2(y_2) \in \operatorname{argmax}_{u_2} \sum_{y_1} c(y_1, y_2, \gamma_1(y_1), u_2)$, so

$$\sum_{y_1} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2)) \geq \frac{1}{|U_2|} \sum_{u_2} \left(\sum_{y_1} c(y_1, y_2, \gamma_1(y_1), u_2) \right)$$

for each $y_2 \in Y_2$. Therefore, the reward achieved by policy γ_1, γ_2 satisfies

$$\begin{aligned}\sum_{y_1} \sum_{y_2} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2)) &\geq \sum_{y_2} \left(\frac{1}{|U_2|} \sum_{u_2} \sum_{y_1} c(y_1, y_2, \gamma_1(y_1), u_2) \right) \\ &= \frac{1}{|U_2|} \sum_{y_1} \left(\sum_{y_2} \sum_{u_2} c(y_1, y_2, \gamma_1(y_1), u_2) \right) \\ &\geq \frac{1}{|U_2|} \sum_{y_1} \sum_{y_2} c(y_1, y_2, \gamma_1^*(y_1), \gamma_2^*(y_2)). \qquad \square\end{aligned}$$

Although the approximation guarantee degrades as the number of available decisions increases, we must keep in mind that even TDP-2 is NP-hard, and in this case we obtain an approximation factor of 1/2.

It is worth noting that Algorithm 3.1, and consequently the proof of Theorem 3.2, has an alternate interpretation in terms of *randomized* decision rules. In this interpretation, γ_2 is initially set to be the randomized decision rule which selects actions randomly according to a uniform distribution, regardless of the value of y_2. We then choose γ_1 to be the decision rule which is optimal with respect to this randomized γ_2. Finally, the randomized γ_2 is replaced with the decision rule which is optimal with respect to the chosen γ_1. While this interpretation of Algorithm 3.1 is slightly more complex conceptually, this interpretation can be used to simplify the proof of Theorem 3.2.

We can show that the bound is tight for this algorithm, and it is achieved on a simple example with $|U_1| = |U_2| = |Y_1| = |Y_2| = 2$. Consider the reward function given in Table 3.2. Here, ϵ is some arbitrarily small constant which is simply introduced to avoid any ambiguity associated with adding a tie-breaking mechanism into the algorithm. The policy produced by the algorithm, $\gamma_1(0) = 0$, $\gamma_1(1) = 0$, $\gamma_2(0) = 1$, $\gamma_2(1) = 1$, achieves $J(\gamma_1, \gamma_2) = 2 + 2\epsilon$. This is true regardless of which variables we

TABLE 3.2
The bound in Theorem 3.2 *is tight for this reward function.*

	$y_1 = 0$ $u_1 = 0$	$y_1 = 0$ $u_1 = 1$	$y_1 = 1$ $u_1 = 0$	$y_1 = 1$ $u_1 = 1$
$y_2 = 0$ $u_2 = 0$	1	0	0	1
$y_2 = 0$ $u_2 = 1$	0	1	$1 + \epsilon$	0
$y_2 = 1$ $u_2 = 0$	0	1	1	0
$y_2 = 1$ $u_2 = 1$	$1 + \epsilon$	0	0	1

choose to marginalize over first. However, the optimal policy $\gamma_1(0) = 0$, $\gamma_1(1) = 1$, $\gamma_2(0) = 0$, $\gamma_2(1) = 1$ achieves $J(\gamma_1^*, \gamma_2^*) = 4 + \epsilon$. Since ϵ can be chosen arbitrarily small, the ratio between the reward achieved by the computed policy and optimal policy can be made arbitrarily close to 1/2. Similar examples can be constructed for larger values of $|U_1|$ and $|U_2|$.

It is worth noting that the policy constructed by Algorithm 3.1 in the previous example is person-by-person optimal. This means that the bound in Theorem 3.2 is still tight even for person-by-person optimal policies obtained by iterating over decision rules using the decision rules produced by Algorithm 3.1 as a starting point.

4. Approximation ratio for multimodular rewards. In this section we consider a class of team decision problems with specially structured rewards and show that our algorithm achieves a tighter approximation ratio on these problems. The problems that we consider here can be thought of as the discrete counterpart of the problems considered by Radner [6]. In particular, Radner considered problems with rewards which are continuous, differentiable, and concave in the decision variables. It was shown that globally optimal policies can be easily computed for these problems. Here we consider rewards which are *multimodular* in discrete decision variables. Multimodularity is an extension of the notion of convexity to functions of discrete variables. Discrete team decision problems with multimodular rewards were studied previously in [9]. The authors attempt to use multimodularity to extend the results of Radner to discrete problems. They show that necessary and sufficient optimality conditions are not assured by multimodularity, but this property still can be exploited when searching for optimal policies. Here we show the class of problems with multimodular rewards is still NP-hard. However, the presence of multimodularity improves the approximation ratio to $1/2 - 1/(2\max\{|U_1|, |U_2|\})$ for the algorithm presented in the previous section. Combining this with the bound proven in the last section, this gives a bound of 1/3 in the worst case which only *improves* as the sizes of U_1 and U_2 increase.

The concept of multimodularity was introduced by Hajek in [2]. In that paper, multimodularity is defined in terms of local properties of a function and the definition we give below is proven later to be a property equivalent to multimodularity. However, here we do not use most of the theory developed for multimodular functions. We use the definition below because it most clearly relates to convexity.

DEFINITION 4.1. *A function $f : \mathbb{Z}^n \to \mathbb{R}$ is said to be* multimodular *iff there exists a concave function $g : \mathbb{R}^n \to \mathbb{R}$ such that $f(x) = g(x)$ for all $x \in \mathbb{Z}^n$.*

It is worth noting that in the existing literature, multimodular functions are typically defined as discrete functions which coincide with *convex* functions. Since

we do not know of a corresponding term for discrete functions which coincide with concave functions, we simply refer to these as multimodular as well. Also, throughout this section we consider functions which are defined on some subset $S \subset \mathbb{Z}^n$. When we say that such a function is multimodular, we mean that there is some concave function which coincides with f everywhere on S.

We can now describe the class of reward functions that we consider in this section. Specifically, we consider rewards $c : Y_1 \times Y_2 \times U_1 \times U_2$ such that for each $y_1 \in Y_1$ and $y_2 \in Y_2$, $c(y_1, y_2, u_1, u_2)$ is multimodular in u_1 and u_2. To make this statement precise, we have $U_1 = \{1, \ldots, |U_1|\}$ and $U_2 = \{1, \ldots, |U_2|\}$. We refer to this class of problems as TDP-M. We can make the following claim regarding the complexity of TDP-M.

THEOREM 4.2. *The problem* TDP-M *is NP-hard.*

Proof. The show this, we just need to show that every instance of TDP-2 is an instance of TDP-M. This is done by showing any function $f : \{0,1\} \times \{0,1\} \to \mathbb{R}$ is multimodular.

For some function $f : \{0,1\} \times \{0,1\} \to \mathbb{R}$, suppose that $f(0,0) + f(1,1) \leq f(0,1) + f(1,0)$. We can define the following piecewise affine continuation on $[0,1] \times [0,1]$:

$$g(x_1, x_2) = \begin{cases} (1-x_1-x_2)f(0,0) + x_1 f(1,0) + x_2 f(0,1) & \text{for } x_1 + x_2 \leq 1, \\ (1-x_2)f(1,0) + (1-x_1)f(0,1) + (x_1+x_2-1)f(1,1) & \text{for } x_1 + x_2 > 1. \end{cases}$$

Verifying that this function is concave on $[0,1] \times [0,1]$ is somewhat tedious, but straightforward. If $f(0,0) + f(1,1) > f(0,1) + f(1,0)$, then we define g as

$$g(x_1, x_2) = \begin{cases} (1-x_2)f(0,0) + (x_2-x_1)f(0,1) + x_1 f(1,1) & \text{for } x_1 \leq x_2, \\ (1-x_1)f(0,0) + (x_1-x_2)f(1,0) + x_2 f(1,1) & \text{for } x_1 > x_2, \end{cases}$$

which is concave on $[0,1] \times [0,1]$. □

Now that we have established that TDP-M is NP-hard, our goal is to show that our algorithm has a tighter approximation ratio for these problems. To ensure the tightest approximation ratio for this class of problems, we make one very minor change to the algorithm.

ALGORITHM 4.3. *Assume, without loss of generality, that* $|U_2| \geq |U_1|$.

1. *Let* $\gamma_1(y_1) \in \operatorname{argmax}_{u_1} \{\sum_{u_2} \sum_{y_2} c(y_1, y_2, u_1, u_2)\}$ *for all* $y_1 \in Y_1$.
2. *Let* $\gamma_2(y_2) \in \operatorname{argmax}_{u_2} \{\sum_{y_1} c(y_1, y_2, \gamma_1(y_1), u_2)\}$ *for all* $y_2 \in Y_2$.

The only difference between this algorithm and Algorithm 3.1 is the order in which the decision rules are computed. As we saw previously, the approximation ratio depends on the number of decisions available to the decision rule which is computed second. For the approximation ratio proven in Theorem 3.2, the quality of the approximation degrades with the number of decisions. For the approximation ratio proven in this section, the quality of the approximation *improves* with the number of decisions.

Before proving approximation ratio for TDP-M, we first need to establish several properties of multimodular functions of a single variable.

LEMMA 4.4. *A function* $f : \mathbb{Z} \to \mathbb{R}$ *is multimodular iff*

$$f(x_2) \geq \frac{x_3 - x_2}{x_3 - x_1} f(x_1) + \frac{x_2 - x_1}{x_3 - x_1} f(x_3)$$

for any $x_1 \leq x_2 \leq x_3$.

Proof. If the inequality is violated for some x_1, x_2, and x_3, then clearly it is violated for any $g : \mathbb{R} \to \mathbb{R}$ which coincides with f. Conversely, if the inequality holds for all $x_1 \le x_2 \le x_3$, then it's easy to verify that the piecewise affine continuation of f,

$$g(x) = (\lceil x \rceil - x) f(\lfloor x \rfloor) + (x - \lfloor x \rfloor) f(\lceil x \rceil),$$

is concave. Here, $\lfloor x \rfloor$ denotes the largest integer less than or equal to x and $\lceil x \rceil$ denotes the smallest integer strictly greater than x. □

When proving Theorem 3.2 in the previous section, we used the fact that the sum of a set of nonnegative numbers is an upper bound on the maximum of these numbers. The key to obtaining a tighter approximation ratio for multimodular rewards is to use the fact that the gap between these two quantities can be quite large when the set of numbers corresponds to the values taken by a multimodular function. This is stated precisely in the following theorem.

THEOREM 4.5. *Suppose $f : \{1, \dots, N\} \to \mathbb{R}_+$ is multimodular. Then*

$$\sum_{i=1}^{N} f(i) \ge \frac{N-1}{2} f(j)$$

for all $j \in \{1, \dots, N\}$.

Proof. Let k be such that $f(k) \ge f(j)$ for all $j \in \{1, \dots, N\}$. We just need to show that $\sum_{i=1}^{N} f(i) \ge \frac{N-1}{2} f(k)$. Since $f(1) \ge 0$ and $f(N) \ge 0$, Lemma 4.4 gives

$$f(i) \ge \frac{i-1}{k-1} f(k) \quad \text{for all } 1 \le i \le k,$$
$$f(i) \ge \frac{N-i}{N-k} f(k) \quad \text{for all } k \le i \le N.$$

If $k = 1$ or $k = N$, we just have one of these inequalities for all $1 \le i \le N$. Using these inequalities,

$$\begin{aligned}\sum_{i=1}^{N} f(i) &\ge \sum_{i=1}^{k} \frac{i-1}{k-1} f(k) + \sum_{k+1}^{N} \frac{N-i}{N-k} f(k) \\ &= \frac{1}{k-1} \left(\frac{1}{2} k(k-1) \right) f(k) + \frac{1}{N-k} \left(\frac{1}{2} (N-k)(N-k-1) \right) f(k) \\ &= \frac{N-1}{2} f(k). \qquad \square\end{aligned}$$

We can now prove the main result of this section.

THEOREM 4.6. *For* TDP-M*, Algorithm* 4.3 *produces a policy γ_1, γ_2 with value satisfying*

$$J(\gamma_1, \gamma_2) \ge \left(\frac{1}{2} - \frac{1}{2 \max\{|U_1|, |U_2|\}} \right) J(\gamma_1^*, \gamma_2^*).$$

Proof. Without any loss of generality, here we will assume that $|U_2| \ge |U_1|$. In the proof of Theorem 3.2, we use the fact that

$$\sum_{u_2} c(y_1, y_2, \gamma_1^*(y_1), u_2) \ge c(y_1, y_2, \gamma_1^*(y_1), \gamma_2^*(y_2)). \tag{4.1}$$

For any fixed y_1 and y_2, $c(y_1, y_2, \gamma_1^*(y_1), u_2)$ is multimodular in u_2. Therefore, we can use Theorem 4.5 to strengthen inequality (4.1) to

$$\sum_{u_2} c(y_1, y_2, \gamma_1^*(y_1), u_2) \geq \frac{|U_2| - 1}{2} c(y_1, y_2, \gamma_1^*(y_1), \gamma_2^*(y_2)).$$

This gives us the bound

$$\begin{aligned}
\sum_{y_1}\sum_{y_2} c(y_1, y_2, \gamma_1(y_1), \gamma_2(y_2)) &\geq \sum_{y_2}\left(\frac{1}{|U_2|}\sum_{u_2}\sum_{y_1} c(y_1, y_2, \gamma_1(y_1), u_2)\right) \\
&\geq \frac{1}{|U_2|}\sum_{y_1}\sum_{y_2}\left(\sum_{u_2} c(y_1, y_2, \gamma_1^*(y_1), u_2)\right) \\
&\geq \left(\frac{1}{2} - \frac{1}{2|U_2|}\right)\sum_{y_1}\sum_{y_2} c(y_1, y_2, \gamma_1^*(y_1), \gamma_2^*(y_2)). \qquad \square
\end{aligned}$$

Note that this bound is not tight for the case where $|U_1| = 2$ or $|U_2| = 2$. This is simply because the bound proven in Theorem 4.4 is trivial and never tight when $N = 2$. Of course in this the case, the bound proven in Theorem 3.2 still holds and is tight. With this in mind, the bound in Theorem 4.6 says that the approximation ratio for any instance of TDP-M is at least 1/3, and approaches 1/2 as the sizes of U_1 and U_2 increase.

5. Conclusions. In this paper we presented an approximation algorithm for the discrete team decision problem [8, 9]. We focused on problems involving two decision makers. For the general discrete team decision problem, the approximation ratio for our algorithm depends on the number of actions available to each decision maker. For the case when at least one decision maker chooses between two actions, which is still NP-hard, we have an approximation ratio of 1/2. We then considered a special case of the discrete team decision problem which can be thought of as the discrete counterpart of the problems considered by Radner [6]. We show that this special class of problems is still NP-hard, but the approximation ratio for our algorithm is significantly tighter on these problems.

REFERENCES

[1] M. R. Garey and D. S. Johnson, *Computers and Intractability—A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.

[2] B. Hajek, *Extremal splittings of point processes*, Math. Oper. Res., 10 (1985), pp. 543–556.

[3] Y. C. Ho, *Team decision theory and information structures*, Proceedings of the IEEE, 68 (1980), pp. 644–654.

[4] J. Krainak, J. Speyer, and S. Marcus, *Static team problems—Part* I: *Sufficient conditions and the exponential cost criterion*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 839–848.

[5] J. Marschak, *Elements for a theory of teams*, Management Science, 1 (1955), pp. 127–137.

[6] R. Radner, *Team decision problems*, Ann. Math. Statist., 33 (1962), pp. 857–881.

[7] J. Tsitsiklis, *Decentralized Detection*, vol. 2, H. V. Poor and J. B. Thomas, eds., JAI Press, Greenwich, CT, 1993, pp. 297–344.

[8] J. Tsitsiklis and M. Athans, *On the complexity of decentralized decision making and detection problems*, IEEE Trans. Automat. Control, 30 (1985), pp. 440–446.

[9] P. R. De Waal and J. H. Van Schuppen, *A class of team problems with discrete action spaces: Optimality conditions based on multimodularity*, SIAM J. Control Optim., 38 (2000), pp. 875–892.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1369–1394

ELLIPSOIDAL TECHNIQUES FOR REACHABILITY UNDER STATE CONSTRAINTS*

A. B. KURZHANSKI† AND P. VARAIYA†

Abstract. The paper presents a scheme to calculate approximations of reach sets and tubes for linear control systems with time-varying coefficients, bounds on the controls, and *constraints on the state*. The scheme provides tight external approximations by ellipsoid-valued tubes. The tubes touch the reach tubes from the outside at *each point* of their boundary so that the surface of the reach tube is totally covered by curves that belong to the approximating tubes. The result is an exact parametric representation of reach tubes through families of external ellipsoidal tubes. The parameters that characterize the approximating ellipsoids are solutions of ordinary differential equations with coefficients given partly in explicit analytical form and partly through the solution of a recursive optimization problem. The scheme combines the calculation of external approximations of infinite sums and intersections of ellipsoids, and suggests an approach to calculate reach sets of hybrid systems.

Key words. reachability, reach sets, control problems, HJB equations, duality theory

AMS subject classifications. 34H05, 34G25, 49L, 49N, 35F, 37N35, 34K99

DOI. 10.1137/S0363012903437605

1. Introduction.

Recent activities in automation of real-time processes have renewed interest in the calculation of reach sets $\mathcal{X}(\tau, t_0, x^0)$ of linear controlled systems of the form

$$\dot{x} = A(t)x + B(t)u, \quad \tau \geq t_0. \tag{1}$$

Typically, there are additional constraints. The most difficult calculations are for state constraints, and for hybrid or switched systems in which the state and dynamics are reset [23, 24, 21, 7]. State constraints are considered in this paper. Hybrid systems are treated in a separate paper [17].

Among methods for reach set calculation are those based on ellipsoidal techniques. Most of the early publications in this area, e.g., [1, 2], determined a *single* (such as minimum volume) ellipsoid that externally approximates the reach set at a *single* point in time.

In a series of papers [12, 13, 15, 16] the authors developed a parametrized family of ellipsoidal approximations, both external and internal, which are

- *tight*—no other ellipsoid in the family can be "squeezed" between the reach set and the approximating ellipsoid;
- *exact*—the intersection of the external ellipsoids and the union of the internal ellipsoids equals the reach set; and
- *recursive*—a single differential equation provides parameters that specify ellipsoids which approximate the reach set for *all* times.

*Received by the editors November 9, 2003; accepted for publication (in revised form) April 10, 2006; published electronically October 6, 2006. This research was supported by National Science Foundation grant ECS-0424445, and grants from the Russian Foundation for Basic Research (03-01-00663) and "Universities of Russia" (UR.03.03.036).

http://www.siam.org/journals/sicon/45-4/43760.html

†Electronics Research Laboratory, University of California at Berkeley, Berkeley, CA 94720-1770 (kurzhans@eecs.berkeley.edu, varaiya@eecs.berkeley.edu).

The approach in these papers is first to convert the problem of calculating the reach set into an equivalent *optimization* problem; then to show that the corresponding value function satisfies a forward Hamilton–Jacobi–Bellman (HJB) equation; and finally to use convex analysis to approximate the level sets of the value function. The same approach is followed here.

The situation in which (1) contains a disturbance term

$$\dot{x} = A(t)x + B(t)u + C(t)v, \quad \tau \geq t_0,$$

and reachability must be guaranteed independent of the disturbance v, requires consideration of *feedback* control. That situation is treated in [14].

This paper deals with reach tubes for control systems (1), with bounds $u(t) \in \mathcal{P}(t)$ on the control, and state constraints $x(t) \in \mathcal{Y}(t)$. As in the case without state constraints [16], the approximating tubes are generated by ellipsoid-valued maps that satisfy a semigroup property and thus form a generalized dynamical system. The ellipsoidal maps are chosen in a certain class, which can be characterized through ordinary differential equations. But, unlike in [16], these differential equations now contain parameters that are the online solutions of a recursive dual optimization problem. The optimizers of the dual problem are the Lagrange multipliers of the state constraints, and may contain generalized Dirac delta functions [27, 28]. However, simpler equations are obtained if we drop the requirement of tightness (resp., exactness).

Section 2 defines the reachability problem for general nonlinear systems. It formulates the associated optimization problem and shows that its value function satisfies an HJB equation. Section 3 uses convex analysis to characterize the reach sets for linear systems with state constraints in terms of the maximum principle. The state constraints introduce a Lagrange multiplier in the adjoint equation. Section 4 derives a different version of the maximum principle [22] that permits a recursive calculation of the reach sets. Section 5 develops the scheme for ellipsoidal approximations. Section 6 exhibits additional properties of the multipliers. Section 7 works out the example of a double integrator. A brief conclusion is presented in section 8. Additional properties of the multipliers necessary for the proofs are borrowed from previous publications and summarized in the appendix.

2. The reachability problem with state constraints. The controlled system is described by the differential equation,

$$\dot{x} = f(t, x, u), \quad t_0 \leq t \leq \tau, \tag{2}$$

in which the state $x \in \mathbb{R}^n$ and the control $u \in \mathbb{R}^m$ are restricted by

$$u(t) \in \mathcal{P}(t), \qquad x(t) \in \mathcal{Y}(t), \tag{3}$$

for all $t \geq t_0$. Here $\mathcal{P}(t), \mathcal{Y}(t)$ are compact set-valued functions in $\mathbb{R}^m, \mathbb{R}^n$, respectively, continuous in the Hausdorff metric. The initial state is restricted by $x(t_0) \in \mathcal{X}^0$, a compact subset of $\mathbb{R}^n$. The function $f(t, x, u)$ is assumed to ensure uniqueness and uniform extension of solutions to any finite interval of time for any $x(t_0) = x^0$, $u(t) \in \mathcal{P}(t)$, $t \geq t_0$.

Definition 2.1. *Given set-valued* positions $\{t_0, \mathcal{X}^0\}$, $\mathcal{X}^0 \cap \mathcal{Y}(t_0) \neq \emptyset$, *the* reach set *(or "attainability domain")* $\mathcal{X}(\tau, t_0, \mathcal{X}^0)$ at time $\tau > t_0$ *is the set*

$$\mathcal{X}[\tau] = \mathcal{X}(\tau, t_0, \mathcal{X}^0)$$

of all states $x[\tau] = x(\tau, t_0, x^0)$, $x^0 = x(t_0) \in \mathcal{X}(t_0) = \mathcal{X}^0$, *that can be reached at time* τ *by system* (2), *from some* $x^0 \in \mathcal{X}^0$, *using all possible controls* u *that ensure constraints* (3). *The set-valued function* $\tau \mapsto \mathcal{X}[\tau] = \mathcal{X}(\tau, t_0, \mathcal{X}^0)$ *is the* reach tube *from* $\{t_0, \mathcal{X}^0\}$.

The basic problem is simply stated.

Problem 2.1. Calculate the reach sets $\mathcal{X}(\tau, t_0, \mathcal{X}^0), \tau \geq t_0$.

A fairly general approach is to relate reach sets to an optimization problem [15]. This could be done by calculating certain *value functions* that may be selected in several ways. Consider first the value function

$$V(\tau, x) = \min_u \left\{ d^2(x(t_0), \mathcal{X}^0) + \int_{t_0}^{\tau} d^2(x(t), \mathcal{Y}(t))dt \right\}, \tag{4}$$

under the restriction $x(\tau) = x$. Here the minimum is over all measurable functions $u(t) \in \mathcal{P}(t)$, $x(t), t \geq t_0$, is the corresponding trajectory, and

$$d^2(x, \mathcal{X}) = \min\{(x - z, x - z) | z \in \mathcal{X}\}$$

is the square of the distance function $d(x, \mathcal{X})$.

LEMMA 2.1. *The following relation is true:*

$$\mathcal{X}(\tau, t_0, \mathcal{X}^0) = \{x : V(\tau, x) \leq 0\}.$$

This follows from the definition of the reach set $\mathcal{X}(\tau, t_0, \mathcal{X}^0)$, which is thus a *level set* of $V(\tau, x)$.

To state the important semigroup property of the value function, we extend definition (4) to more general boundary conditions, namely,

$$V(\tau, x \mid t_0, V(t_0, \cdot)) = \min_u \left\{ V(t_0, \cdot) + \int_{t_0}^{\tau} d^2(x(t), \mathcal{Y}(t))dt \mid u(\cdot) \in \mathcal{P}(\cdot) \right\},$$

under the restriction $x(\tau) = x$. The function $V(t_0, \cdot)$ is any boundary condition. In (4) the boundary condition is $V(t_0, x) = d^2(x, \mathcal{X}^0)$.

THEOREM 2.1. *The value function* $V(\tau, x)$ *satisfies* the principle of optimality, *which has the semigroup form*

$$V(\tau, x \mid t_0, V(t_0, \cdot)) = V(\tau, x \mid t, V(t, \cdot \mid t_0, V(t_0, \cdot))), \quad t_0 \leq t \leq \tau. \tag{5}$$

These properties are established through conventional arguments [5] that also yield similar properties for reach sets. The solution of the reachability problem can be cast in the form of a solution of the "forward" HJB equation that follows from (5). To develop this approach we further assume that the functions $V(t, x)$ and $\varphi(t, x)$ are *differentiable*.

A standard procedure [5] yields

$$V_t(t, x) + \max\{(V_x(t, x), f(t, x, u)) - d^2(x, \mathcal{Y}(t)) \mid u(t) \in \mathcal{P}(t)\} = 0, \quad t_0 \leq t \leq \tau, \tag{6}$$

with boundary condition $V(t_0, x) = d^2(x, \mathcal{X}^0)$. Here V_t, V_x stand for the partial derivatives of $V(t, x)$. Note that the term $d^2(x, \mathcal{Y}(t)) \neq 0$ only outside the state constraint $\mathcal{Y}(t)$.

An alternative scheme relies on the value function

$$V(\tau, x) = \min_u \{\varphi_0(t_0, x(t_0))\}. \tag{7}$$

The minimum in (7) is over all $u(t) \in \mathcal{P}(t),\ t \in [t_0, \tau]$, under restrictions $x(\tau) = x$ and $\varphi(t, x) \le 1,\ t \in [t_0, \tau]$. We will use this formulation for the case when

$$\varphi_0(t_0, x) = (x - x^0, (X^0)^{-1}(x - x^0))^{1/2},\ X^0 = X^{0\prime} > 0,$$

$$\varphi(t, x) = (x - y(t), (Y(t))^{-1}(x - y(t)))^{1/2},\ Y(t) = Y'(t) > 0,$$

so that the initial condition and the state constraints are ellipsoidal:

$$x(t_0) \in \mathcal{E}(x^0, X^0) = \{x : \varphi_0(t_0, x) \le 1\},$$

$$x(t) \in \mathcal{E}(y(t), Y(t)) = \{x : \varphi(t, x(t)) \le 1,\ \ t \in [t_0, \tau]\}.$$

Once $x(\tau; t_0, x(t_0)) = x(\tau) = x$ is fixed, one may figure out whether $x \in \mathcal{X}(\tau, t_0, \mathcal{X}^0)$ by looking at the value $V(\tau, x)$. Then the respective vector $x(t_0) \in \mathcal{E}(x^0, X^0)$ iff there exists a control $u(t),\ t \in [t_0, \tau]$, that ensures $V(\tau, x) \le 1$, so that $x(\tau) \in \mathcal{X}(\tau, t_0, \mathcal{X}^0)$ iff $V(\tau, x) \le 1$.

Therefore the reach set at time τ is

$$\mathcal{X}(\tau, t_0, \mathcal{X}^0) = \{x \mid V(\tau, x) \le 1\},$$

which is thus a level set of $V(\tau, x)$, and the previously stated *semigroup property* of the value function (the principle of optimality) holds again.

The solution of the reachability problem can again be cast in the form of a solution of the forward HJB equation—now somewhat different from (6). Again assume the functions $V(t, x)$ and $\varphi(t, x)$ are *differentiable*.

Denote

$$\mathcal{H}(t, x, V, u) = V_t(t, x) + (V_x(t, x), f(t, x, u)). \tag{8}$$

LEMMA 2.2. *The formal HJB equation is*

$$\max_{u \in \mathcal{P}(t)} \mathcal{H}(t, x, V, u) = 0 \ \textit{if}\ \varphi(t, x) < 1, \tag{9}$$

and

$$\max_{u \in \mathcal{P}_s(t)} \mathcal{H}(t, x, V, u) = 0 \ \textit{if}\ \varphi(t, x) = 1, \tag{10}$$

with

$$\mathcal{P}_s(t) = \mathcal{P}(t) \cap \{u \mid \mathcal{H}(t, x, \varphi, u) \le 0\}, \tag{11}$$

and with the boundary condition

$$V(t_0, x) = \varphi_0(t_0, x). \tag{12}$$

We sketch a proof of Lemma 2.2.

Together with $\varphi(t, x) < 1$, the principle of optimality (5), with $t_0 = \tau - \sigma$, gives

$$0 = \min_u \{V(\tau - \sigma, x(\tau - \sigma)) - V(\tau, x) \mid u \in \mathcal{P}(t),\ t \in [\tau - \sigma, \tau]\} \tag{13}$$

or

$$\max_{u \in \mathcal{P}(t)} \{\mathcal{H}(t, x, V, u)\} = \mathcal{H}(t, x, V, u^0) = 0. \tag{14}$$

With $\varphi(t,x)=1$ we apply the same principle but only through those controls that do not allow the trajectory to move outside the state constraint. These are

$$\mathcal{P}_s=\{u\in\mathcal{P}(t)\cap\{u\mid\mathcal{H}(t,x,\varphi,u)\le 0\}\}.$$

Let $u^0(t,x)$ be the optimal control for the trajectory $x^0(t)$ that starts at $x(\tau)=x$ and minimizes $\varphi_0(t_0,x(t_0))$ under constraints $\varphi(t,x(t))\le 1,\ t\in[t_0,\tau]$.

Note that with $\varphi(t,x(t))=1$ we have two cases: either $\mathcal{H}(t,x(t),\varphi,u^0)=0$, which means the related optimal trajectory runs along the state constraint boundary, or

$$d\varphi(t,x(t))/dt|_{u=u^0}=\mathcal{H}(t,x(t),\varphi,u^0)<0,$$

so the optimal trajectory departs from the boundary, and for $\sigma>0$ we have $\varphi(t+\sigma,x(t+\sigma))<1$. Relations $\mathcal{H}(t,x(t),\varphi,u^0)\equiv 0$ and (14) allow one to find the control u^0 along the state constraint boundary.

If all the operations in (6), (9), (10) result in smooth functions, then these equations may have a classical solution [5]. Otherwise (9), (10) form a symbolic generalized HJB equation, which has to be described in terms of subdifferentials, Dini derivatives, or their equivalents.

However, the typical situation is that V *is not differentiable*. The treatment of (6), (9), (10) then involves the notion of a generalized "viscosity" solution for these equations or their equivalents [5, 30]. One approach is to use the method of characteristics as developed for this type of equation [3, 4, 30]. But it is a fairly complicated procedure, especially in the nonsmooth case for which the method requires additional refinement. Another approach is to look for the solution through level set methods [29].

However, for the specific "linear-convex" problems of this paper the function $V(t,x)$ is indeed differentiable. Moreover, in this case an effective ellipsoidal technique may be applied, which allows one to bypass calculation of solutions to the HJB equation.

3. Linear-convex systems. Consider the linear system

$$\dot{x}=A(t)x+B(t)u,\quad t_0\le t\le t_1, \tag{15}$$

in which $A(t),B(t)$ are continuous, and the system is completely controllable (see [18]). The control set $\mathcal{P}(t)$ is a nondegenerate ellipsoid, $\mathcal{P}(t)=\mathcal{E}(p(t),P(t))$, and

$$\mathcal{E}(p(t),P(t))=\{u\mid(u-p(t),P(t)(u-p(t)))\le 1\}, \tag{16}$$

with $p(t)\in\mathbb{R}^p$ (the center of the ellipsoid) and the symmetric positive definite matrix function $P(t)\in\mathbb{R}^{p\times p}$ (the shape matrix of the ellipsoid) continuous in t. The *support function* of the ellipsoid is

$$\rho(l\mid\mathcal{E}(p(t),P(t)))=\max\{(l,u)\mid u\in\mathcal{E}(p(t),P(t))\}=(l,p(t))+(l,P(t)l)^{1/2}.$$

The *state constraint* is

$$x(t)\in\mathcal{Y}(t),\quad t\in[t_0,t_1], \tag{17}$$

in which $\mathcal{Y}(t)$ is an ellipsoidal-valued function $\mathcal{Y}(t)=\mathcal{E}(y(t),Y(t))$, $Y(t)=Y'(t)>0, Y\in\mathbb{R}^{n\times n}$, and with $y(t),Y(t)$ absolutely continuous. Lastly, $\mathcal{X}^0=\mathcal{E}(x^0,X^0)$ is also an ellipsoid.

In [12] and [15] it is also assumed that the constraints on the controls and initial values are ellipsoidal, which makes the explanations more transparent. However, these methods are applicable to box-valued constraints as well [15, 8].

LEMMA 3.1. *For the linear system* (15), *with convex-valued restrictions* (3), *the reach set* $\mathcal{X}(\tau, t_0, \mathcal{X}^0)$ *at time* τ *is convex and compact.*

Problem 3.1 for calculating the reach set is now formulated as follows.

Problem 3.1. Calculate the support functions $\rho(l \mid \mathcal{X}(\tau, t_0, \mathcal{X}^0)),\ l \in \mathbb{R}^n$.

This is equivalent to solving the following optimal control problem with state constraints:

$$\rho(l \mid \mathcal{X}(\tau, t_0, \mathcal{X}^0)) = \max(l, x(\tau))$$

$$\text{subject to } x(t_0) \in \mathcal{X}^0,\ u(t) \in \mathcal{P}(t),\ x(t) \in \mathcal{Y}(t),\ t_0 \le t \le \tau.$$

For each vector l the solution to the optimal control problem is attained at the terminal point $x^{(0)}(\tau)$ of an *optimal trajectory* $x^{(0)}(t),\ t \in [t_0, \tau]$, which starts from a point $x^{(0)}(t_0) \in \mathcal{X}^0$ determined throughout the solution process.

We will solve Problem 3.1 by calculating the value function (7), namely,

$$V(\tau, x) = \min_u \{ d(x(t_0), \mathcal{E}(x^0, X^0)) \mid x(\tau, x) = x,$$
$$u(t) \in \mathcal{P}(t),\ x(t) \in \mathcal{E}(y(t), Y(t)),\ t \in [t_0, \tau] \},$$

for system (15), using the techniques of convex analysis [11, 9].

To handle the state constraints $\mathcal{X}[\tau]$, one usually imposes the following *constraint qualification.*

Assumption 3.1. There exist a control $u(t) \in \mathcal{P}(t),\ t \in [t_0, \tau]$, $x^0 \in \mathcal{X}^0$, and $\varepsilon > 0$, such that the trajectory $x[t] = x(t, t_0, x^0) = x(t, t_0, x^0 \mid u(\cdot))$ satisfies

$$x(t, t_0, x^0) + \epsilon \mathcal{B}_n(0) \subset \mathcal{Y}(t),\ t \in [t_0, \tau].$$

Here $\mathcal{B}_n(0) = \{z : (z, z) \le 1,\ \ z \in \mathbb{R}^n\}$. Under Assumption 3.1 this gives

$$V(\tau, x) = (l, x) - \mathbf{\Psi}(\tau, t_0, l),$$

where

$$\mathbf{\Psi}(\tau, t_0, l) = \min\{\Psi(\tau, t_0, l, \Lambda(\cdot)) \mid \Lambda(\cdot)\}, \tag{18}$$

$$\Psi(\tau, t_0, l, \Lambda(\cdot)) = s(t_0, \tau, l \mid \Lambda(\cdot))x^0 + (s(t_0, \tau, l \mid \Lambda(\cdot))X^0 s'(t_0, \tau, l \mid \Lambda(\cdot)))^{1/2}$$
$$+ \int_{t_0}^{\tau} \Big(s(t, \tau, l \mid \Lambda(\cdot))B(t)p(t) + (s(t, \tau, l \mid \Lambda(\cdot))B(t)P(t)B'(t)s'(t, \tau, l \mid \Lambda(\cdot)))^{1/2} \Big) dt$$
$$+ \int_{t_0}^{\tau} d\Lambda(t) y(t) + (d\Lambda(t) Y(t) d\Lambda(t))^{1/2}.$$

Here, $s(t, \tau, l \mid \Lambda(\cdot)),\ t \le \tau$, is the *row-vector* solution to the adjoint equation

$$ds = -sA(t)dt + d\Lambda(t),\ \ s(\tau) = l, \tag{19}$$

and $\Lambda(\cdot) \in \mathbf{V}_n[t_0, \tau]$ is the space of n-dimensional functions of bounded variation on $[t_0, \tau]$. We also used the notation

$$\int_{t_0}^{\tau} (d\Lambda(\cdot) Y(t) d\Lambda(\cdot))^{1/2} = \max \left(\int_{t_0}^{\tau} d\Lambda(t) h(t) \mid h(t) \in \mathcal{E}(0, Y(t)), t \in [t_0, \tau] \right).$$

This is the maximum of a Stieltjes integral over continuous functions $h(t) \in \mathcal{E}(0, Y(t))$. The minimum Ψ over $\Lambda(\cdot)$ is reached because of Assumption 3.1.

A direct consequence of formula (18) is the solution to Problem 3.1.

THEOREM 3.1. *Under Assumption* 3.1, *the support function* $\rho(l \mid \mathcal{X}[\tau])$ *is given by*

$$\rho(l \mid \mathcal{X}[\tau]) = \mathbf{\Psi}(\tau, t_0, l) = \min\{\Psi(\tau, t_0, l, \Lambda(\cdot)) \mid \Lambda(\cdot) \in \mathbf{V}_n[t_0, \tau]\}. \tag{20}$$

Let $\Lambda^{(0)}(\cdot) = \arg\min\{\Psi(\tau, t_0, l, \Lambda(\cdot)) \mid \Lambda(\cdot) \in \mathbf{V}_n[t_0, \tau]\}$ be the minimizer of (20) and let $s^{(0)}[t] = s(t, \tau, l \mid \Lambda^{(0)}(\cdot))$ be the solution to (19) with $\Lambda(\cdot) = \Lambda^{(0)}(\cdot)$.

The application of formula (18) indicates a new approach to the next proposition. Alternate derivations of this approach can be found in [11] and [9, pp. 116–121].

THEOREM 3.2 (the "standard" maximum principle under state constraints). *For Problem* 3.1 *under Assumption* 3.1, *the optimal control* $u^{(0)}(t)$, *initial condition* $x^{(0)}$, *and trajectory* $x^{(0)}[t] = x(t, t_0, x^{(0)} | u^{(0)}(\cdot))$ *must satisfy the "maximum principle"*

$$s^{(0)}[t]B(t)u^{(0)}(t) = \max\{s^{(0)}[t]B(t)u \mid u \in \mathcal{E}(p(t), P(t))\}, \tag{21}$$

and the "maximum conditions"

$$\int_{t_0}^{\tau} (d\Lambda^{(0)}(t)y(t) + (d\Lambda^{(0)}(t)Y(t)d\Lambda^{(0)\prime}(t))^{1/2}) = \int_{t_0}^{\tau} d\Lambda^{(0)}(t)x^{(0)}[t] \tag{22}$$

$$= \max\left\{\int_{t_0}^{\tau} (d\Lambda^{(0)}(t)z(t)) \mid z(t) \in \mathcal{E}(y(t), Y(t))\right\},$$

$$s^{(0)}[t_0]x^{(0)}[t_0] = s^{(0)}[t_0]x^0 + (s^{(0)}[t_0], X^0 s^{(0)}[t_0])^{1/2} = \max\{s^{(0)}[t_0]x \mid x \in \mathcal{E}(x^0, X^0)\}. \tag{23}$$

Here we note that the minimum over $\Lambda(\cdot) \in \mathbf{V}_n[t_0, \tau]$ in (20) may be replaced with the minimum over the pair $\{M(t), \lambda(t)\}$, with $d\Lambda(t) = l'M(t)d\lambda(t)$, in which the $n \times n$ matrix $M(t)$ is continuous, and $\lambda(\cdot) \in \mathbf{V}_1[t_0, \tau]$ is a scalar function of bounded variation. Moreover, $M(t)$ may be chosen within a compact set $\mathcal{C}_0$ of continuous functions. This new form of the multiplier $d\Lambda(t)$ is a result of combining the earlier schemes of [6, 10]. We summarize this result as follows.

LEMMA 3.2. *The multiplier* $\Lambda^{(0)}(t)$ *allows the representation*

$$d\Lambda^{(0)}(t) = l'M^{(0)}(t)d\lambda^{(0)}(t), \tag{24}$$

with $M^{(0)}(\cdot) \in \mathcal{C}_0,\ \lambda^{(0)}(\cdot) \in \mathbf{V}_1[t_0, \tau]$.

Denote by $S_M(t, \tau) = S(t, \tau \mid M(\cdot), \lambda(\cdot))$ the solution of the matrix equation

$$dS = -SA(t)dt + M(t)d\lambda(t),\ \ S(\tau) = I. \tag{25}$$

This is a symbolic expression for the linear differential equation whose solution for fixed τ is

$$S_M(t, \tau) = \exp\left(\int_t^{\tau} A(s)ds\right) - \int_t^{\tau} \left(\exp\int_t^{s} A(\xi)d\xi\right)M(s)d\lambda(s),$$

in which the second integral is a Stieltjes integral.

As in the scheme of [10], the single-valued functional $\Psi(\tau, t_0, l, \Lambda(\cdot))$ may be substituted with the set-valued integral

$$R(\tau, t_0, M(\cdot), \lambda(\cdot), \mathcal{X}^0) = S(t_0, \tau \mid M(\cdot), \lambda(\cdot))\mathcal{E}(x^0, X^0)$$
$$+ \int_{t_0}^{\tau} S(t, \tau \mid M(\cdot), \lambda(\cdot))B(t)\mathcal{E}(p(t), P(t))dt + \int_{t_0}^{\tau} d\lambda(t)M(t)\mathcal{E}(y(t), Y(t))dt, \tag{26}$$

which yields the next result whose analytical form differs from that in [10].

LEMMA 3.3. *The following equalities hold:*

$$\mathcal{X}[\tau] = \cap\{R(\tau, t_0, M(\cdot), \lambda(\cdot), \mathcal{X}^0) \mid M(\cdot), \lambda(\cdot)\}$$
$$= \cap\{R(\tau, t_0, M(\cdot), \lambda(\cdot), \mathcal{X}^0) \mid M(\cdot), \lambda \equiv t\},$$

for $M(\cdot) \in \mathcal{C}_0, \lambda \in \mathbf{V}_1[t_0, \tau]$.

Note that $l'M(t)d\lambda(t) = d\Lambda(\cdot)$ is the Lagrange multiplier responsible for the state constraint, with $M(t)$ continuous in t, and $\lambda(t)$, with bounded variation, responsible for the jumps of the multiplier $\Lambda(\cdot)$. Such properties of the multipliers are due to the linearity of the system and the type of the constraints on u, x, taken in this paper (see [9, 6]). In general, the multipliers responsible for the state constraints may be represented through a measure of general type and therefore may contain singular components. This is not the case, however, in the problems of this paper *due to the next assumption which is now required.*

Assumption 3.2. The multipliers $\Lambda^{(0)}(t), \lambda^{(0)}(t)$ do not have singular components.

LEMMA 3.4. *Assumption* 3.2 *holds if the support function* $\rho(q \mid \mathcal{Y}(t))$ *is absolutely continuous in* t*, uniformly in* $q, (q, q) \leq 1$*, and for each* l *the optimal trajectory for Problem* 3.1 *visits the boundary* $\partial\mathcal{Y}(t)$ *for a finite number of intervals.*

Note that Lemma 3.3 gives the reach set at time τ as an intersection of sets $R(\tau, t_0, M(\cdot), \lambda(\cdot), \mathcal{X}^0)$ parametrized by functions $M(\cdot) \in \mathcal{C}_0$. Theorem 3.1 or Lemma 3.3 may be used to calculate the reach set $\mathcal{X}[\tau]$ *for any fixed time* τ.

However, on many occasions our objective is to recursively calculate the whole tube $\mathcal{X}[\tau], \tau \geq t_0$. This means that while solving Problem 2.1 for increasing values of τ, we want a procedure that does not require one to solve the whole problem "afresh" for each new value of τ, but allows the use of the solutions for previous values.

There is a difficulty here, namely, given $M(\cdot), \lambda(\cdot)$, one may observe that in general, $R(\tau, t_0, M(\cdot), \lambda(\cdot), \mathcal{X}^0) \neq R(\tau, t, M(\cdot), \lambda(\cdot), R(t, t_0, M(\cdot), \lambda(\cdot), \mathcal{X}^0))$, which means R given by (26) does not satisfy the semigroup property. Therefore Theorem 3.2 and Lemmas 3.2 and 3.3 have to be modified to meet the recursion requirements. Such a move would yield a modified version of the maximum principle under state constraints for linear systems that eliminates the last difficulty.

4. The modified maximum principle. To achieve the desired results, in the following sections we shall use some properties related to the structure of the optimal controls and the state constraints. These properties and additional assumptions are summarized in the appendix and are typical of the problems under discussion. They are borrowed from previous investigations.

Starting with Assumptions 3.1 and 3.2 and presuming the conditions of Lemma 3.4 are true, let us first restrict Λ in (24) to satisfy the relation $d\Lambda(t) = l'M(t)d\lambda(t)$, in which $\lambda(t)$ is *absolutely continuous.* According to Property A.1 of the appendix this means $d\Lambda(t) = l'M(t)\eta(t)dt$, and $\eta(t) = d\lambda(t)/dt \equiv 0$ for $t \in T_Y = \{t \mid x_l^{(0)}(t) \in int\mathcal{Y}(t)\}$, when $x_l^{(0)}(t)$ is the optimal trajectory for Problem 3.1, for the given l.

Denoting $L(t) = S^{-1\prime}(t,\tau \mid M(\cdot),\eta(\cdot))M(t)\eta(t)$, we may replace

$$dS/dt = -SA(t) + M(t)\eta(t), \quad S(t_0,t_0 \mid M(\cdot),l(\cdot)) = I, \tag{27}$$

whose solution is $S(t,\tau \mid M(\cdot),l(\cdot))$, with

$$dS_L/dt = -S_L(A(t) - L(t)), \quad S_L(\tau,\tau) = I,$$

whose solution is $S_L(t,\tau)$. Then $S(t,\tau \mid M(\cdot),l(\cdot)) \equiv S_L(t,\tau),\ t \in [t_0,\tau]$.

In this case, $R(\tau,t_0,M(\cdot),\eta(\cdot),\mathcal{X}^0)$ transforms into

$$\mathcal{R}(\tau,t_0,L(\cdot)) = S_L(t_0,\tau)\mathcal{E}(x^0,X^0) + \int_{t_0}^{\tau} S_L(t,\tau)(\mathcal{E}(B(t)p(t),B(t)P(t)B'(t))$$

$$+L(t)\mathcal{E}(y(t),Y(t)))dt = \mathcal{X}(\tau,t_0,L(\cdot),\mathcal{X}^0) = \mathcal{X}_L[\tau],$$

and $\mathcal{X}_L[\tau]$ turns out to be the solution to the differential inclusion

$$\dot{x}(t) \in (A(t) - L(t))x + L(t)\mathcal{E}(y(t),Y(t)) + \mathcal{E}(B(t)p(t),B(t)P(t)B'(t)), \tag{28}$$

$$t \geq t_0, \quad x^0 \in \mathcal{E}(x^0,X^0). \tag{29}$$

Here also the compact set $\mathcal{C}_0$ of functions $M(t)$ transforms into a compact set $\mathcal{C}_{00}$ of functions $L(t)$.

We thus arrive at the following important property, which is similar to those proved in [10].

LEMMA 4.1. *The reach set $\mathcal{X}[\tau]$ is the intersection*

$$\mathcal{X}[\tau] = \cap\{\mathcal{X}_L(\tau,t_0,\mathcal{E}(x^0,X^0)) \mid L(\cdot)\} \tag{30}$$

of the "cuts" or "sections" $\mathcal{X}_L(\tau,t_0,\mathcal{E}(x^0,X^0))$ of the reach tubes (solution tubes) $\mathcal{X}_L(\cdot) = \{\mathcal{X}_L[t] : t \geq t_0\}$ of the differential inclusion (28), (29). *The intersection is over all $L(\cdot) \in \mathcal{C}_{00}$, a compact set of continuous matrix functions $L(t), t \in [t_0,\tau]$.*

A further calculation using convex analysis yields the next formula.

THEOREM 4.1. *The support function*

$$\rho(l \mid \mathcal{X}(\tau,t_0,\mathcal{E}(x^0,X^0))) = \inf\{\rho(l \mid \mathcal{X}_L(\tau,t_0,\mathcal{E}(x^0,X^0))) \mid L(\cdot)\}, \tag{31}$$

with

$$\rho(l \mid \mathcal{X}_L(\tau,t_0,\mathcal{E}(x^{(0)},X^0))) = (l,x_L^\star[t]) + (l,G_L(\tau,t_0)X^0G_L'(\tau,t_0)l)^{1/2} \tag{32}$$

$$+\int_{t_0}^{\tau}(l,G_L(\tau,s)B(s)P(s)B'(s)G_L'(\tau,s)l)^{1/2}ds$$

$$+\int_{t_0}^{\tau}(l,G_L(\tau,s)L(s)Y(s)L'(s)G_L'(\tau,s)l)^{1/2}ds.$$

Here

$$x_L^\star[t] = G_L(t,t_0)x^0 + \int_{t_0}^{t} G_L(t,s)B(s)p(s)ds,$$

and $G_L(\tau,s) = S_L(s,\tau)$ is the transition matrix for the homogeneous system

$$\dot{x} = (A(t) - L(t))x\,; \;\; G_L(t,t) = I, \;\; L(\cdot) \in \mathcal{C}_{00}.$$

The significance of the last result is that in this specific problem the support function of the intersection (30) is equal to the *pointwise infimum* (31) of the support functions rather than to their *infimal convolution* as given by general theory [25].

It is not unimportant to specify when the infimum in (31) is attained; that is, it is actually a minimum. Indeed, it may happen that for a given l, the minimum over $L(\cdot)$ is in the class $\mathcal{C}_{00}$ (this, for example, is the case when $\lambda(t)$ turns out to be absolutely continuous, as in the above). But to ensure the minimum is always reached, we have to broaden the class of functions $L(\cdot)$.

To illustrate how to continue the procedure we will assume the following.

Assumption 4.1. For each $l \in \mathbb{R}^n$, the optimal trajectory $x_l^0(t)$ of Problem 3.1 visits the boundary $\partial\mathcal{Y}(t)$ only during one time interval $[t_1,t_2]$, $t_0 \le t_1$, $t_2 \le \tau$.
(The case of finite or countable collection of such intervals is treated in a similar way.)

Then, instead of the product $M(t)l(t)$, in (27) we must deal with multipliers of the form $\mathcal{M}(t) = M(t)l(t) + M_1\delta(t-t_1) + M_2\delta(t-t_2)$, where M_1, M_2 are $n \times n$ matrices.

By introducing a new multiplier $\mathcal{L}(t) = L(t) + L_1\delta(t-t_1) + L_2\delta(t-t_2)$ under transformation $\mathcal{L}(t) = S^{-1}(t,\tau \mid \mathcal{M}(\cdot))\mathcal{M}(t)$, we shall match the formulas for $R(\tau,t_0,\mathcal{M}(\cdot))$ and its transformed version $\mathcal{R}(\tau,t_0,\mathcal{L}(\cdot))$.

Applying the schemes of [6, 10] to the specific case of this paper, it is possible to rewrite the preceding assertions.

LEMMA 4.2. *The support function*

(33)
$$\rho(l \mid \mathcal{X}[\tau]) = \rho(l \mid \mathcal{X}(\tau,t_0,\mathcal{E}(x^0,X^0))) = \min\{\rho(l \mid \mathcal{X}_{\mathcal{L}}(\tau,t_0,\mathcal{E}(x^0,X^0)) \mid \mathcal{L}(\cdot)\},$$

with

$$\rho(l \mid \mathcal{X}_{\mathcal{L}}(\tau,t_0,\mathcal{E}(x^0,X^0)) = \Phi(l,\mathcal{L}(\cdot),\tau,t_0), \tag{34}$$

and

$$\begin{aligned}\Phi(l,\mathcal{L}(\cdot),\tau,t_0) &= (l,x_*[t]) + (l,\mathcal{G}(\tau,t_0)X^0\mathcal{G}'(\tau,t_0)l)^{1/2} \\ &+ \int_{t_0}^{\tau} (l,\mathcal{G}(\tau,s)B(s)P(s)B'(s)\mathcal{G}'(\tau,s)l)^{1/2}ds \\ &+ \int_{t_0}^{\tau} (l,\mathcal{G}(\tau,s)L(s)Y(s)L'(s)\mathcal{G}'(\tau,s)l)^{1/2}ds \\ &+ \int_{t_0}^{\tau} (l,\mathcal{G}(\tau,s)L_1Y(s)L_1'\mathcal{G}'(\tau,s)l)^{1/2}d\chi(s,t_1) \\ &+ \int_{t_0}^{\tau} (l,\mathcal{G}(\tau,s)L_2Y(s)L_2'\mathcal{G}'(\tau,s)l)^{1/2}d\chi(s,t_2).\end{aligned} \tag{35}$$

Here $\mathcal{G}(\tau, s)$ is the transition matrix for the homogeneous system

$$dx = (A(t) - L(t))xdt - L_1 x d\chi(s, t_1) - L_2 x d\chi(s, t_2),$$

namely,

$$\mathcal{G}(\tau, s) = \exp\left(\int_s^\tau (A(\xi) - L(\xi))dt - \int_s^\tau (L_1 d\chi(s, t_1) + L_2 d\chi(s, t_2))\right),$$

so that

$$x(t) = \mathcal{G}(\tau, t_0)x^0,$$

and $\chi(s, t')$ is the step function,

$$\chi(s, t') \equiv 0, s < t';\ \chi(s, t') \equiv 1, s \geq t',\ d\chi(s, t')/ds = \delta(s - t').$$

The vector $x_*[t]$ in (35) may be described by

$$\begin{aligned} dx_*[t] &= ((A(t) - L(t))x_*[t] + B(t)p(t) + L(t)y(t))dt, \\ -L_1(x_*[t] - y(t))d\chi(s, t_1) &- L_2(x_*[t] - y(t))d\chi(s, t_2), x_*[t_0] = x^0. \end{aligned} \tag{36}$$

Remark 4.1. Note that in the exponent $\mathcal{G}(t, s)$ we have the difference of a Riemann integral and a Riemann–Stieltjes integral. On the other hand, the last two integrals in (35) should be interpreted as Lebesgue–Stieltjes integrals [20, 26]. This does not cause additional difficulty since the multipliers L_1, L_2 are among the optimizers in (33).

A result similar to (30) is all the more true for the sets $\mathcal{X}_{\mathcal{L}}(\tau, t_0, \mathcal{E}(x^0, X^0))$.

COROLLARY 4.1. *The following intersection holds:*

$$\mathcal{X}[\tau] = \mathcal{X}(\tau, t_0, \mathcal{E}(x^0, X^0)) = \cap\{\mathcal{X}_{\mathcal{L}}(\tau, t_0, \mathcal{E}(x^0, X^0)) \mid \mathcal{L}(\cdot)\}.$$

The difference between Lemmas 4.1 and 4.2 is that in the former it is not guaranteed that the boundary $\partial\mathcal{X}[\tau]$ is touched at each point by one of the intersecting sets $\mathcal{X}_L(\tau, t_0, \mathcal{E}(x^0, X^0))$, whereas in the latter the boundary $\partial\mathcal{X}[\tau]$ is indeed touched at each point by one of the sets $\mathcal{X}_{\mathcal{L}}(\tau, t_0, \mathcal{E}(x^0, X^0))$. This is because the minimum in (33) is attained.

Under Assumptions 3.1, 3.2, 4.1, and Assumption A.1 of the appendix, the reasoning above leads to the next result.

LEMMA 4.3. *The reach set $\mathcal{X}[t] = \mathcal{X}(t, t_0, \mathcal{E}(x^0, X^0))$ is a convex compact set in $\mathbb{R}^n$ which evolves continuously in t.*

The boundary of the reach set $\mathcal{X}[t]$ has an important characterization. Consider a point x^* on the boundary $\partial\mathcal{X}[\tau]$ of the reach set $\mathcal{X}[\tau] = \mathcal{X}(\tau, t_0, \mathcal{E}(x^0, X^0))$.[1]

Then there exists a *support vector* z^* such that

$$(z^*, x^*) = \rho(z^* \mid \mathcal{X}[\tau]).$$

Let $\mathcal{L}^0(\cdot)$ be the minimizer for the problem (see (33))

$$\rho(z^* \mid \mathcal{X}[\tau]) = \min\{\Phi(z^*, \mathcal{L}(\cdot), \tau, t_0) \mid \mathcal{L}(\cdot)\} = \Phi(z^*, \mathcal{L}^0(\cdot), \tau, t_0). \tag{37}$$

[1]The boundary $\partial\mathcal{X}[\tau]$ of $\mathcal{X}[\tau]$ may be defined as the set $\partial\mathcal{X}[\tau] = \mathcal{X}[\tau] \setminus \text{int}\mathcal{X}[\tau]$. Under the controllability assumption, $\mathcal{X}[\tau]$ has a nonempty interior, $\text{int}\mathcal{X}[\tau] \neq \emptyset$, $\tau > t_0$.

Then the control $u = u^*(t)$, the initial state $x(t_0) = x^{*0} \in \mathcal{E}(x^0, X^0)$, and the corresponding trajectory $x^*(t)$, along which system (15) is transferred from state $x^*(t_0) = x^{*0}$ to $x(\tau) = x^*$, are specified by the following "modified maximum principle."

THEOREM 4.2 (the modified maximum principle under state constraints). *For Problem* 3.1, *suppose state* x^* *is such that*

$$(z^*, x^*) = \rho(z^* \mid \mathcal{X}[\tau]).$$

Then the control $u^*(t)$, *which steers the system* (15) *from* $x^*(t_0) = x^{*0}$ *to* $x^*(\tau) = x^*$ *under constraints* $u(t) \in \mathcal{E}(p(t), P(t))$, $x(t) \in \mathcal{E}(y(t), Y(t))$ *while ensuring*

$$(z^*, x^*) = \max\{(z^*, x) \mid x \in \mathcal{X}[\tau]\},$$

satisfies the following pointwise "maximum principle" for the control ($s \in [t_0, \tau]$)*:*

$$(38)\quad (z^{*\prime}\mathcal{G}^0(\tau, s)B(s), u^*(s)) = \max_u\{(z^{*\prime}\mathcal{G}^0(\tau, s)B(s), u) \mid u \in \mathcal{E}(p(s), P(s))\}$$

$$= (z^*, \mathcal{G}^0(\tau, s)B(s)p(s)) + (z^*, \mathcal{G}^0(\tau, s)B(s)P(s)B'(s)\mathcal{G}^0(\tau, s)z^*)^{1/2},\ s \in [t_0, \tau],$$

and the "maximum conditions" for the system trajectory (pointwise),

$$(39)\quad (z^{*'}\mathcal{G}^0(\tau, s)\mathcal{L}^0(s), x^*(s)) = \max_p\{(z^{*'}\mathcal{G}^0(\tau, s)\mathcal{L}^0(s), p) \mid p \in \mathcal{E}(y(s), Y(s))\}$$

$$= (z^{*'}\mathcal{G}^0(\tau, s)\mathcal{L}^0(s)y(s)) + (z^*, \mathcal{G}^0(\tau, s)\mathcal{L}^0(s)Y(s)\mathcal{L}^{0\prime}(s)\mathcal{G}^0(\tau, s)z^*)^{1/2},$$

and the initial state,

$$(40)\qquad (z^*, \mathcal{G}^0(\tau, t_0)x^{*0}) = \max\{(z^*, x) \mid x \in \mathcal{G}^0(\tau, t_0)\} = \rho(z^* \mid \mathcal{E}(x^0, X^0))$$

$$= (z^*, \mathcal{G}^0(\tau, t_0)x^0) + (z^*, \mathcal{G}^0(\tau, t_0)X^0\mathcal{G}^{0\prime}(\tau, t_0)z^*)^{1/2}.$$

Here $\mathcal{G}^0(\tau, s)$ *stands for the matrix function* $\mathcal{G}(\tau, s)$ *taken for* $\mathcal{L}^0(s)$*—the minimizer of problem* (37).

The function $h(\tau, s)) = l_*{}'\mathcal{G}^0(\tau, s)B(s)$ is taken right-continuous.[2]

Remark 4.2. Suppose we want to solve Problem 3.1, seeking $\rho(l^*(t) \mid \mathcal{X}[t])$ along a curve $l^*(t), t > t_0$. Then, taking $l^*(t) = l'_*(\mathcal{G}^0)^{-1}(t, s)$, one may observe that the *integrands* in functional $\Phi(l^*(t), \mathcal{L}^0(\cdot), t, t_0)$ of (33)–(35) will be independent of t. This property ensures the existence of a *recursive* computational procedure, as indicated in the next section (see also [15]). The modified maximum principle of this section thus allows a solution in *recursive form.* This is not the case for the standard maximum principle.

We now pass to the construction of ellipsoidal approximations for the reach sets.

[2]In the general case, under Assumption 3.2, the optimal trajectory may visit the smooth boundary of the state constraint for a countable set of closed intervals, and the function $L^0_*(\cdot)$ allows not more than a countable set of discontinuities of the first order.

5. External ellipsoids. Despite the linearity of the system, the direct calculation of reach sets is rather difficult. Among effective methods for such calculations are those that rely on ellipsoidal techniques [12, 15]. Indeed, although the initial set $\mathcal{E}(x^0, X^0)$ and the control set $\mathcal{E}(q(t), Q(t))$ are ellipsoids, the reach set $\mathcal{X}[t] = \mathcal{X}(t, t_0, \mathcal{E}(x^0, X^0))$ will of course *not* generally be an ellipsoid.

As shown in [12, 15], in the absence of state constraints the reach set $\mathcal{X}[t]$ may be approximated both externally and internally by ellipsoids $\mathcal{E}_-$ and $\mathcal{E}_+$, with $\mathcal{E}_- \subseteq \mathcal{X}[t] \subseteq \mathcal{E}_+$. Here we deal only with external approximations, but taken under state constraints.

An approximation $\mathcal{E}(x_+, X_+)$ is said to be *tight* if there exists a vector $z \in \mathbb{R}^n$ such that $\rho(z \mid \mathcal{E}(x_+, X_+)) = \rho(z \mid \mathcal{X}[t])$ (the ellipsoid $\mathcal{E}(x_+, X_+)$ *touches* $\mathcal{X}[t]$ along direction z). We shall look for external approximations that are tight, on one hand, and are also recursively computable, on the other.

In order to apply ellipsoidal techniques to state-constrained problems, recall Lemma 4.1 and Corollary 4.1, which indicate how the reach set $\mathcal{X}[t] = \mathcal{X}(t, t_0, \mathcal{X}^0)$ may be presented as an intersection of reach sets $\mathcal{X}_L$ or $\mathcal{X}_\mathcal{L}$ for the system (28) or (36) *without state constraints*. We first study how to approximate sets $\mathcal{X}_L$ by ellipsoids.

To demonstrate the nature of the procedures suggested in this paper we shall introduce the ellipsoidal technique in two stages. First, transform the system coordinates in (15) from x to z according to the formula $z(t) = S(t, t_0)x(t)$, with

$$dS(t, t_0)/dt = -S(t, t_0)A(t), \quad S(t_0, t_0) = I.$$

Then system (15) and state constraint $x(t) \in \mathcal{Y}(t)$ transform into

(41)
$$\dot{z} = S(t, t_0)B(t)u = B_N(t)u, \quad z(t) \in S(t, t_0)\mathcal{E}(y(t), Y(t)), \; z(t_0) \in S(t, t_0)\mathcal{E}(x^0, X^0),$$

while the constraint on control $u \in \mathcal{P}(t) = \mathcal{E}(p(t), P(t))$ remains the same.

Returning to the *old notation*, we will deal with system

(42)
$$\dot{x} = B(t)u, \; x(t_0) \in \mathcal{E}(x^0, X^0), \; y(t) \in \mathcal{Y}(t).$$

Thus, without loss of generality, in the forthcoming formulas we assume $A(t) \equiv 0$.

Problem 5.1. Given a vector function $l^*(t)$, continuously differentiable in t, find external ellipsoids $\mathcal{E}_{L+}[t] = \mathcal{E}(x_L^*(t), X_L^*(t)) \supset \mathcal{X}_L[t]$ such that *for all* $t \geq t_0$, the equalities

$$\rho(l^*(t) \mid \mathcal{X}_L[t]) = \rho(l^*(t) \mid \mathcal{E}_{L+}[t]) = (l^*(t), x_L^*(t))$$

hold, so that the supporting hyperplane for $\mathcal{X}_L[t]$ generated by $l^*(t)$, namely, the plane $(x - x_L^*(t), l^*(t)) = 0$ which touches $\mathcal{X}_L[t]$ at point $x^*(t)$, is also a supporting hyperplane for $\mathcal{E}_{L+}[t]$ and touches it at the same point.

Remark 5.1. Under Assumption 4.1 for the optimal trajectories $x^{(0)}(\cdot)$ of Problem 3.1 the boundary $\partial\mathcal{X}[\tau]$ of the reach set $\mathcal{X}[\tau]$ consists of three types of points, namely (I) those that are reached by $x^{(0)}(\tau)$ *without having visited the boundary of the state constraint*, then (II) those that are reached by $x^{(0)}(\tau)$ *after having visited the boundary of the state constraint*, and (III) those that lie *on the boundary of the state constraint*. Since case (I) has been investigated in detail in [15] and case (III) is trivial, our present investigation actually deals only with case (II).

Apart from Assumptions 3.1 and 3.2, let us assume in this section that the functions $L(\cdot)$ in what follows do not have any delta function components. (The case when such components are present is treated in the next section.)

The solution to Problem 5.1 is given within the following statement.

THEOREM 5.1. *With $l^*(t)$ given, the solution to Problem* 5.1 *is an ellipsoid $\mathcal{E}^*_{L+}[t] = \mathcal{E}_L(x^*_L(t), X^*_L[t])$, in which*

$$X^*_L[t] = \left(\int_{t_0}^{t} (p_u(t,s) + p_Y(t,s))ds + p_0(t,s)\right) \tag{43}$$

$$\times\left(\int_{t_0}^{t} (p_u(t,s))^{-1} G_L(t,s)B(s)P(s)B'(s)G'_L(t,s)ds\right.$$

$$\left. + \int_{t_0}^{t} (p_Y(t,s))^{-1} G_L(t,s)L_t(s)Y(s){L_t}'(s)G'_L(t,s)ds + p_0^{-1}(t)G_L(t,t_0)X^0G'_L(t,t_0)\right),$$

and

$$p_u(t,s) = (l^*(t), G_L(t,s)P(s)G'_L(t,s)l^*(t))^{1/2}, \tag{44}$$

$$p_Y(t,s) = (l^*(t), G_L(t,s)L_t(s)Y(s)L_tG'_L(t,s)l^*(t))^{1/2},$$

$$p_0(t) = (l(t_0), G_L(t,t_0)X^0G'_L(t,t_0)l(t_0))^{1/2},$$

with $x^*_L(t) = x^\star_L[t]$.

This result follows from [12, 15]. Since the calculations have to be made for all t, the parametrizing functions $p_u(t,s), p_Y(t,s), s \in [t_0,t], p_0(t)$ must also formally depend on t.

In other words, relations (43), (44) need to be calculated "afresh" for each t. It may be more convenient for computational purposes to have them in recursive form. As indicated in [15], in the absence of state constraints this could be done by selecting function $l^*(t)$ in an appropriate way. For the case of state constraints we follow Remark 4.2, arriving at the next assumption.

Assumption 5.1. The function $l^*(t)$ is of the form $l^{*'}(t) = l'_*G_L(t_0,t)$, with $l_* \in \mathbb{R}^n$ given.

Under Assumption 5.1 the function $l^*(t)$ is the solution to the differential equation

$$\dot{l}^* = L'(t)l^*,\ l^*(t_0) = l_*. \tag{45}$$

The tight external ellipsoids to $\mathcal{X}_L(t)$, which are

$$\mathcal{E}(x^*_L(t), X^*_L(t)) \supseteq \mathcal{X}_L(t) \supseteq \mathcal{X}[t],$$

may be described applying the formulas of [15, p. 189] to system (28). Each of these ellipsoids will be an external estimate for $\mathcal{X}[t]$.

But to describe the collection of tight ellipsoids for $\mathcal{X}[t]$ we have to ensure the following property:

$$\rho(l|\mathcal{X}[t]) = \min\{(l, x^*_L(t)) + (l, X^*_{L+}(t)l)^{1/2} \mid L(\cdot) \in \mathcal{C}_{00}\}. \tag{46}$$

Under our assumptions *the minimum over $L(\cdot)$* in (33) *is attained for any $l = l^*(t) = G'_L(t_0,t)l^* \in \mathbb{R}^n$*, the minimizers being denoted as $L^t_*(\cdot) \in \mathcal{C}_{00}$.[3]

[3]The same minimum value is also attained here in classes of functions broader than $\mathcal{C}_{00}$.

But prior to moving ahead we have to investigate the following. Suppose element $L_*^t(s)$ is the minimizer of functional

$$\Phi(G'_L(t_0,t)l_*, L(\cdot), t, t_0) = \mathbf{\Phi}(l_*, L(\cdot), t, t_0).$$

The question is, If we minimize function $\Phi(l_*, L(\cdot), t+\sigma, t_0), \sigma > 0$, over a larger interval $[t_0, t+\sigma]$ than before, will the minimizer $L_*^{t+\sigma}(s)$, for the latter problem, taken within $s \in [t_0, t]$, be the same as the minimizer $L_*^t(s), s \in [t_0, t]$ for $\mathbf{\Phi}(l_*, L(\cdot), t, t_0)$?

The answer to this question is given by the next lemma.

LEMMA 5.1. *Taking $l = l^*(t)$ according to Assumption* 5.1, *suppose $L_*^t(s), s \in [t_0, t]$ and $L_*^{t+\sigma}(s), s \in [t_0, t+\sigma], \sigma > 0$, are the minimizers of functionals $\mathbf{\Phi}(l, L(\cdot), t, t_0)$ and $\mathbf{\Phi}(l, L(\cdot), t+\sigma, t_0)$, respectively. Then*

$$L_*^t(s) \equiv L_*^{t+\sigma}(s), \quad s \in [t_0, t].$$

The proof is achieved by contradiction. Note that under Assumption 5.1 and due to Property A.1 of the appendix we may always take $L(t) \equiv 0$ when $t < \tau_1,\ t > \tau_2$, where τ_1, τ_2 are the points of arrival and departure at the state constraint.

Following Assumption 5.1, we proceed further by selecting $L(\cdot)$ to be the minimizer of functional $\Phi(l_*, L(\cdot), t, t_0)$, namely, we now take $L(s) = L_*^t(s) = L_*(s),\ s \in [t_0, t]$, which depends on l_*, s, but, as indicated in Lemma 5.1, does not depend on t.

Let $G_*(t,s)$ denote the transition matrix for the homogeneous system (45) when $L(t) \equiv L_*(t)$.

Then $p_u(t,s), p_Y(t,s), p_0(t)$ of (44) transform into

$$p_u(t,s) = (l_*, G_*(t_0,s)B(s)P(s)B'(s)G'_*(t_0,s)l_*)^{1/2} = p_u(s), \tag{47}$$

$$p_Y(t,s) = (l_*, G_*(t_0,s)L_*(s)Y(s)L_*{}'(s)G'_*(t_0,s)l_*)^{1/2} = p_Y(s),$$

$$p_0(t) = (l_*, X^0 l_*)^{1/2} = p_0;$$

matrix $X_L^*[t]$ transforms into

$$X_+^*[t] = \left(\int_{t_0}^t (p_u(s) + p_Y(s))ds + p_0\right) \tag{48}$$

$$\times\left(\int_{t_0}^t (p_u(s))^{-1} G_*(t_0,s)B(s)P(s)B'(s)G'_*(t_0,s)ds\right.$$

$$\left.+ \int_{t_0}^t (p_Y(s))^{-1} G_*(t_0,s)L_*(s)Y(s)L_*{}'(s)G'_*(t_0,s)ds + p_0^{-1}X^0\right);$$

and the function $\Phi(l^*(t), L_{*t}(\cdot), t, t_0)$ transforms into

$$\Phi(l_*, L_{*t}(\cdot), t, t_0) \tag{49}$$

$$= (l_*, x^*(t)) + (l_*, X^0 l_*)^{1/2} + \int_{t_0}^t (l_*, G_*(t_0,s)B(s)P(s)B'(s)G'_*(t_0,s)l_*)^{1/2}ds$$

$$+ \int_{t_0}^t (l_*, G_*(t_0,s)L_*(s)Y(s)L'_*(s)G'_*(t_0,s)l_*)^{1/2}ds,$$

with $x_L^*(t)$ transformed into

$$x_+^*[t] = x^0 + \int_{t_0}^t G_*(t_0,s)(B(s)p(s) + L_*(s)y(s))ds.$$

We may now differentiate $X_+^*[t], x_+^*[t]$. According to the appendix, the necessary condition (68) for a jump in $L_*(t)$ is not fulfilled, and therefore $L_*(t) \equiv \mathcal{L}^0(t)$.

Denoting

$$\pi_u(t) = p_u(t)\left(\int_{t_0}^t (p_u(s) + p_Y(s))ds + p_0\right)^{-1},$$

$$\pi_Y(t) = p_Y(t)\left(\int_{t_0}^t (p_u(s) + p_Y(s))ds + p_0\right)^{-1}, \tag{50}$$

and differentiating $X_+^*[t], x^*[t]$, we arrive at

$$\dot{X}_+^*[t] = (\pi_u(t) + \pi_Y(t))X_+^*[t] + (\pi_u(t))^{-1}G_*(t_0,t)B(t)P(t)B'(t)G_*{}'(t_0,t) \tag{51}$$
$$+(\pi_Y(t))^{-1}G_*(t_0,t)L_*(t)Y(t)L_*'(t)G_*{}'(t_0,t),$$

$$\dot{x}_+^*[t] = G_*(t_0,t)B(t)p(t) + L_*(t)y(t), \tag{52}$$

$$X_+^*[t_0] = X^0, \;\; x^*[t_0] = x^0. \tag{53}$$

Returning to the function $l^*(t) = G_L'(t_0,t)l_*$ with $G_L'(t,s) = G_*(t,s)$, we have $X_+[t] = G_*(t,t_0)X_+^*[t]G_*(t,t_0)$ and

$$\dot{X}_+[t] = (A(t) - L_*(t))X_+[t] + X_+[t](A(t) - L_*(t))' + G_*(t,t_0)\dot{X}_+^*[t]G_*(t,t_0), \tag{54}$$

so that

$$\dot{X}_+[t] = -L_*(t)X_+[t] - X_+[t]L_*(t) \tag{55}$$
$$+(\pi_u(t) + \pi_Y(t))X_+[t] + (\pi_u(t))^{-1}B(t)P(t)B'(t) + (\pi_Y(t))^{-1}L_*(t)Y(t)L_*{}'(t)$$

and

$$\dot{x}_+ = -L_*(t)x_+ + B(t)p(t) + L_*(t)y(t).$$

Finally, returning to the case $A(t) \neq 0$, we have

$$\mathbf{X}_+[t] = S(t,t_0)X_+[t]S'(t,t_0), \; \mathbf{x}(t) = S(t,t_0)x_+(t),$$

$$\dot{\mathbf{X}}_+[t] = (A(t) - L_*(t))\mathbf{X}_+[t] + \mathbf{X}_+[t](A(t) - L_*(t)) \tag{56}$$
$$+ (\pi_u(t) + \pi_Y(t))\mathbf{X}_+[t] + (\pi_u(t))^{-1}B(t)P(t)B'(t)$$
$$+ (\pi_Y(t))^{-1}L_*(t)Y(t)L_*{}'(t), \; \mathbf{X}_+[t_0] = X^0,$$

and

$$\dot{\mathbf{x}} = (A(t) - L_*(t))\mathbf{x} + B(t)p(t) + L_*(t)y(t), \; \mathbf{x}(t_0) = x^0. \tag{57}$$

Rewriting $\pi_u(t)$, $\pi_Y(t)$ following schemes of [15, p. 189] for these items (see also Property A.4 of the appendix), we have

$$\pi_u(t) = (l^*(t), B(t)P(t)B'(t)l^*(t))^{1/2}/(l^*(t), X_+^*(t)l^*(t))^{1/2}, \tag{58}$$

$$\pi_Y(t) = (l^*(t), L_*(t)Y(t)L_*'(t)l^*(t))^{1/2}/(l^*(t), X_+^*(t)l^*(t))^{1/2}.$$

Summarizing the results we come to the following.

THEOREM 5.2. *With shape matrix $Y(t)$ of the state constraint being nondegenerate and if, under Assumptions* 3.1, 4.1, *and A.*1, *the multipliers $L_*(t)$ responsible for the state constraint have no jumps ($d\Lambda(t) = \lambda(t)dt$), then the external ellipsoids $\mathcal{E}(\mathbf{x}[t], \mathbf{X}[t]) \supseteq \mathcal{X}[t]$ have the form* (56), (57) *with $\mathbf{X}[t_0] = X^0$, $\mathbf{x}[t_0] = x^0$.*

Moreover, with $l^*(t) = G_*{}'(t_0, t)l_*$, $l_* \in \mathbb{R}^n$, the ellipsoids $\mathcal{E}(\mathbf{x}[t], \mathbf{X}[t])$ touch the set $\mathcal{X}[t]$ along the curve $x^*(t)$ which satisfies the condition

$$(l^*(t), x^*(t)) = \rho(l^*(t) \mid \mathcal{X}[t]),\ t \le \tau_1,\ t \ge \tau_2,$$

so that

$$(l^*(t), x^*(t)) = \rho(l^*(t) \mid \mathcal{X}[t]) = \rho(l^*(t) \mid \mathcal{E}(\mathbf{x}[t], \mathbf{X}[t])). \tag{59}$$

On the state constraint boundary, with $t \in [\tau_1, \tau_2]$, the trajectory which satisfies the first equality of (59) may be uniquely determined by the maximum condition (39).

The overall procedure of calculation for Problem 3.1 starting from a point in the interior of the state constraint is as follows.

(i) Consider an array of initial vectors l_* located on the unit ball $\mathcal{B}(0)$ in $\mathbb{R}^n$. Then for each l_* proceed as follows (possibly, in parallel).

(ii) For each l_* construct the array of external ellipsoids $\mathcal{E}_+[\tau]_{l_*}$ for Problem 3.1 of system (15) *without state constraints* following the explicit formulas of [15].

(iii) Construct the array of ellipsoids $\mathcal{E}(\mathbf{x}[t], \mathbf{X}[t])$, namely,

(iii-a) for a given l_* find related $l^*(t) = l_*G_*(t, t_0)$ and, further on, the related curve $x^{(0)}(t)$ for the reachability problem without state constraints. Note that on the interval $[t_0, \tau_1)$, where $L_*(t) \equiv 0$, we have

$$x^{(0)}(t) = x^*(t) + X_+(t)l^*(t)(l^*(t), X_+(t)l^*(t))^{1/2},$$

where the explicit relations for elements $x^*(t), X_+(t)$ of this formula are indicated in Property A.4 of the appendix. Follow curve $x^{(0)}(t)$ until the state constraint boundary is reached at time, say, τ_1;

(iii-b) for $l^*(\tau_1)$ solve the optimization problem (33) over $L(\cdot)$. Use Lemma A.1 of the appendix for additional information on optimizer $L_*(\cdot)$, taking τ_2 as a parameter which varies within interval $[\tau_1, \tau]$;

(iii-c) given optimizer $L_*(\cdot)$, which depends on the selected l_*, solve equations (56), (57), determining the related external ellipsoid $\mathcal{E}_{l_*}(\mathbf{x}[t], \mathbf{X}[t])$.

(iv) In view of Remark 5.1, the intersection

$$\cap\{\mathcal{E}_+[\tau]_{l_*} | l_* \in \mathcal{B}(0)\} \cap \{\mathcal{E}_{l_*}(\mathbf{x}[t], \mathbf{X}[t]) \mid l_* \in \mathcal{B}(0)\} \cap \mathcal{E}(y(\tau), Y(\tau)) = \mathcal{X}_a[\tau] \supseteq \mathcal{X}[\tau]$$

is an external approximation of $\mathcal{X}[\tau]$. Loosely speaking, the greater is the number of vectors l_* used as starting points of the calculations, the more accurate will be the approximation $\mathcal{X}_a$.

Remark 5.2. If it is not necessary to find the tight external approximations for the exact reach set $\mathcal{X}[\tau]$, but a conservative estimate of $\mathcal{X}[\tau]$ using *only one* ellipsoid would suffice, then there is no need to solve the optimization problem (33) in L, since the ellipsoids $\mathcal{E}(\mathbf{x}[t], \mathbf{X}[t])$, defined by (56), (57), happen to be external estimates of $\mathcal{X}[\tau]$ *for all* $L(\cdot)$. One may single out one of these through some appropriate criterion.

An obviously conservative estimate may also be obtained by intersecting the reach set without state constraints with the ellipsoid that defines the state constraint. The objective of this paper, however, is to indicate the calculation of the *exact reach set under state constraints*, especially when the state constraint produces a reach set different from the one without state constraints.

We now pass to the more general case, in which the multipliers may have delta function components.

6. Generalized multipliers. In this section, Assumptions 3.1 and A.1 and the conditions of Lemma 3.4 are taken to be true together with Assumption 3.2. This allows us to consider ellipsoidal approximations without the additional requirement of Property A.2 in the appendix.

Now functional $\Phi(l, \mathcal{L}(\cdot), \tau, t_0)$ has the form (35), and its minimizer for a given $l = l^*(\tau)$ is of the form

$$\mathcal{L}^0(t) = L^0_*(t) + L^0_1\delta(t-\tau_1) - L^0_2\delta(t-\tau_2)$$

for all $l^*(\tau)$ with $L^0_*(t)$ absolutely continuous and $L^0_*(t) \equiv 0$ whenever $x^*(t) \in \mathrm{int}\mathcal{Y}(t)$. Then the respective transition matrix is $\mathcal{G}(t,s) = \mathcal{G}^0_*(t,s)$ and $l^*(t) = \mathcal{G}^{0\prime}_*(t,s)l^*$.

Here $\mathcal{L}^0(\cdot)$ may be interpreted as the generalized derivative of a generalized Lagrange multiplier $\mathbf{\Lambda^0}(\cdot)$ similar to $\Lambda^0(\cdot)$ of section 3. $\mathbf{\Lambda^0}(\cdot)$ is piecewise absolutely right-continuous, with possible jumps at points τ_1, τ_2 and possible jumps at t_0 and/or τ (when it happens that $t_0 = \tau_1$ and/or $\tau = \tau_2$). $\mathbf{\Lambda^0}(t) \equiv \mathrm{const}$, whenever $x^*(t) \in \mathrm{int}\mathcal{Y}(t)$.

Following the reasoning of the previous section, we may derive equations for the approximating external ellipsoids similar to (56), (57). The necessary prerequisites for such a move are similar to Lemma 5.1 and Theorems 5.1 and 5.2.

Thus, we come to equations for the tight external ellipsoids:

(60)
$$dx^* = ((A(t) - L^0_*(t))x^* + (B(t)p(t) + L^0_*(t)y(t))dt - \sum_{i=1}^{2} L^0_i(x^* - y(t))d\chi_i(t,\tau_i),$$

(61)
$$dX_+[t] = ((A(t) - L^0_*(t))X_+[t] + X_+[t](A'(t) - L^{0\prime}_*(t))dt$$
$$+ \left(\pi_u(t) + \pi_Y(t) + \sum_{i=1}^{k}\pi_i\right)X_+[t]dt + (\pi_u(t))^{-1}B(t)P(t)B'(t)dt$$
$$+(\pi_Y(t))^{-1}L^0_*(t)Y(t)L^{0\prime}_*(t)dt + \sum_{i=1}^{2}\pi_i^{-1}L^0_iY(t)L^{0\prime}_i d\chi(t,\tau_i),$$
$$X_+[t_0] = X^0, \quad x^*(t_0) = x^0.$$

The $\pi_i > 0$ are additional parametrizing coefficients similar to $\pi_Y(t)$.

The last equation may be interpreted as before (see Remark 4.1).

THEOREM 6.1. *Under Assumptions* 3.2, *A*.1, 5.1, *and those of Lemma* 3.4 *with* $l^*(t) = {G_*'}^0(t_0,t)l_*,\ l_* \in \mathbb{R}^n$, *the external ellipsoids* $\mathcal{E}(x^*(t), X_+^*[t]) \supseteq \mathcal{X}[t]$ *have the form* (52), (55) *with* $X_+^*[t_0] = X^0,\ \ x^*(t_0) = x^0$. *One may select the parametrizing coefficients* $\pi_u, \pi_{Y(t)}, p_i$ *so that the ellipsoids* $\mathcal{E}(x^*(t), X_+^*[t])$ *touch set* $\mathcal{X}[t]$ *along the curve* $x^*(t)$, *which satisfies condition* (59).

If the optimal trajectories visit the boundary of the state constraint $m > 1$ times, then the reasoning is the same as in the above, and the sums in (60), (61) will have $2m$ terms.

Remark 6.1. The results of this paper remain true if the state constraint is applied to a system output $z(t) = Tx(t),\ z \in \mathbb{R}^k,\ k < n$, rather than to the whole state vector x. The state constraint $\mathcal{Y}$ is then given by an elliptical cylinder in $\mathbb{R}^n$. The proofs may be achieved either by directly following the scheme of this paper or by substituting the elliptical cylinder with an ellipsoid having $n-k$ axes of size r followed by the limit transition $r \to \infty$.

7. Example. The system is the double integrator:

$$\dot{x}_1 = x_2,\ \ \dot{x}_2 = u,\ \ x(0) = 0, \tag{62}$$

under constraints

$$|u| \le \mu, \quad |x_2| \le \nu. \tag{63}$$

We wish to find the reach set $\mathcal{X}[\tau]$ at time $\tau > 0$.

The state constraint may be treated either directly or as a limit as $\epsilon \to 0$ of the ellipsoid $\varepsilon^2 x_1^2 + x_2^2 \le \nu^2$.

The treatment of this example in [9, p. 119] indicated the following:

1. The boundary of the reach set $\mathcal{X}[\tau]$ at each time τ consists of three types of points as follows:
   (a) points reached without visiting the boundary of the state constraint;
   (b) points reached after visiting the boundary of the state constraint for some time;
   (c) points on the boundary of the state constraint.
2. Each optimal trajectory $x^{(0)}(t)$, for given $l \in \mathbb{R}^2$, visits the state constraint not more than once.
3. The necessary conditions for a jump of the multiplier responsible for the state constraint are not fulfilled, so there are no jumps.
4. Assumption A.1 of the appendix is fulfilled.

Case (1a) is treated according to [15, p. 202], while case (1c) is trivial. Therefore we concentrate on case (1b).

The exact parametric equations for case (1b) may be derived, following [9, pp. 119–121], by minimizing over η the functional

$$\rho(l|\mathcal{X}[\tau]) = \min\left\{\mu\int_0^\tau |s_2(t)|d\xi + \nu\int_0^\tau |\eta(t)|d\xi - (l,x) \mid \eta(\cdot)\right\}, \tag{64}$$

where $s_2(t) = -l_1\xi + l_2 - \int_0^t \eta(\xi)d\xi$ is the solution to the adjoint equation

$$\dot{s_1} = 0,\ \dot{s_2} = -s_1 + \eta(t), s_1(\tau) = l_1, s_2(\tau) = l_2.$$

This gives parametric equations for the boundary $\partial\mathcal{X}[\tau]$:

$$x_1 = -\mu(t_1 - \sigma)^2 + c, \;\; x_2 = \mu\sigma + d, \; \tau_1 \leq \sigma \leq \tau.$$

Here $\tau_1 = \nu/\mu < \tau$ is the first instance when the trajectory reaches the boundary of the state constraint, while σ is a parameter that indicates the instance of τ_2 when the trajectory leaves this boundary. Also $c = \nu(\tau - t_0) - \nu^2/2\mu; d = \nu - \mu\tau$. Here we have used Assumption A.1 of the appendix, which implies that optimizer $\eta^0(t)$ satisfies relations

$$\eta^0(t) \equiv -l_1, \; t \in [\tau_1, \tau_2], \;\; \eta^0(t) \equiv 0, \; t \notin [\tau_1, \tau_2],$$

and $s_1(t) \equiv -l_1$.

A typical trajectory for case (b) is when it reaches the boundary of the state constraint at time τ_1 and then runs along the boundary and leaves it at time τ_2. Later it runs toward the boundary $\partial\mathcal{X}[\tau]$ while staying in the interior of the constraint.

For the interval $[t_0, \tau]$, the adjoint system given above, when written in matrix form (27), is

$$\dot{S} = -SA(t) + M(t)\eta(t). \tag{65}$$

To solve the problem through a numerical procedure involving external ellipsoids, using recursive formula and following section 6, transform the adjoint equation (65) of the above into the modified form

$$\dot{S}_L = -S_L(A(t) - L(t)), S(\tau) = I.$$

Taking $l_* = l_*(0)$, we will have $l_*' M(t)\eta^0(t) = (0,1)\eta^0(t)$ when $t \in [\tau_1, \tau_2]$. This yields $M(t) = M_0(t)$ depending on η^0. Direct calculation then allows one to find for each l_* the relation for the optimizer $L_*(t)$ of (31). This is

$$L_*(t) = S_{M_0}(t)^{-1} M_0(t), \; t \in [\tau_1, \tau_2],$$

with $L_*(t) \equiv 0$ when $t < \tau_1, t > \tau_2$. Here $S_M(t)$ is the solution to (65) with initial condition $S_M(\tau) = I$ and $M(t) \equiv 0, \; t, \tau_1, t > \tau_2$.

Recalling the ellipsoidal equations, we have

$$\dot{X} = (A(t) - L(t))X + X(A(t) - L(t))' \tag{66}$$

$$+(\pi_u(t) + \pi_Y(t))X + \pi_u^{-1}(t)BP(t)B' + \pi_Y^{-1}(t)L(t)Y(t)L(t).$$

Here one should take $L(t) = L_*(t)$ and π_u, π_Y according to (58).

The algorithm for this example involves the following steps:

1. For the given starting direction l_* find time $\tau_1 = \nu/\mu$ of the first exit (i.e., encounter with the boundary of the state constraint).
2. Solve extremal problem (64), determining $l_1 = -\eta^0 = \text{const}$ and l_2, for each $\tau_2 = \sigma \in [\tau_1, \tau]$.
3. Construct the ellipsoidal approximations for the system without constraint $t \leq \tau_1$, taking $L(t) \equiv 0$. Denote these as $\mathcal{E}(0, X_l^+(\tau))$.
4. For each $\tau_2 = \sigma \in [\tau_1, \tau]$ construct the ellipsoidal approximation following (66).

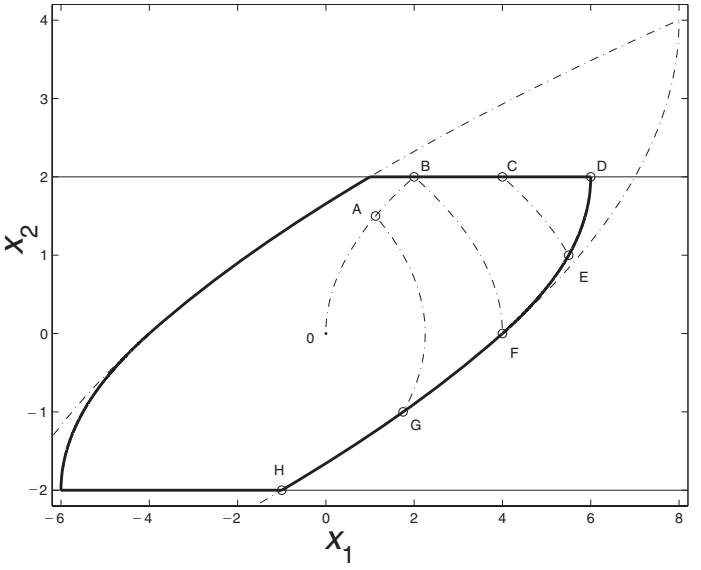


FIG. 1. *Reach set: structure of the boundary.*

The part of the boundary related to case (b) is described by the boundary of the intersection $\cap\{\mathcal{E}(0, X(\tau)) \mid L(\cdot)\}$, where $L(\cdot)$ actually is reduced to the parameter $\sigma = \tau_2$. Denote this as $\mathcal{E}(0, X^+_\sigma(\tau))$.

The total reach set is the intersection of the approximations for each group (a), (b), and (c), namely,

$$\mathcal{X}[\tau] = \mathcal{X}(\tau, 0, 0) = \bigcap_l \mathcal{E}(0, X^+_l(\tau)) \bigcap_\sigma \mathcal{E}(0, X^+_\sigma(\tau)) \bigcap \mathcal{E}(0, Y).$$

Figure 1 shows the structure of the reach set boundary, given here by a thick line. The system trajectory OAG reaches its end point G without visiting the boundary of the state constraint. Point G is attained at time τ. It lies on the reach set boundary of type (a) (see section 7). Trajectory OBF visits the boundary of the state constraint at only one point $\tau_1 = \tau_2 < \tau$. Its end point F is a point of both types (a) and (b). Trajectory OBCE lies on the boundary of the state constraint during the interval $[\tau_1, \tau_2)$ with $\tau_1 < \tau_2 < \tau$, and its end point E lies on the reach set boundary of type (b). Finally, trajectory OBD reaches the boundary of the state constraint at time $\tau_2 < \tau$ and lies on it until the final time τ, so that point D is of both types (b) and (c). Thus, the part of the reach set boundary along points BDEFGH consists of segments BD (of type (c)), DF (of type (b)), and FGH (of type (a)).

Figure 2 shows the structure of the generalized Lagrange multiplier $\eta(t)$ for the case of trajectory $x_2(t)$ of type OBCE shown in the top of the figure. Here $\eta \equiv 0$ on intervals $[0, \tau_1)$ and $[\tau_2, \tau]$ and $\eta \equiv \text{const}$ on $[\tau_1, \tau_2)$. Note that the related trajectory $x_2(t)$ runs along the boundary of the state constraint during the interval $[\tau_1, \tau_2)$, where $\eta(t) \neq 0$. Points τ_1, τ_2 potentially could have δ-functions (indicated by question marks) as components of η. But the necessary conditions for the existence of such components are not fulfilled in this example.

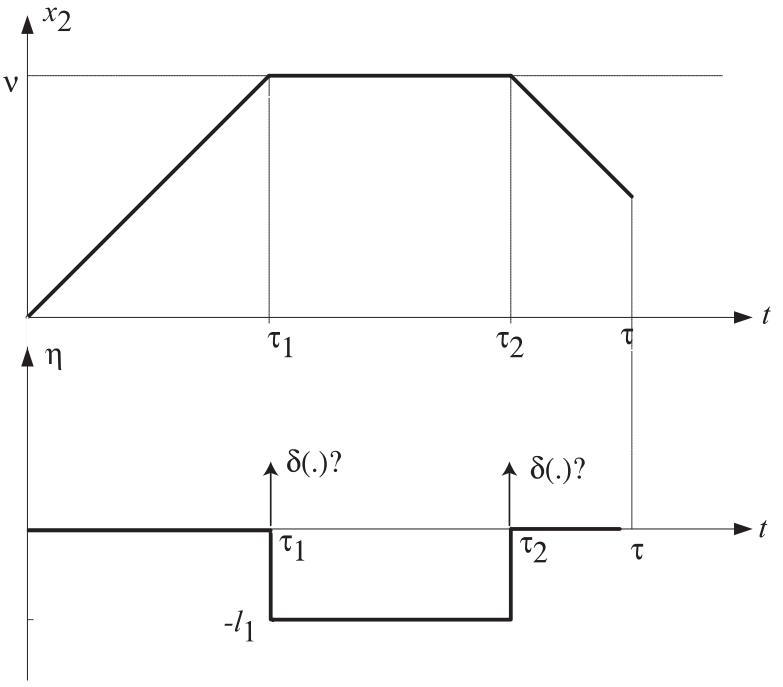


FIG. 2. *Structure of multiplier* η.

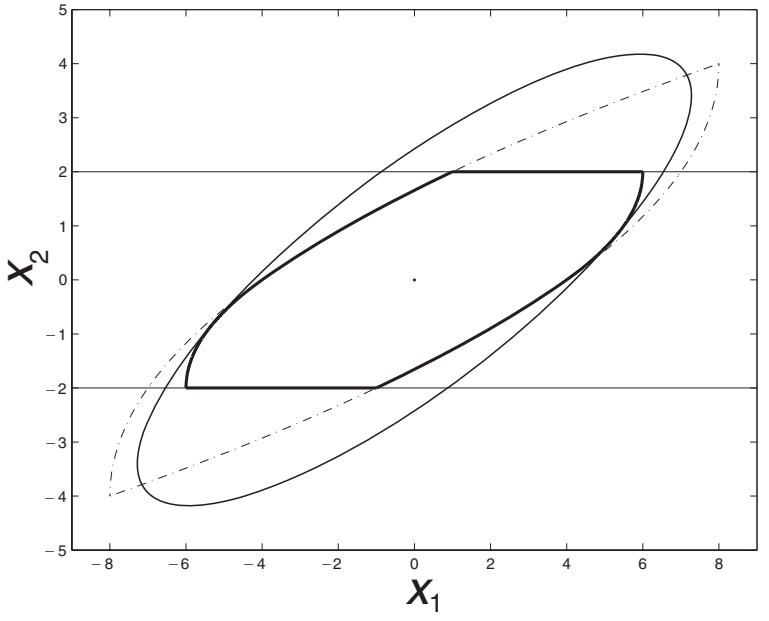


FIG. 3. *External ellipsoid at boundary point of type* (b).

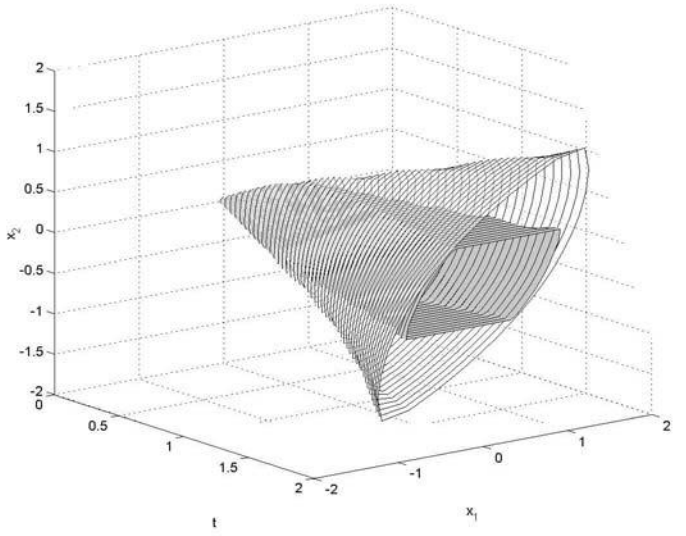


FIG. 4. *Reach tubes with and without state constraints.*

Figure 3 shows an external ellipsoid (thin line) for the boundary of the reach set under state constraints (thick line), indicating its difference from the reach set without state constraints (dash-dotted line). Figure 4 illustrates the final reach tube with and without state constraints for this example.

8. Conclusion. In this paper we presented an ellipsoidal technique to calculate reach sets for linear systems with constraints on the control and state. The suggested scheme introduces parametrized families of tight ellipsoidal-valued tubes that approximate the exact reach tube from above, touching the tube along specially selected "good" curves [31] that cover the entire exact tube. This leads to recursive relations that compared to other approaches, simplify calculations. The proofs rely on a special "recursive" version of the maximum principle under state constraints.

Appendix. In this appendix we present some additional facts useful for proving the main assertions of this paper.

One difficulty in solving the control problem under state constraints (Problem 3.1) is to determine the set of times $\{t \,:\, x(t) \in \partial\mathcal{Y}(t)\}$. This is a union of closed intervals during which the optimal trajectory is on the boundary of the state constraint.

Here are some helpful facts. For Problem 3.1 under Assumptions 3.1 and 3.2, the Properties A.1–A.3 given below are modifications of those shown in [9, section 6, pp. 94–120], and [11, 6].

Property A.1. For any $l \in \mathbb{R}^n$, the minimizer of (20) is $\Lambda^{(0)}(t) = \text{const}$ and in (24) the corresponding minimizer is $\lambda^{(0)}(t) = \text{const}$ during time intervals for which $x(t) \in \text{int}\mathcal{Y}(t)$, the interior of $\mathcal{Y}(t)$.

Since $l'M^{(0)}(t)d\lambda^{(0)}(t) = d\Lambda^{(0)}(t)$, we may track whether the trajectory is on the boundary of $\mathcal{Y}(t)$ by the multiplier $\lambda^{(0)}(t)$. Thus, we need not be interested in values of $M^{(0)}(t)$ when the trajectory is not on the boundary $\partial\mathcal{Y}(t)$.

Property A.2. Suppose $l \in \mathbb{R}^n$ is given and $x^{(0)}(t)$ is the solution of Problem 3.1. For the function $\Lambda^{(0)}(t)$ to have a jump at t^*, under the smoothness conditions on the state constraint of Lemma 3.4, t^* must be a time of arrival or departure from the boundary of the tube $\mathcal{E}(y(t), Y(t)),\ t_0 \le t \le \tau$, and the trajectory $x^{(0)}(t)$ must be *tangent* to the tube $\mathcal{E}(y(t), Y(t))$ at t^* [9, p. 106]. Thus, $x^{(0)}(t)$ is differentiable at t^* and

$$(\chi, \dot{x}^{(0)}(t^* - 0)) = (\chi, \dot{x}^{(0)}(t^* + 0)) = 0. \tag{67}$$

Here χ is the support vector to the state constraint $\mathcal{Y}(t^*)$ at $x^{(0)}(t^*)$. In general, if the jump is $\Lambda(t^* + 0) - \Lambda(t^* - 0) = \chi$, then the necessary condition for such a jump is

$$(\chi, \dot{x}^{(0)}(t^* + 0) - \dot{x}^{(0)}(t^* - 0)) \ge 0. \tag{68}$$

For example, if $x^{(0)}(t^* - \sigma), \sigma > 0$, lies inside the interior of the constraint set and $x(t^* + \sigma)$ lies on the boundary, then

$$(\chi, \dot{x}^{(0)}(t^* + 0) - \dot{x}^{(0)}(t^* - 0)) \le 0,$$

and (68) will be fulfilled only if this inner product equals zero.

We also need the following assumption.

Assumption A.1. For Problem 3.1, with given l, there exists no control $u^*(s)$ that satisfies the maximum principle

$$l' S(s, \tau \mid M^{(0)}(\cdot), \lambda^{(0)}(\cdot)) B(s) u^*(s)$$

$$= \max\{l' S(s, \tau \mid M^{(0)}(\cdot), \lambda^{(0)}(\cdot)) B(s) u \mid u \in \mathcal{E}(p(s), P(s))\}$$

for $\{s \mid l' S(s, \tau \mid M^{(0)}(\cdot), \lambda^{(0)}(\cdot)) B(s) \ne 0\}$ and at the same time ensures for these values of s that the corresponding trajectory $x^*(s) \in \partial\mathcal{Y}(s)$.

Property A.3. Assumption A.1 holds.

This means that if the control $u^{(0)}(s)$ is determined by the maximum principle, with $h^{(0)}(\tau, s) = l' S(s, \tau \mid M^{(0)}(\cdot), \lambda^{(0)}(\cdot)) B(s) \ne 0$, and therefore attains its extremal values under given hard bounds, then this control cannot also keep the corresponding trajectory $x^{(0)}(s)$ along the boundary $\partial\mathcal{Y}(s)$. In other words, in this case *the maximum principle is degenerate along the state constraint*, i.e., $h^{(0)}(\tau, s) \equiv 0$ and does not help to find the control when the trajectory lies on the boundary $\partial\mathcal{Y}(s)$.

This assumption excludes the case when the solution of Problem 3.1 *without state constraints* also happens to solve the same problem with state constraints.

LEMMA A.1. *Under Assumption A.1 applied to Problem* 3.1, *the function*

$$h^{(0)}(\tau, s) = l' S(s, \tau \mid M^{(0)}(\cdot), \lambda^{(0)})(\cdot) B(s) \equiv 0$$

whenever $x^{(0)}(s) \in \partial\mathcal{Y}(s)$.

Property A.4. The equations for tight external ellipsoids *without state constraints* are explained in detail in [15, pp. 188–190, 198, 199]. Here is the basic result.

The equality

$$\rho(l^*(t) \mid \mathcal{X}[\tau]) = (l^*(t), x^{(0)}(t)) + (l^*(t), X_+(t) l^*(t))^{1/2}$$

is true for all $t \ge t_0$. Here

$$x^{(0)}(t) = x^*(t) + X_+^*(t) l^*(t) (l^*(t), X_+^*(t) l^*(t))^{1/2},$$

with

$$\dot{x}^* = A(t)x^* + B(t)p(t),\ x^*(t_0) = x^0,$$

$$\dot{X}_+^* = A(t)X_+ + X_+A'(t) + \pi(t)X_+ + \pi^{-1}(t)P(t),\ X_+(t_0) = X^0,$$

$$l^*(t) = -A(t)'l^*(t),\ l^*(t_0) = l_*,$$

and

$$\pi(t) = (l^*(t), B(t)P(t)B'(t)l^*(t))^{1/2}/(l^*(t), X_+(t)l^*(t))^{1/2}. \tag{69}$$

This indicates that by solving the given explicit equations we may find the support function $\rho(l^*(t)|\mathcal{X}[\tau])$ *for all* $l_* \in \mathcal{B}(0)$, the unit ball in $\mathbb{R}^n$, so that the related vectors $l^*(t) = l_*'S(t_0, t)$ would cover all the directions in $\mathbb{R}^n$ and, after being normalized, the set of all the vectors $l^*(t)/(l^*(t), l^*(t))^{1/2}$ would also coincide with $\mathcal{B}(0)$.

Acknowledgments. The authors thank M. N. Kirilin for the computer illustrations.

REFERENCES

[1] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[2] F. L. Chernousko, *State Estimation for Dynamic Systems*, CRC Press, Boca Raton, FL, 1994.

[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 1, Wiley, New York, 1953.

[4] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, Wiley, New York, 1962.

[5] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[6] M. I. Gusev and A. B. Kurzhanski, *Optimization of controlled systems with restrictions.* I, Differ. Uravn., 7 (1971), pp. 1591–1602 (in Russian).

[7] T. A. Henzinger, P. W. Kopke, A. Puri, and P. Varaiya, *What's decidable about hybrid automata*?, in Proceedings of the 27th Annual Symposium on Theory of Computing, ACM, New York, 1995, pp. 373–382.

[8] M. N. Kirilin and A. B. Kurzhanski, *Ellipsoidal techniques for readability problems under nonellipsoidal constraints*, in Proceedings of the 5th IFAC Symposium on Nonlinear Control Systems, Elsevier Science, Ltd., Dordrecht, 2002, pp. 735–740.

[9] A. B. Kurzhanski, *Control and Observation Under Uncertainty Conditions*, Nauka, Moscow, 1977 (in Russian).

[10] A. B. Kurzhanski and T. F. Filippova, *On the theory of trajectory tubes—A mathematical formalism for uncertain dynamics, viability and control*, in Advances in Nonlinear Dynamics and Control. A Report from Russia, Birkhäuser Boston, Boston, MA, 1993, pp. 122–188.

[11] A. B. Kurzhanski and Yu. S. Osipov, *On control problems with restricted coordinates*, J. Appl. Math. Mech., 33 (1989), pp. 682–695.

[12] A. B. Kurzhanski and I. Valyi, *Ellipsoidal Calculus for Dynamics and Control*, Birkhäuser Boston, Boston, MA, 1997.

[13] A. B. Kurzhanski and P. Varaiya, *Ellipsoidal techniques for reachability analysis: Internal approximation*, Systems Control Lett., 41 (2000), pp. 201–211.

[14] A. B. Kurzhanski and P. Varaiya, *On reachability under uncertainty*, SIAM J. Control Optim., 41 (2002), pp. 181–216.

[15] A. B. Kurzhanski and P. Varaiya, *On ellipsoidal techniques for reachability analysis.* I. *External approximations*, Optim. Methods Softw., 17 (2002), pp. 177–206.

[16] A. B. Kurzhanski and P. Varaiya, *Dynamic optimization for reachability problems*, J. Optim. Theory Appl., 108 (2002), pp. 227–251.

[17] A. B. Kurzhanski and P. Varaiya, *Ellipsoidal techniques in reachability analysis for hybrid systems*, in New Directions and Applications in Control Theory, W. P. Dayawansa, A. Lindquist, and Y. Zhou, eds., Lecture Notes in Control and Inform. Sci. 321, Springer, Berlin, 2005, pp. 193–205.
[18] E. B. Lee and L. Marcus, *Foundations of Optimal Control Theory*, Wiley, New York, 1967.
[19] G. Leitmann, *Optimality and reachability via feedback control*, in Dynamic Systems and Mycrophysics, A. Blaquiere and G. Leitmann, eds., Academic Press, New York, 1982.
[20] E. J. Mcshane, *Integration*, Princeton University Press, Princeton, NJ, 1947.
[21] G. L. Pappas and S. Sastry, *Straightening out differential inclusions*, Systems Control Lett., 35 (1998), pp. 79–85.
[22] L. S. Pontryagin, V. G. Boltyanski, R. V. Gamkrelidze, and E. F. Mischenko, *Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.
[23] A. Puri, V. Borkar, and P. Varaiya, *ϵ-approximations of differential inclusions*, in Hybrid Systems, R. Alur, T. A. Henzinger, and E. D. Sonntag, eds., Lecture Notes in Comput. Sci. 1201, Springer, Berlin, 1996, pp. 109–123.
[24] A. Puri and P. Varaiya, *Decidability of hybrid systems with rectangular inclusions*, in Proceedings CAV-94, D. Dill, ed., Lecture Notes in Comput. Sci. 1066, Springer, Berlin, 1996, pp. 95–104.
[25] R. T. Rockafellar, *Convex Analysis*, 2nd ed., Princeton University Press, Princeton, NJ, 1999.
[26] S. Saks, *The Theory of the Integral*, Dover, New York, 1964.
[27] L. Schwartz, *La théorie des distributions*, I, Hermann, Paris, 1950.
[28] L. Schwartz, *La théorie des distributions*, II, Hermann, Paris, 1951.
[29] J. A. Sethian, *Level-Set Methods and Fast-Marching Methods*, Cambridge University Press, Cambridge, UK, 1999.
[30] A. I. Subbotin, *Generalized Solutions of First-Order PDE's*, Birkhäuser Boston, Boston, MA, 1995.
[31] P. Varaiya, *Reach set computation using optimal control*, in Proceedings of KIT Workshop on Verification of Hybrid Systems, Verimag, Grenoble, 1998, pp. 377–383.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1395–1446

GLOBAL CARLEMAN INEQUALITIES FOR PARABOLIC SYSTEMS AND APPLICATIONS TO CONTROLLABILITY*

ENRIQUE FERNÁNDEZ-CARA† AND SERGIO GUERRERO†

Abstract. This paper has been conceived as an overview on the controllability properties of some relevant (linear and nonlinear) parabolic systems. Specifically, we deal with the null controllability and the exact controllability to the trajectories. We try to explain the role played by the observability inequalities in this context and the need of global Carleman estimates. We also recall the main ideas used to overcome the difficulties motivated by nonlinearities. First, we considered the classical heat equation with Dirichlet conditions and distributed controls. Then we analyze recent extensions to other linear and semilinear parabolic systems and/or boundary controls. Finally, we review the controllability properties for the Stokes and Navier–Stokes equations that are known to date. In this context, we have paid special attention to obtaining the necessary Carleman estimates. Some open questions are mentioned throughout the paper. We hope that this unified presentation will be useful for those researchers interested in the field.

Key words. Carleman inequalities, parabolic equations, unique continuation, observability, null controllability

AMS subject classifications. 35B37, 70Q05, 93B05, 93B07

DOI. 10.1137/S0363012904439696

Introduction. The goal of this paper is to provide a panorama of an important subfield of control theory: the null controllability analysis of parabolic equations and systems.

The main contributions to this area are due to O. Yu. Imanuvilov, who popularized the use of global Carleman estimates in the context of null controllability. Many arguments from [36], [30], [39], and [37] will be reproduced here. Another relevant contributor has been E. Zuazua, who was able to deduce global controllability results for some nonlinear systems for the first time in [56].

The controllability of partial differential equations has been the object of intensive research during the last few decades. In 1978, Russell [53] made a survey of the most relevant results that were available in the literature at that time. In that paper, the author described a number of different tools that were developed to address controllability problems, often inspired and related to other subjects concerning partial differential equations: multipliers, moment problems, nonharmonic Fourier series, etc. More recently, J.-L. Lions introduced the so-called *Hilbert uniqueness method* (for instance, see [44], [45], [46]). That was the starting point of a fruitful period on the subject.

It would be impossible to present here all the relevant results that have been proved in this area. We will thus reduce our scope drastically, considering only null controllability problems for parabolic equations and systems.

In order to get an idea, let us consider the simplest case of the linear heat equation

*Received by the editors January 14, 2004; accepted for publication (in revised form) April 10, 2006; published electronically October 12, 2006. This work was partially supported by D.G.E.S. (Spain), grant BFM2003-06446.

http://www.siam.org/journals/sicon/45-4/43969.html

†Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Aptdo. 1160, 41080 Sevilla, Spain (cara@numer.us.es, sguerrero@us.es).

with Dirichlet boundary conditions and distributed controls:

$$\begin{cases} y_t - \Delta y = v1_{\mathcal{O}} & \text{in} \quad Q = \Omega \times (0,T), \\ y = 0 & \text{on} \quad \Sigma = \partial\Omega \times (0,T), \\ y(0) = y^0 & \text{in} \quad \Omega. \end{cases} \tag{0.1}$$

Here, $\Omega \subset \mathbf{R}^N$ is a bounded domain of class C^2, $\mathcal{O} \subset \Omega$ is a nonempty open subset, $1_{\mathcal{O}}$ is the characteristic function of $\mathcal{O}$, and T is a given positive time. We assume that the initial state y^0 is given in $L^2(\Omega)$ and try to find a control $v \in L^2(\mathcal{O} \times (0,T))$ such that the associated state $y = y(x,t)$ possesses a desired behavior at time $t = T$.

Roughly speaking, systems of parabolic type like (0.1) are characterized by *non-reversibility*, the *dissipativity* of the solutions (i.e., the fact that *energy* is lost along the trajectories), and the *regularizing effect*.

In accordance with this last property, it is not possible to lead the solutions of (0.1) exactly to *every* final state in a Sobolev space like $L^2(\Omega)$ or $H^m(\Omega)$, unless the control is exerted *brutally*, i.e., unless we are in the trivial and uninteresting situation $\mathcal{O} = \Omega$.

On the contrary, it may be interesting to investigate whether it is possible to drive the solutions to (0.1) (acting on a small set $\mathcal{O} \subset\subset \Omega$) exactly to a state on a trajectory. For instance, it is completely meaningful to search for controls v such that the associated states y satisfy

$$y(T) = 0 \quad \text{in} \quad \Omega. \tag{0.2}$$

If such controls exist, we say that (0.1) is *null controllable* at time T.

Observe that the null controllability of a parabolic system is a very useful property from the viewpoint of applications. Indeed, it permits us to reach in a finite amount of time a state that is "natural" for the system. All the work has to be done for $t \in [0,T]$. Afterwards, no additional effort is needed to get a satisfactory situation.

It is also easy to see that, for a linear system of the kind above, null controllability is equivalent to *exact controllability to the trajectories*, i.e., to the following property: for any $y^0 \in L^2(\Omega)$ and any function $\overline{y}$ satisfying

$$\begin{cases} \overline{y}_t - \Delta \overline{y} = 0 & \text{in} \quad Q, \\ \overline{y} = 0 & \text{on} \quad \Sigma \end{cases}$$

and

$$\overline{y} \in C^0([0,T]; L^2(\Omega)),$$

there exist controls v such that

$$y(T) = \overline{y}(T) \quad \text{in} \quad \Omega.$$

It will be explained below that the null controllability of a linear parabolic system is, roughly speaking, equivalent to the *observability* of the associated adjoint states.

More precisely, for each $\varphi^0 \in L^2(\Omega)$, let us consider the so-called *adjoint system*

$$\begin{cases} -\varphi_t - \Delta \varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega. \end{cases} \tag{0.3}$$

Then (0.1) is null controllable with controls in $L^2(\mathcal{O}\times(0,T))$ if and only if there exists $C>0$ such that

$$\|\varphi(0)\|^2_{L^2(\Omega)} \le C \iint_{\mathcal{O}\times(0,T)} |\varphi|^2\,dx\,dt \qquad \forall \varphi^0 \in L^2(\Omega). \tag{0.4}$$

At present, the most powerful tools to prove inequalities like (0.4) for general parabolic systems are *global Carleman estimates*. They have the form

$$\iint_{\Omega\times(0,T)} \rho^2\,|\varphi|^2\,dx\,dt \le C \iint_{\mathcal{O}\times(0,T)} \rho^2\,|\varphi|^2\,dx\,dt, \tag{0.5}$$

where $\rho=\rho(x,t)$ is continuous and strictly positive for $t\in(0,T)$. The previous considerations motivate the need for establishing such estimates.

It is also natural to ask if null controllability and/or exact controllability to the trajectories hold in the framework of semilinear or even genuinely nonlinear parabolic systems. For instance, we can be interested in controls and states satisfying

$$\begin{cases} y_t - \Delta y + f(y) = v1_{\mathcal{O}} & \text{in} \quad Q,\\ y=0 & \text{on} \quad \Sigma,\\ y(0)=y^0 & \text{in} \quad \Omega \end{cases} \tag{0.6}$$

and (0.2), where the function $f:\mathbf{R}\mapsto\mathbf{R}$ is given and, for instance, we assume that $f\in C^1(\mathbf{R})$ and $f(0)=0$.

For systems of this kind, a classical approach relies on the use of *fixed point arguments*. The idea is to find a fixed point of the mapping $z\mapsto y$, where y is, together with some v, a solution to the linearized system

$$\begin{cases} y_t - \Delta y + \dfrac{f(z)}{z}\,y = v1_{\mathcal{O}} & \text{in} \quad Q,\\ y=0 & \text{on} \quad \Sigma,\\ y(0)=y^0 & \text{in} \quad \Omega \end{cases} \tag{0.7}$$

satisfying (0.2). The existence of v is again equivalent to an observability inequality which is implied by appropriate Carleman estimates similar to (0.5).

The fixed point approach can be applied whenever, among other things, the way the constants arising in Carleman estimates depend on the coefficients of the linearized systems is known in detail. Therefore, a careful analysis of (0.5) is needed.

The theoretical results we present below can be applied in many contexts. For instance, they can serve as tools for

- controlling the temperature of a medium eventually influenced by transport and/or chemical effects;
- controlling the velocity field of a real fluid (like those modeled by the Navier–Stokes equations), etc.

In the following sections, we will explain how these ideas can be applied to several parabolic equations and systems. A similar (but different) overview on this topic is [6], where the author is also concerned with the stabilization of parabolic systems.

An important aspect of the controllability of time-dependent systems that will not be treated in this paper is the analysis and obtainment of numerical approximations. Of course, the most interesting question concerning the null controllability of (0.1) and (0.6) is *how to construct* a control with the desirable properties.

Several ideas that reproduce the theoretical arguments at a finite-dimensional level were given, for instance, in [34]. Unfortunately, the proposed algorithms seem to be still too expensive. We simply point out here that the controllability analysis we recall in section 2, based on the solution of a linear fourth order problem, could provide some promising ideas.

There are many other important subjects related to the controllability of parabolic equations that will not be visited here. Let us mention some of them:

- The analysis of the existence of *insensitizing controls* in the sense of [47] or, more generally, the null controllability of cascade systems. This has been the objective of considerable work in recent years (see, for instance, [12], [9], and [26]).
- The controllability of parabolic systems with memory. This includes interesting families of linear viscoelastic fluids. Some partial results in this field are given in [7] and [14].
- The control of coupled systems of parabolic-hyperbolic type, like a combination of heat and wave equations, the system of thermoelasticity, etc. Some of these problems have been treated in [57], [42], and [43].
- The control of fluid-solid interaction systems. This seems to be unexplored at present. For very simple models, some results are given in [13].
- The control of nondeterministic systems. Some advances were made in the context of linear problems. See, for instance, [8]. In this field, the main open question turns out to be the obtainment of controllability results for nonlinear systems, due to the lack of compactness.

The plan of this paper is the following. In section 1, we consider the classical heat equation and some variants. Essentially, we deal there with heat equations with *strong* solutions. We recall the proof of a basic Carleman estimate and present some applications to semilinear systems.

Section 2 is devoted to the null controllability analysis of other more complicated parabolic equations involving *weak* and/or *very weak* solutions. It will be shown that the proof of the related Carleman inequalities is much more involved and needs some extra work. We will also indicate how this can be used to deduce the exact controllability to the trajectories of more general semilinear systems.

Finally, section 3 is concerned with systems of the Stokes kind. Here, the main difficulties are caused by the pressure. We will recall a recent argument leading to an appropriate but not completely satisfactory Carleman inequality. We will also explain how this leads to the null controllability of Stokes-like systems and to the local exact controllability to the trajectories of the Navier–Stokes equations.

1. Null controllability of the linear heat equation and applications.

1.1. Null controllability and observability. Let us consider again the simplest case of the linear heat equation with Dirichlet boundary conditions and distributed control with support in a small set:

$$\begin{cases} y_t - \Delta y = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega. \end{cases} \tag{1.1}$$

We assume that $\mathcal{O} \subset\subset \Omega$ is a nonempty (small) open subset and T is a given positive time. In what follows, $n(x)$ will stand for the outward unit normal vector at the point $x \in \partial\Omega$. We assume that the initial state y^0 is given in $L^2(\Omega)$, and we try to

find a control $v \in L^2(\mathcal{O} \times (0,T))$ such that the associated state $y = y(x,t)$ possesses a desired behavior at time $t = T$.

Recall that under these assumptions, system (1.1) has a unique *weak* solution y satisfying

$$y \in L^2(0,T; H_0^1(\Omega)) \cap C^0([0,T]; L^2(\Omega))$$

that depends continuously on y^0 and v.

Our interest is to provide an answer to the following questions.

Question 1: *Null controllability.* For every $y^0 \in L^2(\Omega)$, can a control $v \in L^2(\mathcal{O} \times (0,T))$ be found such that $y(T) = 0$ in Ω?

This question will be answered affirmatively in this section. To this end, we will introduce, for each $\varphi^0 \in L^2(\Omega)$, the adjoint system

$$\begin{cases} -\varphi_t - \Delta\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega, \end{cases} \tag{1.2}$$

and we will try to answer the following auxiliary question.

Question 2: *Observability inequality.* Can we find a constant $C > 0$ such that, for each $\varphi^0 \in L^2(\Omega)$, the associated solution of (1.2) satisfies

$$\|\varphi(0)\|^2_{L^2(\Omega)} \leq C \iint_{\mathcal{O}\times(0,T)} |\varphi|^2 \, dx \, dt \,? \tag{1.3}$$

An affirmative answer to Question 2 implies an affirmative answer to Question 1. This is what is proved in the following result.

THEOREM 1.1. *The observability inequality* (1.3) *implies the null controllability of* (1.1).

Proof. We divide the proof into two steps. First, we build a sequence of controls $v_\varepsilon \in L^2(\mathcal{O} \times (0,T))$ with $\varepsilon > 0$ which provide the *approximate controllability* of (1.1) (see (1.7)). Second, we pass to the limit when ε tends to zero and we conclude.

Step 1. Let $y^0 \in L^2(\Omega)$ and $\varepsilon > 0$ be given. Let us introduce the functional J_ε, with

$$J_\varepsilon(\varphi^0) = \frac{1}{2} \iint_{\mathcal{O}\times(0,T)} |\varphi|^2 \, dx \, dt + \varepsilon \|\varphi^0\|_{L^2(\Omega)} + (\varphi(0), y^0)_{L^2(\Omega)} \tag{1.4}$$

for every $\varphi^0 \in L^2(\Omega)$. Here, φ is the solution of (1.2) associated to the initial condition φ^0. Using (1.3), it is not difficult to check that J_ε is strictly convex, continuous, and coercive in $L^2(\Omega)$, so it possesses a unique minimum $\varphi^0_\varepsilon \in L^2(\Omega)$, whose associated solution is denoted by φ_ε. Let us now introduce the control $v_\varepsilon = \varphi_\varepsilon 1_\mathcal{O}$, and let us denote by y_ε the solution of (1.1) associated to v_ε.

Let y_1 be the final state of the solution to (1.1) with vanishing control. Let us remark that the unique interesting case to be studied turns out to be when $\|y_1\|_{L^2(\Omega)} > \varepsilon$, since this is equivalent to $\varphi^0_\varepsilon \neq 0$. See [19] for more details. Under this assumption, we can differentiate the functional J_ε at φ^0_ε and obtain a necessary condition for J_ε to reach a minimum at φ^0_ε, say,

$$\iint_{\mathcal{O}\times(0,T)} \varphi_\varepsilon \, \varphi \, dx \, dt + \varepsilon \left(\frac{\varphi^0_\varepsilon}{\|\varphi^0_\varepsilon\|_{L^2(\Omega)}}, \varphi^0 \right)_{L^2(\Omega)} + (\varphi(0), y^0)_{L^2(\Omega)} = 0 \tag{1.5}$$

for every $\varphi^0 \in L^2(\Omega)$.

Using this and (1.3) for $\varphi^0 = \varphi^0_\varepsilon$, we obtain $\|v_\varepsilon\|_{L^2(\mathcal{O}\times(0,T))} \leq \sqrt{C}\,\|y^0\|_{L^2(\Omega)}$, where C is the observability constant of (1.3).

Since systems (1.1) and (1.2) are in duality, we have

$$\iint_{\mathcal{O}\times(0,T)} \varphi_\varepsilon\,\varphi\,dx\,dt = (y_\varepsilon(T),\varphi^0)_{L^2(\Omega)} - (y^0,\varphi(0))_{L^2(\Omega)}, \tag{1.6}$$

which, combined with (1.5), yields

$$\|y_\varepsilon(T)\|_{L^2(\Omega)} \leq \varepsilon. \tag{1.7}$$

Step 2. Since the sequence $\{v_\varepsilon\}$ is bounded in $L^2(\mathcal{O}\times(0,T))$, it possesses a (weakly) convergent subsequence to certain $v \in L^2(\mathcal{O}\times(0,T))$. Using classical parabolic estimates we deduce that, at least for a subsequence,

$$y_\varepsilon \to y \quad \text{weakly in } L^2(0,T;H^1_0(\Omega)) \cap H^1(0,T;H^{-1}(\Omega)), \tag{1.8}$$

where y is the solution of (1.1) with control v. In particular, this gives weak convergence for $\{y_\varepsilon(t)\}\,(t \in [0,T])$ in $L^2(\Omega)$ so we have $y(T) = 0$.

Remark 1. At this point, some comments must be made:

1. We have proved that (1.3) implies null controllability with a control that satisfies

$$\|v\|_{L^2(\mathcal{O}\times(0,T))} \leq \sqrt{C}\,\|y^0\|_{L^2(\Omega)}, \tag{1.9}$$

where C is the observability constant.

Conversely, if we have null controllability with controls $v \in L^2(\mathcal{O}\times(0,T))$ that satisfy

$$\|v\|_{L^2(\mathcal{O}\times(0,T))} \leq \sqrt{C}\,\|y^0\|_{L^2(\Omega)} \tag{1.10}$$

for some constant $C > 0$, then it can be checked that we have (1.3) with the same constant C.

2. It is possible to present a similar argument in a general frame. Let us consider three Hilbert spaces U, H, E and two linear continuous operators $L \in \mathcal{L}(U;E)$ and $M \in \mathcal{L}(H;E)$. Then we have

$$\|M^*\varphi^0\|_H \leq C\|L^*\varphi^0\|_{U'} \quad \forall \varphi^0 \in E' \tag{1.11}$$

for some positive constant C if and only if $R(M) \subset R(L)$ and, moreover,

$$\forall y^0 \in H,\ \exists v \in U \text{ such that } Lv = My^0,\ \|v\|_U \leq C\|y^0\|_H. \tag{1.12}$$

Properties of this kind have been established and analyzed for the first time in the framework of control theory in [52]. They have been successfully used in many different contexts in recent years.

3. Because of linearity, the null controllability of (1.1) is equivalent to the exact controllability to (uncontrolled) trajectories. In other words, (1.1) is null controllable if and only if, for every $\overline{y}^0 \in L^2(\Omega)$, we can find a control $v \in L^2(\mathcal{O}\times(0,T))$ such that $y(T) = \overline{y}(T)$ in Ω, where y is the solution of (1.1) associated to v, and $\overline{y}$ satisfies

$$\begin{cases} \overline{y}_t - \Delta\overline{y} = 0 & \text{in } Q, \\ \overline{y} = 0 & \text{on } \Sigma, \\ \overline{y}(0) = \overline{y}^0 & \text{in } \Omega. \end{cases} \tag{1.13}$$

4. In general, for any $y^1 \in L^2(\Omega)$, it is not reasonable to look for a control $v \in L^2(\mathcal{O} \times (0,T))$ such that $y(T) = y^1$ in Ω (recall that $\mathcal{O} \subset\subset \Omega$). Indeed, due to the regularizing effect, every solution to (1.1) is space-analytic in $\Omega \setminus \overline{\mathcal{O}}$ at time T, which is not necessarily the case for y^1. Thus, the *exact controllability problem* is out of order for the heat equation.

In what follows, for the reasons stated above, we will focus on the proof of (1.3). This will rely on other (previous) inequalities for the solutions to (1.2), known as *global Carleman inequalities*. Let us mention that a general global Carleman inequality adopts the form

$$\iint_{\Omega\times(0,T)} \rho^2 \,|\varphi|^2 \,dx\,dt \le C \iint_{\mathcal{O}\times(0,T)} \rho^2 \,|\varphi|^2 \,dx\,dt, \tag{1.14}$$

where $\rho = \rho(x,t)$ is a continuous and strictly positive weight function. For a function ρ satisfying $\rho > 0$ in $\overline{\Omega} \times (0,T)$, we will be able to deduce (1.14) and then estimates like

$$\iint_{\Omega\times(T/4,3T/4)} |\varphi|^2 \,dx\,dt \le C \iint_{\mathcal{O}\times(0,T)} |\varphi|^2 \,dx\,dt. \tag{1.15}$$

This, together with the dissipation properties of the solutions of (1.2), will lead to (1.3). More details are given below.

Remark 2. To our knowledge, the first (boundary) controllability results for the heat equation were obtained by Egorov in the early sixties; see [17]. Later, an important general principle was established by Russell in [52]: if T_0 is sufficiently large and the wave system

$$\begin{cases} y_{tt} - \Delta y = v1_{\mathcal{O}} & \text{in} \quad \Omega \times (0,T_0), \\ y = 0 & \text{on} \quad \partial\Omega \times (0,T_0), \\ y(0) = y^0, \quad y_t(0) = y^1 & \text{in} \quad \Omega \end{cases} \tag{1.16}$$

is *exactly controllable* at time T_0, then (1.1) is null controllable at time T for all $T > 0$.

This principle has been revisited by several authors. In particular, an "abstract" version is given in [4].

Using the *method of moments*, the exact controllability of (1.16) can be established for large T_0 when, for instance, $\overline{\mathcal{O}}$ is a neighborhood of $\partial\Omega$. Thus, one has null controllability for (1.1) in this case.

Remark 3. In the particular case of the classical heat equation (1.1), there is another way to prove (1.3), based on the spectral decomposition of the solutions. This approach was introduced by Lebeau and Robbiano [41] and led to the null controllability of the heat equation with controls supported by general subdomains $\mathcal{O} \subset\subset \Omega$ for the first time.

1.2. A global Carleman inequality for the linear heat equation and its consequences.

LEMMA 1.2. *Let $\omega \subset\subset \Omega$ be a nonempty open subset. Then there exists $\eta^0 \in C^2(\overline{\Omega})$ such that $\eta^0 > 0$ in Ω, $\eta^0 = 0$ on $\partial\Omega$, and $|\nabla\eta^0| > 0$ in $\overline{\Omega \setminus \omega}$.*

A proof of this lemma can be found in [30]. A much easier proof can be developed when Ω is star shaped with respect to a point $x^0 \in \omega$.

Let ω be a nonempty open set satisfying $\omega \subset\subset \mathcal{O}$ and let us set

$$\begin{aligned} \alpha(x,t) &= \frac{e^{2\lambda m\|\eta^0\|_\infty} - e^{\lambda(m\|\eta^0\|_\infty+\eta^0(x))}}{t(T-t)}, \\ \xi(x,t) &= \frac{e^{\lambda(m\|\eta^0\|_\infty+\eta^0(x))}}{t(T-t)} \end{aligned} \tag{1.17}$$

for $(x,t) \in Q$, where η^0 is the function furnished by Lemma 1.2 for this ω and $m > 1$. Weight functions of this kind were first introduced by Imanuvilov. See [30] for a systematic use of them.

Now we arrive at the main result of this section.

LEMMA 1.3. *There exist three constants $\lambda_1 = C(\Omega,\mathcal{O}) \geq 1$, $s_1 = C(\Omega,\mathcal{O})(T+T^2)$, and $C_1(\Omega,\mathcal{O})$ such that, for any $\lambda \geq \lambda_1$ and any $s \geq s_1$, the following inequality holds:*

$$\begin{aligned} & s^{-1}\iint_Q e^{-2s\alpha}\,\xi^{-1}(|q_t|^2+|\Delta q|^2)\,dx\,dt + s\,\lambda^2\iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt \\ & \quad + s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \leq C_1\Big(\iint_Q e^{-2s\alpha}\,|q_t+\Delta q|^2\,dx\,dt \\ & \quad + s^3\,\lambda^4\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\Big) \end{aligned} \tag{1.18}$$

for all $q \in C^2(\overline{Q})$ with $q = 0$ on Σ.

In what follows, $C(\Omega,\mathcal{O})$ or simply C will denote a generic constant depending only on Ω and $\mathcal{O}$.

Before giving the proof of Lemma 1.3, we will deduce the observability inequality (1.3) (and, accordingly, the null controllability of (1.1)) from this result. This can be made in three steps.

Step 1. From the density of the smooth functions in the space where the solutions of (1.2) with $\varphi^0 \in L^2(\Omega)$ live, we first observe that the Carleman inequality above is verified by all of them. Thus, fixing $\lambda = \lambda_1$, we get

$$\iint_Q e^{-2s\alpha}\,t^{-3}(T-t)^{-3}\,|\varphi|^2\,dx\,dt \leq C\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,t^{-3}(T-t)^{-3}\,|\varphi|^2\,dx\,dt \tag{1.19}$$

for all $s \geq s_1$.

Step 2. Using the inequalities

$$e^{-2s_1\alpha}\,t^{-3}(T-t)^{-3} \geq e^{-2C(\Omega,\mathcal{O})(1+1/T)}\frac{1}{T^6} \quad \text{in} \quad \Omega\times(T/4,3T/4) \tag{1.20}$$

and

$$e^{-2s_1\alpha}\,t^{-3}(T-t)^{-3} \leq e^{-C(\Omega,\mathcal{O})(1+1/T)}\frac{1}{T^6} \quad \text{in} \quad \Omega\times(0,T), \tag{1.21}$$

we get

$$\iint_{\Omega\times(T/4,3T/4)} |\varphi|^2\,dx\,dt \leq C(\Omega,\mathcal{O},T)\iint_{\mathcal{O}\times(0,T)} |\varphi|^2\,dx\,dt, \tag{1.22}$$

with a constant $C(\Omega, \mathcal{O}, T)$ of the form $e^{C(\Omega,\mathcal{O})(1+1/T)}$.

Step 3. From the equation verified by φ, we readily obtain

$$\|\varphi(0)\|^2_{L^2(\Omega)} \leq \frac{2}{T}\int_{T/4}^{3T/4} \|\varphi(t)\|^2_{L^2(\Omega)}\, dt. \tag{1.23}$$

This, together with (1.22), gives the observability inequality (1.3).

In view of Theorem 1.1, we have the following.

THEOREM 1.4. *The system* (1.1) *is null controllable with controls* $v \in L^2(\mathcal{O} \times (0,T))$ *that verify*

$$\|v\|_{L^2(\mathcal{O}\times(0,T))} \leq C\|y^0\|_{L^2(\Omega)}, \tag{1.24}$$

where the constant C *is of the form* $e^{C(\Omega,\mathcal{O})(1+1/T)}$.

Let us now give the proof of the Carleman estimate (1.18). The proof we present here is based on the ideas in [30]. We will try to keep an explicit dependence of all the constants with respect to the parameters s and λ and the final time T. This follows the spirit of [27] and will be crucial for the analysis of similar nonlinear problems.

Proof of Lemma 1.3. For simplicity, we will divide the proof into three steps.

Step 1. *Change of variables and plan of what follows.* In this step, we set the differential equation satisfied by a new function ψ, which will be q up to a weight function.

Thus, let us introduce the new functions $\psi = e^{-s\alpha}\, q$ and $g = e^{-s\alpha}\, f$, where we have denoted $f = q_t + \Delta q$. Then we easily obtain that

$$M_1\psi + M_2\psi = g_{s,\lambda}, \tag{1.25}$$

where

$$\begin{aligned} M_1\psi &= -2s\,\lambda^2\,|\nabla\eta^0|^2\,\xi\,\psi - 2s\,\lambda\,\xi\,\nabla\eta^0\cdot\nabla\psi + \psi_t, \\ M_2\psi &= s^2\,\lambda^2\,|\nabla\eta^0|^2\,\xi^2\,\psi + \Delta\psi + s\,\alpha_t\,\psi, \end{aligned} \tag{1.26}$$

and

$$g_{s,\lambda} = g + s\,\lambda\,\Delta\eta^0\,\xi\,\psi - s\,\lambda^2\,|\nabla\eta^0|^2\,\xi\,\psi. \tag{1.27}$$

To simplify the notation, we will denote by $(M_i\psi)_j$ $(1 \leq i \leq 2,\ 1 \leq j \leq 3)$ the jth term in the expression of $M_i\psi$ given in (1.26).

Observe that, at first sight, it seems "natural" to put $(M_1\psi)_1$ on the right-hand side of (1.25). However, we have decided to keep it on the left, because it will serve to produce a positive term on $|\nabla\psi|^2$ (see the scalar product $((M_1\psi)_1, (M_2\psi)_2)_{L^2(Q)}$ below).

With the previous notation, we have from (1.25)

$$\|M_1\psi\|^2_{L^2(Q)} + \|M_2\psi\|^2_{L^2(Q)} + 2\sum_{i,j=1}^{3} ((M_1\psi)_i, (M_2\psi)_j)_{L^2(Q)} = \|g_{s,\lambda}\|^2_{L^2(Q)}. \tag{1.28}$$

In the following steps, we will see that the definition we have made of α makes $2(M_1\psi, M_2\psi)_{L^2(Q)}$ positive up to several terms that can be controlled whenever we make an appropriate choice of the parameters s and λ.

More precisely, in the second step we will make the computations of the double products $2(M_1\psi, M_2\psi)_{L^2(Q)}$. This will give an inequality with two global terms of

$|\psi|^2$ and $|\nabla\psi|^2$ on the left-hand side, while two local terms of $|\psi|^2$ and $|\nabla\psi|^2$ will appear on the right-hand side (see (1.57)). In the third step we will add two terms (involving ψ_t and $\Delta\psi$) to the left of (1.57). This will help us to eliminate the local term containing $\nabla\psi$ that appears on the right-hand side and will provide a Carleman inequality for the function ψ (see (1.63)). Finally, we will turn back to the original function q and deduce the inequality (1.18).

Step 2. *First estimates*. In this step, we will develop the nine terms appearing in $(M_1\psi, M_2\psi)_{L^2(Q)}$. For this, we will integrate by parts several times with respect to the space and time variables, so derivatives of the weight functions will be involved. Actually, we will use the estimates

$$\begin{aligned} &\partial_i\alpha = -\partial_i\xi = -\lambda\,\partial_i\eta^0\,\xi \le C\,\lambda\,\xi, \\ &\alpha_t = -(T-2t)\frac{e^{2\lambda m\|\eta^0\|_\infty} - e^{\lambda(m\|\eta^0\|_\infty+\eta^0)}}{t^2(T-t)^2} \le C\,T\,\xi^2, \end{aligned} \tag{1.29}$$

where C is a positive constant depending only on Ω and $\mathcal{O}$.

The last inequality follows from the fact that

$$e^{2\lambda m\|\eta^0\|_\infty} \le e^{2\lambda(m\|\eta^0\|_\infty+\eta^0)} \quad \text{in } \Omega. \tag{1.30}$$

First, we have

$$((M_1\psi)_1, (M_2\psi)_1)_{L^2(Q)} = -2s^3\,\lambda^4 \iint_Q |\nabla\eta^0|^4\,\xi^3\,|\psi|^2\,dx\,dt = A. \tag{1.31}$$

Then

$$\begin{aligned} ((M_1\psi)_2, (M_2\psi)_1)_{L^2(Q)} &= -2s^3\,\lambda^3 \iint_Q |\nabla\eta^0|^2\,\xi^3\,(\nabla\eta^0\cdot\nabla\psi)\,\psi\,dx\,dt \\ &= 3s^3\,\lambda^4 \iint_Q |\nabla\eta^0|^4\,\xi^3\,|\psi|^2\,dx\,dt \\ &\quad + s^3\,\lambda^3 \iint_Q \Delta\eta^0\,|\nabla\eta^0|^2\,\xi^3\,|\psi|^2\,dx\,dt \\ &\quad + 2s^3\,\lambda^3 \sum_{i,j=1}^N \iint_Q \partial_i\eta^0\,\partial_{ij}\eta^0\,\partial_j\eta^0\,\xi^3\,|\psi|^2\,dx\,dt \\ &= B_1 + B_2 + B_3. \end{aligned} \tag{1.32}$$

We clearly have that $A+B_1$ is a positive term. As a consequence of the properties of η^0 (see Lemma 1.2), we have

$$\begin{aligned} A + B_1 = s^3\,\lambda^4 \iint_Q |\nabla\eta^0|^4\,\xi^3\,|\psi|^2\,dx\,dt \ge C\,s^3\,\lambda^4 \iint_Q \xi^3\,|\psi|^2\,dx\,dt \\ - C\,s^3\,\lambda^4 \iint_{\omega\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt = \widetilde{A} - \widetilde{B} \end{aligned} \tag{1.33}$$

for some $C = C(\Omega, \mathcal{O})$. The first of these last two integrals ($\widetilde{A}$) will stay on the left-hand side, while the second one ($\widetilde{B}$) will go to the right-hand side.

The terms B_2 and B_3 are absorbed by $\widetilde{A}$ by simply taking $\lambda \ge C$, since $\eta^0 \in C^2(\overline{\Omega})$.

We also have

$$\begin{aligned}((M_1\psi)_3,(M_2\psi)_1)_{L^2(Q)} &= s^2\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi^2\,\psi_t\,\psi\,dx\,dt\\ &= -s^2\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,\xi_t\,|\psi|^2\,dx\,dt,\end{aligned}\tag{1.34}$$

which, by virtue of (1.29), is bounded by

$$C\,s^2\,\lambda^2\,T\iint_Q \xi^3\,|\psi|^2\,dx\,dt.\tag{1.35}$$

This term can also be absorbed if we take $\lambda \geq 1$ and $s \geq C(\Omega,\mathcal{O})\,T$.

Consequently, we have

$$\begin{aligned}(M_1\psi,(M_2\psi)_1)_{L^2(Q)} &= ((M_1\psi)_1+(M_1\psi)_2+(M_1\psi)_3,(M_2\psi)_1)_{L^2(Q)}\\ &\geq C\,s^3\,\lambda^4\iint_Q \xi^3\,|\psi|^2\,dx\,dt\\ &\quad - C\,s^3\,\lambda^4\iint_{\omega\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt\end{aligned}\tag{1.36}$$

for any $\lambda \geq C(\Omega,\mathcal{O})$ and $s \geq C(\Omega,\mathcal{O})\,T$.

On the other hand, we have

$$\begin{aligned}((M_1\psi)_1,(M_2\psi)_2)_{L^2(Q)} &= -2s\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,\Delta\psi\,\psi\,dx\,dt\\ &= 2s\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,|\nabla\psi|^2\,dx\,dt\\ &\quad + 4s\,\lambda^2\sum_{i,j=1}^N\iint_Q \partial_i\eta^0\,\partial_{ij}\eta^0\,\xi\,\partial_j\psi\,\psi\,dx\,dt\\ &\quad + 2s\,\lambda^3\iint_Q |\nabla\eta^0|^2\,\xi\,(\nabla\eta^0\cdot\nabla\psi)\,\psi\,dx\,dt\\ &= C_1+C_2+C_3.\end{aligned}\tag{1.37}$$

We will keep C_1 on the left-hand side. For C_2 and C_3, we have

$$C_2 \leq C\,s\,\lambda^4\iint_Q \xi\,|\psi|^2\,dx\,dt + C\,s\iint_Q \xi\,|\nabla\psi|^2\,dx\,dt\tag{1.38}$$

and

$$C_3 \leq C\,s^2\,\lambda^4\iint_Q \xi^2\,|\psi|^2\,dx\,dt + C\,\lambda^2\iint_Q |\nabla\psi|^2\,dx\,dt.\tag{1.39}$$

Therefore, taking $s \geq C\,T^2$, we find that

$$\begin{aligned}C_1+C_2+C_3 &\geq 2s\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,|\nabla\psi|^2\,dx\,dt\\ &\quad - C\,s^2\,\lambda^4\iint_Q \xi^2\,|\psi|^2\,dx\,dt - C\iint_Q (s\,\xi+\lambda^2)\,|\nabla\psi|^2\,dx\,dt.\end{aligned}\tag{1.40}$$

We also have

$$\begin{aligned}((M_1\psi)_2,(M_2\psi)_2)_{L^2(Q)} &= -2s\,\lambda\iint_Q \xi\,(\nabla\eta^0\cdot\nabla\psi)\,\Delta\psi\,dx\,dt\\ &= -2s\,\lambda\iint_\Sigma \frac{\partial\eta^0}{\partial n}\,\xi\,\left|\frac{\partial\psi}{\partial n}\right|^2 d\sigma\,dt\\ &\quad + 2s\,\lambda\sum_{i,j=1}^N\iint_Q \partial_{ij}\eta^0\,\xi\,\partial_i\psi\,\partial_j\psi\,dx\,dt\\ &\quad + 2s\,\lambda^2\iint_Q \xi\,|\nabla\eta^0\cdot\nabla\psi|^2\,dx\,dt\\ &\quad + s\,\lambda\iint_Q \xi\,\nabla\eta^0\cdot\nabla|\nabla\psi|^2\,dx\,dt\\ &= D_1+D_2+D_3+D_4.\end{aligned}\tag{1.41}$$

Let us remark that D_3 is a positive term. Furthermore,

$$D_2\le C\,s\,\lambda\iint_Q \xi\,|\nabla\psi|^2\,dx\,dt.\tag{1.42}$$

After some additional computations we also see that

$$\begin{aligned}D_4 &= s\,\lambda\iint_Q \xi\,\nabla\eta^0\cdot\nabla|\nabla\psi|^2\,dx\,dt = s\,\lambda\iint_\Sigma \xi\,\frac{\partial\eta^0}{\partial n}\,\left|\frac{\partial\psi}{\partial n}\right|^2 d\sigma\,dt\\ &\quad - s\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,|\nabla\psi|^2\,dx\,dt - s\,\lambda\iint_Q \Delta\eta^0\,\xi\,|\nabla\psi|^2\,dx\,dt\\ &= D_{41}+D_{42}+D_{43}.\end{aligned}\tag{1.43}$$

By virtue of the properties satisfied by η^0, we notice that $D_1+D_{41}\ge 0$ and that D_{43} can be bounded in the same way as D_2.

Consequently,

$$\begin{aligned}D_1+D_2+D_3+D_4 &\ge -s\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,|\nabla\psi|^2\,dx\,dt\\ &\quad - C\,s\,\lambda\iint_Q \xi\,|\nabla\psi|^2\,dx\,dt.\end{aligned}\tag{1.44}$$

Additionally, we find that

$$((M_1\psi)_3,(M_2\psi)_2)_{L^2(Q)} = \iint_Q \psi_t\,\Delta\psi\,dx\,dt = 0.\tag{1.45}$$

From (1.40)–(1.45), we deduce that

$$\begin{aligned}(M_1\psi,(M_2\psi)_2)_{L^2(Q)} &= ((M_1\psi)_1+(M_1\psi)_2+(M_1\psi)_3,(M_2\psi)_2)_{L^2(Q)}\\ &\ge s\,\lambda^2\iint_Q |\nabla\eta^0|^2\,\xi\,|\nabla\psi|^2\,dx\,dt\\ &\quad - C\,s^2\,\lambda^4\iint_Q \xi^2\,|\psi|^2\,dx\,dt\\ &\quad - C\iint_Q (s\,\lambda\,\xi+\lambda^2)\,|\nabla\psi|^2\,dx\,dt\end{aligned}\tag{1.46}$$

for $\lambda\ge 1$ and $s\ge C\,T^2$.

Hence, we have the following for $\lambda \geq C(\Omega, \mathcal{O})$ and $s \geq C(\Omega, \mathcal{O})\, T^2$:

$$\begin{aligned}(M_1\psi, (M_2\psi)_2)_{L^2(Q)} \geq{}& C\, s\, \lambda^2 \iint_Q \xi\, |\nabla\psi|^2\, dx\, dt \\ &- C\, s^2\, \lambda^4 \iint_Q \xi^2\, |\psi|^2\, dx\, dt \\ &- C\, s\, \lambda^2 \iint_{\omega\times(0,T)} \xi\, |\nabla\psi|^2\, dx\, dt.\end{aligned} \tag{1.47}$$

Let us now consider the scalar product

$$\begin{aligned}((M_1\psi)_1, (M_2\psi)_3)_{L^2(Q)} &= -2s^2\, \lambda^2 \iint_Q |\nabla\eta^0|^2\, \alpha_t\, \xi\, |\psi|^2\, dx\, dt \\ &\leq C\, s^2\, \lambda^2\, T \iint_Q \xi^3\, |\psi|^2\, dx\, dt.\end{aligned} \tag{1.48}$$

Obviously, this is absorbed by $\widetilde{A}$ if we take $\lambda \geq 1$ and $s \geq C\, T$.

Furthermore,

$$\begin{aligned}((M_1\psi)_2, (M_2\psi)_3)_{L^2(Q)} &= -2s^2\, \lambda \iint_Q \alpha_t\, \xi\, (\nabla\eta^0 \cdot \nabla\psi)\, \psi\, dx\, dt \\ &= s^2\, \lambda^2 \iint_Q \alpha_t\, |\nabla\eta^0|^2\, \xi\, |\psi|^2\, dx\, dt \\ &\quad + s^2\, \lambda \iint_Q \nabla\alpha_t \cdot \nabla\eta^0\, \xi\, |\psi|^2\, dx\, dt \\ &\quad + s^2\, \lambda \iint_Q \alpha_t\, \Delta\eta^0\, \xi\, |\psi|^2\, dx\, dt.\end{aligned} \tag{1.49}$$

From (1.29), one can easily check that the previous three terms can be bounded (if $\lambda \geq 1$) by

$$C\, s^2\, \lambda^2\, T \iint_Q \xi^3\, |\psi|^2\, dx\, dt. \tag{1.50}$$

Thus, we have

$$((M_1\psi)_2, (M_2\psi)_3)_{L^2(Q)} \geq -C\, s^2\, \lambda^2\, T \iint_Q \xi^3\, |\psi|^2\, dx\, dt. \tag{1.51}$$

Finally, we have

$$\begin{aligned}((M_1\psi)_3, (M_2\psi)_3)_{L^2(Q)} &= s \iint_Q \alpha_t\, \psi_t\, \psi\, dx\, dt \\ &= -\frac{1}{2} s \iint_Q \alpha_{tt}\, |\psi|^2\, dx\, dt \leq C\, s\, T^2 \iint_Q \xi^3\, |\psi|^2\, dx\, dt,\end{aligned} \tag{1.52}$$

since

$$\alpha_{tt} \leq C\, \xi^2 (1 + T^2\, \xi) \leq C\, T^2\, \xi^3. \tag{1.53}$$

From (1.48)–(1.52), we deduce for $\lambda \geq C(\Omega, \mathcal{O})$ and $s \geq C(\Omega, \mathcal{O})\, T$ that

$$\begin{aligned}(M_1\psi, (M_2\psi)_3)_{L^2(Q)} &= ((M_1\psi)_1 + (M_1\psi)_2 + (M_1\psi)_3, (M_2\psi)_3)_{L^2(Q)} \\ &\geq -C\ s^3\,\lambda^2 \iint_Q \xi^3\,|\psi|^2\,dx\,dt.\end{aligned} \tag{1.54}$$

Taking into account (1.36), (1.47), and (1.54), we obtain

$$\begin{aligned}(M_1\psi, M_2\psi)_{L^2(Q)} &\geq C \iint_Q (s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt \\ &\quad - C \iint_{\omega\times(0,T)} (s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt\end{aligned} \tag{1.55}$$

for any $\lambda \geq C(\Omega, \mathcal{O})$ and $s \geq C(\Omega, \mathcal{O})(T + T^2)$. Using (1.28), this gives

$$\begin{aligned}&\|M_1\psi\|^2_{L^2(Q)} + \|M_2\psi\|^2_{L^2(Q)} + \iint_Q (s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt \\ &\quad \leq C\left(\|g_{s,\lambda}\|^2_{L^2(Q)} + \iint_{\omega\times(0,T)} (s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt\right) \\ &\quad \leq C\bigg(\iint_Q e^{-2s\alpha}\,|f|^2\,dx\,dt + s^2\,\lambda^4 \iint_Q \xi^2\,|\psi|^2\,dx\,dt \\ &\qquad + s\lambda^2 \iint_{\omega\times(0,T)} \xi\,|\nabla\psi|^2\,dx\,dt + s^3\,\lambda^4 \iint_{\omega\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt\bigg).\end{aligned} \tag{1.56}$$

Thus, we also have

$$\begin{aligned}&\|M_1\psi\|^2_{L^2(Q)} + \|M_2\psi\|^2_{L^2(Q)} + \iint_Q (s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt \\ &\quad \leq C\left(\iint_Q e^{-2s\alpha}\,|f|^2\,dx\,dt + s\,\lambda^2 \iint_{\omega\times(0,T)} \xi\,|\nabla\psi|^2\,dx\,dt\right. \\ &\qquad \left. + s^3\,\lambda^4 \iint_{\omega\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt\right)\end{aligned} \tag{1.57}$$

for $\lambda \geq C(\Omega, \mathcal{O})$ and $s \geq C(\Omega, \mathcal{O})(T + T^2)$.

Step 3. *Indirect estimates and conclusion.* The final step will be to add integrals of $|\Delta\psi|^2$ and $|\psi_t|^2$ to the left-hand side of (1.57). This can be made using the expressions of $M_i\psi$ $(i = 1, 2)$. Indeed, from (1.26) we have

$$\begin{aligned}s^{-1} \iint_Q \xi^{-1}\,|\psi_t|^2\,dx\,dt &\leq C\bigg(s\,\lambda^2 \iint_Q \xi\,|\nabla\psi|^2\,dx\,dt \\ &\quad + s\,\lambda^4 \iint_Q \xi\,|\psi|^2\,dx\,dt + \|M_1\psi\|^2_{L^2(Q)}\bigg)\end{aligned} \tag{1.58}$$

and

$$\begin{aligned}s^{-1} \iint_Q \xi^{-1}\,|\Delta\psi|^2\,dx\,dt &\leq C\bigg(s^3\,\lambda^4 \iint_Q \xi^3\,|\psi|^2\,dx\,dt \\ &\quad + s\,T^2 \iint_Q \xi^3\,|\psi|^2\,dx\,dt + \|M_2\psi\|^2_{L^2(Q)}\bigg)\end{aligned} \tag{1.59}$$

for $s \geq C\,T^2$. Accordingly, we deduce from (1.57) that

$$\begin{aligned}
&\iint_Q (s^{-1}\,\xi^{-1}(|\psi_t|^2 + |\Delta\psi|^2) + s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt \\
&\qquad \leq C\left(\iint_Q e^{-2s\alpha}\,|f|^2\,dx\,dt + s\,\lambda^2 \iint_{\omega\times(0,T)} \xi\,|\nabla\psi|^2\,dx\,dt\right. \\
&\qquad\qquad \left. +\ s^3\,\lambda^4 \iint_{\omega\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt\right)
\end{aligned} \tag{1.60}$$

for any $\lambda \geq C(\Omega, \mathcal{O})$ and $s \geq C(\Omega, \mathcal{O})(T + T^2)$.

We are now ready to eliminate the second integral on the right-hand side. To this end, let us introduce a function $\theta = \theta(x)$, with

$$\theta \in C_c^2(\mathcal{O}), \quad \theta \equiv 1 \quad \text{in} \quad \omega, \quad 0 \leq \theta \leq 1, \tag{1.61}$$

and let us make some computations. We have

$$\begin{aligned}
s\,\lambda^2 \iint_{\omega\times(0,T)} \xi\,|\nabla\psi|^2\,dx\,dt &\leq\ s\,\lambda^2 \iint_{\mathcal{O}\times(0,T)} \theta\,\xi\,|\nabla\psi|^2\,dx\,dt \\
&= -s\,\lambda^2 \iint_{\mathcal{O}\times(0,T)} \theta\,\xi\,\Delta\psi\,\psi\,dx\,dt \\
&\quad - s\,\lambda^2 \iint_{\mathcal{O}\times(0,T)} \xi\,(\nabla\theta\cdot\nabla\psi)\,\psi\,dx\,dt \\
&\quad - s\,\lambda^3 \iint_{\mathcal{O}\times(0,T)} \theta\,\xi\,(\nabla\eta^0\cdot\nabla\psi)\,\psi\,dx\,dt \\
&\leq\ \varepsilon\,s^{-1} \iint_{\mathcal{O}\times(0,T)} \xi^{-1}\,|\Delta\psi|^2\,dx\,dt \\
&\quad + C\left(s^3\,\lambda^4 \iint_{\mathcal{O}\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt + s\,\lambda^4 \iint_{\mathcal{O}\times(0,T)} \xi\,|\psi|^2\,dx\,dt\right)
\end{aligned} \tag{1.62}$$

for a small enough constant $\varepsilon = \varepsilon(\Omega, \mathcal{O}) > 0$ and where we have used the fact that $\lambda \geq 1$. Hence, we can eliminate the integral of $|\nabla\psi|^2$ on the right-hand side of (1.60), paying the price of having $|\psi|^2$ in $\mathcal{O}\times(0,T)$. From (1.60) and this remark, we deduce that

$$\begin{aligned}
&\iint_Q (s^{-1}\,\xi^{-1}(|\psi_t|^2 + |\Delta\psi|^2) + s\,\lambda^2\,\xi\,|\nabla\psi|^2 + s^3\,\lambda^4\,\xi^3\,|\psi|^2)\,dx\,dt \\
&\qquad \leq C\left(\iint_Q e^{-2s\alpha}\,|f|^2\,dx\,dt + s^3\,\lambda^4 \iint_{\mathcal{O}\times(0,T)} \xi^3\,|\psi|^2\,dx\,dt\right)
\end{aligned} \tag{1.63}$$

for $\lambda \geq C(\Omega, \mathcal{O})$ and $s \geq C(\Omega, \mathcal{O})(T + T^2)$.

We finally turn back to our original function, which was given by $q = e^{s\alpha}\psi$. For

the moment, we have

$$\begin{aligned} & s^{-1}\iint_Q \xi^{-1}\,|\psi_t|^2\,dx\,dt + s^{-1}\iint_Q \xi^{-1}\,|\Delta\psi|^2\,dx\,dt \\ & \quad + s\,\lambda^2\iint_Q \xi\,|\nabla\psi|^2\,dx\,dt + s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \\ & \quad \le C\left(\iint_Q e^{-2s\alpha}\,|f|^2\,dx\,dt + s^3\,\lambda^4\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\right). \end{aligned} \tag{1.64}$$

Using that

$$\nabla q = e^{s\alpha}(\nabla\psi - s\,\lambda\,\nabla\eta^0\,\xi\,\psi), \tag{1.65}$$

we find

$$\begin{aligned} s\,\lambda^2\iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt \le C\,s\,\lambda^2\iint_Q \xi\,|\nabla\psi|^2\,dx\,dt \\ + C\,s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt. \end{aligned} \tag{1.66}$$

Consequently, we can add the previous integral of $|\nabla q|^2$ to the left-hand side of (1.64):

$$\begin{aligned} & s^{-1}\iint_Q \xi^{-1}\,|\psi_t|^2\,dx\,dt + s^{-1}\iint_Q \xi^{-1}\,|\Delta\psi|^2\,dx\,dt \\ & \quad + s\,\lambda^2\iint_Q \xi\,|\nabla q|^2\,dx\,dt + s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \\ & \quad \le C\left(\iint_Q e^{-2s\alpha}\,|f|^2\,dx\,dt + s^3\,\lambda^4\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\right). \end{aligned} \tag{1.67}$$

For Δq, we use the identity

$$\Delta\psi = e^{-s\alpha}(\Delta q + s\,\lambda\,\Delta\eta^0\,\xi\,q + s\,\lambda^2\,|\nabla\eta^0|^2\,\xi\,q + 2s\,\lambda\,\xi\,\nabla\eta^0\cdot\nabla q + s^2\,\lambda^2\,|\nabla\eta^0|^2\,\xi^2\,q) \tag{1.68}$$

and obtain

$$\begin{aligned} s^{-1}\iint_Q e^{-2s\alpha}\,\xi^{-1}\,|\Delta q|^2\,dx\,dt \le C\Bigg(s^{-1}\iint_Q \xi^{-1}\,|\Delta\psi|^2\,dx\,dt \\ + s\,\lambda^2\iint_Q e^{-2s\alpha}\,\xi\,|q|^2\,dx\,dt + s\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi\,|q|^2\,dx\,dt \\ + s\,\lambda^2\iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt + s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\Bigg). \end{aligned} \tag{1.69}$$

Finally, for q_t we get

$$\begin{aligned} s^{-1}\iint_Q e^{-2s\alpha}\,\xi^{-1}\,|q_t|^2\,dx\,dt \le C\Bigg(s^{-1}\iint_Q \xi^{-1}\,|\psi_t|^2\,dx\,dt \\ + s\,T^2\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\Bigg), \end{aligned} \tag{1.70}$$

where we have used the identity $q_t = e^{s\alpha}(\psi_t + s\,\alpha_t\,\psi)$. Thus, taking $\lambda \geq 1$ and $s \geq C(\Omega, \mathcal{O})(T + T^2)$, we are able to introduce all the terms involving $|\Delta q|^2$ and $|q_t|^2$ on the left-hand side of (1.64). This gives (1.18) and concludes the proof of Lemma 1.3.

Remark 4. It would be very interesting to know whether the powers of s and λ arising in (1.18) are optimal. In other words, would it be possible to deduce an inequality like (1.18) with higher powers of s and λ accompanying the terms on the left-hand side? A positive answer to this would lead to the controllability of nonlinear parabolic systems more general than those considered in subsections 1.3.2 and 2.2.1.

1.3. Applications and generalizations. There are many other similar systems for which null controllability properties can be analyzed as before. Let us mention some of them.

1.3.1. General parabolic linear systems. We will now try to apply the Carleman inequality obtained in the previous subsection to systems where the heat equation has a zero order term and a first order term in the divergence form. More precisely, we would like to prove the null controllability of the system

$$\begin{cases} y_t - \Delta y + \nabla \cdot (y\,B(x,t)) + a(x,t)\,y = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{1.71}$$

where $y^0 \in L^2(\Omega)$, $a \in L^\infty(Q)$, and $B \in L^\infty(Q)^N$.

In this case, the associated adjoint system is the following:

$$\begin{cases} -\varphi_t - \Delta\varphi - B(x,t)\cdot\nabla\varphi + a(x,t)\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega. \end{cases} \tag{1.72}$$

A result similar to Theorem 1.1 holds for systems (1.71) and (1.72). Consequently, what we have to do is prove an observability inequality for the solutions to (1.72).

In this situation, Lemma 1.3 provides a first estimate of the kind (1.18) (with q replaced by φ), with two additional terms on the right-hand side, namely,

$$C\iint_Q e^{-2s\alpha}\,|a\,\varphi|^2\,dx\,dt \quad \text{and} \quad C\iint_Q e^{-2s\alpha}\,|B\cdot\nabla\varphi|^2\,dx\,dt. \tag{1.73}$$

But these can be absorbed by the left-hand side if we take s large enough. Indeed, it suffices to take

$$s \geq C(\Omega,\mathcal{O})\,T^2(\|a\|_\infty^{2/3} + \|B\|_\infty^2) \tag{1.74}$$

to have

$$\begin{aligned} &C\left(\iint_Q e^{-2s\alpha}\,|a\,\varphi|^2\,dx\,dt + \iint_Q e^{-2s\alpha}\,|B\cdot\nabla\varphi|^2\,dx\,dt\right) \\ &\quad \leq \frac{1}{2}s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|\varphi|^2\,dx\,dt + \frac{1}{2}s\,\lambda^2\iint_Q e^{-2s\alpha}\,\xi\,|\nabla\varphi|^2\,dx\,dt. \end{aligned} \tag{1.75}$$

Arguing as we did when we proved the observability inequality for the solutions to (1.2), we easily find

$$\iint_{\Omega\times(T/4,3T/4)} |\varphi|^2\,dx\,dt \leq C(\Omega,\mathcal{O},T,a,B)\iint_{\mathcal{O}\times(0,T)} |\varphi|^2\,dx\,dt, \tag{1.76}$$

with a constant

$$C(\Omega,\mathcal{O},T,a,B)=\exp\{C(1+1/T+\|a\|_\infty^{2/3}+\|B\|_\infty^2)\}. \tag{1.77}$$

On the other hand, the dissipativity of φ yields

$$\|\varphi(0)\|^2_{L^2(\Omega)}\le \exp\{C\,T(\|a\|_\infty+\|B\|_\infty^2)\}\,\|\varphi(t)\|^2_{L^2(\Omega)} \tag{1.78}$$

for all $t\in(T/4,3T/4)$. Combining (1.76) and (1.78), we see that

$$\|\varphi(0)\|^2_{L^2(\Omega)}\le e^{C(1+1/T+\|a\|_\infty^{2/3}+T\,\|a\|_\infty+(1+T)\|B\|_\infty^2)}\iint_{\mathcal{O}\times(0,T)}|\varphi|^2\,dx\,dt \tag{1.79}$$

for all the solutions of (1.72) associated to final data $\varphi^0\in L^2(\Omega)$. As mentioned above (see Theorem 1.1), this implies the null controllability of (1.71).

THEOREM 1.5. *System* (1.71) *is null controllable, with controls* v *satisfying*

$$\|v\|_{L^2(\mathcal{O}\times(0,T))}\le C(\Omega,\mathcal{O},T,a,B)\|y^0\|_{L^2(\Omega)}, \tag{1.80}$$

where

$$C(\Omega,\mathcal{O},T,a,B)=e^{C(\Omega,\mathcal{O})(1+1/T+\|a\|_\infty^{2/3}+T\|a\|_\infty+(1+T)\|B\|_\infty^2)}. \tag{1.81}$$

Remark 5. In a recent paper, Seidman [54] looked at the qualitative asymptotic behavior of the constants found in estimates of the kind (1.80) with respect to relevant parameters. In particular, the blow-up of $C(\Omega,\mathcal{O},T,a,B)$ as $T\to 0$ has been analyzed in some particular cases.

The next step will be to establish the null controllability for a general linear heat equation system with first order terms and coefficients in $L^\infty(Q)$, i.e., an equation of the form

$$y_t-\Delta y+B(x,t)\cdot\nabla y+a(x,t)\,y=v1_{\mathcal{O}}, \tag{1.82}$$

with $a\in L^\infty(Q)$ and $B\in L^\infty(Q)^N$. A different Carleman inequality is now required since, in the corresponding adjoint equation, there is the term $-\nabla\cdot(\varphi\,B(x,t))$, which belongs to $L^2(0,T;H^{-1}(\Omega))$. This will be treated in section 2 using arguments from [39].

1.3.2. Extension to some semilinear systems. We will now consider possible extensions of the null controllability results above to nonlinear problems of the form

$$\begin{cases} y_t-\Delta y+f(y)=v1_{\mathcal{O}} & \text{in} \quad Q,\\ y=0 & \text{on} \quad \Sigma,\\ y(0)=y^0 & \text{in} \quad \Omega, \end{cases} \tag{1.83}$$

where

$$f\in C^1(\mathbf{R})\ \text{(for instance)},\quad f(0)=0,\quad \text{and}\quad |f(s)|\le C(1+|s|). \tag{1.84}$$

The classical strategy used in [36] and [30] relies on the introduction of a mapping

$$z\in L^2(Q)\mapsto y_z\in L^2(Q), \tag{1.85}$$

where y_z is, together with v_z, a solution of

$$\begin{cases} (y_z)_t - \Delta y_z + g(z)y_z = v_z 1_{\mathcal{O}} & \text{in} \quad Q, \\ y_z = 0 & \text{on} \quad \Sigma, \\ y_z(0) = y^0 & \text{in} \quad \Omega \end{cases} \tag{1.86}$$

such that

$$y_z(T) = 0 \tag{1.87}$$

and (1.80) holds for v_z with $a = g(z)$ and $B \equiv 0$. Here, g is given by

$$g(s) = \begin{cases} \dfrac{f(s)}{s} & \text{if} \quad s \neq 0, \\ f'(0) & \text{if} \quad s = 0. \end{cases} \tag{1.88}$$

Using the previous results and arguing as in [36], it can be seen that this mapping can be correctly defined. It can also be proved that it is continuous and compact. Furthermore, in view of (1.84), the estimate (1.80) written for v_z, and the classical estimates for y_z, it maps the whole space $L^2(Q)$ into a ball. Hence, from Schauder's theorem, we deduce that this mapping possesses a fixed point and, consequently, the following holds.

THEOREM 1.6. *Under conditions* (1.84) *on* f, *system* (1.83) *is null controllable, with controls* v *satisfying*

$$\|v\|_{L^2(\mathcal{O}\times(0,T))} \leq \exp\{C(\Omega,\mathcal{O})(1 + 1/T + \|g\|_\infty^{2/3} + T\|g\|_\infty)\}\|y^0\|_{L^2(\Omega)}, \tag{1.89}$$

where g *is given by* (1.88).

This fixed point approach for the solution of controllability problems for nonlinear systems was introduced in [56] in the context of the controllability of the wave equation. Later, it was successfully applied to the semilinear heat equation in [36], [19], [30], [20], and [28].

The next case corresponds to a superlinear f, i.e., to a function not satisfying the last condition in (1.84). In this context, it is much more complicated to estimate the norms of v_z and y_z, since g is not necessarily uniformly bounded. However, something can still be made in some particular cases. More precisely, the following result is proved in [28].

THEOREM 1.7. *Assume that*

$$\begin{cases} f \in C^1(\mathbf{R}) \quad (\textit{for instance}), \quad f(0) = 0, \quad \textit{and} \\ \dfrac{|f(s)|}{s\,\log^{3/2}(1+|s|)} \longrightarrow 0 \quad \textit{as } |s| \to +\infty. \end{cases} \tag{1.90}$$

Then (1.83) *is null controllable with controls* $v \in L^\infty(\mathcal{O} \times (0,T))$.

Remark 6. It is also proved in [28] that, in general, when

$$|f(s)| \sim |s|\log^\beta(1+|s|) \tag{1.91}$$

for some $\beta > 2$, (1.83) is not null controllable. However, it is unknown whether or not the system (1.83) is null controllable when this is satisfied with $\frac{3}{2} < \beta \leq 2$.

1.3.3. Another boundary conditions. Let us consider the following system, where the heat equation is completed with linear Robin (or Fourier) boundary conditions:

$$\begin{cases} y_t - \Delta y = v1_{\mathcal{O}} & \text{in} \quad Q, \\ \dfrac{\partial y}{\partial n} + a(x,t)\, y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{1.92}$$

where $a \in L^\infty(\Sigma)$ and $y^0 \in L^2(\Omega)$. Recall that under these assumptions there exists a unique solution y to (1.92), with

$$y \in L^2(0,T; H^1(\Omega)) \cap C^0([0,T]; L^2(\Omega)).$$

Arguing as above, it is readily seen that the null controllability of (1.92) is implied by the observability of the adjoint system

$$\begin{cases} -\varphi_t - \Delta \varphi = 0 & \text{in} \quad Q, \\ \dfrac{\partial \varphi}{\partial n} + a(x,t)\, \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega. \end{cases} \tag{1.93}$$

An observability inequality for the solutions to (1.93) is proved in [30], under the additional assumption $a_t \in L^\infty(\Sigma)$. As a consequence, one also has the null controllability of (1.92) whenever a fulfills this hypothesis. For some partial results concerning a nonlinear version of (1.92), see [15].

An improvement of these results has been obtained in the more recent work [21]. See section 2 for further details.

1.3.4. Another controllability problems. Another interesting problem is the null controllability of the heat equation with boundary controls, which means the null controllability of the system

$$\begin{cases} y_t - \Delta y = 0 & \text{in} \quad Q, \\ y = v1_\gamma & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{1.94}$$

when γ is a (small) part of $\partial\Omega$. In this case, the adjoint system is again (1.2) but the required observability inequality is somewhat different. More precisely, we need the estimate

$$\|\varphi(0)\|^2_{L^2(\Omega)} \le C \iint_{\gamma\times(0,T)} \left|\frac{\partial \varphi}{\partial n}\right|^2 d\sigma\, dt \tag{1.95}$$

for all solutions to (1.2).

This can be obtained as a consequence of the following Carleman estimates, similar to (1.18) (see [30]):

$$\begin{aligned} &\iint_Q e^{-2s\widetilde{\alpha}} \big((s\,\widetilde{\xi})^{-1}(|q_t|^2 + |\Delta q|^2) + s\,\lambda^2\,\widetilde{\xi}\,|\nabla q|^2 + s^3\,\lambda^4\,\widetilde{\xi}^3\,|q|^2\big)\, dx\, dt \\ &\quad \le C\left(\iint_Q e^{-2s\widetilde{\alpha}}\,|q_t + \Delta q|^2\, dx\, dt + s\iint_{\gamma\times(0,T)} e^{-2s\widetilde{\alpha}}\,\widetilde{\xi}\left|\frac{\partial q}{\partial n}\right|^2 d\sigma\, dt\right). \end{aligned}$$

Here, $\widetilde{\alpha} = \widetilde{\alpha}(x,t)$ and $\widetilde{\xi} = \widetilde{\xi}(x,t)$ are appropriate weight functions, similar to α and ξ, respectively.

Using (1.95) and arguments like those in the proof of Theorem 1.1, one can easily deduce the null controllability of (1.94).

1.3.5. Other more general partial differential equations. The previous null controllability results can also be extended to more general equations with sufficiently regular coefficients. For instance, for the system

$$\begin{cases} y_t - \partial_i(a_{ij}(x,t)\,\partial_j y) = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{1.96}$$

where the coefficients $a_{ij} \in C^2(\overline{Q})$ are uniformly elliptic, it is proved in [30] that null controllability is achieved. This result has recently been improved in [51]; see subsection 2.2.3 below.

2. Null controllability of other more general parabolic problems.

2.1. The case of the linear heat equation with general zero and first order terms. In this subsection, we are going to prove the null controllability, with distributed controls, of the linear heat equation with zero and first order terms and coefficients in $L^\infty(Q)$, i.e.,

$$\begin{cases} y_t - \Delta y + B(x,t)\cdot\nabla y + a(x,t)\,y = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{2.1}$$

with $y^0 \in L^2(\Omega)$, $a \in L^\infty(Q)$, and $B \in L^\infty(Q)^N$. To achieve this, we will first prove an observability inequality for the associated adjoint system

$$\begin{cases} -\varphi_t - \Delta\varphi - \nabla\cdot(\varphi\,B(x,t)) + a(x,t)\,\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega. \end{cases} \tag{2.2}$$

More precisely, we will prove that

$$\|\varphi(0)\|^2_{L^2(\Omega)} \le C \iint_{\mathcal{O}\times(0,T)} |\varphi|^2\,dx\,dt \tag{2.3}$$

for all $\varphi^0 \in L^2(\Omega)$ for some $C = C(\Omega, \mathcal{O}, T, a, B)$. To this end we will use the following lemma, which provides a suitable Carleman inequality (see [39]).

LEMMA 2.1. *There exist $\lambda_2 = C(\Omega,\mathcal{O})$, $s_2 = C(\Omega,\mathcal{O})(T+T^2)$, and $C_2 = C_2(\Omega,\mathcal{O}) > 0$ such that, for any $\lambda \ge \lambda_2$ and any $s \ge s_2$, one has*

$$\begin{aligned} &s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt + s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \\ &\quad\le C_2\Bigg(s^3\,\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \\ &\qquad + \iint_Q e^{-2s\alpha}\,|F_0|^2\,dx\,dt + s^2\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi^2\,|F|^2\,dx\,dt\Bigg) \end{aligned} \tag{2.4}$$

for all functions $q \in C^1(\overline{Q})$ *with* $q = 0$ *on* Σ *and* $q_t + \Delta q = F_0 + \nabla \cdot F$, *where* $F_0 \in L^2(Q)$, $F \in L^2(Q)^N$. *Here,* α *and* ξ *are the functions defined just before Lemma* 1.3.

For the moment, let us assume that Lemma 2.1 is true and let us deduce an observability inequality of the kind (2.3) for the solutions to (2.2). Fixing $\lambda = \lambda_2$ and applying Lemma 2.1 to φ, we have

$$\begin{aligned} &s^3 \iint_Q e^{-2s\alpha}\, \xi^3\, |\varphi|^2\, dx\, dt + s \iint_Q e^{-2s\alpha}\, \xi\, |\nabla \varphi|^2\, dx\, dt \\ &\quad \le C \left(s^3 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\, \xi^3\, |\varphi|^2\, dx\, dt + \|a\|_\infty^2 \iint_Q e^{-2s\alpha}\, |\varphi|^2\, dx\, dt \right. \\ &\qquad \left. + s^2\, \|B\|_\infty^2 \iint_Q e^{-2s\alpha}\, \xi^2 |\varphi|^2\, dx\, dt \right) \end{aligned} \tag{2.5}$$

for any $s \ge s_2$. Consequently, we also have

$$\iint_Q e^{-2s\alpha}\, t^{-3}(T-t)^{-3}\, |\varphi|^2\, dx\, dt \le C \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\, t^{-3}(T-t)^{-3}\, |\varphi|^2\, dx\, dt \tag{2.6}$$

for all $s \ge C\,(T + T^2 + T^2\,(\|a\|_\infty^{2/3} + \|B\|_\infty^2))$, whence it is easy to deduce that

$$\iint_{\Omega\times(T/4,3T/4)} |\varphi|^2\, dx\, dt \le e^{C(1+1/T+\|a\|_\infty^{2/3}+\|B\|_\infty^2)} \iint_{\mathcal{O}\times(0,T)} |\varphi|^2\, dx\, dt. \tag{2.7}$$

On the other hand, we also have (1.78) for the solutions to (2.2). This, together with (2.7), leads to the desired observability inequality (2.3). Thus, we have the following.

THEOREM 2.2. *System* (2.1) *is null controllable with controls* v *satisfying* (1.80) *and* (1.81).

Following [39], let us now give the proof of the Carleman inequality (2.4).

Proof of Lemma 2.1. Let $q = q(x,t)$ satisfy the hypothesis stated above and set $q^0 = q|_{t=T}$. Then q can be viewed as a solution by transposition of the backwards system

$$\begin{cases} q_t + \Delta q = F_0 + \nabla \cdot F & \text{in} \quad Q, \\ q = 0 & \text{on} \quad \Sigma, \\ q(T) = q^0 & \text{in} \quad \Omega. \end{cases} \tag{2.8}$$

In other words, for each $G \in L^2(0,T; H^{-1}(\Omega))$ we must have

$$\int_0^T \langle G(t), q(t)\rangle\, dt = -\int_0^T \langle F_0(t) + \nabla \cdot F(t), z(t)\rangle\, dt + (q^0, z(T))_{L^2(\Omega)}, \tag{2.9}$$

where z is the solution of the linear problem

$$\begin{cases} z_t - \Delta z = G & \text{in} \quad Q, \\ z = 0 & \text{on} \quad \Sigma, \\ z(0) = 0 & \text{in} \quad \Omega. \end{cases} \tag{2.10}$$

Here, $\langle \cdot, \cdot \rangle$ denotes the usual duality product between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Let s and λ be as in Lemma 1.3 and let us introduce the following fourth order problem, which will be justified below:

$$\begin{cases} \mathcal{L}(e^{-2s\alpha}\mathcal{L}^*p) + s^3\,\lambda^4\,e^{-2s\alpha}\,\xi^3\,q = -s^3\,\lambda^4\,e^{-2s\alpha}\,\xi^3\,p1_{\mathcal{O}} & \text{in} \quad Q, \\ p = 0, \quad e^{-2s\alpha}\mathcal{L}^*p = 0 & \text{on} \quad \Sigma, \\ (e^{-2s\alpha}\mathcal{L}^*p)(0) = (e^{-2s\alpha}\mathcal{L}^*p)(T) = 0 & \text{in} \quad \Omega. \end{cases} \tag{2.11}$$

Here, we have used the notation $\mathcal{L}q \equiv q_t - \Delta q, \quad \mathcal{L}^*q \equiv -q_t - \Delta q$. The partial differential problem (2.11) possesses exactly one (weak) solution p with

$$\iint_Q e^{-2s\alpha}(\xi^{-1}(|p_t|^2 + |\Delta p|^2) + \xi\,|\nabla p|^2 + \xi^3\,|p|^2)\,dx\,dt < +\infty. \tag{2.12}$$

Indeed, let P_0 be the linear space

$$P_0 = \{z \in C^2(\overline{Q}) : z = 0 \text{ on } \Sigma\} \tag{2.13}$$

and let us set

$$\begin{cases} \kappa(p,p') = \displaystyle\iint_Q e^{-2s\alpha}\,\mathcal{L}^*p\,\mathcal{L}^*p'\,dx\,dt \\ \qquad\qquad + s^3\,\lambda^4 \displaystyle\iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,p\,p'\,dx\,dt \quad \forall p, p' \in P_0 \end{cases} \tag{2.14}$$

and

$$l(p) = -s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,q\,p\,dx\,dt \quad \forall p \in P_0. \tag{2.15}$$

Then $\kappa(\cdot\,,\cdot)$ is a positive and symmetric bilinear form in P_0.

Let P be the completion of P_0 for the norm $\|p\|_P = (\kappa(p,p))^{1/2}$. Then P is a Hilbert space for the scalar product $\kappa(\cdot\,,\cdot)$ and, in view of the Carleman inequality (1.18), we have that the functions in P satisfy (2.12). It is also clear from (1.18) that l is a continuous linear form on P:

$$|l(p)| \le C\left(s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\right)^{1/2} \|p\|_P \quad \forall p \in P. \tag{2.16}$$

Consequently, from the Lax–Milgram lemma, the following variational equation possesses exactly one solution $p \in P$:

$$\kappa(p,p') = l(p') \quad \forall p' \in P. \tag{2.17}$$

It is not difficult to see that the unique solution to (2.17) also solves the fourth order problem (2.11) in the distributional sense.

Of course, the space P and the function $p \in P$ depend on the choice we have made of s and λ.

Now, let us set

$$\widehat{z} = -e^{-2s\alpha}\,\mathcal{L}^*p, \quad \widehat{u} = s^3\,\lambda^4\,e^{-2s\alpha}\,\xi^3\,p1_{\mathcal{O}}. \tag{2.18}$$

It is readily seen from (2.11) that $\widehat{z}$ is, together with $\widehat{u}$, a solution to the null controllability problem

$$\begin{cases} \widehat{z}_t - \Delta \widehat{z} = s^3\,\lambda^4\, e^{-2s\alpha}\,\xi^3\, q + \widehat{u} 1_{\mathcal{O}} & \text{in} \quad Q, \\ \widehat{z} = 0 & \text{on} \quad \Sigma, \\ \widehat{z}(0) = \widehat{z}(T) = 0 & \text{in} \quad \Omega. \end{cases} \tag{2.19}$$

For the moment, we will assume that there exist positive $\widetilde{s} = \widetilde{s}(\Omega, \mathcal{O})$ and $\widetilde{\lambda} = \widetilde{\lambda}(\Omega, \mathcal{O})$, such that

$$\begin{aligned} & s^{-3}\lambda^{-4} \iint_{\mathcal{O}\times(0,T)} e^{2s\alpha}\,\xi^{-3} |\hat{u}|^2\,dx\,dt + \iint_Q e^{2s\alpha}\,|\hat{z}|^2\,dx\,dt \\ & \quad + s^{-2}\lambda^{-2} \iint_Q e^{2s\alpha}\,\xi^{-2} |\nabla \hat{z}|^2\,dx\,dt \le C\,s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \end{aligned} \tag{2.20}$$

for all $s \ge \widetilde{s}(T + T^2)$ and $\lambda \ge \widetilde{\lambda}$.

Let us then prove (2.4). First, we will get an estimate for the second term on the left-hand side of (2.4). In view of (2.9) written for $G = s^3\,\lambda^4\, e^{-2s\alpha}\,\xi^3\, q + \widehat{u} 1_{\mathcal{O}}$, we have

$$s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt = -\iint_{\mathcal{O}\times(0,T)} q\,\widehat{u}\,dx\,dt + \int_0^T \langle F_0(t) + \nabla \cdot F(t), \widehat{z}(t)\rangle\,dt. \tag{2.21}$$

Therefore,

$$\begin{aligned} & s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \\ & \le \left(\int_0^T \int_{\mathcal{O}} e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \right)^{1/2} \left(\int_0^T \int_{\mathcal{O}} e^{2s\alpha}\,\xi^{-3}\,|\widehat{u}|^2\,dx\,dt \right)^{1/2} \\ & \quad + \left(\iint_Q e^{-2s\alpha} |F_0|^2\,dx\,dt \right)^{1/2} \left(\iint_Q e^{2s\alpha} |\widehat{z}|^2\,dx\,dt \right)^{1/2} \\ & \quad + \left(\iint_Q e^{-2s\alpha}\,\xi^2 |F|^2\,dx\,dt \right)^{1/2} \left(\iint_Q e^{2s\alpha}\,\xi^{-2} |\nabla \widehat{z}|^2\,dx\,dt \right)^{1/2}. \end{aligned} \tag{2.22}$$

And now, using (2.20), we see that

$$\begin{aligned} & s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \le C \Bigg(s^3\,\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \\ & \quad + \iint_Q e^{-2s\alpha}\,|F_0|^2\,dx\,dt + s^2\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi^2\,|F|^2\,dx\,dt \Bigg) \end{aligned} \tag{2.23}$$

for $s \ge \widehat{s}(T + T^2)$ and $\lambda \ge \widehat{\lambda}$.

Let us now get an estimate for the first term on the left-hand side of (2.4). To this end, we multiply by $s\,\lambda^2\, e^{-2s\alpha}\,\xi\, q$ the equation satisfied by q and integrate in space

and time. This gives

$$
\begin{aligned}
s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt + s\frac{\lambda^2}{2}\iint_Q \left((e^{-2s\alpha}\,\xi)_t - \Delta(e^{-2s\alpha}\,\xi)\right)|q|^2\,dx\,dt \\
= s\,\lambda^2 \iint_Q e^{-2s\alpha}(\xi\,F_0\,q - \xi\,F\cdot\nabla q)\,dx\,dt \\
- s\,\lambda^2 \iint_Q F\cdot\nabla(e^{-2s\alpha}\,\xi) q\,dx\,dt.
\end{aligned}
\tag{2.24}
$$

Then we use the estimates

$$
\begin{aligned}
&|(e^{-2s\alpha}\,\xi)_t| \le C\,e^{-2s\alpha}\,(s\,T\,\xi^3 + T\,\xi^2) \le C\,s^2\,e^{-2s\alpha}\,\xi^3, \\
&|\Delta(e^{-2s\alpha}\,\xi)| \le C\,s^2\,\lambda^2\,e^{-2s\alpha}\,\xi^3
\end{aligned}
\tag{2.25}
$$

for $s \ge C(T + T^2)$,

$$
\begin{aligned}
s\,\lambda^2 \left|\iint_Q e^{-2s\alpha}\,\xi\,F_0\,q\,dx\,dt\right| &\le C\left(s^3\lambda^4 \iint_Q e^{-2s\alpha}\xi^3|q|^2\,dx\,dt \right. \\
&\qquad \left. + s^{-1}\iint_Q e^{-2s\alpha}\xi^{-1}|F_0|^2\,dx\,dt\right), \\
s\,\lambda^2 \left|\iint_Q e^{-2s\alpha}\,\xi\,F\cdot\nabla q\,dx\,dt\right| &\le C\,s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|F|^2\,dx\,dt \\
&\qquad + \frac{1}{2}s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt,
\end{aligned}
\tag{2.26}
$$

and

$$
\begin{aligned}
&s\,\lambda^2 \left|\iint_Q F\cdot\nabla(e^{-2s\alpha}\,\xi) q\,dx\,dt\right| \\
&\le C\left(s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|F|^2\,dx\,dt + s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\right),
\end{aligned}
\tag{2.27}
$$

to obtain from (2.24) the following:

$$
\begin{aligned}
s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|\nabla q|^2\,dx\,dt \le C\left(s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt \right. \\
\left. + \iint_Q e^{-2s\alpha}\,|F_0|^2\,dx\,dt + s\,\lambda^2 \iint_Q e^{-2s\alpha}\,\xi\,|F|^2\,dx\,dt\right)
\end{aligned}
\tag{2.28}
$$

for $s \ge C(T + T^2)$. This, together with (2.23), provides (2.4).

Let us finally prove (2.20).

We first multiply the equation in (2.11) by p, so we have

$$
\chi(p,p) = -s^3\,\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,q\,p\,dx\,dt,
\tag{2.29}
$$

which, combined with the Carleman inequality (1.18) proved in Lemma 1.3, provides the desired inequality for the two first terms on the left-hand side of (2.20).

To estimate the first order term, we multiply by $s^{-2}\lambda^{-2}e^{2s\alpha}\,\xi^{-2}\,\widehat{z}$ the equation verified by $\widehat{z}$. Then we integrate by parts with respect to the space variable and get

$$\begin{aligned} &s^{-2}\lambda^{-2}\iint_Q e^{2s\alpha}\,\xi^{-2}\,\widehat{z}\,\widehat{z}_t\,dx\,dt + s^{-2}\lambda^{-2}\iint_Q e^{2s\alpha}\,\xi^{-2}\,|\nabla\widehat{z}|^2\,dx\,dt \\ &\quad - 2s^{-1}\lambda^{-1}\iint_Q e^{2s\alpha}\,\xi^{-1}\nabla\eta^0\cdot\nabla\widehat{z}\,\widehat{z}\,dx\,dt \\ &\quad - 2s^{-2}\lambda^{-1}\iint_Q e^{2s\alpha}\,\xi^{-2}\nabla\eta^0\cdot\nabla\widehat{z}\,\widehat{z}\,dx\,dt \\ &\quad = s\,\lambda^2\iint_Q \xi\,q\,\widehat{z} + s^{-2}\lambda^{-2}\iint_{\mathcal{O}\times(0,T)} e^{2s\alpha}\,\xi^{-2}\,\widehat{u}\,\widehat{z}\,dx\,dt. \end{aligned} \tag{2.30}$$

This time, let us integrate by parts in the first term with respect to the time variable:

$$\begin{aligned} &s^{-2}\lambda^{-2}\iint_Q e^{2s\alpha}\,\xi^{-2}\,\widehat{z}\,\widehat{z}_t\,dx\,dt \\ &\quad = -\frac{1}{2}s^{-2}\lambda^{-2}\iint_Q (e^{2s\alpha}\,\xi^{-2})_t\,|\widehat{z}|^2\,dx\,dt \\ &\quad \le C\,s^{-1}\lambda^{-2}\,T\iint_Q e^{2s\alpha}\,|\widehat{z}|^2\,dx\,dt \le C\iint_Q e^{2s\alpha}\,|\widehat{z}|^2\,dx\,dt \end{aligned} \tag{2.31}$$

for $s \ge C\,T$ and $\lambda \ge 1$.

Finally, we use Young's inequality for the other terms of (2.30) and obtain

$$\begin{aligned} &- 2s^{-1}\lambda^{-1}\iint_Q e^{2s\alpha}\,\xi^{-1}\nabla\eta^0\cdot\nabla\widehat{z}\,\widehat{z}\,dx\,dt \\ &- 2s^{-2}\lambda^{-1}\iint_Q e^{2s\alpha}\,\xi^{-2}\nabla\eta^0\cdot\nabla\widehat{z}\,\widehat{z}\,dx\,dt \\ &\quad \le C\iint_Q e^{2s\alpha}\,|\widehat{z}|^2\,dx\,dt + \frac{1}{2}s^{-2}\lambda^{-2}\iint_Q e^{2s\alpha}\,\xi^{-2}\,|\nabla\widehat{z}|^2\,dx\,dt, \end{aligned} \tag{2.32}$$

$$s\,\lambda^2\iint_Q \xi\,q\,\widehat{z} \le C\left(\iint_Q e^{2s\alpha}\,|\widehat{z}|^2\,dx\,dt + s^3\,\lambda^4\iint_Q e^{-2s\alpha}\,\xi^3\,|q|^2\,dx\,dt\right), \tag{2.33}$$

and

$$\begin{aligned} &s^{-2}\lambda^{-2}\iint_{\mathcal{O}\times(0,T)} e^{2s\alpha}\,\xi^{-2}\,\widehat{u}\,\widehat{z}\,dx\,dt \\ &\quad \le C\left(\iint_Q e^{2s\alpha}\,|\widehat{z}|^2\,dx\,dt + s^{-3}\lambda^{-4}\iint_{\mathcal{O}\times(0,T)} e^{2s\alpha}\,\xi^{-3}\,|\widehat{u}|^2\,dx\,dt\right) \end{aligned} \tag{2.34}$$

for $s \ge C\,T^2$.

As a conclusion, we deduce (2.20) directly from (2.30).

Remark 7. In the same spirit of Remark 4, it would be very interesting to know whether or not the powers of s and λ in (2.4) are optimal.

2.2. Null controllability of other semilinear problems and further comments. With the help of Lemma 2.1, Theorem 2.2, and some variants, we can now prove the null controllability of other nonlinear parabolic systems. This is the goal of this section.

2.2.1. Null controllability of the semilinear heat equation with nonlinearities in the zero and first order terms. In this paragraph, we will deduce the null controllability of the system

$$\begin{cases} y_t - \Delta y + f(y, \nabla y) = v 1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{2.35}$$

where we suppose $y^0 \in H_0^1(\Omega) \cap W^{1,\infty}(\Omega)$, $f \in C^1(\mathbf{R} \times \mathbf{R}^N)$, $f(0,0) = 0$, and

$$f(s,p) = g(s,p)s + G(s,p) \cdot p \quad \forall (s,p) \in \mathbf{R} \times \mathbf{R}^N \tag{2.36}$$

for some functions g and G satisfying $|g| \leq C$ and $|G| \leq C$.

To this end, we will combine Theorem 2.2 and a fixed point argument.

Thus, let us consider the map

$$z \in L^2(0,T; H_0^1(\Omega)) \mapsto y_z \in L^2(0,T; H_0^1(\Omega)), \tag{2.37}$$

where (y_z, v_z) is the solution of

$$\begin{cases} (y_z)_t - \Delta y_z + g(z, \nabla z)\, y_z + G(z, \nabla z) \cdot \nabla y_z = v_z 1_{\mathcal{O}} & \text{in} \quad Q, \\ y_z = 0 & \text{on} \quad \Sigma, \\ y_z(0) = y^0, \quad y_z(T) = 0 & \text{in} \quad \Omega, \end{cases} \tag{2.38}$$

constructed as in the previous subsection. We know that

$$\|v_z\|_{L^2(\mathcal{O} \times (0,T))} \leq K \, \|y^0\|_{L^2(\Omega)}, \tag{2.39}$$

with a constant K of the form

$$K = \exp\left\{ C \left(1 + \frac{1}{T} + \|g\|_\infty^{2/3} + T \, \|g\|_\infty + (1+T)\|G\|_\infty^2 \right) \right\}. \tag{2.40}$$

Now, one can check from parabolic regularity and the fact that g and G are uniformly bounded that Kakutani's fixed point theorem is applicable. For convenience, let us state this result.

THEOREM 2.3. *Let Z be a Banach space and let $\Lambda : Z \mapsto Z$ be a set-valued mapping satisfying the following assumptions:*

1. *$\Lambda(z)$ is a nonempty closed convex set of Z for every $z \in Z$.*
2. *There exists a nonempty convex compact set $K \subset Z$ such that $\Lambda(K) \subset K$.*
3. *Λ is upper-hemicontinuous in Z, i.e., for each $\sigma \in Z'$ the single-valued mapping*

$$z \mapsto \sup_{y \in \Lambda(z)} \langle \sigma, y \rangle_{Z',Z} \tag{2.41}$$

is upper-semicontinuous.

Then Λ possesses a fixed point in the set K, i.e., there exists $z \in K$ such that $z \in \Lambda(z)$.

For the proof, see, for instance, [3]. We then have the following result.

THEOREM 2.4. *Under the previous assumptions on y^0 and f,* (2.35) *is null controllable with controls v satisfying*

$$\|v\|_{L^2(\mathcal{O}\times(0,T))} \leq K \|y^0\|_{L^2(\Omega)}, \tag{2.42}$$

where the constant K is given by (2.40).

Now, we pretend to extend this result to the (superlinear) case where we only have $f \in C^1(\mathbf{R}\times\mathbf{R}^N)$ and $f(0,0) = 0$. This time, to apply similar arguments, one must work in the space $L^\infty(0,T;W^{1,\infty}(\Omega))$ instead of $L^2(0,T;H_0^1(\Omega))$. In fact, we have the following result (see [16]).

THEOREM 2.5. *Assume that $f \in C^1(\mathbf{R}\times\mathbf{R}^N)$, $f(0,0) = 0$, and* (2.36) *holds with*

$$\frac{g(s,p)}{\log^{3/2}(1+|s|+|p|)} \to 0 \quad \textit{and} \quad \frac{G(s,p)}{\log^{1/2}(1+|s|+|p|)} \to 0 \tag{2.43}$$

as $|s| \to +\infty$. Then (2.35) *is null controllable with controls in $L^\infty(\mathcal{O}\times(0,T))$.*

Remark 8. As in Theorem 1.7, it is unknown whether system (2.35) is controllable when

$$|g(s,p)| \sim \log^\beta(1+|s|+|p|) \quad \text{and/or} \quad |G(s,p)| \sim \log^\gamma(1+|s|+|p|) \tag{2.44}$$

with $\beta > \frac{3}{2}$, $\gamma > \frac{1}{2}$.

Let us now address other typical controllability problems in this context.

(A) Going back to system (1.83), with no hypothesis on the growth of f, the null controllability can be achieved if f possesses the "good sign," that is to say, if $f(s)\,s \geq 0$ for all $s \in \mathbf{R}$.

In this situation, it can be proved that, for any $\rho > 0$, there exists a time $T = T(\Omega,\mathcal{O},f,\rho) > 0$ such that, for every $y^0 \in L^2(\Omega)$ with $\|y^0\|_{L^2(\Omega)} \leq \rho$, there exists a control $v \in L^2(\mathcal{O}\times(0,T))$ whose associated solution y verifies $y(T) = 0$ (see [2], [5], and [28] for more details).

(B) Using methods similar to those above, the exact controllability to the trajectories can be deduced as well. For instance, the following result is proved in [16].

THEOREM 2.6. *Let us assume that $y^0 \in H_0^1(\Omega)\cap W^{1,\infty}(\Omega)$, $f \in C^1(\mathbf{R}\times\mathbf{R}^N)$,*

$$\frac{1}{\log^{3/2}(1+|s|+|p|)} \left| \int_0^1 \frac{\partial f}{\partial s}(s_0+\lambda s, p_0+\lambda p)\, d\lambda \right| \to 0 \tag{2.45}$$

and

$$\frac{1}{\log^{1/2}(1+|s|+|p|)} \left| \int_0^1 \frac{\partial f}{\partial p_i}(s_0+\lambda s, p_0+\lambda p)\, d\lambda \right| \to 0 \tag{2.46}$$

for $1 \leq i \leq N$ as $|(s,p)| \to +\infty$, uniformly in $(s_0,p_0) \in K$ for every compact set $K \subset \mathbf{R}\times\mathbf{R}^N$. Let $y_*^0 \in H_0^1(\Omega)\cap W^{1,\infty}(\Omega)$ and $v_* \in L^\infty(\mathcal{O}\times(0,T))$ be given and let $y_* \in C([0,T];W^{1,\infty}(\Omega))$ be the corresponding solution to (2.35). *Then there exists $v \in L^\infty(\mathcal{O}\times(0,T))$ such that its associated solution y verifies $y(T) = y_*(T)$.*

2.2.2. Heat equations with Fourier boundary conditions. In this paragraph, we will recall some recent results concerning the null and exact controllability to the trajectories of the following partial differential systems:

$$\begin{cases} y_t - \Delta y + B \cdot \nabla y + a\, y = v 1_{\mathcal{O}} & \text{in} \quad Q, \\ \dfrac{\partial y}{\partial n} + \beta\, y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega \end{cases} \tag{2.47}$$

and

$$\begin{cases} y_t - \Delta y + F(y, \nabla y) = v 1_{\mathcal{O}} & \text{in} \quad Q, \\ \dfrac{\partial y}{\partial n} + f(y) = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{2.48}$$

where $y^0 \in L^2(\Omega)$,

$$a \in L^\infty(Q), \quad B \in L^\infty(Q)^N, \quad \beta \in L^\infty(\Sigma), \tag{2.49}$$

and (for instance)

$$F \in C^1(\mathbf{R} \times \mathbf{R}^N), \quad f \in C^1(\mathbf{R}). \tag{2.50}$$

Let us first deduce a null controllability result for (2.47). To this end, we will need, as before, an observability estimate for the solutions to the adjoint system associated to (2.47).

Thus, let us consider the backwards system

$$\begin{cases} -\varphi_t - \Delta\varphi - \nabla \cdot (\varphi\, B(x,t)) + a(x,t)\, \varphi = 0 & \text{in} \quad Q, \\ (\nabla\varphi + \varphi\, B(x,t)) \cdot n + \beta(x,t)\, \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega, \end{cases} \tag{2.51}$$

where $\varphi^0 \in L^2(\Omega)$. For the solutions to these problems, one has the following Carleman inequality (see [21]):

$$\iint_Q e^{-2s\zeta}\, \kappa^3\, |\varphi|^2\, dx\, dt \le C \iint_{\mathcal{O}\times(0,T)} e^{-2s\zeta}\, \kappa^3\, |\varphi|^2\, dx\, dt \tag{2.52}$$

for appropriate positive weight functions κ and ζ (similar to those defined in section 1) depending on a parameter λ and a suitable choice of s and λ. The proof of (2.52) can be achieved using duality arguments similar to those presented in the proof of Lemma 2.1.

From (2.52), one can easily deduce an observability inequality for the solutions to (2.51). This yields the next null controllability result.

THEOREM 2.7. *Let* (2.49) *be fulfilled. Then, for all* $y^0 \in L^2(\Omega)$, *there exists a control* $v \in L^\infty(\mathcal{O} \times (0,T))$ *such that the associated solution to* (2.47) *verifies*

$$y(T) = 0 \quad \textit{in} \quad \Omega. \tag{2.53}$$

Finally, the same arguments used in [28], based on the explicit dependence of the constant C in (2.52) with respect to T and the L^∞-norms of the coefficients a, B, and β, lead to an exact controllability result to the trajectories for the nonlinear system (2.48).

THEOREM 2.8. *Let* (2.50) *be fulfilled and let* $\overline{y}$ *be a solution of the nonlinear system* (2.48) *with* $v \equiv 0$ *such that*

$$\overline{y} \in L^2(0,T;H^1(\Omega)) \cap L^\infty(Q), \quad \overline{y}(0) \in L^\infty(\Omega).$$

Moreover, assume that the functions F *and* f *verify*

$$\lim_{|s|\to\infty} \frac{|F(s,p)-F(r,p)|}{|s-r|\log^{3/2}(1+|s-r|)} = 0 \tag{2.54}$$

uniformly in $(r,p) \in \mathbf{R}\times\mathbf{R}^N$,

$$\begin{cases} \forall L>0,\ \exists M>0 \ \text{ such that} \\ |F(s,p)-F(r,p)| \le M|s-r|, \quad |F(s,p)-F(s,q)| \le M|p-q| \\ \forall (s,r,p,q) \in [-L,L]^2 \times \mathbf{R}^N \times \mathbf{R}^N, \end{cases} \tag{2.55}$$

and

$$\lim_{|s|\to\infty} \frac{|f(s)-f(r)|}{|s-r|\log^{1/2}(1+|s-r|)} = 0 \tag{2.56}$$

uniformly in $r \in \mathbf{R}$. *Then there exist controls* $v \in L^\infty(\mathcal{O}\times(0,T))$ *and associated solutions* y *to* (2.48) *such that* $y(T) = \overline{y}(T)$.

For the proof of this result and other related questions, see [22].

2.2.3. Null controllability of some variants of the heat system. Here, we are interested in solving the null controllability problem of more general parabolic systems, where various coefficients appear in the differential equation.

We first consider the case of a general second order differential operator in the heat system—more precisely, the following problem:

$$\begin{cases} y_t - \displaystyle\sum_{i,j=1}^N \partial_i(a_{ij}\,\partial_j y) = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{2.57}$$

where $y^0 \in L^2(\Omega)$ and the coefficients $a_{ij} = a_{ji} \in W^{1,\infty}(Q)\,(1 \le i,j \le N)$ form a uniformly elliptic matrix. For the solutions to the adjoint system associated to (2.57) (which is in fact the same), we find a global Carleman inequality in the recent paper [51]. Of course, this leads to the null controllability of (2.57).

Analogously as we found in Theorem 2.2, this allows us to deduce a null controllability result for (2.57) in a more general case.

THEOREM 2.9. *Let us assume that, in* (2.57), *the coefficients* a_{ij} *are of the form* $a_{ij} = a_{ij}^0 + \varepsilon\, h_{ij}$ *with* $a_{ij}^0 = a_{ji}^0 \in W^{1,\infty}(Q)$ *uniformly elliptic,* $(T-t)^{-1/2}h_{ij} \in L^\infty(Q)$ *and* ε *small enough, depending on* Ω, $\mathcal{O}$, T, a_{ij}^0, *and* h_{ij}. *Then, for any* $T>0$, *system* (2.57) *is null controllable at time* T.

For the proof of Theorem 2.9, it suffices to combine the previously mentioned result for general $a_{ij} \in W^{1,\infty}(Q)$ with Lemma 2.1 adapted to this situation.

The second case we consider is the following:

$$\begin{cases} y_t - \displaystyle\sum_{i,j=1}^{N} a_{ij}\,\partial_{ij} y = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{2.58}$$

with $y^0 \in L^2(\Omega)$. To prove the null controllability of (2.58), we can use a recent Carleman inequality which can be found in [23] for heat equations with right-hand sides in $L^2(0,T;H^{-2}(\Omega))$. This is given in the following result.

LEMMA 2.10. *Let us consider the following problem:*

$$\begin{cases} -z_t - \Delta z = \displaystyle\sum_{i,j=1}^{N} \partial_{ij} H_{ij} & \textit{in} \quad Q, \\ z = 0 & \textit{on} \quad \Sigma, \\ z(T) = z^0 & \textit{in} \quad \Omega, \end{cases} \tag{2.59}$$

where $z_0 \in H^{-1}(\Omega)$ and the functions H_{ij} verify $H_{ij} \in L^2(Q)$ with

$$\sum_{j=1}^{N} \partial_j H_{ij} \in L^2(Q) \quad \textit{and} \quad \sum_{j=1}^{N} H_{ij}\, n_j = 0. \tag{2.60}$$

There exist constants $C_3(\Omega,\mathcal{O}) > 0$, $\lambda_3 = C(\Omega,\mathcal{O})$, and $s_3 = C(\Omega,\mathcal{O})(T+T^2)$ such that

$$\begin{aligned} s^3\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|z|^2\,dx\,dt \le C_3 \Bigg(& s^3\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|z|^2\,dx\,dt \\ & + s^4\lambda^4 \sum_{i,j=1}^{N} \iint_Q e^{-2s\alpha}\,\xi^4\,|H_{ij}|^2\,dx\,dt \Bigg) \end{aligned} \tag{2.61}$$

for all $\lambda \ge \lambda_3$ and $s \ge s_3$.

Let us introduce the adjoint system to (2.58):

$$\begin{cases} -\varphi_t - \partial_{ij}(a_{ij}\,\varphi) = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega. \end{cases} \tag{2.62}$$

A simple application of Lemma 2.10 gives the following result.

THEOREM 2.11. *Let T be an arbitrary positive time. Let us suppose that $a_{ij} = \delta_{ij} + \varepsilon\, h_{ij}\,(i,j = 1,\dots,N)$ with $(T-t)^{-1/2} h_{ij} \in L^\infty(Q)$ and $\varepsilon > 0$ small enough, depending on Ω, $\mathcal{O}$, and T. Then* (2.58) *is null controllable at time T.*

Let us finally consider the control problem

$$\begin{cases} b\,y_t - \Delta y = v1_{\mathcal{O}} & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega, \end{cases} \tag{2.63}$$

whose adjoint system is the following:

$$\begin{cases} -(b\,\varphi)_t - \Delta\varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega. \end{cases} \tag{2.64}$$

In order to analyze the null controllability of (2.63), we need a Carleman inequality for heat equations with right-hand sides in $H^{-1}(0,T;L^2(\Omega))$. Its proof can also be found in [23].

LEMMA 2.12. *Consider the system*

$$\begin{cases} -z_t - \Delta z = G_t\,Q, \\ z = 0\,\Sigma, \\ z(T) = z^0\,\Omega, \end{cases} \tag{2.65}$$

where $z^0 \in H^{-1}(\Omega)$ *and* $G \in L^2(Q)$ *with* $G \in C_w([0,T];H^{-1}(\Omega))$. *Then there exist* $C_4(\Omega,\mathcal{O}) > 0$, $\lambda_4 = C(\Omega,\mathcal{O})$, *and* $s_4 = C(\Omega,\mathcal{O})(T+T^2)$ *such that*

$$\begin{aligned} s^3\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^3\,|z|^2\,dx\,dt \le C_4 \Bigg(& s^3\lambda^4 \iint_{\mathcal{O}\times(0,T)} e^{-2s\alpha}\,\xi^3\,|z|^2\,dx\,dt \\ & + s^4\lambda^4 \iint_Q e^{-2s\alpha}\,\xi^4\,|G|^2\,dx\,dt \Bigg) \end{aligned} \tag{2.66}$$

for all $\lambda \ge \lambda_4$ *and* $s \ge s_4$.

From this Carleman inequality, one can readily deduce the following result.

THEOREM 2.13. *Let* T *be a positive time and let us assume that* $b = 1 + \varepsilon\,h$ *with* $(T-t)^{-1/2}h \in L^\infty(Q)$ *and* $\varepsilon > 0$ *small enough, depending on* Ω, $\mathcal{O}$, *and* T. *Then system* (2.63) *is null controllable at time* T.

Remark 9. The proofs of Lemmas 2.10 and 2.12 are based on the duality arguments presented in Lemma 2.1. Observe in particular that the solutions to systems (2.58) and (2.63) belong only to $L^2(Q)$ and must be defined by transposition.

3. Null controllability of systems of the Stokes kind. In this section, $N = 2$ or $N = 3$. The controllability properties of the Navier–Stokes system have been the subject of intensive research these last years. The question was first considered by Lions in [48], where *approximate controllability* was conjectured. This was followed by several papers, where various partial (positive) answers were furnished. See [18], [10], [11], and [50]. Concerning null controllability and exact controllability to the trajectories, the first local results were given in [29].

For completeness, we will recall briefly one of the main results in [10]. This paper deals with the approximate controllability of the two-dimensional Navier–Stokes equations with *Navier-slip* boundary conditions on the lateral boundary $\partial\Omega \times (0,T)$. Thus, let us assume that the state equation is

$$\begin{cases} y_t + (y\cdot\nabla)y - \Delta y + \nabla p = v1_{\mathcal{O}}, \quad \nabla\cdot y = 0 & \text{in} \quad Q, \\ y\cdot n = 0, \quad \nabla\times y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega \end{cases} \tag{3.1}$$

(where $\Omega \subset \mathbf{R}^2$). The general fact the author makes good use of is the treatment of $(y\cdot\nabla)y$ as the leading term, while $-\Delta y$ is considered as a "disturbance."

Roughly speaking, the following is proved: Let y^0 and y^1 be two prescribed velocity fields. Then the fluid can be driven from y^0 at time $t = 0$ to a velocity field at time $t = T$ which is arbitrarily close to y^1. Here, "arbitrarily close" means that, for instance, the distance of the velocity field to y^1 in $W^{-1,\infty}(\Omega)$ and the distance of the vorticity to $\nabla \times y_1$ in all spaces $L^\infty(K)$ ($K \subset \Omega$ is an arbitrary compact set) can be made arbitrarily small. This is proved for any $T > 0$, for any regular Ω, y^0, and y^1, and for all nonempty open sets $\mathcal{O} \subset \Omega$.

The proof is as follows:

1. For each $\alpha > 0$, it is possible to find v_α and a couple (y_α, p_α) such that y_α is a gradient outside $\mathcal{O} \times (0,T)$,

$$\begin{cases} y_{\alpha,t} + (y_\alpha \cdot \nabla) y_\alpha - \Delta y_\alpha + \nabla p_\alpha = v_\alpha 1_{\mathcal{O}}, \quad \nabla \cdot y_\alpha = 0 & \text{in} \quad Q, \\ y_\alpha \cdot n = 0, \quad \nabla \times y_\alpha = 0 & \text{on} \quad \Sigma, \\ y_\alpha(0) = y_\alpha(T) = 0 & \text{in} \quad \Omega \end{cases} \tag{3.2}$$

and, furthermore, the linearized Euler system at y_α, that is to say,

$$\begin{cases} z_t + (y_\alpha \cdot \nabla) z + (z \cdot \nabla) y_\alpha + \nabla \pi = w 1_{\mathcal{O}}, \quad \nabla \cdot z = 0 & \text{in} \quad Q, \\ z \cdot n = 0, \quad \nabla \times z = 0 & \text{on} \quad \Sigma, \end{cases} \tag{3.3}$$

is α-controllable in the following sense: For any given z_0 and z_d of class C^∞, there exists a control w such that (3.3) possesses at least one solution bounded in $C^3(\overline{Q})$ independently of α and satisfying

$$z(0) = z_0 \ \text{ in } \ \Omega$$

and

$$z(T) = z_d \ \text{ in } \ \{\, x \in \Omega;\ \text{dist}\,(x, \partial\Omega) \geq \alpha \,\}$$

(notice we are saying that, for every α, this property is satisfied *for any* z_0 and z_d).

2. Let us set $v(x,t) = 0$ in (3.1) for $t \in [0, (1-\delta)T]$ (δ will be determined below). This defines y and p without ambiguity in $\Omega \times [0, (1-\delta)T]$ and, in particular, we can speak of $y((1-\delta)T)$. In $[(1-\delta)T, T]$, we do the following:

- First, v_α, y_α, and p_α are rescaled. In view of the α-controllability of (3.3), we introduce a first control function $\widetilde{v}$:

$$\widetilde{v}(x,t) \equiv \frac{1}{\delta} v_\alpha\left(x, \frac{1}{\delta}(t - (1-\delta)T)\right) + w\left(x, \frac{1}{\delta}(t - (1-\delta)T)\right).$$

  The associate state is

$$(\widetilde{y}, \widetilde{p}) \equiv \frac{1}{\delta}(y_\alpha, p_\alpha)\left(x, \frac{1}{\delta}(t - (1-\delta)T)\right) + (z, \pi)\left(x, \frac{1}{\delta}(t - (1-\delta)T)\right).$$

  Here, w and (z, π) are perturbations depending on α (which will be also fixed below), the initial state $y((1-\delta)T)$, and the desired state y_d. In order to drive (3.1) to a final state close to y_d, it is natural (at least formally) to look for a control close to $\widetilde{v}$ for $t \in [(1-\delta)T, T]$.
- We introduce a second control function by modifying $\widetilde{v}$ just as needed. Thus, we solve the following problem:

$$\begin{cases} y_t + (y \cdot \nabla) y - \Delta y + \nabla p = v 1_{\mathcal{O}} + (\nabla \times \overline{y})(y - \widetilde{y})^\perp, \quad \nabla \cdot y = 0 & \text{in} \quad Q, \\ y \cdot n = 0, \quad \nabla \times y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega. \end{cases}$$

Here, we have introduced the function

$$\overline{y}(x,t) = \frac{1}{\delta} y_\alpha \left(x, \frac{1}{\delta}(t-(1-\delta)T) \right), \quad (a_1,a_2)^\perp = (-a_2,a_1).$$

Since $\nabla \times \overline{y}$ vanishes outside $\mathcal{O} \times (0,T)$, it is clear that (y,p) solves (3.1) with

$$v = \widetilde{v} + (\nabla \times \overline{y})(y - \widetilde{y})^\perp.$$

Now, the task is reduced to show that for every $\varepsilon > 0$, there exist positive α and δ such that

$$\|y(T) - y_d\|_{W^{-1,\infty}} \leq \varepsilon. \tag{3.4}$$

This can be achieved in the following way. Let us set $R = y - y^0$, $\omega = \nabla \times R$. Then

$$\partial_t \omega + (R + y^0) \cdot \nabla \omega - \Delta \omega = -(R+z) \cdot \nabla(\nabla \times z) + \Delta(\nabla \times z)$$

in $\Omega \times ((1-\delta)T, T)$. Furthermore, $R(x,(1-\delta)T) \equiv 0$ and z and all its derivatives of order ≤ 3 are uniformly bounded. This leads, at first time, to a pointwise estimate of ω and then to an estimate of $R(\cdot,T)$ in $W^{-1,\infty}(\Omega)$ when α and δ are sufficiently small. Consequently, for any given $\varepsilon > 0$, there exist $\alpha^0 > 0$ and $\eta : (0,\alpha^0) \mapsto \mathbf{R}_+$ such that whenever $0 < \alpha < \alpha^0$ and $0 < \delta < \eta(\alpha)$, one has (3.4) (for further details, see [10]).

This method has several limitations. Thus, the boundary conditions have to be of the Navier-slip type. In practice, this is equivalent to prescribing the values on the boundary of the stream function and the vorticity. At present, it is unknown how the method has to be modified in order to keep its validity in the context of Dirichlet (no-slip) boundary conditions. The case in which Ω is a manifold without boundary is considered in [11].

An extension of this method to Boussinesq systems in spatial dimensions $N = 2$ and $N = 3$ has been given in [31]. For the analysis of the similar three-dimensional situation for the Euler equations, see [33].

In the remainder of this section, we will be concerned with the null controllability of Stokes-like and Navier–Stokes systems completed with Dirichlet boundary conditions. Let us recall the definition of the following spaces, which are usual in the analysis of Stokes systems:

$$\mathcal{V} = \{ z \in C_0^\infty(\Omega)^N : \nabla \cdot z = 0 \text{ in } \Omega \}$$

and

$$H \text{ (resp., } V) = \text{the adherence of } \mathcal{V} \text{ in } L^2(\Omega)^N \text{ (resp., } H_0^1(\Omega)^N).$$

We have

$$H = \{ z \in L^2(\Omega)^N : \nabla \cdot z = 0 \text{ in } \Omega,\ z \cdot n = 0 \text{ on } \partial\Omega \} \tag{3.5}$$

and

$$V = \{ z \in H_0^1(\Omega)^N : \nabla \cdot z = 0 \text{ in } \Omega \}. \tag{3.6}$$

For instance, see [55].

3.1. Null controllability of Stokes systems. We are first interested in proving the null controllability of some systems of the Stokes kind with distributed control. In other words, we would like to be able to find v, y, and p such that

$$\begin{cases} y_t - \Delta y + \nabla \cdot (y \otimes \overline{y}) + \nabla \cdot (\overline{y} \otimes y) + \nabla p = v 1_{\mathcal{O}} & \text{in} \quad Q, \\ \nabla \cdot y = 0 & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega \end{cases} \tag{3.7}$$

and

$$y(T) = 0 \quad \text{in } \Omega \tag{3.8}$$

for any $y^0 \in H$. Here, we will assume that the following regularity properties hold:

$$\overline{y} \in L^\infty(Q)^N, \quad \overline{y}_t \in L^2(0,T; L^\sigma(\Omega))^N, \tag{3.9}$$

with $\sigma > 1$ if $N = 2$ and $\sigma > 6/5$ if $N = 3$.

Let us remark that (3.7) can be viewed as the linearized Navier–Stokes system, so we are actually trying to prove the exact controllability to the trajectories of those equations. In fact, this will serve to deduce the local exact controllability to the trajectories of the Navier–Stokes equations.

A previous result of this kind was proved in [37] under stronger hypotheses on the trajectory $\overline{y}$. In what follows, we will sketch the arguments presented in [24], which are also based on those in [37].

Concerning the null controllability of (3.7), we will first recall a Carleman inequality for the backwards system

$$\begin{cases} -\varphi_t - \Delta \varphi - (D\varphi)\,\overline{y} + \nabla \pi = 0 & \text{in} \quad Q, \\ \nabla \cdot \varphi = 0 & \text{in} \quad Q, \\ \varphi = 0 & \text{on} \quad \Sigma, \\ \varphi(T) = \varphi^0 & \text{in} \quad \Omega, \end{cases} \tag{3.10}$$

where $D\varphi = \nabla \varphi + \nabla \varphi^t$ and, consequently, the ith component of $(D\varphi)\,\overline{y}$ is

$$((D\varphi)\,\overline{y})_i = \sum_{j=1}^{N} (\partial_j \varphi^i + \partial_i \varphi^j)\,\overline{y}_j. \tag{3.11}$$

To this end, let us introduce the following weight functions:

$$\begin{aligned} &\beta(x,t) = \frac{e^{5/4\lambda k \|\eta\|_\infty} - e^{\lambda(k\|\eta\|_\infty + \eta(x))}}{t^4 (T-t)^4}, \quad \gamma(x,t) = \frac{e^{\lambda(k\|\eta\|_\infty + \eta(x))}}{t^4(T-t)^4}, \\ &\widehat{\beta}(t) = \min_{x \in \overline{\Omega}} \beta(x,t), \quad \beta^*(t) = \max_{x \in \overline{\Omega}} \beta(x,t), \\ &\widehat{\gamma}(t) = \max_{x \in \overline{\Omega}} \gamma(x,t), \quad \gamma^*(t) = \min_{x \in \overline{\Omega}} \gamma(x,t). \end{aligned} \tag{3.12}$$

Here, $k > 4$ is a fixed number and η is the function given by Lemma 1.2 associated to $\omega = \mathcal{O}_5 \subset\subset \mathcal{O}$. Then we have the following.

LEMMA 3.1. *Let us assume that* (3.9) *is satisfied. Then for every* $\varphi^0 \in H$ *there exist three positive constants* $\overline{s}$, $\overline{\lambda}$, *and* C *depending only on* Ω *and* $\mathcal{O}$, *such that*

$$\begin{aligned} &s^{-1}\iint_Q e^{-2s\beta}\,\gamma^{-1}\,|\varphi_t|^2\,dx\,dt + s^{-1}\iint_Q e^{-2s\beta}\,\gamma^{-1}\,|\Delta\varphi|^2\,dx\,dt \\ &+ s\lambda^2\iint_Q e^{-2s\beta}\,\gamma\,|\nabla\varphi|^2\,dx\,dt + s^3\lambda^4\iint_Q e^{-2s\beta}\,\gamma^3\,|\varphi|^2\,dx\,dt \\ &\le C\,s^{16}\lambda^{40}(1+T^2)\iint_{\mathcal{O}\times(0,T)} e^{-8s\widehat{\beta}+6s\beta^*}\,\widehat{\gamma}^{16}\,|\varphi|^2\,dx\,dt \end{aligned} \tag{3.13}$$

for all $\lambda \ge \overline{\lambda}(1 + \|\overline{y}\|_\infty + \|\overline{y}_t\|^2_{L^2(L^\sigma)} + e^{\overline{\lambda}T\|\overline{y}\|^2_\infty})$ *and* $s \ge \overline{s}(T^4 + T^8)$.

The proof of Lemma 3.1 is given in the next subsection. Notice that the hypotheses imposed above over $\overline{y}$ are not completely satisfactory. In fact, it is expected (and desirable) that an inequality like (3.13) hold for any $\overline{y} \in L^\infty(Q)^N$, but for the moment this is unknown.

Now, from this Carleman inequality, one can deduce the null controllability of (3.7) in a classical way.

THEOREM 3.2. *Let us assume that* (3.9) *holds. Then for any* $T > 0$, *the linear system* (3.7) *is null controllable with controls* $v \in L^2(\mathcal{O}\times(0,T))^N$ *at time* T.

After some additional work, one can also deduce some local controllability results for the Navier–Stokes equations. Specifically, we mean local exact controllability to the trajectories. Thus, let us introduce the Navier–Stokes system

$$\begin{cases} y_t - \Delta y + \nabla\cdot(y\otimes y) + \nabla p = v1_{\mathcal{O}} & \text{in} \quad Q, \\ \nabla\cdot y = 0 & \text{in} \quad Q, \\ y = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega. \end{cases} \tag{3.14}$$

We have the following.

THEOREM 3.3. *Let* $(\overline{y},\overline{p})$ *be a solution to the Navier–Stokes problem*

$$\begin{cases} \overline{y}_t - \Delta\overline{y} + \nabla\cdot(\overline{y}\otimes\overline{y}) + \nabla\overline{p} = 0 & \textit{in} \quad Q, \\ \nabla\cdot\overline{y} = 0 & \textit{in} \quad Q, \\ \overline{y} = 0 & \textit{on} \quad \Sigma, \\ \overline{y}(0) = \overline{y}^0 & \textit{in} \quad \Omega \end{cases} \tag{3.15}$$

verifying (3.9), *and let us introduce the space* E, *with*

$$E = H \quad \textit{if} \quad N = 2 \qquad \textit{and} \qquad E = H\cap L^4(\Omega)^3 \quad \textit{if} \quad N = 3. \tag{3.16}$$

Then there exists $\varepsilon > 0$ *such that for each* $y^0 \in E$ *satisfying* $\|\overline{y}^0 - y^0\|_E \le \varepsilon$, *there exist controls* $v \in L^2(\mathcal{O}\times(0,T))^N$ *and associated states* (y,p) *such that one has* (3.14) *and*

$$y(T) = \overline{y}(T) \quad \textit{in} \quad \Omega. \tag{3.17}$$

This result can be proved by arguing as follows. We deduce in a first step a null controllability result for (3.7) with suitable right-hand side f.

More precisely, let us set $Ly = y_t - \Delta y + \nabla \cdot (\overline{y} \otimes y + y \otimes \overline{y})$ and let us introduce the spaces E_N, with

$$\begin{aligned} E_2 = \{\, (y,v) : {} & e^{2s\widehat{b}-sb^*}\widehat{\gamma}^{-15/4}y,\ e^{4s\widehat{b}-3sb^*}\widehat{\gamma}^{-8}v1_\omega \in L^2(Q)^2, \\ & e^{sb^*/2}(\gamma^*)^{-1/4}y \in L^2(0,T;V) \cap L^\infty(0,T;H), \\ & \exists p : e^{sb^*}(\gamma^*)^{-1/2}(Ly + \nabla p - v1_\omega) \in L^2(0,T;H^{-1}(\Omega)^2)\,\} \end{aligned}$$

and

$$\begin{aligned} E_3 = \{\, (y,v) : {} & e^{2s\widehat{b}-sb^*}\widehat{\gamma}^{-15/4}y,\ e^{4s\widehat{b}-3sb^*}\widehat{\gamma}^{-8}v1_\omega \in L^2(Q)^3, \\ & e^{sb^*/2}(\gamma^*)^{-1/4}y \in L^2(0,T;V) \cap L^\infty(0,T;H), \\ & e^{sb^*/2}(\gamma^*)^{-1/4}y \in L^4(0,T;L^{12}(\Omega)^3), \\ & \exists p : e^{sb^*}(\gamma^*)^{-1/2}(Ly + \nabla p - v1_\omega) \in L^2(0,T;W^{-1,6}(\Omega)^3)\,\}, \end{aligned}$$

where the new weight functions b, b^*, etc., are given by

$$b(x,t) = \frac{e^{5/4\lambda m\|\eta^0\|_\infty} - e^{\lambda(m\|\eta^0\|_\infty + \eta^0(x))}}{\ell(t)^4},$$

$$\widehat{b}(t) = \min_{x\in\overline{\Omega}} b(x,t), \quad b^*(t) = \max_{x\in\overline{\Omega}} b(x,t),$$

$$\gamma(x,t) = \frac{e^{\lambda(m\|\eta^0\|_\infty + \eta^0(x))}}{\ell(t)^4},$$

$$\widehat{\gamma}(t) = \max_{x\in\overline{\Omega}} \gamma(x,t), \quad \gamma^*(t) = \min_{x\in\overline{\Omega}} \gamma(x,t).$$

Here, we have introduced

$$\ell(t) = \begin{cases} T^2/4 & \text{for} \quad 0 < t < T/2, \\ t(T-t) & \text{for} \quad T/2 < t < T. \end{cases}$$

We then have the following.

PROPOSITION 3.4. *Assume that $\overline{y}$ satisfies* (3.9) *and*

- $y^0 \in H$, $e^{sb^*}(\gamma^*)^{-1/2}f \in L^2(0,T;H^{-1}(\Omega)^2)$ *if* $N = 2$;
- $y^0 \in L^4(\Omega)^3 \cap H$, $e^{sb^*}(\gamma^*)^{-1/2}f \in L^2(0,T;W^{-1,6}(\Omega)^3)$ *if* $N = 3$.

Then there exists a control $v \in L^2(\omega \times (0,T))^N$ such that if y is the associated solution to (3.7)*, we have $(y,v) \in E_N$.*

Notice that this is actually a null controllability result for (3.7). Indeed, if $(y,v) \in E_N$, we have in particular that $y(T) = 0$ in Ω.

The rest of the proof of Theorem 3.3 relies on an appropriate *inverse mapping theorem*. More precisely, we use the following result from [1].

PROPOSITION 3.5. *Let X, F be two Banach spaces and let $\mathcal{A} : X \mapsto F$ satisfy $\mathcal{A} \in C^1(X;F)$. Assume that $e_0 \in X$, $\mathcal{A}(e_0) = h_0$, and $\mathcal{A}'(e_0) : X \mapsto F$ is an epimorphism. Then there exists $\delta > 0$ such that for every $h \in F$ satisfying $\|h - h_0\|_F < \delta$, there exists a solution of the equation*

$$\mathcal{A}(e) = h, \quad e \in X.$$

Let us consider the mapping $\mathcal{A} : X \mapsto F$, given by

$$\mathcal{A}(y,v) = (Ly + (y\cdot\nabla)y + \nabla p - v1_\omega, y(\cdot,0)) \quad \forall (y,v) \in X,$$

where $X = E_N$ and

$$F = \begin{cases} L^2(e^{sb^*}(\gamma^*)^{-1/2}; H^{-1}(\Omega)^2) \times H & \text{if } N = 2, \\ L^2(e^{sb^*}(\gamma^*)^{-1/2}; W^{-1,6}(\Omega)^3) \times (L^4(\Omega)^3 \cap H) & \text{if } N = 3. \end{cases}$$

From the definition of E_N, one can easily check that $\mathcal{A}$ is well defined and satisfies $\mathcal{A} \in C^1(X; F)$. Furthermore, the identity

$$\mathrm{Im}(\mathcal{A}'(0,0)) = F$$

is equivalent to the result stated in Proposition 3.4. Therefore, we can apply Proposition 3.5 to $\mathcal{A}$ with $e_0 = (0,0)$ and $h_0 = (0,0)$. This ends the proof of Theorem 3.3.

For more details concerning the local exact controllability of (3.14), see [24]; see also [25].

3.2. Proof of the Carleman inequality. In this subsection, we sketch the proof of Lemma 3.1. This is contained in a more general case covered by the results in [24]. Throughout the proof, C will denote a generic positive constant that may depend on Ω and $\mathcal{O}$ and can change its value from one line to the next.

For easier comprehension, we will divide the proof in several steps.

3.2.1. Carleman estimate for the heat equation and first estimate of the pressure. Let us apply the Carleman inequality (1.18) to each component φ^i with right-hand side

$$-\partial_i \pi + ((D\varphi)\,\overline{y})_i \quad (1 \le i \le N). \tag{3.18}$$

Introducing the notation

$$\begin{aligned} I(s,\lambda;\varphi) = {} & s^{-1} \iint_Q e^{-2sb}\,\gamma^{-1}\,(|\varphi_t|^2 + |\Delta\varphi|^2)\,dx\,dt \\ & + s\lambda^2 \iint_Q e^{-2s\beta}\,\gamma\,|\nabla\varphi|^2\,dx\,dt + s^3\lambda^4 \iint_Q e^{-2s\beta}\,\gamma^3\,|\varphi|^2\,dx\,dt, \end{aligned} \tag{3.19}$$

one has

$$\begin{aligned} I(s,\lambda;\varphi) \le C \Bigg(& s^3\lambda^4 \iint_{\mathcal{O}_5\times(0,T)} e^{-2s\beta}\,\gamma^3\,|\varphi|^2\,dx\,dt \\ & + \iint_Q e^{-2s\beta}\,(|\nabla\pi|^2 + |(D\varphi)\,\overline{y}|^2)\,dx\,dt \Bigg) \end{aligned} \tag{3.20}$$

for $\lambda \ge C$ and $s \ge C(T^7 + T^8)$.

Taking $\lambda \ge C\,\|\overline{y}\|_\infty$ and $s \ge C\,T^8$, one can eliminate the term $|(D\varphi)\,\overline{y}|^2$ on the right-hand side, whence we obtain

$$I(s,\lambda;\varphi) \le C \left(s^3\lambda^4 \iint_{\mathcal{O}_5\times(0,T)} e^{-2s\beta}\,\gamma^3\,|\varphi|^2\,dx\,dt + \iint_Q e^{-2s\beta}\,|\nabla\pi|^2\,dx\,dt \right) \tag{3.21}$$

for $\lambda \ge C(1 + \|\overline{y}\|_\infty)$ and $s \ge C(T^4 + T^8)$.

Let us now present an estimate of the last integral. To this end, we will apply the main result in [38] to the elliptic equation

$$-\Delta\pi(t) = -\nabla \cdot ((D\varphi)\,\overline{y})(t) \quad \text{in} \quad \Omega \tag{3.22}$$

for each $t \in (0,T)$. This provides

(3.23)
$$\int_\Omega e^{2\tau e^{\lambda\eta}} |\nabla\pi(t)|^2\,dx \le C\bigg(\int_{\mathcal{O}_5} e^{2\tau e^{\lambda\eta}} (|\nabla\pi(t)|^2 + \tau^2\lambda^2\, e^{2\lambda\eta}\, |\pi(t)|^2)dx + \tau \int_\Omega e^{2\tau e^{\lambda\eta}}\, e^{\lambda\eta}\, |(D\varphi)\,\overline{y}(t)|^2\,dx + \tau^{1/2}\, e^{2\tau}\, \|\pi(t)\|^2_{H^{1/2}(\partial\Omega)}\bigg)$$

for λ, $\tau \ge C$.

One can readily see that the local term of $|\nabla\pi|^2$ can be eliminated by arguing as in the proof of Lemma 1.3, so we obtain

(3.24)
$$\int_\Omega e^{2\tau e^{\lambda\eta}} |\nabla\pi(t)|^2\,dx \le C\bigg(\tau^2\lambda^2 \int_{\mathcal{O}_4} e^{2\tau e^{\lambda\eta}}\, e^{2\lambda\eta}\, |\pi(t)|^2 dx + \tau \int_\Omega e^{2\tau e^{\lambda\eta}}\, e^{\lambda\eta}\, |(D\varphi)\,\overline{y}(t)|^2\,dx + \tau^{1/2}\, e^{2\tau}\, \|\pi(t)\|^2_{H^{1/2}(\partial\Omega)}\bigg)$$

for an open subset $\mathcal{O}_4$ verifying $\mathcal{O}_5 \subset\subset \mathcal{O}_4 \subset\subset \mathcal{O}$.

Next, we put

$$\tau = \frac{s}{t^4(T-t)^4} e^{\lambda k\|\eta\|_\infty}, \tag{3.25}$$

we multiply (3.24) by

$$exp\left\{-2s\frac{e^{5/4\lambda k\|\eta\|_\infty}}{t^4(T-t)^4}\right\}, \tag{3.26}$$

and we integrate with respect to t in $(0,T)$. This yields

$$\begin{aligned}\iint_Q e^{-2s\beta}\, |\nabla\pi|^2\,dx\,dt \le C\bigg(& s^2\lambda^2 \iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta}\, \gamma^2\, |\pi|^2\,dx\,dt \\ & + s\iint_Q e^{-2s\beta}\, \gamma\, |(D\varphi)\,\overline{y}|^2\,dx\,dt \\ & + s^{1/2}\int_0^T e^{-2s\beta^*}\, (\gamma^*)^{1/2}\, \|\pi(t)\|^2_{H^{1/2}(\partial\Omega)}\,dt\bigg)\end{aligned} \tag{3.27}$$

for $\lambda \ge C$ and $s \ge C\,T^8$.

Let us now estimate the term containing the trace of π following the ideas of [37]. Thus, let us introduce the functions

$$\varphi^* = s^{1/4}\, e^{-s\beta^*}\, (\gamma^*)^{1/4}\, \varphi \quad \text{and} \quad \pi^* = s^{1/4}\, e^{-s\beta^*}\, (\gamma^*)^{1/4}\, \pi, \tag{3.28}$$

which satisfy

$$\begin{cases} -\varphi^*_t - \Delta\varphi^* + \nabla\pi^* = (D\varphi^*)\,\overline{y} - s^{1/4}\,(e^{-s\beta^*}\,(\gamma^*)^{1/4})_t\,\varphi & \text{in} \quad Q, \\ \nabla\cdot\varphi^* = 0 & \text{in} \quad Q, \\ \varphi^* = 0 & \text{on} \quad \Sigma, \\ \varphi^*(T) = 0 & \text{in} \quad \Omega. \end{cases} \tag{3.29}$$

Using well-known regularity properties of the Stokes system, we find

(3.30)
$$\begin{aligned}\|\pi^*\|^2_{L^2(0,T;H^{1/2}(\partial\Omega))} &\le C\left(\iint_Q (|D\varphi^*\,\overline{y}|^2 + s^{1/2}|(e^{-s\beta^*}\,(\gamma^*)^{1/4})_t|^2\,|\varphi|^2)\,dx\,dt\right)\\ &\le C\left(s^{1/2}\iint_Q e^{-2s\beta^*}\,(\gamma^*)^{1/2}\,|(D\varphi)\,\overline{y}|^2\,dx\,dt\right.\\ &\qquad\left. + s^{5/2}\,T^2\iint_Q e^{-2s\beta^*}\,(\gamma^*)^3\,|\varphi|^2\,dx\,dt\right)\end{aligned}$$

for $s \ge C\,T^8$. This, used in combination with (3.27), leads to the inequality

$$\begin{aligned}\iint_Q e^{-2s\beta}\,|\nabla\pi|^2\,dx\,dt \le C\Bigg(&s^2\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta}\,\gamma^2\,|\pi|^2\,dx\,dt\\ &+ s\iint_Q e^{-2s\beta}\,\gamma\,|(D\varphi)\,\overline{y}|^2\,dx\,dt\\ &+ s^{5/2}\,T^2\iint_Q e^{-2s\beta}\,\gamma^3\,|\varphi|^2\,dx\,dt\Bigg),\end{aligned} \tag{3.31}$$

which must hold for $\lambda \ge C$ and $s \ge C\,T^8$.

We finally insert this inequality into (3.21) and absorb the last two terms on the right-hand side:

$$I(s,\lambda;\varphi) \le C\,s^2\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta}(s\lambda^2\,\gamma^3\,|\varphi|^2 + \gamma^2\,|\pi|^2)\,dx\,dt \tag{3.32}$$

for $\lambda \ge C(1 + \|\overline{y}\|_\infty)$ and $s \ge C(T^4 + T^8)$.

3.2.2. Local estimate of the pressure. The goal of this paragraph is to estimate the local integral of π appearing in (3.32). This is the most complicated part of the proof. Let us assume that we have chosen π such that

$$\int_{\mathcal{O}_4} \pi(t)\,dx = 0 \quad \forall t \in (0,T). \tag{3.33}$$

Then from the definition of the weight functions and the Poincaré–Wirtinger inequality, we have

$$\begin{aligned}s^2\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta}\,\gamma^2\,|\pi|^2\,dx\,dt &\le s^2\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\widehat{\beta}}\,\widehat{\gamma}^2\,|\pi|^2\,dx\,dt\\ &\le C\,s^2\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\widehat{\beta}}\,\widehat{\gamma}^2\,|\nabla\pi|^2\,dx\,dt.\end{aligned} \tag{3.34}$$

Recall that $\widehat{\beta} = \widehat{\beta}(t)$ and $\widehat{\gamma} = \widehat{\gamma}(t)$ are given as follows:

$$\widehat{\beta}(t) = \min_{\overline{\Omega}} \beta(x,t), \quad \widehat{\gamma}(t) = \max_{\overline{\Omega}} \gamma(x,t). \tag{3.35}$$

From the differential equation in (3.10), we deduce that in order to estimate the last integral in (3.34), it suffices to bound locally the Laplacian and the time derivative of the velocity field φ.

A local estimate of $\Delta\varphi$. Let us first introduce two open sets $\mathcal{O}_2$ and $\mathcal{O}_3$, with

$$\mathcal{O}_4 \subset\subset \mathcal{O}_3 \subset\subset \mathcal{O}_2 \subset\subset \mathcal{O}, \tag{3.36}$$

and a function $\rho \in \mathcal{D}(\mathcal{O}_2)$, with $\rho \equiv 1$ in $\mathcal{O}_3$.

Next, we put

$$u(x,t) = \widehat{p}(t)\,\rho(x)\,\Delta\varphi(x,T-t) \quad \forall (x,t) \in \mathbf{R}^N \times (0,T), \tag{3.37}$$

where we have denoted by $\widehat{p}$ the function $\widehat{p}(t) = s\lambda\, e^{-s\widehat{\beta}}\,\widehat{\gamma}$.

From (3.10), we obtain

$$(\Delta\varphi(T-t))_t - \Delta(\Delta\varphi(T-t)) = f \quad \text{in } Q, \tag{3.38}$$

with

$$f(t) = \Delta((D\varphi)\,\overline{y})(T-t) - \nabla(\nabla\cdot((D\varphi)\,\overline{y})(T-t)). \tag{3.39}$$

This provides the following heat problem for the function u:

$$\begin{cases} u_t - \Delta u = F & \text{in} \quad \mathbf{R}^N \times (0,T), \\ u(0) = 0 & \text{in} \quad \mathbf{R}^N, \end{cases} \tag{3.40}$$

with

$$F(t) = \widehat{p}\,\rho\, f + \widehat{p}'\,\rho\,\Delta\varphi(T-t) - 2\widehat{p}\,\nabla\rho\cdot\nabla\Delta\varphi(T-t) - \widehat{p}\,\Delta\rho\,\Delta\varphi(T-t) \tag{3.41}$$

for all $t \in (0,T)$. Then we rewrite F in the form $F = F_1 + F_2$, with

$$\begin{aligned} F_1(t) = {} & \widehat{p}\,\Delta(\rho((D\varphi)\,\overline{y})(T-t)) - \widehat{p}\,\nabla(\nabla\cdot(\rho((D\varphi)\,\overline{y})(T-t))) \\ & + \widehat{p}'\,\Delta(\rho\,\varphi(T-t)) \end{aligned} \tag{3.42}$$

and

$$\begin{aligned} F_2(t) = {} & -2\widehat{p}\,\nabla\rho\cdot\nabla((D\varphi)\overline{y})(T-t) - \widehat{p}\,\Delta\rho\,((D\varphi)\,\overline{y})(T-t) \\ & + \widehat{p}\,\nabla(\nabla\rho\cdot((D\varphi)\,\overline{y})(T-t)) + \widehat{p}\,\nabla\rho\cdot(\nabla\cdot((D\varphi)\,\overline{y})(T-t)) \\ & - 2\widehat{p}'\,\nabla\rho\cdot\nabla\varphi(T-t) - \widehat{p}'\,\Delta\rho\,\varphi(T-t) - 2\widehat{p}\,\nabla\rho\cdot\nabla\Delta\varphi(T-t) \\ & - \widehat{p}\,\Delta\rho\,\Delta\varphi(T-t), \end{aligned} \tag{3.43}$$

and we define u^i $(i = 1,2)$ as the solution to

$$\begin{cases} u^i_t - \Delta u^i = F_i & \text{in} \quad \mathbf{R}^N \times (0,T), \\ u^i(0) = 0 & \text{in} \quad \mathbf{R}^N. \end{cases} \tag{3.44}$$

An estimate of u^1. Let us remark that $F_1 \in L^2(0,T;H^{-2}(\mathbf{R}^N)^N)$, while we would like to estimate the L^2-norm of u^1. Accordingly, we have to define u^1 like a solution by transposition. More precisely, u^1 is the unique function in $L^2(\mathbf{R}^N \times (0,T))^N$ verifying

$$\begin{aligned} \iint_{\mathbf{R}^N\times(0,T)} u^1 \cdot k\,dx\,dt = {} & \iint_{\mathbf{R}^N\times(0,T)} \widehat{p}\,\rho((D\varphi)\,\overline{y}(T-t))\cdot\Delta z\,dx\,dt \\ & - \iint_{\mathbf{R}^N\times(0,T)} \widehat{p}\,\rho((D\varphi)\,\overline{y}(T-t))\cdot\nabla(\nabla\cdot z)\,dx\,dt \\ & + \iint_{\mathbf{R}^N\times(0,T)} \widehat{p}'\,\rho\,\varphi(T-t)\cdot\Delta z\,dx\,dt \end{aligned} \tag{3.45}$$

for any $k \in L^2(\mathbf{R}^N \times (0,T))^N$, where z is the solution of

$$\begin{cases} -z_t - \Delta z = k & \text{in} \quad \mathbf{R}^N \times (0,T), \\ z(T) = 0 & \text{in} \quad \mathbf{R}^N. \end{cases} \tag{3.46}$$

Then we get the following directly from (3.45):

$$\begin{aligned} &\iint_{\mathbf{R}^N \times (0,T)} |u^1|^2 \, dx \, dt \\ &\quad \le C \left(\iint_{\mathbf{R}^N \times (0,T)} |\widehat{p} \, \rho \, (D\varphi) \, \overline{y}|^2 \, dx \, dt + \iint_{\mathbf{R}^N \times (0,T)} |\widehat{p}' \, \rho \, \varphi|^2 \, dx \, dt \right) \\ &\quad \le C \left(\iint_{\mathcal{O}_2 \times (0,T)} |\widehat{p}|^2 \, |(D\varphi) \, \overline{y}|^2 \, dx \, dt + \iint_{\mathcal{O}_2 \times (0,T)} |\widehat{p}'|^2 \, |\varphi|^2 \, dx \, dt \right). \end{aligned} \tag{3.47}$$

An estimate of u^2. This time we must take into account the fact that the function F_2 belongs to the space $L^2(0,T;H^{-1}(\mathbf{R}^N)^N)$. But it is also essential to remember that supp $F_2 \subset \mathcal{O}_2 \setminus \overline{\mathcal{O}}_3$, since derivatives of ρ appear everywhere. This allows us to write, for $x \in \mathcal{O}_4$,

$$u^2(x,t) = \int_0^t \int_{\mathcal{O}_2 \setminus \overline{\mathcal{O}}_3} G(x-y, t-s) \, F_2(y,s) \, dy \, ds, \tag{3.48}$$

where G is the fundamental solution of the heat equation, i.e.,

$$G(x,t) = \frac{1}{(4\pi t)^{N/2}} e^{-\frac{|x|^2}{2t}} \quad \forall (x,t) \in \mathbf{R}^N \times \mathbf{R}_+. \tag{3.49}$$

Now, we integrate by parts with respect to y in (3.48), taking all the space derivatives out of $(D\varphi)\,\overline{y}$ and φ. This gives the following equality:

$$u^2(x,t) = \int_0^t \int_{\mathcal{O}_2 \setminus \overline{\mathcal{O}}_3} \sum_{\alpha \in I, \beta \in J} D_y^\alpha G(x-y, t-s) \, D_y^\beta \rho(y) \, w_{\alpha,\beta}(y,s) \, dy \, ds, \tag{3.50}$$

where

$$w_{\alpha,\beta}(y,s) = (A_{\alpha,\beta} \, \widehat{p}(s) + B_{\alpha,\beta} \, \widehat{p}'(s)) \varphi(y,s) + C_{\alpha,\beta} \, \widehat{p}(s) \, ((D\varphi) \, \overline{y})(y,s) \tag{3.51}$$

for appropriate constants $A_{\alpha,\beta}$, $B_{\alpha,\beta}$, and $C_{\alpha,\beta}$,

$$I \subset \{\gamma \in \mathrm{N}^N : |\gamma| \le 3\} \quad \text{and} \quad J \subset \{\gamma \in \mathrm{N}^N : 1 \le |\gamma| \le 4\}. \tag{3.52}$$

Next, we denote by d the quantity $d = \operatorname{dist}(\partial\mathcal{O}_3, \partial\mathcal{O}_4)$, we fix $\delta \in (0,d)$, and we observe that

$$|D_y^\alpha G(x-y, t-s)| \le C \, \exp\left(-\frac{\delta^2}{2(t-s)} \right) \quad \forall x \in \mathcal{O}_4, \, \forall y \in \mathcal{O}_2 \setminus \overline{\mathcal{O}}_3. \tag{3.53}$$

Therefore, we deduce from (3.50) that

$$|u^2(x,t)| \le C \int_0^t \int_{\mathcal{O}_2 \setminus \overline{\mathcal{O}}_3} e^{-\frac{\delta^2}{2(t-s)}} \, |w(y,s)| \, dy \, ds \quad \forall (x,t) \in \mathcal{O}_4 \times (0,T), \tag{3.54}$$

where

$$w(y,s) = (\widehat{p}(s) + \widehat{p}\,'(s))\varphi(y,s) + \widehat{p}(s)\,((D\varphi)\,\overline{y})(y,s). \tag{3.55}$$

Integrating with respect to x and t in $\mathcal{O}_4 \times (0,T)$, we find that

$$\begin{aligned}\iint_{\mathcal{O}_4\times(0,T)} |u^2|^2\,dx\,dt &\le C\,T\int_0^T\!\!\int_0^t e^{-\delta^2/(t-s)}\,\|w(y,s)\|^2_{L^2(\mathcal{O}_2)}\,ds\,dt \\ &= C\,T\int_0^T (f_1 * f_2)(t)\,dt,\end{aligned} \tag{3.56}$$

where

$$f_1(t) = e^{-\delta^2/t} \quad \text{and} \quad f_2(t) = \|w(t)\|^2_{L^2(\mathcal{O}_2)}. \tag{3.57}$$

Notice that $f_1, f_2 \in L^1(0,T)$. As a consequence of Young's inequality, we have

$$\begin{aligned}\iint_{\mathcal{O}_4\times(0,T)} |u^2|^2\,dx\,dt &\le C\,T\iint_{\mathcal{O}_2\times(0,T)} |w|^2\,dx\,dt \\ &\le C\,T\bigg(\iint_{\mathcal{O}_2\times(0,T)} (|\widehat{p}|^2 + |\widehat{p}\,'|^2)|\varphi|^2\,dx\,dt \\ &\qquad + \iint_{\mathcal{O}_2\times(0,T)} |\widehat{p}|^2\,|(D\varphi)\,\overline{y}|^2\,dx\,dt\bigg).\end{aligned} \tag{3.58}$$

Combining this with (3.47), we obtain the local estimate of $\Delta\varphi$:

$$\begin{aligned}s^2\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\widehat{\beta}}\,\widehat{\gamma}^2\,|\Delta\varphi|^2\,dx\,dt &\le C(1+T)\bigg(\iint_{\mathcal{O}_2\times(0,T)} (|\widehat{p}|^2 + |\widehat{p}\,'|^2)|\varphi|^2\,dx\,dt \\ &\qquad + \iint_{\mathcal{O}_2\times(0,T)} |\widehat{p}|^2\,|(D\varphi)\,\overline{y}|^2\,dx\,dt\bigg).\end{aligned} \tag{3.59}$$

A local estimate of $\boldsymbol{\varphi_t}$. Let us first introduce the weight function p, with

$$p(t) = s^{15/4}\,e^{-2s\widehat{\beta}+s\beta^*}\,\widehat{\gamma}^{15/4}. \tag{3.60}$$

By setting $(\psi,\theta) = (p\,\varphi, p\,\pi)$, we have

$$\begin{cases} -\psi_t - \Delta\psi - (D\psi)\,\overline{y} + \nabla\theta = -p'\,\varphi & \text{in } Q, \\ \nabla\cdot\psi = 0 & \text{in } Q, \\ \psi = 0 & \text{on } \Sigma, \\ \psi(T) = 0 & \text{in } \Omega. \end{cases} \tag{3.61}$$

Recall that the term we have to estimate is

$$\begin{aligned}\iint_{\mathcal{O}_4\times(0,T)} |\widehat{p}|^2\,|\varphi_t|^2\,dx\,dt &= \iint_{\mathcal{O}_4\times(0,T)} p^{-2}\,|\widehat{p}|^2\,|p\,\varphi_t|^2\,dx\,dt \\ &= s^{-11/2}\lambda^2\iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-11/2}\,|\psi_t - p'\,\varphi|^2\,dx\,dt.\end{aligned} \tag{3.62}$$

Hence, we will try to estimate the (weighted) L^2-norm of the time derivative of ψ. To this end, we integrate by parts with respect to t:

$$\begin{aligned} & s^{-11/2}\lambda^2 \iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-11/2}\,|\psi_t|^2\,dx\,dt \\ &= \frac{1}{2}s^{-11/2}\lambda^2 \iint_{\mathcal{O}_4\times(0,T)} (e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-11/2})_{tt}\,|\psi|^2\,dx\,dt \\ &\quad -s^{-11/2}\lambda^2 \iint_{\mathcal{O}_4\times(0,T)} e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-11/2}\,\psi_{tt}\,\psi\,dx\,dt. \end{aligned} \tag{3.63}$$

One can easily check that

$$s^{-11/2}\lambda^2(e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-11/2})_{tt} \le C\,s^{-7/2}\lambda^2\,T^2\,e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-3} \tag{3.64}$$

for $s \ge C\,T^8$.

Therefore, we must concentrate on estimating the last term in (3.63). Now, we first denote by p^* the function

$$p^* = s^{-11/2}\lambda^{-4}\,e^{-2s\beta^*+2s\widehat{\beta}}\,\widehat{\gamma}^{-11/2} \tag{3.65}$$

and then use Hölder's inequality:

$$\begin{aligned} \lambda^6 \iint_{\mathcal{O}_4\times(0,T)} p^*\,\psi_{tt}\,\psi\,dx\,dt &\le \lambda^6\,\|p^*\,\psi_{tt}\|_{L^2(0,T;L^q(\mathcal{O}_4))}\,\|\psi\|_{L^2(0,T;L^{q'}(\mathcal{O}_4))} \\ &\le 2\|p^*\,\psi_{tt}\|^2_{L^2(0,T;L^q(\mathcal{O}_4))} \\ &\quad + 2\lambda^{12}\|\psi\|^2_{L^2(0,T;L^{q'}(\mathcal{O}_4))}, \end{aligned} \tag{3.66}$$

where q is chosen lower than σ but greater than 1 in dimension $N = 2$ and greater than $6/5$ in dimension $N = 3$, and q' is the conjugate exponent of q (recall that $\sigma > 1$ if $N = 2$ and $\sigma > 6/5$ if $N = 3$).

Let us first deal with the second term on the right-hand side of (3.66). Let $\rho \in C^2(\mathcal{O}_3)$ be a cut-off function satisfying

$$\text{supp}\ \rho \subset \mathcal{O}_3 \quad \text{and} \quad \rho \equiv 1 \text{ in } \mathcal{O}_4. \tag{3.67}$$

Then

$$\begin{aligned} \|\psi\|^2_{L^2(L^{q'}(\mathcal{O}_4))} &\le \|\Delta(\rho\,\psi)\|^2_{L^2(L^2(\mathcal{O}_3))} \\ &= \|\Delta\rho\,\psi + 2\nabla\rho\cdot\nabla\psi + \rho\,\Delta\psi\|^2_{L^2(L^2(\mathcal{O}_3))}. \end{aligned} \tag{3.68}$$

Following the same steps as in the previous section for obtaining the local estimate of $\Delta\varphi$, after some long computations, one can readily get an estimate of the last term on the right-hand side of (3.68). Consequently, we have

$$\|\psi\|^2_{L^2(0,T;L^{q'}(\mathcal{O}_4))} \le C(1+\|\overline{y}\|^2_\infty)(1+T)\iint_{\mathcal{O}_1\times(0,T)} (|\psi|^2 + |\nabla\psi|^2 + |p'\varphi|^2)\,dx\,dt, \tag{3.69}$$

where $\mathcal{O}_1$ is an open set verifying

$$\mathcal{O}_2 \subset\subset \mathcal{O}_1 \subset\subset \mathcal{O}. \tag{3.70}$$

Finally, we treat the norm of $p^*\psi_{tt}$ in (3.66).

Observing the problem verified by $(p^*\,\psi_t, p^*\,\theta_t)$, one can obtain the next a priori estimate for ψ_{tt}:

$$\begin{aligned}\|p^*\,\psi_{tt}\|_{L^2(L^q)} \le (1+\|\overline{y}\|_\infty)\,e^{C(1+T\|\overline{y}\|_\infty^2)}\,\big(\|p^*\,p''\,\varphi\|_{L^2(L^2)}\\ +\|p^*\,p'\,\varphi_t\|_{L^2(L^2)}+\|(p^*)'\,\psi_t\|_{L^2(L^2)}+\|p^*\,(D\psi)\,\overline{y}_t\|_{L^2(L^q)}\big)\,.\end{aligned}\tag{3.71}$$

Consequently, from (3.62)–(3.66) and (3.69)–(3.71), we have

$$\begin{aligned}&\iint_{\mathcal{O}_4\times(0,T)}|\widehat{p}|^2\,|\varphi_t|^2\,dx\,dt\\ &\quad\le(1+\|\overline{y}\|_\infty^2)\,e^{C(1+T\|\overline{y}\|_\infty^2)}\,\Big(\lambda^{12}(1+T)(\|p\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))}\\ &\quad+\|p\,\nabla\varphi\|^2_{L^2(L^2(\mathcal{O}_1))}+\|p'\varphi\|^2_{L^2(L^2(\mathcal{O}_1))})+\|p^*\,p''\,\varphi\|^2_{L^2(L^2)}\\ &\quad+\|p^*\,p'\,\varphi_t\|^2_{L^2(L^2)}+\|(p^*)'\,\psi_t\|^2_{L^2(L^2)}+\|p^*\,(D\psi)\,\overline{y}_t\|^2_{L^2(L^q)}\Big)\,.\end{aligned}\tag{3.72}$$

An estimate of $p^*\,(D\psi)\,\overline{y}_t$. Let $\ell\in(1,+\infty)$ be given. In this section, we will prove that for $\overline{y}\in L^\infty(Q)^N$, we have $p^*\,\psi\in L^\infty(0,T;W^{1,\ell}(\Omega))^N$ for all $\ell<\infty$, and we will get an estimate of its norm. Once this is established, we will directly obtain an estimate of $p^*\,(D\psi)\,\overline{y}_t$ in $L^2(L^q)$ (taking ℓ large enough).

Thus, let us look at the system fulfilled by $(p^*\,\psi, p^*\,\theta)$:

$$\begin{cases}-(p^*\,\psi)_t-\Delta(p^*\,\psi)-D(p^*\,\psi)\,\overline{y}+\nabla(p^*\,\theta)=-(p^*)'\,\psi-p^*\,p'\,\varphi & \text{in } Q,\\ \nabla\cdot(p^*\,\psi)=0 & \text{in } Q,\\ p^*\,\psi=0 & \text{on } \Sigma,\\ (p^*\,\psi)(T)=0 & \text{in } \Omega.\end{cases}\tag{3.73}$$

We can first view ψ as a function in

$$L^2(0,T;H^2(\Omega)^N)\cap L^\infty(0,T;V).\tag{3.74}$$

Then, using interpolation inequalities, we have that $D(p^*\,\psi)\,\overline{y}\in L^{k_1}(L^{k_2})$, with $2<k_1<4$ and $3<k_2<6$ (which will be chosen below) verifying

$$\frac{4/3}{k_1}+\frac{2}{k_2}=1.\tag{3.75}$$

Using well-known regularity properties of the Stokes system (see [32]), we obtain from (3.73) that

$$p^*\,\psi\in L^{k_1}(0,T;W^{2,k_2}(\Omega)^N)\tag{3.76}$$

and

$$\|p^*\,\psi\|_{L^{k_1}(W^{2,k_2})}\le C(\|(p^*)'\,\psi+p^*\,p'\,\varphi\|_{L^{k_1}(L^{k_2})}+\|\overline{y}\|_\infty\|\nabla(p^*\,\psi)\|_{L^{k_1}(L^{k_2})}.\tag{3.77}$$

We can now perform a bootstrap argument again based on a priori estimates of the Stokes system but this time with a right-hand side in $L^{k_1}(L^\ell)$ with

$$\frac{2}{k_1}+\frac{6}{\ell}=1.\tag{3.78}$$

This determines the value of k_1 and, consequently, of k_2 as well. This yields $\nabla(p^*\,\theta) \in L^{k_1}(L^\ell)$ and

$$\|\nabla(p^*\,\theta)\|_{L^{k_1}(L^\ell)} \le C(\|(p^*)'\,\psi + p^*\,p'\,\varphi\|_{L^{k_1}(L^l)} + \|\overline{y}\|_\infty\,\|\nabla(p^*\,\psi)\|_{L^{k_1}(L^\ell)}). \tag{3.79}$$

Combining this last inequality and (3.77), we get

$$\begin{aligned}\|\nabla(p^*\,\theta)\|_{L^{k_1}(L^\ell)} \le C(1+\|\overline{y}\|_\infty^2)(&\|(p^*+(p^*)')\,\Delta\psi\|_{L^2(L^2)}\\ &+ \|((p^*+(p^*)')\,\psi)_t\|_{L^2(L^2)} + \|p^*\,p'\,\Delta\varphi\|_{L^2(L^2)}\\ &+ \|(p^*\,p'\,\varphi)_t\|_{L^2(L^2)}).\end{aligned} \tag{3.80}$$

This tells us that we can view (3.73) as a system of N heat equations with right-hand sides in $L^{k_1}(0,T;L^\ell(\Omega))$. Then, using the representation of the solution of heat systems in terms of the semigroup, we deduce that

$$\|p^*\,\psi(t)\|_{W^{1,\ell}(\Omega)} \le \int_0^t (t-s)^{-1/2}\|B(s)\|_{L^\ell(\Omega)}\,ds \quad \forall t \in (0,T), \tag{3.81}$$

where $B(s) = -(p^*)'\,\psi - p^*\,p'\,\varphi + D(p^*\,\psi)\,\overline{y} - \nabla(p^*\,\theta)$. See [40] for more details.

Since $k_1 > 2$, from Young's inequality we see that $\|p^*\,\psi\|_{W^{1,\ell}(\Omega)} \in L^\infty(0,T)$ and

$$\|p^*\,\psi\|_{L^\infty(W^{1,\ell})} \le (1-k_1'/2)^{-1/k_1'}\,T^{-1/2+1/k_1'}\,\|B\|_{L^{k_1}(L^\ell)}. \tag{3.82}$$

Hence, from (3.80), we obtain the estimate

$$\begin{aligned}\|p^*\,\psi\|_{L^\infty(W^{1,\ell})} \le C\,T^{-1/2+1/k_1'}(1+\|\overline{y}\|_\infty^2)(&\|(p^*+(p^*)')\,\Delta\psi\|_{L^2(L^2)}\\ &+ \|((p^*+(p^*)')\,\psi)_t\|_{L^2(L^2)} + \|p^*\,p'\,\Delta\varphi\|_{L^2(L^2)} + \|(p^*\,p'\,\varphi)_t\|_{L^2(L^2)})\end{aligned} \tag{3.83}$$

and, consequently, with a proper choice of ℓ, we have

$$\begin{aligned}\|p^*\,D\psi\,\overline{y}_t\|_{L^2(L^q)} \le \|p^*\,(D\psi)\|_{L^\infty(L^\ell)}\,\|\overline{y}_t\|_{L^2(L^\sigma)} &\le C\,\|\overline{y}_t\|_{L^2(L^\sigma)}\,T^{-1/2+1/k_1'}(1+\|\overline{y}\|_\infty^2)\\ &\Big(\|(p^*+(p^*)')\,\Delta\psi\|_{L^2(L^2)} + \|((p^*+(p^*)')\,\psi)_t\|_{L^2(L^2)}\\ &+ \|p^*\,p'\,\Delta\varphi\|_{L^2(L^2)} + \|(p^*\,p'\,\varphi)_t\|_{L^2(L^2)}\Big).\end{aligned} \tag{3.84}$$

Finally, we put (3.84) together with (3.72) to obtain the desired local estimate of φ_t:

$$\begin{aligned}&\iint_{\mathcal{O}_4\times(0,T)} |\widehat{p}|^2\,|\varphi_t|^2\,dx\,dt\\ &\le (1+\|\overline{y}\|_\infty^6)\,\|\overline{y}_t\|_{L^2(L^\sigma)}^2\,e^{C(1+T\|\overline{y}\|_\infty^2)}\Big(\lambda^{12}(1+T)(\|p\,\varphi\|_{L^2(L^2(\mathcal{O}_1))}^2\\ &\quad + \|p\,\nabla\varphi\|_{L^2(L^2(\mathcal{O}_1))}^2 + \|p'\,\varphi\|_{L^2(L^2(\mathcal{O}_1))}^2) + T^{-1+2/k_1'}(\|p^*\,p''\,\varphi\|_{L^2(L^2)}^2\\ &\quad + \|(p^*)'\,p'\,\varphi\|_{L^2(L^2)}^2 + \|p^*\,p'\,\varphi_t\|_{L^2(L^2)}^2 + \|p^*\,p'\,\Delta\varphi\|_{L^2(L^2)}^2\\ &\quad + \|p^*\,\psi_t\|_{L^2(L^2)}^2 + \|(p^*)'\,\psi_t\|_{L^2(L^2)}^2 + \|(p^*)'\,\psi\|_{L^2(L^2)}^2\\ &\quad + \|(p^*)''\,\psi\|_{L^2(L^2)}^2 + \|p^*\,\Delta\psi\|_{L^2(L^2)}^2 + \|(p^*)'\,\Delta\psi\|_{L^2(L^2)}^2)\Big).\end{aligned} \tag{3.85}$$

3.2.3. Conclusion. In this subsection we will end the proof of Lemma 3.1. We will first see that the global integrals in (3.85) can be absorbed by the terms in $I(s,\lambda;\varphi)$, and then we will combine the local estimates to end the proof.

Notice that the next bounds hold for the weight functions:

$$|p^*p'| + |(p^*)'p| \le C\,T\,s^{-3/4}\lambda^{-4}e^{-s\beta^*}\widehat{\gamma}^{-1/2}, \tag{3.86}$$

$$|(p^*)'p'| + |(p^*)''p| + |p^*p''| \le C\,T^2\,s^{1/4}\lambda^{-4}e^{-s\beta^*}(\gamma^*)^{3/4}. \tag{3.87}$$

Combining this and (3.85), we get

$$\begin{aligned}
&\iint_{\mathcal{O}_4\times(0,T)} |\widehat{p}|^2\,|\varphi_t|^2\,dx\,dt \\
&\le (1+\|\overline{y}\|_\infty^6)\,\|\overline{y}_t\|^2_{L^2(L^\sigma)} e^{C(1+T\,\|\overline{y}\|_\infty^2)}\Bigg(\lambda^{12}(1+T)(\|p\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))} \\
&\quad + \|p'\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))} + \|p\,\nabla\varphi\|^2_{L^2(L^2(\mathcal{O}_1))}) \\
&\quad + s^3\lambda^{-8}\iint_Q e^{-2s\beta^*}(\gamma^*)^3|\varphi|^2\,dx\,dt \\
&\quad + s^{-1}\lambda^{-8}\iint_Q e^{-2s\beta^*}\widehat{\gamma}^{-1}(|\varphi_t|^2 + |\Delta\varphi|^2)\,dx\,dt\Bigg)
\end{aligned} \tag{3.88}$$

for $s \ge C(T^4+T^8)$.

Taking now $\lambda \ge C(1+\|\overline{y}\|_\infty + \|\overline{y}_t\|^2_{L^2(L^\sigma)} + e^{C\,T\,\|\overline{y}\|_\infty^2})$, we obtain

$$\begin{aligned}
\iint_{\mathcal{O}_4\times(0,T)} |\widehat{p}|^2\,|\varphi_t|^2\,dx\,dt &\le C\,\lambda^{20}(1+T)(\|p\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))} \\
&\quad + \|p'\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))} + \|p\,\nabla\varphi\|^2_{L^2(L^2(\mathcal{O}_1))}) + \varepsilon\,I(s,\lambda;\varphi).
\end{aligned} \tag{3.89}$$

Let us finally combine (3.89), (3.32), (3.34), and (3.59). This yields

$$\begin{aligned}
I(s,\lambda;\varphi) \le C\,\lambda^{20}(1+T)\Big(&\|p\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))} + \|p'\,\varphi\|^2_{L^2(L^2(\mathcal{O}_1))} \\
&+ \|p\,\nabla\varphi\|^2_{L^2(L^2(\mathcal{O}_1))}\Big)
\end{aligned} \tag{3.90}$$

for $s \ge C(T^4+T^8)$ and $\lambda \ge C(1+\|\overline{y}\|_\infty + \|\overline{y}_t\|^2_{L^2(L^\sigma)} + e^{C\,T\,\|\overline{y}\|_\infty^2})$.

It remains only to absorb the last norm in (3.90). This can be achieved using localization arguments. Indeed, we have

$$\begin{aligned}
\iint_{\mathcal{O}_1\times(0,T)} |p|^2\,|\nabla\varphi|^2\,dx\,dt \le C\Bigg(&s\lambda^{20}(1+T)\iint_{\mathcal{O}\times(0,T)} e^{2s\beta^*}|p|^4\,\widehat{\gamma}\,|\varphi|^2\,dx\,dt \\
&+ \varepsilon\,s^{-1}\lambda^{-20}(1+T)^{-1}\iint_{\mathcal{O}\times(0,T)} e^{-2s\beta^*}\widehat{\gamma}^{-1}|\Delta\varphi|^2\,dx\,dt\Bigg)
\end{aligned} \tag{3.91}$$

for a constant $\varepsilon = \varepsilon(\Omega,\mathcal{O}) > 0$ small enough.

In conclusion, we get from (3.90)

$$I(s,\lambda;\varphi) \le C\, s^{16}\lambda^{40}(1+T^2) \iint_{\mathcal{O}\times(0,T)} e^{-8s\widehat{\alpha}+6s\alpha^*}\, \widehat{\xi}^{16}\, |\varphi|^2\, dx\, dt \tag{3.92}$$

for $s \ge C(T^4+T^8)$ and $\lambda \ge C(1+\|\overline{y}\|_\infty + \|\overline{y}_t\|^2_{L^2(L^\sigma)} + e^{C\,T\,\|\overline{y}\|^2_\infty})$, which is exactly (3.13).

3.3. Some additional comments. In this subsection, we will indicate several recent results. The first two concern the controllability of the three-dimensional Stokes and Navier–Stokes systems with two scalar controls. The third one is related to the null controllability of the Stokes system with slip boundary conditions.

3.3.1. Three-dimensional Stokes and Navier–Stokes systems and two scalar controls. Let us consider system (3.7) with $N = 3$. What we would like to do now is to control acting over a (small) part of Ω just in two directions. More precisely, if we denote the control by $v = (v_1, v_2, v_3)$, we intend to drive the system (3.7) to zero at time $t = T$ with the restriction

$$v_1 = 0 \quad \text{or} \quad v_2 = 0 \quad \text{or} \quad v_3 = 0. \tag{3.93}$$

In the context of approximate controllability, this question was considered (and solved) in [49] for cylindrical domains Ω.

We will have to impose an additional restriction on the control domain $\mathcal{O}$, namely,

$$\exists x^0 \in \partial\Omega,\ \exists \varepsilon > 0 \text{ such that } \overline{\mathcal{O}} \cap \partial\Omega \supset B(x^0;\varepsilon) \cap \partial\Omega. \tag{3.94}$$

The following result holds.

THEOREM 3.6. *Let T be an arbitrary positive time. Let* (3.9) *and* (3.94) *be fulfilled. Then, for every $y^0 \in H$, there exists a two-dimensional L^2 control v such that the solution to* (3.7) *satisfies*

$$y(T) = 0 \quad \text{in} \quad \Omega. \tag{3.95}$$

Let us indicate briefly the idea of the proof. It is not restrictive to assume that

$$B(x^0;\varepsilon) \cap \overline{\Omega} \subset \overline{\mathcal{O}}$$

and, for instance,

$$n_3(x) \ne 0 \quad \forall x \in B(x^0;\varepsilon) \cap \partial\Omega.$$

In view of the general arguments proved in section 1, the key point is to get a Carleman inequality of the form

$$s^{-1} \iint_Q \rho^2 |\varphi|^2\, dx\, dt \le C \iint_{\mathcal{O}\times(0,T)} \rho^2 \left(|\varphi_1|^2 + |\varphi_2|^2\right)\, dx\, dt \tag{3.96}$$

for the solutions to (3.10).

Let us set $\mathcal{O}_0 = B(x^0;\varepsilon) \cap \Omega$. We start from (3.13), which must hold for some $C > 0$. In $\mathcal{O}_0$, we have

$$\varphi_3(x,t) = \int_{l(x,t)} (\partial_1\varphi_1 + \partial_2\varphi_2)(x_1, x_2, \xi_3, t)\, d\xi_3, \tag{3.97}$$

where $l(x,t)$ is an appropriate segment in $\mathcal{O}_0$ that starts from (x,t), is parallel to e_3, and ends at a point on $\partial\Omega$.

Using (3.97) in (3.13), in view of Fubini's formula, Hölder's inequality, and standard localization arguments, together with the particular definition of the set $\mathcal{O}_0$, one can find a Carleman inequality with local terms of $|\varphi_1|^2$ and $|\varphi_2|^2$ on the right-hand side. This and the dissipativity of system (3.10) imply, in a usual way, an observability inequality for the associated solutions with only these two terms on the right-hand side, which yields the result stated in Theorem 3.6.

As a consequence of Theorem 3.6, following again the arguments developed in [37] and [24], one can deduce the local exact controllability to the trajectories of the Navier–Stokes system with two scalar controls:

THEOREM 3.7. *Let $(\overline{y}, \overline{p})$ be a solution of* (3.15) *verifying* (3.9) *and $\overline{y}(0) \in E = H \cap L^4(\Omega)^3$. There exists $\varepsilon > 0$ such that, for each $y^0 \in E$ satisfying*

$$\|y^0 - \overline{y}(0)\|_{L^4} \leq \varepsilon,$$

one can find a two-dimensional L^2 control v and an associated solution (y, p) to (3.14) *for which we have*

$$y(T) = \overline{y}(T) \quad \text{in } \Omega. \tag{3.98}$$

3.3.2. Stokes systems with slip boundary conditions. Let us introduce the following system:

$$\begin{cases} y_t - \Delta y + \nabla p = v 1_{\mathcal{O}} & \text{in} \quad Q, \\ \nabla \cdot y = 0 & \text{in} \quad Q, \\ y \cdot n = 0, \quad ((Dy) \cdot \vec{n})_{tg} + (A(x,t)\, y)_{tg} = 0 & \text{on} \quad \Sigma, \\ y(0) = y^0 & \text{in} \quad \Omega. \end{cases} \tag{3.99}$$

Here, we have used the notation a_{tg} to denote the tangential component of any vector $a \in \mathbf{R}^N$:

$$a_{tg} = a - (a \cdot n)n. \tag{3.100}$$

We will impose the following assumptions on A:

$$\begin{aligned} &A \in H^{(3-\delta)/2}(0, T; H^{1/2}(\partial\Omega)^{N\times N}), \\ &A \in H^{1-\delta}(0, T; H^{-1+N/2+2\delta'}(\partial\Omega)^{N\times N}), \end{aligned} \tag{3.101}$$

where $0 < \delta < 1/2$ arbitrarily close to $1/2$ and $\delta < \delta' < 1/2$. Then we have the next result from [35].

THEOREM 3.8. *Let us assume that A is a uniformly elliptic matrix verifying $A \in L^\infty(\Sigma)^{N\times N}$ and* (3.101). *Then, for each $y^0 \in H$, there exists $v \in L^2(\mathcal{O}\times(0,T))^N$ such that the solution to* (3.99) *satisfies*

$$y(T) = 0 \quad \text{in } \Omega. \tag{3.102}$$

Remark 10. It would be interesting to know if this result still holds, for general uniformly elliptic $A \in L^\infty(\Sigma)^{N\times N}$, without the hypotheses (3.101). However, this seems to be a difficult question.

For similar systems for the Stokes or Navier–Stokes equations involving nonlinear slip boundary conditions, the situation is much more complicated. See [35] for more details.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Contemp. Soviet Math., Consultants Bureau, New York, 1987.

[2] S. Anita and V. Barbu, *Null controllability of nonlinear convective heat equations*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 157–173.

[3] J.-P. Aubin, *L'analyse non linéaire et ses motivations économiques*, Masson, Paris, 1984.

[4] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials. The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[5] V. Barbu, *Exact controllability of the superlinear heat equation*, Appl. Math. Optim., 42 (2000), pp. 73–89.

[6] V. Barbu, *Controllability of parabolic and Navier-Stokes equations*, Sci. Math. Jpn., 56 (2002), pp. 143–211.

[7] V. Barbu and M. Iannelli, *Controllability of the heat equation with memory*, Differential Integral Equations, 13 (2003), pp. 1393–1412.

[8] V. Barbu and G. Tessitore, *Null controllability of stochastic heat equations with a multiplicative noise*, Appl. Math. Optim., 47 (2000), pp. 97–120.

[9] O. Bodart, M. González-Burgos, and R. Pérez-García, *Insensitizing controls for a semilinear heat equation with a superlinear nonlinearity*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 677–682.

[10] J.-M. Coron, *On the controllability of the* 2*-D incompressible Navier-Stokes equations with the Navier slip boundary conditions*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 35–75.

[11] J.-M. Coron and A. V. Fursikov, *Global exact controllability of the* 2*-D Navier-Stokes equations on a manifold without boundary*, Russian J. Math. Phys., 4 (1996), pp. 1–19.

[12] L. de Teresa, *Insensitizing controls for a semilinear heat equation*, Comm. Partial Differential Equations, 25 (2000), pp. 39–72.

[13] A. Doubova and E. Fernández-Cara, *Some control results for simplified one-dimensional models of fluid-solid interaction*, Math. Models Methods Appl. Sci., 15 (2005), pp. 783–824.

[14] A. Doubova, E. Fernández-Cara, and M. González-Burgos, *Controllability results for linear viscoelastic fluids of the Maxwell and Jeffreys kinds*, C. R. Acad. Sci. Paris, 331 (2000), pp. 537–542.

[15] A. Doubova, E. Fernández-Cara, and M. González-Burgos, *On the controllability of the heat equation with nonlinear boundary Fourier conditions*, J. Differential Equations, 196 (2004), pp. 385–417.

[16] A. Doubova, E. Fernández-Cara, M. González-Burgos, and E. Zuazua, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient*, SIAM J. Control Optim., 41 (2002), pp. 798–819.

[17] Yu. V. Egorov, *Some problems in the theory of optimal control*, Z. Vycisl. Mat. i Mat. Fiz., 3 (1963), pp. 887–904 (in Russian).

[18] C. Fabre, *Uniqueness results for Stokes equations and their consequences in linear and nonlinear control problems*, ESAIM Control Optim. Calc. Var., 1 (1995/96), pp. 267–302.

[19] C. Fabre, J.-P. Puel, and E. Zuazua, *Approximate controllability of the semilinear heat equation*, Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[20] E. Fernández-Cara, *Null controllability of the semilinear heat equation*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 87–103.

[21] E. Fernández-Cara, M. González-Burgos, S. Guerrero, and J.-P. Puel, *Null controllability of the heat equation with Fourier Fourier boundary conditions: The linear case*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 442–465.

[22] E. Fernández-Cara, M. González-Burgos, S. Guerrero, and J.-P. Puel, *Exact controllability to the trajectories of the heat equation with Fourier boundary conditions: The semilinear case*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 466–483.

[23] E. Fernández-Cara and S. Guerrero, *Global Carleman estimates for solutions of parabolic systems defined by transposition and some applications to controllability*, AMRX Appl. Math. Res. Express, Volume 2006 (2006), article 75090; available online from http://www.hindawi.com/GetArticle.aspx?doi=10.1155/AMRX/2006/75090.

[24] E. Fernández-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel, *Local exact controllability to the trajectories of the Navier-Stokes equations*, J. Math. Pures Appl., 83 (2004), pp. 1501–1542.

[25] E. Fernández-Cara, S. Guerrero, O. Yu. Imanuvilov, and J.-P. Puel, *Some controllability results for the* N*-dimensional Navier–Stokes and Boussinesq systems with* $N-1$ *scalar controls*, SIAM J. Control Optim., 45 (2006), pp. 146–173.

[26] E. Fernández-Cara and L. de Teresa, *Null controllability of a cascade system of parabolic-hyperbolic equations*, Discrete Contin. Dyn. Syst., 11 (2004), pp. 699–714.

[27] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[28] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 583–616.

[29] A. V. Fursikov and O. Yu. Imanuvilov, *Local exact controllability of the Navier-Stokes equations*, C. R. Acad. Sci. Paris Sér. I Math., 323 (1996), pp. 275–280.

[30] A. V. Fursikov and O. Yu. Imanuvilov, *Controllability of Evolution Equations*, Lecture Note Series 34, Research Institute of Mathematics, Seoul National University, Seoul, 1996.

[31] A. V. Fursikov and O. Yu. Imanuvilov, *Exact controllability of the Navier-Stokes and Boussinesq equations*, Uspekhi Mat. Nauk, 54 (1999), pp. 93–146 (in Russian); English translation in Russian Math. Surveys, 54 (1999), pp. 565–618.

[32] Y. Giga and H. Sohr, *Abstract L^p estimates for the Cauchy problem with applications to the Navier-Stokes equations in exterior domains*, J. Funct. Anal., 102 (1991), pp. 72–94.

[33] O. Glass, *An addendum to a J. M. Coron theorem concerning the controllability of the Euler system for 2D incompressible inviscid fluids*, J. Math. Pures Appl. (9), 80 (2001), pp. 845–877.

[34] R. Glowinski and J.-L. Lions, *Exact and approximate controllability for distributed parameter systems*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 1995, pp. 159–333.

[35] S. Guerrero, *Local exact controllability to the trajectories of the Navier-Stokes system with nonlinear Navier-slip boundary conditions*, ESAIM Control Optim. Calc. Var., 12 (2006), pp. 484–544.

[36] O. Yu. Imanuvilov, *Controllability of parabolic equations*, Sb. Math., 186 (1995), pp. 109–132 (in Russian).

[37] O. Yu. Imanuvilov, *Remarks on exact controllability for the Navier-Stokes equations*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 39–72.

[38] O. Yu. Imanuvilov and J.-P. Puel, *Global Carleman estimate for weak elliptic non homogeneous Dirichlet problem*, Int. Math. Res. Not., 16 (2003), pp. 883–913.

[39] O. Yu. Imanuvilov and M. Yamamoto, *Carleman estimate for a parabolic equation in a Sobolev space of negative order and its applications*, in Control of Nonlinear Distributed Parameter Systems, Lecture Notes in Pure and Appl. Math. 218, Marcel Dekker, New York, 2001, pp. 113–137.

[40] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, translated from the Russian by S. Smith, Translations of Mathematical Monographs 23, AMS, Providence, RI, 1967.

[41] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[42] G. Lebeau and E. Zuazua, *Null-controllability of a system of linear thermoelasticity*, Arch. Rational Mech. Anal., 141 (1998), pp. 297–329.

[43] G. Lebeau and E. Zuazua, *Decay rates for the three-dimensional linear system of thermoelasticity*, Arch. Rational Mech. Anal., 148 (1999), pp. 179–231.

[44] J.-L. Lions, *Exact controllability, stabilizability and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[45] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, Tome* 1, Rech. Math. Appl. 8, Masson, Paris, 1988.

[46] J.-L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, Tome* 2, Rech. Math. Appl. 9, Masson, Paris, 1988.

[47] J.-L. Lions, *Quelques notions dans l'analyse et le contrôle de systèmes à données incomplètes*, in Proceedings of the 11th Congress on Differential Equations and Applications/First Congress on Applied Mathematics, University of Málaga, Málaga, 1989, pp. 43–54.

[48] J.-L. Lions, *Are there connections between turbulence and controllability?*, in Analysis and Optimization of Systems, A. Bensoussan and J.-L. Lions, eds., Lecture Notes Control and Inform. Sci. 144, Springer-Verlag, Berlin, 1990.

[49] J.-L. Lions and E. Zuazua, *A generic uniqueness result for the Stokes system and its control theoretical consequences*, in Partial Differential Equations and Applications, P. Marcellini, G. Talenti, and E. Visentini, eds., Lecture Notes in Pure and Appl. Math. 177, Marcel Dekker, New York, 1996, pp. 221–235.

[50] J.-L. Lions and E. Zuazua, *Exact boundary controllability of Galerkin's approximations of Navier-Stokes equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 605–621.

[51] J.-P. Puel, *Controllability of Partial Differential Equations*, Lectures in the Universidade Federal do Rio de Janeiro, Brazil, 2002.

[52] D. L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Int. Math. Res. Not., 52 (1973), pp. 189–221.

[53] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[54] Th. I. Seidman, *On uniform null-controllability and blow-up estimates*, in Control Theory of Partial Differential Equations, Lect. Notes Pure Appl. Math. 242, Chapman & Hall/CRC, Boca Raton, FL, 2005, pp. 213–227.

[55] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, 2nd ed., North-Holland, Amsterdam, 1977.

[56] E. Zuazua, *Exact boundary controllability for the semilinear wave equation*, in Nonlinear Partial Differential Equations and Their Applications, Vol. 10, Pitman Res. Notes Math. Ser. 220, Longman, Harlow, UK, 1991, pp. 357–391.

[57] E. Zuazua, *Controllability of the linear system of thermoelasticity*, J. Math. Pures Appl., 74 (1995), 303–346.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1447–1466

PHASE-FIELD RELAXATION OF TOPOLOGY OPTIMIZATION WITH LOCAL STRESS CONSTRAINTS*

MARTIN BURGER† AND ROMAN STAINKO‡

Abstract. We introduce a new relaxation scheme for structural topology optimization problems with local stress constraints based on a phase-field method. In the basic formulation we have a PDE-constrained optimization problem, where the finite element and design analysis are solved simultaneously. The starting point of the relaxation is a reformulation of the material problem involving linear and 0-1 constraints only. The 0-1 constraints are then relaxed and approximated by a Cahn–Hilliard-type penalty in the objective functional, which yields convergence of minimizers to 0-1 designs as the penalty parameter decreases to zero. A major advantage of this kind of relaxation opposed to standard approaches is a uniform constraint qualification that is satisfied for any positive value of the penalization parameter. The relaxation scheme yields a large-scale optimization problem with a high number of linear inequality constraints. We discretize the problem by finite elements and solve the arising finite-dimensional programming problems by a primal-dual interior point method. Numerical experiments for problems with local stress constraints based on different criteria indicate the success and robustness of the new approach.

Key words. topology optimization, local stress constraints, phase-field methods, interior-point methods, one-shot methods

AMS subject classifications. 74P05, 74P10, 74P15, 90C51, 74S05

DOI. 10.1137/05062723X

1. Introduction.

Topology optimization denotes problems of finding optimal material distributions in given design domains subject to certain criteria and, possibly, satisfying several additional constraints. In the last two decades, advances in homogenization, optimization theory, and numerical analysis as well as new engineering approaches caused topology optimization techniques to become a standard tool of engineering design (cf. [5, 18] for an overview), in particular in structural mechanics. In structural optimization there are two design-constraint combinations of particular importance, namely the maximization of material stiffness at given mass and the minimization of mass while keeping a certain stiffness. The formulation of the first combination as the so-called minimal compliance problem has become standard, and seems to be well understood with respect to its mathematical properties (cf., e.g., [2, 8, 26]), and various successful numerical techniques have been proposed (cf., e.g., [3, 7, 20, 27, 31]). The treatment of the second problem is, by far, less understood, although this problem characterizes one of the oldest mechanical requirements in structural design, e.g., in aeronautics minimization of mass without material failure under loading is an important issue. Until now there seems to have been no approach capable of computing reliable (global) optima with respect to local stress constraints within reasonable computational effort.

*Received by the editors March 21, 2005; accepted for publication (in revised form) April 16, 2006; published electronically October 12, 2006. This work was supported by the Austrian Science Fund FWF through SFB F 013, subgrants 08 (MB) and 09 (RS).
http://www.siam.org/journals/sicon/45-4/62723.html
†Institut für Industriemathematik, Johannes Kepler Universität, Altenbergerstr. 69, A 4040 Linz, Austria (martin.burger@jku.at).
‡Spezialforschungsbereich SFB F 013 Numerical and Symbolic Scientific Computing, Altenbergerstr. 69, A 4040 Linz, Austria (roman.stainko@jku.at).

The main source of difficulties in this problem is a lack of constraint qualification in the feasible set defined by the local stress constraints, which already appear for simple truss structures (cf. [24, 29]). Moreover, there are several complications for specific methods, e.g., convergence issues of homogenized stress criteria for material interpolation schemes (cf. [5, 17]).

We start by describing the main mathematical setup used in what follows. By $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ we denote the design domain, which we assume to be sufficiently regular. The function $\mathbf{u} : \Omega \subset \mathbb{R}^d \to \mathbb{R}^d$ denotes the elastic displacement, the strain

$$\mathbf{e}_{ij} := \frac{1}{2}\left(\frac{\partial \mathbf{u}_i}{\partial x_j} + \frac{\partial \mathbf{u}_j}{\partial x_i}\right)$$

and the stress σ, determined from the strain via a standard linearly elastic relation

$$\sigma_{ij} = \sum_{k,\ell} C_{ijk\ell}\mathbf{e}_{k\ell}$$

with a suitable symmetric positive definite elasticity tensor $(C_{ijk\ell})$; i.e.,

$$\sum_{i,j,k,\ell} C_{ijk\ell}\mathbf{e}_{ij}\mathbf{e}_{k\ell} \geq \eta \sum_{ij} \mathbf{e}_{ij}^2$$

for some $\eta > 0$. Below we shall abbreviate the stress-strain relation as $\sigma = \mathbf{C} : \mathbf{e}$. The scaled density of the material is denoted by $\chi : \Omega \to \{0, 1\}$, which we normalize to $\chi(\mathbf{x}) = 1$ if there is material at the point $\mathbf{x}$ and $\chi(\mathbf{x}) = 0$ otherwise. Then the stress constrained topology optimization problem is given by

$$\begin{aligned}
\int_\Omega \chi \, \mathrm{d}\mathbf{x} \quad &\to \quad \min_{\chi \in \{0,1\} \text{ a.e.}}, \\
\operatorname{div} \sigma = 0 \quad &\text{in } \{\chi = 1\}, \\
\sigma = \mathbf{C} : \mathbf{e} \quad &\text{in } \Omega, \\
\mathbf{u} = \mathbf{0} \quad &\text{on } \Gamma_D \subset \partial\Omega, \\
\sigma \cdot \mathbf{n} = \mathbf{t} \quad &\text{on } \Gamma_N \subset \partial\Omega, \\
\sigma \cdot \mathbf{n} = \mathbf{0} \quad &\text{on } \partial\{\chi = 1\} \cup (\partial\Omega - \Gamma_D - \Gamma_N), \\
\sigma^{\min} \leq \sigma \leq \sigma^{\max} \quad &\text{in } \{\chi = 1\}, \\
\mathbf{u}^{\min} \leq \mathbf{u} \leq \mathbf{u}^{\max} \quad &\text{in } \Omega,
\end{aligned}$$

where $\mathbf{n}$ denotes the outward unit normal and $\mathbf{t}$ is a traction force applied on a part Γ_N of the boundary $\partial\Omega$. In the current geometrical setup, the material is kept fixed on Γ_D, while the remaining part $\partial\Omega - \Gamma_D - \Gamma_N$ does not need to represent material boundaries. Moreover, we have ignored body forces for simplicity, but they could be incorporated by adding a right-hand side to the divergence equation.

The matrices $\sigma^{\min}$ and $\sigma^{\max}$ denote the allowed minimal and maximal value for the local stresses, respectively. We shall call this criterion total stress constraints. Alternatively, the case of von Mises stress constraints is of interest, where the local constraints on σ are replaced by

$$\Phi(\sigma) \leq \Phi^{\max}, \tag{1.1}$$

where the von Mises stress is denoted via the functional $\Phi : \mathbb{R}^{d\times d} \to \mathbb{R}$ given by

$$\Phi(\sigma) = \sqrt{\frac{\sum_{i,j}(\lambda_i - \lambda_j)^2}{2}},$$

where $\lambda_j, j = 1, \dots, d$, are the principal stresses (the eigenvalues of σ). Note that for $d = 2$, the case we are focusing on, we simply have

$$\Phi(\sigma) = |\lambda_1 - \lambda_2| = \sqrt{(\sigma_{11} - \sigma_{22})^2 + 4\sigma_{12}^2}.$$

A frequently used approach to overcome the difficulties with the missing constraint qualification is the so-called epsilon-relaxation approach (cf. [15, 24]), originally used in the optimization of truss structures, which perturbs the stress criterion by some small parameter ϵ. In the topology optimization of continuum structures, the ϵ-relaxation can be combined with standard material interpolation schemes and finite element discretization in order to compute approximations of solutions, an approach carried out by Duysinx and Bendsøe [17]. A drawback of the ϵ-relaxation approach is the fact that the constraint qualification is not uniform for positive ϵ and there could still be cases where the Slater qualifications is violated (see [17] for a discrete setting). Recently, it has been shown for such examples by Stolpe and Svanberg [29] that the trajectories of minimizers of the ϵ-relaxed approach can have a discontinuity for arbitrarily small ϵ, and as a consequence it is difficult (or even impossible) to compute reliable minimizers of topology optimization problems with local stress constraints using this approach.

Due to the well-known ill-posedness of topology optimization problems (see [26]), we add a perimeter penalization to the objective functional; i.e., we minimize

$$J^\gamma(\chi) = \gamma \int_\Omega \chi \, \mathrm{d}\mathbf{x} + |\chi|_{BV} \tag{1.2}$$

for a (large) parameter $\gamma > 0$, where

$$|\chi|_{BV} := \sup_{\substack{\mathbf{g} \in C_0^\infty(\Omega;\mathbb{R}^d) \\ \|\mathbf{g}\|_\infty \le 1}} \int_\Omega \operatorname{div} \mathbf{g} \; \chi \, \mathrm{d}\mathbf{x}.$$

The additional perimeter term equals the length of the curve $\partial\{\chi = 1\}$ for $d = 2$ and the area of the surface $\partial\{\chi = 1\}$ for $d = 3$. The boundedness of the perimeter regularizes the topology optimization problem; in particular it excludes checkerboard effects and mesh-dependent results (cf. [19, 23]).

In this paper we use a different approach to the relaxation of the local constraints. The starting point of our analysis is a reformulation of the equality constraints describing the elastic equilibrium and the local inequality constraints for stresses and displacements into a system of linear inequality constraints, as recently proposed by Stolpe and Svanberg [30]. This reformulation is approximate at the continuum level but exact for finite element discretizations with suitable parameter choice. The main difficulty is that the arising problem also involves 0-1 constraints in addition to the linear inequalities. The computational effort of methods for the global minimization of these mixed linear programming problems grows fast with the number of degrees of freedom (DOFs) in the discretization, so that the problem could be solved only for very coarse discretizations so far (cf. [28, 30]). Instead of solving mixed linear programming problems, we propose using a phase-field relaxation of the reformulated problem. The phase-field relaxation consists in using an interpolated material density ρ, similar to material interpolation schemes. In addition, a Cahn–Hilliard-type penalization functional (see [14]) of the form

$$P^\epsilon(\rho) = \frac{\epsilon}{2}\int_\Omega |\nabla\rho|^2 \,\mathrm{d}\mathbf{x} + \frac{1}{\epsilon}\int_\Omega W(\rho)\,\mathrm{d}\mathbf{x} \tag{1.3}$$

is used to approximate the perimeter, where $W : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is a scalar function with exactly two minimizers at 0 and 1 satisfying $W(0) = W(1) = 0$. The second term of the penalty functional ensures that the values of the material density ρ converge to 0 or 1 as $\epsilon \to 0$, while the first term controls the perimeter of level sets of ρ. Due to a famous result by Modica and Mortola [22] (cf. also [1, 21]), minimizers of P^ϵ with fixed volume $\int_\Omega \rho \,\mathrm{d}\mathbf{x}$ converge to minimizers χ of the perimeter at fixed volume over functions satisfying $\chi \in \{0, 1\}$ almost everywhere in Ω. This convergence arises in the framework of Γ-convergence (see [10] and the references therein), which ensures, in particular, convergence of minimizers. Together with finite element discretization, we shall use a continuation $\epsilon \to 0$ and solve the arising discretized problems by an interior-point method. Obviously, the phase-field approach is not the only possibility to relax the reformulated constraints; one could, e.g., use a direct relaxation without further approximations, standard material interpolation schemes, or level set methods (cf., e.g., [12]), which are closely related to phase-field methods (see [4]). However, the phase-field relaxation incorporates a variety of advantages with respect to such approaches:

- In contrast to a direct relaxation to a continuous density variable (and also in contrast to material interpolation schemes), the phase-field method still provides geometric information. In particular, one can expect $\{\rho > \eta\}$ (with $\eta \ll \frac{1}{2}$ small) to be a superset of the limit $\{\chi = 1\}$ and $\{\rho < 1 - \eta\}$ to be a subset (for ϵ sufficiently small). Moreover, geometric quantities such as mean curvature can be approximated in terms of derivatives of ρ for small ϵ.
- With the phase-field relaxation one can still use the density linearly in the constraints, which is not true for material interpolation (e.g., in SIMP one has ρ^p, $p > 1$) or for level set methods (where the unknown is a signed distance function to some boundary, and in the relaxation one usually takes an application of a smoothed Heaviside function). The additional nonlinearity does not only complicate the constraints but might also destroy constraint qualifications.
- The parameter ϵ can be used for continuation. For ϵ being large, the functional P^ϵ is strictly convex, so that one can compute global optima for arbitrary initial values. When decreasing ϵ, the minimizer of the previous step can be expected to provide a good initial guess for the next step carried out with a smaller ϵ.

To our knowledge, the phase-field approach in topology optimization was first introduced by Bourdin and Chambolle [9] for a design problem with design-dependent loads, another type of problem where standard material interpolation schemes encounter difficulties. The approach has recently been applied to minimal compliance-type problems by Wang and Zhou [33].

The remainder of this paper is organized as follows: in section 2 we review the constraint reformulation due to [30] and extend the approach to an approximate reformulation of the continuous problem. In section 3 we introduce the phase-field relaxation and analyze its basic properties. The finite element discretization yielding linearly constrained programming problems is discussed in section 4 and the solution of these programming problems by interior-point methods in section 5. Finally, we present numerical results obtained for local as well as for von Mises stress constraints.

2. Reformulation of constraints. In the following we consider a reformulation of constraints on subsets of locally bounded stresses; i.e.,

$$\beta|\sigma_{ij}| \le 1 \qquad \text{in } \Omega, \qquad i,j = 1,\dots,d, \tag{2.1}$$

for some (small) $\beta > 0$ such that $\beta \max_{i,j} \sigma_{ij}^{\max} < 1$ and $\beta \min_{i,j} \sigma_{ij}^{\min} > -1$ (i.e., the original total stress constraints are more restrictive for $\chi = 1$). Condition (2.1) is an additional restriction on the design set, which will be needed for the reformulations below. Since the original stress constraint is stronger than (2.1) in regions with material, the additional constraint just states that the stresses should continue in a reasonable (in particular finite) way in regions without material. The total constraints for the displacement $\mathbf{u}$, the stress σ, and the density χ are given by

$$\begin{aligned}
\operatorname{div} \sigma &= 0 && \text{in } \{\chi = 1\},\\
\sigma &= \mathbf{C} : \mathbf{e}(\mathbf{u}) && \text{in } \Omega,\\
\mathbf{u} &= \mathbf{0} && \text{on } \Gamma_D \subset \partial\Omega,\\
\sigma \cdot \mathbf{n} &= \mathbf{t} && \text{on } \Gamma_N \subset \partial\Omega,\\
\sigma \cdot \mathbf{n} &= \mathbf{0} && \text{on } \partial\{\chi = 1\} \cup (\partial\Omega - \Gamma_D - \Gamma_N),\\
\chi &\in \{0,1\} && \text{a.e. in } \Omega,\\
\Phi^{\min} \le \Phi(\sigma) &\le \Phi^{\max} && \text{in } \{\chi = 1\},\\
\mathbf{u}^{\min} \le \mathbf{u} &\le \mathbf{u}^{\max} && \text{in } \Omega.
\end{aligned} \tag{2.2}$$

We shall reformulate the constraints (2.1), (2.2) as linear inequality constraints for the case of total stress and von Mises stress constraints.

2.1. Reformulation of total stress constraints. We start with the reformulation in the case of total stress constraints, i.e., $\Phi(\sigma) = \sigma$. For this sake we introduce the approximate constraint sets

$$\begin{aligned}
\mathcal{C}^\beta := \{&(\chi,\mathbf{u},\sigma) \in BV(\Omega;[0,1]) \times H^1(\Omega;\mathbb{R}^d) \times L^\infty(\Omega;\mathbb{R}^{d\times d}) \mid\\
&(\chi,\mathbf{u},\sigma) \text{ satisfies (2.1), (2.2)}\}
\end{aligned}$$

and an additional artificial stress variable $\mathbf{s} \in L^\infty(\Omega;\mathbb{R}^{d\times d})$.

Let $(\chi,\mathbf{u},\sigma) \in \mathcal{C}^\beta$, and let $\mathbf{s} = \sigma$ if $\chi = 1$ and $\mathbf{s} = \mathbf{0}$ if $\chi = 0$, i.e., $\mathbf{s} = \chi\sigma$. Then the constraints

$$-(1-\chi)\mathbf{1} \le \beta(\sigma - \mathbf{s}) \le (1-\chi)\mathbf{1} \tag{2.3}$$

with the matrix $\mathbf{1} = (1)_{ij}$ and

$$\sigma^{\min}\chi \le \mathbf{s} \le \sigma^{\max}\chi \tag{2.4}$$

are satisfied. Vice versa, assume that

$$(\chi,\mathbf{u},\sigma,\mathbf{s}) \in BV(\Omega;[0,1]) \times H^1(\Omega;\mathbb{R}^d) \times L^\infty(\Omega;\mathbb{R}^{d\times d}) \times L^\infty(\Omega;\mathbb{R}^{d\times d})$$

fulfills (2.3), (2.4) and $\chi \in \{0,1\}$ almost everywhere. Then, for $\chi = 0$ (2.4) implies $\mathbf{s} = 0$ and (2.3) gives $-1 \le \beta\sigma_{ij} \le 1$. For $\chi = 1$ (2.3) implies $\mathbf{s} = \sigma$ and (2.4) becomes

$$\sigma^{\min}\chi \le \sigma \le \sigma^{\max}\chi.$$

Moreover, since either $\mathbf{s} = \sigma$ or $\mathbf{s} = \mathbf{0}$, we obtain that $\operatorname{div} \mathbf{s} = 0$ almost everywhere in Ω.

Due to the above arguments we conclude that $(\chi, \mathbf{u}, \sigma) \in \mathcal{C}^\beta$ if and only if there exists $\mathbf{s}$ such that

$$
\begin{aligned}
\operatorname{div} \mathbf{s} &= 0 && \text{in } \Omega,\\
\sigma &= \mathbf{C} : \mathbf{e}(\mathbf{u}) && \text{in } \Omega,\\
\mathbf{u} &= \mathbf{0} && \text{on } \Gamma_D \subset \partial\Omega,\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{t} && \text{on } \Gamma_N \subset \partial\Omega,\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{0} && \text{on } \partial\Omega - \Gamma_D - \Gamma_N,\\
-(1-\chi)\mathbf{1} \le \beta(\sigma - \mathbf{s}) &\le (1-\chi)\mathbf{1} && \text{in } \Omega,\\
\chi &\in \{0,1\} && \text{a.e. in } \Omega,\\
\sigma^{\min}\chi \le \mathbf{s} &\le \sigma^{\max}\chi && \text{in } \Omega,\\
\mathbf{u}^{\min} \le \mathbf{u} &\le \mathbf{u}^{\max} && \text{in } \Omega.
\end{aligned}
\tag{2.5}
$$

We mention that the conditions $\operatorname{div} \mathbf{s} = 0$, $\mathbf{s}\cdot\mathbf{n} = \mathbf{t}$ on Γ_D, and $\mathbf{s}\cdot\mathbf{n} = \mathbf{0}$ on $\partial\Omega - \Gamma_D - \Gamma_N$ have to be understood in a weak sense, namely, as $\int_\Omega \mathbf{s} : \mathbf{e}(\mathbf{v})\, d\mathbf{x} = \int_{\Gamma_N} \mathbf{v}\cdot\mathbf{t}\, d\mathbf{a}$ for all $\mathbf{v} \in \{\mathbf{v} \in H^1(\Omega;\mathbb{R}^d) \mid \mathbf{v}|_{\Gamma_D} = \mathbf{0}\}$.

Note that (except $\chi \in \{0,1\}$) the constraints (2.5) are linear with respect to the new vector of unknowns $(\chi, \mathbf{u}, \sigma, \mathbf{s})$; in particular all constraints are formulated on Ω and not on the unknown set $\{\chi = 1\}$. We would like to mention that the drawback of the reformulation is an increase in the number of unknowns and a high number of inequality constraints. On the other hand, this higher number of unknowns and constraints seems to be a reasonable price for the linear reformulation of the complicated original constraints.

2.2. Reformulation of von Mises stress constraints. In the following we discuss the reformulation of the inequalities in the case of von Mises stress constraints for dimension $d = 2$ (similar computations are possible for $d = 3$). Since both sides of the constraint (1.1) are positive, we can square them, and since the constraint must hold only for $\chi = 1$, it can be written equivalently as

$$\chi(\sigma_{11} - \sigma_{22})^2 + 4\chi\sigma_{12}^2 \le \chi(\Phi^{\max})^2.$$

A more conservative version of the von Mises criterion (see [30]) is given by

$$\chi|\sigma_{11} - \sigma_{22}| + 2\chi|\sigma_{12}| \le \chi\Phi^{\max}. \tag{2.6}$$

As above we introduce an artificial stress variable $\mathbf{s} = \chi\sigma$, and since $\chi^2 = \chi$, we can reformulate the von Mises constraint as

$$(\mathbf{s}_{11} - \mathbf{s}_{22})^2 + 4(\mathbf{s}_{12})^2 \le \chi(\Phi^{\max})^2.$$

The reformulation of the von Mises stress constraints on the constraint set $\mathcal{C}^\beta$ is given by

$$
\begin{aligned}
\operatorname{div} \mathbf{s} &= 0 && \text{in } \Omega,\\
\sigma &= \mathbf{C} : \mathbf{e}(\mathbf{u}) && \text{in } \Omega,\\
\mathbf{u} &= \mathbf{0} && \text{on } \Gamma_D \subset \partial\Omega,\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{t} && \text{on } \Gamma_N \subset \partial\Omega,\\
\mathbf{s}\cdot\mathbf{n} &= \mathbf{0} && \text{on } \partial\Omega - \Gamma_D - \Gamma_N,\\
-(1-\chi)\mathbf{1} \le \beta(\sigma - \mathbf{s}) &\le (1-\chi)\mathbf{1} && \text{in } \Omega,\\
(\mathbf{s}_{11} - \mathbf{s}_{22})^2 + 4(\mathbf{s}_{12})^2 &\le \chi(\Phi^{\max})^2 && \text{in } \Omega,\\
\chi &\in \{0,1\} && \text{a.e. in } \Omega,\\
\mathbf{u}^{\min} \le \mathbf{u} &\le \mathbf{u}^{\max} && \text{in } \Omega.
\end{aligned}
\tag{2.7}
$$

Since the von Mises stress is nonnegative, we do not state a lower bound $\Phi^{\min}$ on the stresses in (2.7) (we could choose the redundant constraint $\Phi^{\min} = 0$). Note that in this case, $\chi = 0$ in the von Mises stress constraint does not imply directly $\mathbf{s} = \mathbf{0}$ but only $\mathbf{s}_{11} - \mathbf{s}_{22} = 0$ and $\mathbf{s}_{12} = 0$. However, since $\mathbf{s}$ is divergence-free in addition, we may conclude for those special stresses that $\nabla \mathbf{s}_{11} = \mathbf{0}$. Hence, $\mathbf{s}$ is of the form

$$\mathbf{s} = \chi\sigma + (1-\chi)\begin{pmatrix} C & 0 \\ 0 & C \end{pmatrix}$$

for any constant $C \in \mathbb{R}$. Since such artificial stresses will not change the von Mises stress, and since $\mathbf{s}$ does not have a physical meaning for $\chi = 0$, the additional terms will not play a major role (we are basically free to define $\mathbf{s}$ arbitrarily in $\{\chi = 0\}$). The reformulation (2.7) involves convex quadratic constraints but only with respect to the stress variable $\mathbf{s}$, which still yields constraint qualification after relaxation and discretization.

In this paper we rather use the conservative von Mises stress criterion (2.6), which can actually be reformulated into linear inequalities. Again we use $\mathbf{s} = \chi\sigma$ to obtain

$$|\mathbf{s}_{11} - \mathbf{s}_{22}| + 2|\mathbf{s}_{12}| \le \chi\Phi^{\max}.$$

Moreover, we introduce functions p and q such that $p \ge 0$, $q \ge 0$ and

$$-p \le \mathbf{s}_{11} - \mathbf{s}_{22} \le p, \qquad -q \le \mathbf{s}_{12} \le q.$$

From these constraints we obtain that

$$|\mathbf{s}_{11} - \mathbf{s}_{22}| \le p, \qquad |\mathbf{s}_{12}| \le q.$$

Finally, we impose the constraint

$$p + 2q \le \chi\Phi^{\max}$$

and obtain as a consequence the conservative von Mises stress for $\mathbf{s}$. On the other hand, if the constraint is satisfied by $\mathbf{s}$, the variables

$$\begin{aligned} p &= \max\{-(\mathbf{s}_{11} - \mathbf{s}_{22}), \mathbf{s}_{11} - \mathbf{s}_{22}\}, \\ q &= \max\{-\mathbf{s}_{12}, \mathbf{s}_{12}\} \end{aligned}$$

satisfy the new linear constraints, and thus we conclude equivalence to the original conservative von Mises constraints. The reformulation of the conservative von Mises stress constraints on the constraint set $\mathcal{C}^\beta$ is given by

$$\begin{aligned} \operatorname{div} \mathbf{s} = 0 \quad & \text{in } \Omega, \\ \sigma = \mathbf{C} : \mathbf{e}(\mathbf{u}) \quad & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} \quad & \text{on } \Gamma_D \subset \partial\Omega, \\ \mathbf{s} \cdot \mathbf{n} = \mathbf{t} \quad & \text{on } \Gamma_N \subset \partial\Omega, \\ \mathbf{s} \cdot \mathbf{n} = \mathbf{0} \quad & \text{on } \partial\Omega - \Gamma_D - \Gamma_N, \\ -(1-\chi)\mathbf{1} \le \beta(\sigma - \mathbf{s}) \le (1-\chi)\mathbf{1} \quad & \text{in } \Omega, \\ p \ge 0, \quad q \ge 0 \quad & \text{in } \Omega, \\ -p \le \mathbf{s}_{11} - \mathbf{s}_{22} \le p \quad & \text{in } \Omega, \\ -q \le \mathbf{s}_{12} \le q \quad & \text{in } \Omega, \\ p + 2q \le \chi\Phi^{\max} \quad & \text{in } \Omega, \\ \chi \in \{0,1\} \quad & \text{a.e. in } \Omega, \\ \mathbf{u}^{\min} \le \mathbf{u} \le \mathbf{u}^{\max} \quad & \text{in } \Omega. \end{aligned}$$

3. Phase-field relaxation. We now turn our attention to the relaxation of the stress constrained topology optimization problem. For this sake we replace the indicator function χ by a density $\rho : \Omega \to [0,1]$ and approximate the perimeter term in the regularized objective functional J^γ (1.2) by the Cahn–Hilliard term P^ϵ (1.3). The resulting relaxation in the case of total stress constraints is given by

$$
\begin{aligned}
M(\rho) = \gamma \int_\Omega \rho \, \mathrm{d}\mathbf{x} + P^\epsilon(\rho) \quad &\to \quad \min, \\
\operatorname{div} \mathbf{s} = 0 \quad &\text{in } \Omega, \\
\sigma = \mathbf{C} : \mathbf{e}(\mathbf{u}) \quad &\text{in } \Omega, \\
\mathbf{u} = \mathbf{0} \quad &\text{on } \Gamma_D, \\
\mathbf{s} \cdot \mathbf{n} = \mathbf{t} \quad &\text{on } \Gamma_N, \\
\mathbf{s} \cdot \mathbf{n} = \mathbf{0} \quad &\text{on } \partial\Omega - \Gamma_D - \Gamma_N, \\
\rho = 1 \quad &\text{on } \Gamma_N, \\
-(1-\rho)\mathbf{1} \le \beta(\sigma - \mathbf{s}) \le (1-\rho)\mathbf{1} \quad &\text{in } \Omega, \\
\sigma^{\min}\rho \le \mathbf{s} \le \sigma^{\max}\rho \quad &\text{in } \Omega, \\
0 \le \rho \le 1 \quad &\text{a.e. in } \Omega, \\
\mathbf{u}^{\min} \le \mathbf{u} \le \mathbf{u}^{\max} \quad &\text{in } \Omega.
\end{aligned}
\tag{3.1}
$$

The function space setting for this relaxation is given by

$$(\rho, \mathbf{u}, \sigma, \mathbf{s}) \in (H^1(\Omega) \cap L^\infty(\Omega)) \times (H^1(\Omega;\mathbb{R}^d) \cap L^\infty(\Omega;\mathbb{R}^d)) \times L^\infty(\Omega;\mathbb{R}^{d\times d})^2.$$

Note that, in addition to the original constraint reformulation (2.5), we have added a Dirichlet boundary condition for ρ on Γ_N, which is well defined in the sense of traces of functions in $H^1(\Omega)$. The reasoning for adding this condition is as follows: in the original constraint we have $\mathbf{s} \cdot \mathbf{n} = \mathbf{t} \neq \mathbf{0}$ on Γ_N. Hence, there exists a small open neighborhood of Γ_N in Ω, where $\chi = 1$ (since otherwise $\chi = 0$ would imply $\mathbf{s} = \mathbf{0}$ and thus $\mathbf{s} \cdot \mathbf{n} = \mathbf{0}$ on Γ_N). In the relaxed formulation, the trace of ρ is positive by analogous arguments but not necessarily equal to one. Thus, the additional constraint will not change the limit of the constraint set, but, on the other hand, it restricts the relaxation and simplifies the analysis of the relaxed problem.

In a similar way we can state a relaxed formulation of the problem with conservative von Mises stress constraints ($d = 2$) as

$$
\begin{aligned}
M(\rho) = \gamma \int_\Omega \rho \, \mathrm{d}\mathbf{x} + P^\epsilon(\rho) \quad &\to \quad \min, \\
\operatorname{div} \mathbf{s} = 0 \quad &\text{in } \Omega, \\
\sigma = \mathbf{C} : \mathbf{e}(\mathbf{u}) \quad &\text{in } \Omega, \\
\mathbf{u} = \mathbf{0} \quad &\text{on } \Gamma_D, \\
\mathbf{s} \cdot \mathbf{n} = \mathbf{t} \quad &\text{on } \Gamma_N, \\
\mathbf{s} \cdot \mathbf{n} = \mathbf{0} \quad &\text{on } \partial\Omega - \Gamma_D - \Gamma_N, \\
\rho = 1 \quad &\text{on } \Gamma_N, \\
-(1-\rho)\mathbf{1} \le \beta(\sigma - \mathbf{s}) \le (1-\rho)\mathbf{1} \quad &\text{in } \Omega, \\
p \ge 0, \quad q \ge 0 \quad &\text{in } \Omega, \\
-p \le \mathbf{s}_{11} - \mathbf{s}_{22} \le p \quad &\text{in } \Omega, \\
-q \le \mathbf{s}_{12} \le q \quad &\text{in } \Omega, \\
p + 2q \le \chi\Phi^{\max} \quad &\text{in } \Omega, \\
0 \le \rho \le 1 \quad &\text{a.e. in } \Omega, \\
\mathbf{u}^{\min} \le \mathbf{u} \le \mathbf{u}^{\max} \quad &\text{in } \Omega,
\end{aligned}
\tag{3.2}
$$

with variables

$$(\rho, \mathbf{u}, \sigma, \mathbf{s}, p, q) \in (H^1(\Omega) \cap L^\infty(\Omega)) \times (H^1(\Omega; \mathbb{R}^d) \times L^\infty(\Omega; \mathbb{R}^d)) \\ \times L^\infty(\Omega; \mathbb{R}^{d\times d})^2 \times L^\infty(\Omega)^2,$$

with the notation $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$. In our discussion below we will focus on total stress constraints, i.e., the relaxed optimization problem (3.1), but analogous reasoning is possible for von Mises stress constraints.

The basis of the relaxation idea is the Γ-convergence of problem (3.1) to (2.5) as ϵ tends to zero (respectively, Γ-convergence of (3.2) to (2.7) for the conservative von Mises constraints). This can be derived from the Γ-convergence of the perimeter term P^ϵ (cf. [21, 22]) to which only a continuous functional is added (compare Proposition 6.2.1 in [16]).

3.1. Structure of the relaxed problem. In the following we further examine the structure of the relaxed problem (3.1). Due to the term $W(\rho)$ in the Cahn–Hilliard penalty, we have to expect the objective functional to be nonconvex, in particular for small ϵ, when minimizers are forced to take values close to 0 or 1. For large ϵ, the first term of the Cahn–Hilliard energy dominates, and thus the optimization problem is convex.

THEOREM 3.1. *Let $W \in C^2([0,1])$. Then there exists $\epsilon_0 > 0$ dependent on Ω only, such that the objective functional $\rho \mapsto M(\rho) = \gamma \int_\Omega \rho \, d\mathbf{x} + P^\epsilon(\rho)$ is convex for all $\epsilon > \epsilon_0$.*

Proof. The objective functional is twice continuously differentiable with derivatives

$$M'(\rho)\psi = \gamma \int_\Omega \psi \, \mathrm{d}\mathbf{x} + \epsilon \int_\Omega \nabla\rho \cdot \nabla\psi \, \mathrm{d}\mathbf{x} + \frac{1}{\epsilon} \int_\Omega W'(\rho)\psi \, \mathrm{d}\mathbf{x}$$

and

$$M''(\rho)(\psi_1, \psi_2) = \epsilon \int_\Omega \nabla\psi_1 \cdot \nabla\psi_2 \, \mathrm{d}\mathbf{x} + \frac{1}{\epsilon} \int_\Omega W''(\rho)\psi_1\psi_2 \, \mathrm{d}\mathbf{x}.$$

Due the constraint $0 \le \rho \le 1$, we obtain $|W''(\rho)| \le W_0$, where $W_0 \in \mathbb{R}$ is the maximum of W'' in the interval $[0,1]$. Moreover, since $\rho = 1$ on Γ_N, admissible variations satisfy $\psi = 0$ on Γ_N, and due to a Friedrich-type inequality, there exists a constant $C_P > 0$ such that

$$\int_\Omega \psi^2 \, \mathrm{d}\mathbf{x} \le C_P \int_\Omega |\nabla\psi|^2 \, \mathrm{d}\mathbf{x}$$

for all admissible variations $\psi \in H^1(\Omega)$ with $\psi = 0$ on Γ_N. Hence,

$$M''(\rho)(\psi, \psi) \ge \left(\epsilon - \frac{W_0 C_P}{\epsilon}\right) \int_\Omega |\nabla\psi|^2 \, \mathrm{d}\mathbf{x}.$$

Consequently, for $\epsilon \ge \epsilon_0 := \sqrt{W_0 C_P}$, M is convex. Since all constraints are linear, the relaxed optimization problem (3.1) is convex. ∎

So far, we have not discussed possible choices for the function W. Commonly used in phase-field simulations of phase-transition problems (e.g., in the Allen–Cahn and Cahn–Hilliard equation; cf., e.g., [4, 13]) is the *double-well potential*

$$W(r) = r^2(1-r)^2, \qquad r \in \mathbb{R}.$$

Recently, the so-called *double-obstacle potential*

$$W(r) = r(1-r), \qquad r \in [0,1], \tag{3.3}$$

has received further attention (see [6]). In the case of evolutions such as the Allen–Cahn equation, the use of the double-obstacle potential is rather a computational complication, since $0 \leq \rho \leq 1$ has to be enforced (in contrast to the evolution with the double-well potential), and the PDE has to be reformulated as a variational inequality. In our case, the choice of the double-obstacle potential (3.3) seems more attractive, since we enforce the bound constraints on ρ anyway (within our system of inequality constraints), and the double-obstacle problem causes a polynomial nonlinearity of lower degree than the double-well potential. Using (3.3) in P^ϵ, we observe that (3.1) and (3.2) are quadratic optimization problems subject to linear constraints.

3.2. Existence of solutions. We now investigate the existence of solutions of the relaxed problem (3.1). For this sake we introduce the Banach spaces of functions with essentially bounded strain

$$BS^\infty(\Omega) := \{\mathbf{u} \in L^\infty(\Omega;\mathbb{R}^d) \mid \mathbf{e}(\mathbf{u}) \in L^\infty(\Omega;\mathbb{R}^{d\times d})\}$$

and square-integrable strain

$$BS^2(\Omega) := \{\mathbf{u} \in L^2(\Omega;\mathbb{R}^d) \mid \mathbf{e}(\mathbf{u}) \in L^2(\Omega;\mathbb{R}^{d\times d})\}$$

with norms

$$\|\mathbf{u}\|_{BS^\infty} := \max\{\|\mathbf{u}\|_\infty, \|\mathbf{e}(\mathbf{u})\|_\infty\}$$

and

$$\|\mathbf{u}\|_{BS^2} := \sqrt{\|\mathbf{u}\|_2^2 + \|\mathbf{e}(\mathbf{u})\|_2^2}.$$

One can verify by standard arguments that $BS^\infty(\Omega)$ is a Banach space (including all elements of the Sobolev space $W^{1,\infty}(\Omega;\mathbb{R}^d)$) and that $BS^2(\Omega)$ is a Hilbert space with scalar product

$$\langle \mathbf{u}, \mathbf{v}\rangle_{BS^2} := \langle \mathbf{u}, \mathbf{v}\rangle_{L^2} + \langle \mathbf{e}(\mathbf{u}), \mathbf{e}(\mathbf{v})\rangle_{L^2}.$$

As usual for weak solutions of PDEs, we understand the equality constraints on $\mathbf{s}$ in a standard weak sense, i.e.,

$$\int_\Omega \mathbf{s} : \nabla\Psi \, \mathrm{d}\mathbf{x} = \int_{\Gamma_N} \mathbf{t}\cdot\Psi \, \mathrm{d}a \qquad \forall\ \Psi \in W^{1,2}(\Omega;\mathbb{R}^d), \Psi|_{\Gamma_D} = 0.$$

Similarly, we interpret the stress-strain relation in an L^2-sense, i.e.,

$$\int_\Omega \left[\sigma : \Psi - \Psi : \mathbf{C} : \mathbf{e}(\mathbf{u})\right] \, \mathrm{d}\mathbf{x} = 0 \qquad \forall\ \Psi \in L^2(\Omega;\mathbb{R}^{d\times d}).$$

We start the analysis with a lower semicontinuity property.

LEMMA 3.1. *Let W be defined by* (3.3). *Then the functional $M : H^1(\Omega) \to \mathbb{R}$ is sequentially weakly lower semicontinuous.*

Proof. Due to the compact embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$, the linear functional $\rho \mapsto \gamma \int_\Omega \rho \, \mathrm{d}\mathbf{x}$ and the quadratic functional $\rho \mapsto \frac{1}{\epsilon}\int_\Omega W(\rho) \, \mathrm{d}\mathbf{x}$ are weakly continuous.

Together with the sequential weak lower semicontinuity of the square of the norm in Hilbert spaces applied to the third term in M, we obtain the assertion. ∎

Besides lower semicontinuity, a fundamental ingredient for the existence of solutions is compactness in appropriate topologies. In order to obtain some weak compactness, we examine the boundedness of the constraint set.

LEMMA 3.2. *Let $\epsilon > 0$, and let*

$$(\rho, \mathbf{u}, \sigma, \mathbf{s}) \in L^\infty(\Omega) \times BS^\infty(\Omega) \times L^\infty(\Omega; \mathbb{R}^{d\times d})^2$$

satisfy the constraints in (3.1). *Then $(\rho, \mathbf{u}, \sigma, \mathbf{s})$ lies in a bounded set with respect to the corresponding norms.*

Proof. From the bound constraints $0 \le \rho \le 1$ we immediately conclude that ρ lies in the unit ball of $L^\infty(\Omega)$. Consequently, we deduce that

$$\min\left\{\mathbf{0}, \sigma^{\min}\right\} \le \mathbf{s} \le \max\left\{\sigma^{\max}, \mathbf{0}\right\},$$

and hence $\mathbf{s}$ is bounded in the norm of $L^\infty(\Omega; \mathbb{R}^{d\times d})$. Due to

$$\mathbf{s} - \frac{1-\rho}{\beta}\mathbf{1} \le \sigma \le \mathbf{s} + \frac{1-\rho}{\beta}\mathbf{1}$$

we further conclude the boundedness of σ in the norm of $L^\infty(\Omega; \mathbb{R}^{d\times d})$. Finally, the bound constraints on $\mathbf{u}$ imply its boundedness in the norm of $L^\infty(\Omega; \mathbb{R}^d)$, and together with the stress-strain relation and the positive definiteness we may conclude the boundedness of $\mathbf{u}$ in the norm of $BS^\infty(\Omega)$. ∎

With these preliminary results we can provide an existence result for the relaxed topology optimization problem for arbitrary positive ϵ.

THEOREM 3.2. *Let $\epsilon > 0$, $\beta > 0$, and let W be defined by* (3.3). *Moreover, let the admissible set defined by the constraints in* (3.1) *be nonempty. Then there exists a solution*

$$(\rho, \mathbf{u}, \sigma, \mathbf{s}) \in (H^1(\Omega) \cap L^\infty(\Omega)) \times BS^\infty(\Omega) \times L^\infty(\Omega; \mathbb{R}^{d\times d})^2$$

of the constrained optimization problem (3.1).

Proof. For admissible densities $\rho \ge 0$, the objective functional M is bounded below by zero, and hence the infimum m_0 of M on the admissible set is finite. Hence, we can find a minimizing sequence

$$(\rho^n, \mathbf{u}^n, \sigma^n, \mathbf{s}^n) \in (H^1(\Omega) \cap L^\infty(\Omega)) \times BS^\infty(\Omega) \times L^\infty(\Omega; \mathbb{R}^{d\times d})^2$$

such that $M(\rho^n) \to m_0$. Since $M(\rho^n)$ converges, the sequence is bounded in particular, and since

$$\frac{2}{\epsilon} M(\rho^n) \ge \int_\Omega |\nabla \rho^n|^2 \mathrm{d}\mathbf{x},$$

we obtain boundedness of ρ^n in $H^1(\Omega)$. Due to Lemma 3.2 and standard precompactness results for bounded sets in weak or weak-* topologies, we can extract a subsequence (again denoted by the superscript n) such that

$$\begin{aligned}
\rho^n &\to \hat{\rho} \text{ weak in } H^1(\Omega) \text{ and weak-* in } L^\infty(\Omega),\\
\mathbf{u}^n &\to \hat{\mathbf{u}} \text{ weak in } BS^2(\Omega) \text{ and weak-* in } L^\infty(\Omega; \mathbb{R}^d),\\
\sigma^n &\to \hat{\sigma} \text{ weak-* in } L^\infty(\Omega; \mathbb{R}^{d\times d}),\\
\mathbf{s}^n &\to \hat{\mathbf{s}} \text{ weak-* in } L^\infty(\Omega; \mathbb{R}^{d\times d}).
\end{aligned}$$

Due to closedness of simple bounds with respect to weak-* convergence in L^∞, we can conclude that the limit $(\hat\rho, \hat{\mathbf{u}}, \hat\sigma, \hat{\mathbf{s}})$ satisfies all the inequalities in (3.1). Moreover, since for $\Psi \in W^{1,2}(\Omega; \mathbb{R}^d)$ we have in particular $\nabla\Psi \in L^1(\Omega)$, we may conclude that

$$\int_{\Gamma_N} \mathbf{t} \cdot \Psi \, \mathrm{d}a = \int_\Omega \mathbf{s}^n : \nabla\Psi \, \mathrm{d}\mathbf{x} \to \int_\Omega \mathbf{s} : \nabla\Psi \, \mathrm{d}\mathbf{x}$$

due to weak-* convergence in L^∞. Hence, $\hat{\mathbf{s}}$ satisfies the associated equality constraints. From the weak convergence of $\mathbf{u}^n$ in $BS^2(\Omega)$ we conclude that $\hat{\mathbf{u}}$ satisfies the stress-strain relation, and hence $(\hat\rho, \hat{\mathbf{u}}, \hat\sigma, \hat{\mathbf{s}})$ is in the admissible set. With the sequential lower semicontinuity from Lemma 3.1 we finally obtain that, thus, $(\hat\rho, \hat{\mathbf{u}}, \hat\sigma, \hat{\mathbf{s}})$ is a solution of the optimization problem (3.1). ∎

4. Discretization. In the following we consider the discretization of the relaxed problems for $\Omega \subset \mathbb{R}^2$, detailing the analysis again for the case of (3.1). For simplicity (and motivated by the typical choices of design domains), we assume that Ω is of polygonal shape.

Our aim is to construct a finite element approximation on a triangular grid; i.e., we decompose $\Omega = \bigcup_{T\in\mathcal{T}^h} \overline{T}$ for a suitable family $\mathcal{T}^h$ of triangles satisfying standard regularity conditions (see [11]). The parameter $h > 0$ denotes the grid size (equal to the maximal diameter of triangles in $\mathcal{T}^h$). We shall use two different discrete subspaces, namely the H^1-subspace of linear elements (for the density ρ and displacement components $\mathbf{u}_i$)

$$\mathcal{V}^h := \{\varphi \in C(\Omega) \mid \varphi \text{ is affinely linear in } T \;\forall\, T \in \mathcal{T}^h\}$$

and the L^∞-subspace of constant elements (for the stress components σ_{ij} and $\mathbf{s}_{ij}$)

$$\mathcal{W}^h := \{\varphi \in L^\infty(\Omega) \mid \varphi \text{ is constant in } T \;\forall\, T \in \mathcal{T}^h\}.$$

Note that for $\mathbf{u} \in \mathcal{V}^h \times \mathcal{V}^h \subset BS^\infty(\Omega)$, we obtain $\frac{\partial \mathbf{u}_i}{\partial \mathbf{x}_j} \in \mathcal{W}^h$.

The equality constraints are discretized by a standard finite element approach; i.e., we look for $\mathbf{s} \in (\mathcal{W}^h)^{2\times 2}$, $\sigma \in (\mathcal{W}^h)^{2\times 2}$, and $\mathbf{u} \in \mathcal{V}^h \times \mathcal{V}^h$ satisfying

$$\int_\Omega \mathbf{s} : \nabla\Psi \, \mathrm{d}\mathbf{x} = \int_{\Gamma_N} \mathbf{t} \cdot \Psi \, \mathrm{d}a \qquad \forall\, \Psi \in \mathcal{V}^h \times \mathcal{V}^h, \Psi|_{\Gamma_D} = 0$$

and

$$\int_\Omega \left[\sigma : \Psi - \Psi : \mathbf{C} : \mathbf{e}(\mathbf{u})\right] \, \mathrm{d}\mathbf{x} = 0 \qquad \forall\, \Psi \in (\mathcal{W}^h)^{2\times 2}.$$

The bound constraints on the displacement and density $\rho \in \mathcal{V}^h$ can be enforced directly; for piecewise linear functions, the constraints hold if and only if they hold in all nodes of the grid.

Finally, we need to discretize the inequality constraints involving both stress variables and the density. Since the components of σ and $\mathbf{s}$ are in a different subspace than ρ, the discretization is not straightforward. In particular, we cannot pose local constraints in the grid nodes or on edges, since functions in $\mathcal{W}^h$ are discontinuous over the edges. Consequently, the more promising approach is to interpret the inequality constraints as constraints in $\mathcal{W}^h$. For this sake we introduce the discrete projection operator (with respect to the L^2-norm) $\mathcal{P}^h : \mathcal{V}^h \to \mathcal{W}^h$,

$$(\mathcal{P}^h v)|_T := \frac{1}{|T|} \int_T v \, \mathrm{d}\mathbf{x} \qquad \forall\, T \in \mathcal{T}^h, \forall\, v \in \mathcal{V}^h.$$

The discretized formulation of constraints in $\mathcal{W}^h$ can then be formulated as

$$\begin{aligned} -(1-\mathcal{P}^h\rho)\mathbf{1} \le \beta(\sigma - \mathbf{s}) \le (1-\mathcal{P}^h\rho)\mathbf{1} \qquad &\text{in } \Omega, \\ \sigma^{\min}\mathcal{P}^h\rho \le \mathbf{s} \le \sigma^{\max}\mathcal{P}^h\rho \qquad &\text{in } \Omega. \end{aligned}$$

Let the coefficients $\mathbf{R} \in \mathbb{R}^N$, $\mathbf{U} \in \mathbb{R}^{2N}$, $\mathbf{S} \in \mathbb{R}^{3M}$, and $\mathbf{\Sigma} \in \mathbb{R}^{3M}$ contain the coefficients R_j, U_j^i, S_k^{ij}, and Σ_k^{ij} (using the symmetry $S_k^{ij} = S_k^{ji}$ and $\Sigma_k^{ij} = \Sigma_k^{ji}$) of the finite element functions of ρ, $\mathbf{u}$, $\mathbf{s}$, and σ, respectively. The discretized problem can now be written equivalently as a quadratic programming problem with linear constraints for the unknown $(\mathbf{R}, \mathbf{U}, \mathbf{\Sigma}, \mathbf{S}) \in \mathbb{R}^{3N+6M}$:

$$\begin{aligned} \gamma\mathbf{E}^T\mathbf{R} + \frac{\epsilon}{2}\mathbf{R}^T\mathbf{K}\mathbf{R} + \frac{1}{\epsilon}\mathbf{R}^T(\mathbf{E} - \mathbf{M}\mathbf{R}) \quad &\to \quad \min, \\ \left(\mathbf{I} - \mathbf{L}_1^T\mathbf{L}_1\right)\mathbf{B}^T\mathbf{S} &= \mathbf{T}, \\ \mathbf{D}\mathbf{\Sigma} - \mathbf{C}\mathbf{B}\mathbf{U} &= \mathbf{0}, \\ \mathbf{L}_1\mathbf{U} &= \mathbf{0}, \\ \mathbf{L}_2\mathbf{R} &= \mathbf{1}, \\ -\mathbf{Q}(\mathbf{1} - \mathbf{P}\mathbf{R}) \le \beta(\Sigma - \mathbf{S}) &\le \mathbf{Q}(\mathbf{1} - \mathbf{P}\mathbf{R}), \\ \boldsymbol{\sigma}^{\min}\mathbf{Q}\mathbf{P}\mathbf{R} \le \mathbf{S} &\le \boldsymbol{\sigma}^{\max}\mathbf{Q}\mathbf{P}\mathbf{R}, \\ \mathbf{0} \le \mathbf{R} &\le \mathbf{1}, \\ \mathbf{u}^{\min} \le \mathbf{U} &\le \mathbf{u}^{\max}. \end{aligned} \tag{4.1}$$

In the objective functional, $\mathbf{E} \in \mathbb{R}^N$ is a vector representing the coefficients of the constant function 1 with respect to the basis functions ϕ_j. $\mathbf{K} \in \mathbb{R}^{N\times N}$ is a stiffness matrix arising from the finite element discretization of the negative Laplacian in $\mathcal{V}^h$, and $\mathbf{M} \in \mathbb{R}^{N\times N}$ is a mass matrix for the identity in $\mathcal{V}^h$. In the discretized formulation of the constraints, the matrix $\mathbf{B}^T \in \mathbb{R}^{3M\times 2N}$ is the discretization of the divergence operator (restricted to symmetric stress tensors), $\mathbf{C} \in \mathbb{R}^{3M\times 3M}$ represents the stress-strain relation, $\mathbf{D}$ is a (diagonal) mass matrix for the identity in $\mathcal{W}^h$, and $\mathbf{T} \in \mathbb{R}^{2N}$ is a discrete representation of the traction force. The matrices $\mathbf{L}_1 \in \mathbb{R}^{2N_1\times 2N}$ and $\mathbf{L}_2 \in \mathbb{R}^{N_2\times N}$ with entries 0 or 1 realize the boundary conditions, where N_1 is the number of node points on Γ_D and N_2 is the number of node points on Γ_N. $\boldsymbol{\sigma}^{\min}$ and $\boldsymbol{\sigma}^{\max}$ are diagonal matrices, representing the corresponding entries of $\sigma^{\min}$ and $\sigma^{\max}$. Finally, $\mathbf{Q} \in \mathbb{R}^{3M\times M}$ is an extension matrix, and $\mathbf{P} \in \mathbb{R}^{M\times N}$ is the matrix representation of the projection operator $\mathcal{P}^h$.

The above reasoning shows that after discretization we end up with a linearly constrained quadratic programming problem for the variable

$$\mathbf{X} = (\mathbf{R}, \mathbf{U}, \mathbf{\Sigma}, \mathbf{S}) \in \mathbb{R}^{3N+6M}$$

with $2N + 3M + 2N_1 + N_2$ equality, $12M$ inequality constraints, and $6N$ bound constraints. Note that $2N_1$ equalities corresponding to the divergence constraint for nodal points in Γ_D are actually in the form $0 = 0$ and can be eliminated. In addition, we can eliminate the components of $\mathbf{U}$ corresponding to nodal points on Γ_D and the corresponding bound constraints (we have to assume $\mathbf{u}^{\min} \le \mathbf{0} \le \mathbf{u}^{\max}$ in order to obtain feasible points anyway). Moreover, we can eliminate the components of $\mathbf{R}$ corresponding to nodal values on Γ_N and the corresponding bound constraints. Since all values of $\mathbf{\Sigma}$ are determined by the corresponding equality constraints, we remove these constraints and replace $\mathbf{\Sigma}$ by $\mathbf{C}\mathbf{B}\mathbf{U}$ in the inequality constraints. We consequently end up with a smaller programming problem with $3N + 3M - 2N_1 - N_2$ unknowns, $2N - 2N_1$ equality constraints, $12M$ inequality constraints, and $6N - 4N_1 - 2N_2$ bound constraints.

The existence of solutions for the reduced programming problem can be verified in an analogous way to the infinite-dimensional situation under the assumption that there exists a feasible point.

4.1. Constraint qualification. In order to obtain constraint qualification for the linear constraints, it suffices to show that the equality constraints are linearly independent and that there exists a feasible point satisfying all inequalities strictly. Note again that the feasible set does not depend on the relaxation parameter ϵ, and hence the constraint qualification is always uniform with respect to the relaxation.

The linear independence of the equality constraints can be verified by standard reasoning for finite element discretizations, and we therefore turn our attention to the inequality constraints. In order to verify constraint qualification, we shall use a natural assumption, namely, that the stress and displacement obtained from a design domain completely filled with material satisfy the displacement and stress constraints strictly. This assumption is natural, since one expects the maximal stress and displacement to increase for decreasing mass. Thus, if the constraints are active at maximal mass already, it is quite unlikely to find an optimal design with lower mass anyway, or, in other words, the constraints are too severe to compute a different optimum. In mathematical terms, the assumption can be formulated as follows: Let $\mathbf{R}^1 = \mathbf{1}$, and let $\mathbf{U}^1$ and $\mathbf{S}^1 = \mathbf{\Sigma}^1$ be the solutions of the corresponding elasticity problem:

$$(\mathbf{I} - \mathbf{L}_1^T\mathbf{L}_1)\mathbf{B}^T\mathbf{S}^1 = \mathbf{T}, \qquad \mathbf{D}\mathbf{\Sigma}^1 - \mathbf{C}\mathbf{B}\mathbf{U}^1 = \mathbf{0}, \qquad \mathbf{L}_1\mathbf{U}^1 = \mathbf{0},$$

which can be shown to be uniquely defined from standard finite element theory. Then we assume that

$$\boldsymbol{\sigma}^{\min}\mathbf{Q}\mathbf{P}\mathbf{R}^1 < \mathbf{S}^1 < \boldsymbol{\sigma}^{\max}\mathbf{Q}\mathbf{P}\mathbf{R}^1, \qquad \mathbf{u}^{\min} < \mathbf{U}^1 < \mathbf{u}^{\max}, \tag{4.2}$$

where $<$ means strict inequality for each component. Then we obtain the following result.

THEOREM 4.1. *Let* (4.2) *be satisfied, and let* $\beta > 0$ *be sufficiently small. Then the constraint set in* (4.1) *with the elimination of variables and constraints as explained above satisfies the Slater condition.*

Proof. As noticed above, it suffices to find a feasible point satisfying the constraints strictly. For this sake we choose a density $\mathbf{R}^\eta = \eta\mathbf{1}$ with $0 < \eta < 1$. Then $\|\mathbf{R}^1 - \mathbf{R}^\eta\| = N(1-\eta)$; i.e., the distance to $\mathbf{R}^1$ becomes arbitrarily small as $\eta \to 1$. Because of continuity, we can find $0 < \eta < 1$, and a solution $\mathbf{U}^\eta$, $\mathbf{S}^\eta = \eta\Sigma^\eta$ of

$$(\mathbf{I} - \mathbf{L}_1^T\mathbf{L}_1)\mathbf{B}^T\mathbf{S}^\eta = \mathbf{T}, \qquad \mathbf{D}\mathbf{\Sigma}^\eta - \mathbf{C}\mathbf{B}\mathbf{U}^\eta = \mathbf{0}, \qquad \mathbf{L}_1\mathbf{U}^\eta = \mathbf{0},$$

such that

$$\boldsymbol{\sigma}^{\min}\mathbf{Q}\mathbf{P}\mathbf{R}^\eta < \mathbf{S}^\eta < \boldsymbol{\sigma}^{\max}\mathbf{Q}\mathbf{P}\mathbf{R}^\eta, \qquad \mathbf{u}^{\min} < \mathbf{U}^\eta < \mathbf{u}^{\max}$$

holds. Moreover, we have

$$\mathbf{0} < \mathbf{R}^\eta = \eta\mathbf{1} < \mathbf{1}$$

and

$$\begin{aligned} -\mathbf{Q}(1-\eta)\mathbf{1} &= -\mathbf{Q}(\mathbf{1} - \mathbf{P}\mathbf{R}^\eta) < \beta(\mathbf{\Sigma}^\eta - \mathbf{S}^\eta) \\ &= \beta(1-\eta)\mathbf{\Sigma}^\eta < \mathbf{Q}(\mathbf{1} - \mathbf{P}\mathbf{R}^\eta) = \mathbf{Q}(1-\eta)\mathbf{1}, \end{aligned}$$

provided β is sufficiently small. Hence, all (reduced) inequality constraints are satisfied strictly by $(\mathbf{R}^\eta, \mathbf{U}^\eta, \mathbf{S}^\eta, \Sigma^\eta)$, which implies the assertion. ∎

5. Solution of the discretized problem.

5.1. Continuation in ϵ. As motivated in the introduction and in section 3.1 the problem (4.1) will be solved for a decreasing sequence of ϵ. As $\epsilon \to 0$, in analogy to Γ-convergence of the perimeter functional, we expect convergence of the sequence of the minimum solutions to a final solution. Moreover, since the double obstacle term is of leading order in ϵ, such a final solution will have a sharp interface between material ($\{\rho{=}1\}$) and void ($\{\rho{=}0\}$).

To achieve this we will use a continuation method such that we choose a decreasing sequence $\{\epsilon^l\}$ with $\epsilon^l \to 0$ for $l = 0, \ldots, L$, where L describes the total number of continuation levels. The corresponding optimization problems are then solved by an interior-point method, as described in the next section. Between the levels ϵ can be reduced, e.g., like $\epsilon^{l+1} := \delta\epsilon^l$ with $0 < \delta < 1$ or like $\epsilon^{l+1} := (\epsilon^0)^l$ if $0 < \epsilon^0 < 1$. If we decrease ϵ too slowly, we may expect from theory and observations from numerical tests that the final solution is not changed, but we end up with a possibly unnecessary high number of levels L. On the other hand, if ϵ is decreased too quickly, the optimization process might get stuck in some undesired local minimum, since the objective functional $M(\rho)$ is turned from convex to concave too quickly.

5.2. Adaption of the problem to IPOPT. We solve the problem (4.1) with IPOPT, which is a freely available optimization code realizing a primal-dual interior-point method. IPOPT, implemented by A. Wächter and L. T. Biegler, is able to solve problems of the following form:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \\ \text{s.t.} \quad & c_{\mathcal{E}}(\mathbf{x}) = \mathbf{0}, \\ & \mathbf{x}^{\min} \le \mathbf{x} \le \mathbf{x}^{\max}. \end{aligned}$$

More information about the implementation of IPOPT can be found in Wächter and Biegler [32]. Tailoring the problem to a black-box optimizer is not the most efficient approach for the numerical solution. However, the aim of this paper is rather to develop and test the general relaxation approach instead of constructing efficient optimization methods.

General nonlinear programming problems with inequality constraints can be written in the above framework using slack variables. So we reformulate (4.1) in the above form by introducing some vector $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_4) \in \mathbb{R}^{12M}$ of slack variables, leading to

$$\begin{aligned} \gamma\mathbf{E}^T\mathbf{R} + \frac{\epsilon}{2}\mathbf{R}^T\mathbf{K}\mathbf{R} + \frac{1}{\epsilon}\mathbf{R}^T(\mathbf{E} - \mathbf{M}\mathbf{R}) \quad &\to \quad \min_{(\mathbf{R},\mathbf{U},\mathbf{S},\mathbf{Z})\in\mathbb{R}^{3N+15M}}, \\ \left(\mathbf{I} - \mathbf{L}_1^T\mathbf{L}_1\right)\mathbf{B}^T\mathbf{S} - \mathbf{T} &= \mathbf{0}, \\ -\mathbf{Q}(\mathbf{1} - \mathbf{P}\mathbf{R}) - \beta(\mathbf{C}\mathbf{B}\mathbf{U} - \mathbf{S}) + \mathbf{Z}_1 &= \mathbf{0}, \\ \beta(\mathbf{C}\mathbf{B}\mathbf{U} - \mathbf{S}) - \mathbf{Q}(\mathbf{1} - \mathbf{P}\mathbf{R}) + \mathbf{Z}_2 &= \mathbf{0}, \\ \boldsymbol{\sigma}^{\min}\mathbf{Q}\mathbf{P}\mathbf{R} - \mathbf{S} + \mathbf{Z}_3 &= \mathbf{0}, \\ \mathbf{S} - \boldsymbol{\sigma}^{\max}\mathbf{Q}\mathbf{P}\mathbf{R} + \mathbf{Z}_4 &= \mathbf{0}, \\ \mathbf{0} \le \mathbf{R} &\le \mathbf{1}, \\ \mathbf{u}^{\min} \le \mathbf{U} &\le \mathbf{u}^{\max}, \\ \mathbf{0} &\le \mathbf{Z}. \end{aligned} \tag{5.1}$$

Finally we solve a programming problem with $3N + 15M - 2N_1 - N_2$ unknowns, $2N + 12M - 2N_1 - N_2$ equality constraints, and $6N + 12M - 4N_1 - 2N_2$ bound constraints.

A similar discrete programming problem for the problem with the conservative von Mises stress constraints (3.2) can be deduced in an analogous way.

6. Numerical results. For the numerical examples we have chosen two simple examples where the global optimal designs are known on very coarse grids for maximum safety factor problems (see [30]); nevertheless, the topological structure of the final beams for the minimal mass problems are similar. Since we solve perimeter constrained optimization problems, we can expect the results to be independent of the mesh size, which was indeed supported by our numerical tests. For the sake of simplicity Young's modulus and Poisson's ratio of the given material are $E = 1N/m^2$ and $\nu = 0.3$. We used the plain strain model for the computations, and also, for simplicity, all structures have a unit thickness of $1m$ and are loaded with half of the unit load. Reasonable bounds for the displacements $\mathbf{u}$ and the stresses σ are provided by the unique solutions $\overline{\mathbf{u}}$ and $\overline{\sigma}$ of the corresponding elasticity problem when the whole design domain Ω is filled with material $\rho = 1$. Then the displacement bounds are, e.g., set to

$$u_i^{\max} = -u_i^{\min} = 2\max\{|\overline{u}_i(\mathbf{x})| \ : \ \mathbf{x} \in \Omega\}, \qquad i = 1, 2,$$

and the stress bounds to

$$\begin{aligned} \sigma_{11}^{\max} = \sigma_{22}^{\max} &= \max\{|\overline{\sigma}_{11}(\mathbf{x})|, |\overline{\sigma}_{22}(\mathbf{x})| \ : \ \mathbf{x} \in \Omega\}, \\ \sigma_{12}^{\max} = \sigma_{21}^{\max} &= \max\{|\overline{\sigma}_{12}(\mathbf{x})|, |\overline{\sigma}_{21}(\mathbf{x})| \ : \ \mathbf{x} \in \Omega\} \end{aligned}$$

with $\sigma^{\min} = -\sigma^{\max}$. The von Mises stress bound is given by

$$\Phi^{\max} = \max\{\Phi(\sigma(\mathbf{x})) \ : \ \mathbf{x} \in \Omega\}.$$

All numerical examples are performed on a PC using a 2.4 GHz Intel CPU and 2 GB memory. For the mesh generation and the finite element part of the computations the software package NETGEN/NGSolve was used. The optimization part was done using the interior-point code IPOPT. As IPOPT is used as a "black-box," we did not adjust its linear solver and its stopping criterion to our needs. So we stop the optimization process per continuation level if the stopping criterion of IPOPT is fulfilled with a tolerance of 10^{-5} or a maximum number of iterations is reached. For an approximation of the Hessian of the Lagrange functional a BFGS routine is used. In both examples the scaling parameter γ of the mass term in the objective is set to $\gamma = 2$, and we start the ϵ-continuation for $L = 5$ levels with $\epsilon^0 = 1$. Between the levels ϵ is reduced like $\epsilon^{l+1} = 0.25\epsilon^l$.

We want to point out that these numerical examples just show the potential of this solution approach. Since there has been no emphasis on efficiency so far, the computational times are far from being optimal. To decrease ϵ in a more sophisticated way (e.g., in combination with the continuation of μ, the barrier parameter of the interior-point formulation), a proper stopping criterion and a linear solver with optimal complexity for the IPOPT package would decrease the runtimes significantly. So far most of the CPU time is spent in solving linear systems in IPOPT.

6.1. A short beam example. As a first example we treat the problem shown in Figure 6.1. There the load condition, bearings, and geometry are illustrated with a design domain of dimension $2m \times 1m$. Here we consider stress constraints with

TABLE 6.1
Computational data from the short beam example.

Level	Obj. evaluations	Time (h)	Iterations	ϵ
0	629	3.48	150	1
1	1350	6.00	225	0.25
2	1208	7.26	300	0.0625
3	307	3.36	300	0.015625
4	309	3.62	300	0.003906

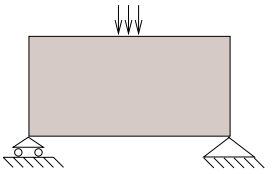


FIG. 6.1. *A short beam example.*

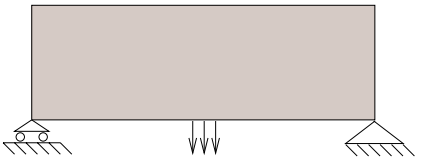


FIG. 6.2. *A long beam example.*

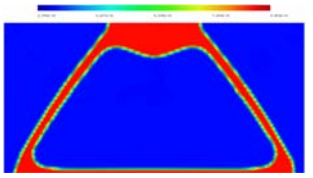


FIG. 6.3. *Optimal material distribution.*

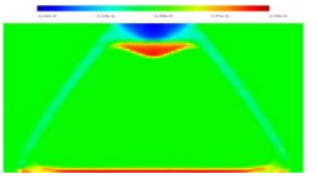


FIG. 6.4. σ_{11} *of the optimal design.*

respect to local stresses. The corresponding bound constraints are

$$\mathbf{u}^{\max} = \begin{pmatrix} 0.525 \\ 0.525 \end{pmatrix} m, \qquad \sigma^{\max} = \begin{pmatrix} \sigma_{11}^{\max} \\ \sigma_{22}^{\max} \\ \sigma_{12}^{\max} \end{pmatrix} = \begin{pmatrix} 0.65 \\ 0.65 \\ 0.262 \end{pmatrix} \frac{N}{m^2}$$

with $\mathbf{u}^{\min} = -\mathbf{u}^{\min}$ and $\sigma^{\min} = -\sigma^{\max}$. A mesh with 14182 elements is used for the optimization process, so we finally end up with 234532 unknowns. In more details we have 7260, 14182, 42546, and 170184 DOFs for the density, displacements, stresses, and slacks, respectively. The total number of equality constraints is 184725 and 191986 for the bound constraints. The overall computational time for the 5 levels is about 23.5 hours, and the volume of the final optimal design is 0.356 m^3 (17.8% of $|\Omega|$). We can see some more detailed information about the short beam computation in Table 6.1. The optimal design and the σ_{11} stress distribution we show in Figures 6.3 and 6.4.

6.2. A long beam example. For the second example we consider the load condition, bearings, and geometry shown in Figure 6.2, where the dimensions of the design domain are $3m \times 1m$. Here we choose to calculate an optimal design with respect to bounded conservative von Mises stress, and again we list the corresponding bound constraints:

$$\mathbf{u}^{\max} = \begin{pmatrix} 0.99 \\ 0.99 \end{pmatrix} m, \qquad \Phi^{\max} = 0.78 \frac{N}{m^2}$$

with $\mathbf{u}^{\min} = -\mathbf{u}^{\min}$. For the discretization and the optimization process we use a mesh with 17291 elements, which results in 355098 unknowns and 8849, 17698,

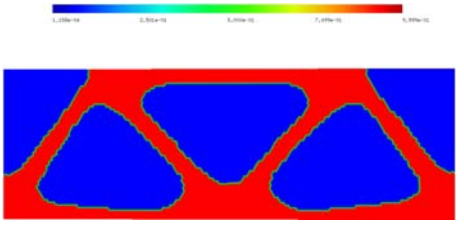


FIG. 6.5. *Optimal material distribution.*

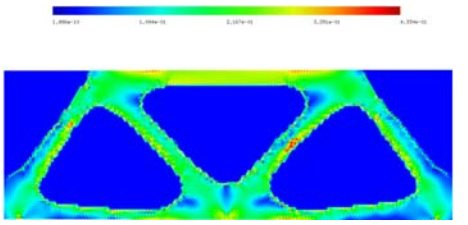
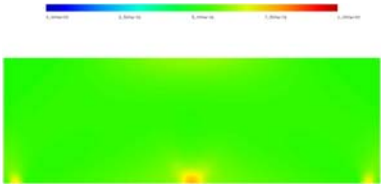
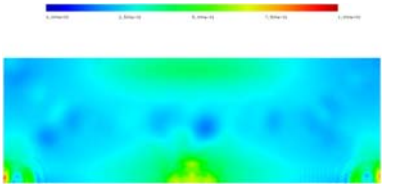


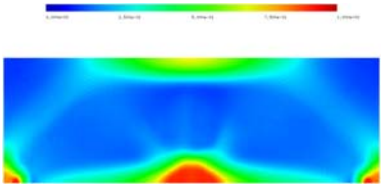
FIG. 6.6. *von Mises stress distribution.*



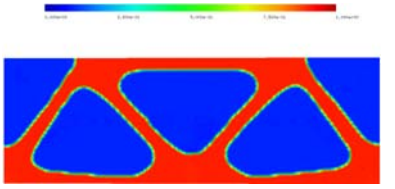
Level 0, $\epsilon = 1$



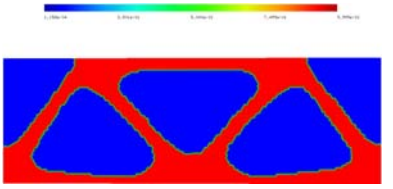
Level 1, $\epsilon = 0.25$.



Level 2, $\epsilon = 0.0625$



Level 3, $\epsilon = 0.015625$.



Level 4, $\epsilon = 0.003906$.

FIG. 6.7. *ϵ-continuation over 5 levels.*

51873, 51873, and 224784 DOFs for the density, displacements, stresses, conservative von Mises approximation, and slacks, respectively. Here we end up with a total number of 242503 equality constraints and 251352 bound constraints. In Figure 6.5 we see the optimal design of the problem and in Figure 6.6 the corresponding von Mises stress distribution. Moreover, we show in Figure 6.7 a sequence of optimal designs. For solving 5 levels it takes about 26.5 hours, and the volume is reduced to 1.1 m^3 (36.73% of $|\Omega|$). More details about the ϵ-continuation are given in Table 6.2.

TABLE 6.2
Computational data from the long beam example.

Level	Obj. evaluations	Time (h)	Iterations	ϵ
0	769	3.72	150	1
1	1294	6.41	225	0.25
2	1254	9.65	300	0.0625
3	302	5.23	300	0.015625
4	88	1.51	59	0.003906

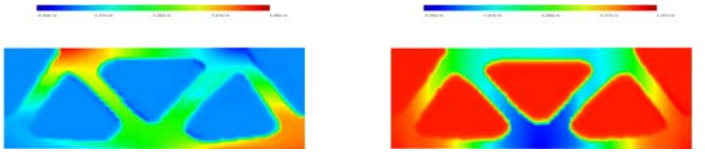


FIG. 6.8. *x- and y-displacements of the optimal design.*

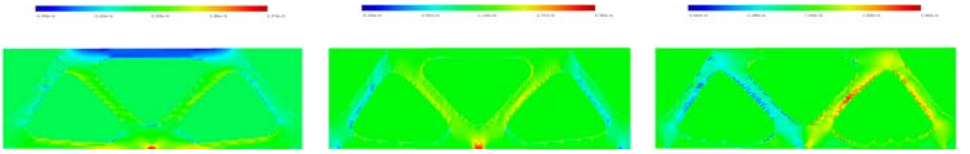


FIG. 6.9. *σ_{11}, σ_{22}, and σ_{12} distribution of the optimal design.*

Finally, in Figures 6.8 and 6.9 we show the displacement and stress distribution of the optimal design. Here we can see that the corresponding displacements and stresses are only nonzero over the parts of the domain occupied with material.

7. Conclusions. A new method for solving structural topology optimization problems with local stress constraints has been presented. The reformulation of the problem and the phase-field relaxation leads to a parameter-dependent family of large-scale optimization problems satisfying uniform constraint qualifications. Using parameter continuation it is possible to compute optima in a robust way with reasonable effort (e.g., compared to mixed linear programming techniques).

So far no particular emphasis has been laid on the efficient solution of the discretized problems for very large number of unknowns (as appearing, e.g., in three-dimensional applications), but there is a lot of potential to speed up the solution techniques, which will be investigated in future research.

Acknowledgments. Useful and stimulating discussions as well as various links to literature are acknowledged to Wolfram Mühlhuber (Audi AG, Ingolstadt), Benjamin Hackl (RICAM, Linz), Karl Kunisch (University Graz and RICAM, Linz), Martin Bendsøe (Danish Technical University, Lyngby), Blaise Bourdin (Louisiana State University, Baton Rouge), and Don Goldfarb (Columbia University, New York). The interior-point code IPOPT (see http://www.coin-or.org/Ipopt/) was used with permission from Andreas Wächter, whom we also thank for kind support. Mesh generation and finite element tools were used from the software package NETGEN/NGSolve developed by Joachim Schöberl at the University of Linz (see http://www.hpfem.jku.at).

REFERENCES

[1] G. ALBERTI, *Variational models for phase transitions, an approach via Γ-convergence*, in Calculus of Variations and Partial Differential Equations, L. Ambrosio, N. Dancer, G. Buttazzo, A. Marino, and M. K. V. Murthy, eds., Springer, Berlin, 2000, pp. 95–114.

[2] G. ALLAIRE, *Shape Optimization by the Homogenization Method*, Springer, New York, 2002.

[3] G. ALLAIRE, F. JOUVE, AND A. M. TOADER, *Structural optimization using sensitivity analysis and a level-set method*, J. Comput. Phys., 194 (2004), pp. 363–393.

[4] G. BARLES, H. M. SONER, AND P. E. SOUGANIDIS, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[5] M. P. BENDSØE AND O. SIGMUND, *Topology Optimization*, Springer, Berlin, 2003.

[6] J. F. Blowey and C. M. Elliott, *A phase-field model with a double obstacle potential*, in Motion by Mean Curvature and Related Topics, G. Buttazzo and A. Visintin, eds., de Gruyter, Berlin, 1994, pp. 1–22.
[7] T. Borvall and J. Petersson, *Topology optimization using regularized intermediate density control*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4911–4928.
[8] B. Bourdin, *Filters in topology optimization*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 2143–2158.
[9] B. Bourdin and A. Chambolle, *Design-dependent loads in topology optimization*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 19–48.
[10] A. Braides, Γ-*Convergence for Beginners,* Oxford University Press, Oxford, UK, 2002.
[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods,* 2nd ed., Springer, New York, 2002.
[12] M. Burger and S. Osher, *A survey on level set methods for inverse problems and optimal design*, European J. Appl. Math., 16 (2005), pp. 263–301.
[13] G. Caginalp and E. Socolovsky, *Phase field computations of single-needle crystals, crystal growth, and motion by mean curvature*, SIAM J. Sci. Comput., 15 (1994), pp. 106–126.
[14] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system.* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.
[15] G. Cheng and X. Guo, ϵ-*relaxed approach in structural topology optimization*, Struct. Multidisc. Optim., 13 (1997), pp. 258–266.
[16] G. Dal Maso, *An Introduction to* Γ-*Convergence*, Birkhäuser, Boston, 1993.
[17] P. Duysinx and M. P. Bendsøe, *Topology optimization of continuum structures with local stress constraints*, Internat. J. Numer. Methods Engrg., 43 (1998), pp. 1453–1478.
[18] H. A. Eschenauer and N. Olhoff, *Topology optimization of continuum structures: A review*, Appl. Mech. Rev., 54 (2001), pp. 331–390.
[19] R. B. Haber, C. S. Jog, and M. P. Bendsøe, *A new approach to variable topology shape design using a constraint on perimeter*, Struct. Multidisc. Optim., 11 (1996), pp. 1–12.
[20] B. Maar and V. Schulz, *Interior point multigrid methods for topology optimization*, Struct. Multidisc. Optim., 19 (2000), pp. 214–224.
[21] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal. 98 (1987), pp. 123–142.
[22] L. Modica and S. Mortola, *Un esempio di* Γ-*convergenza*, Boll. Un. Mat. Ital. B(5), 14 (1977), pp. 285–299.
[23] J. Petersson, *Some convergence results in perimeter-controlled topology optimization*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 123–140.
[24] G. I. N. Rozvany, *Difficulties in truss topology optimization with stress, local buckling and system stability constraints*, Struct. Multidisc. Optim., 11 (1996), pp. 213–217.
[25] G. I. N. Rozvany, *On design-dependent constraints and singular topologies*, Struct. Multidisc. Optim., 21 (2001), pp. 164–172.
[26] O. Sigmund and J. Petersson, *Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies, and local minima*, Struct. Multidisc. Optim., 16 (1998), pp. 68–75.
[27] R. Stainko, *An adaptive multilevel approach to minimal compliance topology optimization*, Comm. Numer. Methods Engrg., 22 (2006), pp. 109–118.
[28] M. Stolpe, *Global optimization of minimum weight truss topology problems with stress, displacement, and local buckling constraints using branch-and-bound*, Internat. J. Numer. Methods Engrg., 61 (2004), pp. 1270–1309.
[29] M. Stolpe and K. Svanberg, *On the trajectories of the epsilon-relaxation approach for stress-constrained truss topology optimization*, Struct. Multidisc. Optim., 25 (2003), pp. 62–64.
[30] M. Stolpe and K. Svanberg, *Modelling topology optimization problems as linear mixed* 0–1 *programs*, Internat. J. Numer. Methods Engrg., 57 (2003), pp. 723–739.
[31] K. Svanberg, *The method of moving asymptotes—a new method for structural optimization*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 359–373.
[32] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57.
[33] M. Y. Wang and S. W. Zhou, *Phase field: A variational method for structural topology optimization*, CMES Comput. Model. Eng. Sci., 6 (2004), pp. 547–566.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1467–1507

STABILITY AND STABILIZATION OF MULTIDIMENSIONAL INPUT/OUTPUT SYSTEMS*

ULRICH OBERST†

Abstract. In this paper we discuss stability and stabilization of continuous and discrete multidimensional input/output (IO) behaviors (of dimension r) which are described by linear systems of complex partial differential (resp., difference) equations with constant coefficients, where the signals are taken from various function spaces, in particular from those of polynomial-exponential functions. Stability is defined with respect to a disjoint decomposition of the r-dimensional complex space into a stable and an unstable region, with the standard stable region in the one-dimensional continuous case being the set of complex numbers with negative real part. A rational function is called stable if it has no poles in the unstable region. An IO behavior is called stable if the characteristic variety of its autonomous part has no points in the unstable region. This is equivalent to the stability of its transfer matrix and an additional condition. The system is called stabilizable if there is a compensator IO system such that the output feedback system is well-posed and stable. We characterize stability and stabilizability and construct all stabilizing compensators of a stabilizable IO system (parametrization). The theorems and proofs are new but essentially inspired and influenced by and related to the stabilization theorems concerning multidimensional IO maps as developed, for instance, by Bose, Guiver, Shankar, Sule, Xu, Lin, Ying, Zerz, and Quadrat and, of course, the seminal papers of Vidyasagar, Youla, and others in the one-dimensional case. In contrast to the existing literature, the theorems and proofs of this paper do not need or employ the so-called fractional representation approach, i.e., various matrix fraction descriptions of the transfer matrix, thus avoiding the often lengthy matrix computations and seeming to be of interest even for one-dimensional systems (at least to the author). An important mathematical tool, new in systems theory, is Gabriel's localization theory which, only in the case of ideal-convex (Shankar, Sule) unstable regions, coincides with the usual one. Algorithmic tests for stability, stabilizability, and ideal-convexity, and the algorithmic construction of stabilizing compensators, are addressed but still encounter many difficulties; see in particular the open problems listed by Xu et al.

Key words. stability, stabilization, multidimensional system, behavior, stable transfer matrix

AMS subject classifications. 93D15, 93D25, 93C20, 93C35

DOI. 10.1137/050639004

1. Introduction.

Stabilization theory is a part of control theory and usually involves the following ingredients [7, p. 60].

1. *Stability*: Select the class of admissible systems and define and characterize the stable systems in this class.
2. *Stabilizability*: Determine which admissible systems can be stabilized by output feedback.
3. *Stabilization*: Construct a stabilizing compensator for a given stabilizable system.
4. *Parametrization*: Classify or construct all stabilizing compensators for a given stabilizable system.

In this paper we discuss these problems for continuous and discrete multidimensional input/output (IO) behaviors, which are described by linear systems of complex partial differential equations on $\mathbb{R}^r$ (resp., difference equations on $\mathbb{N}^r$) with constant

*Received by the editors August 26, 2005; accepted for publication (in revised form) May 12, 2006; published electronically October 16, 2006.
http://www.siam.org/journals/sicon/45-4/63900.html
†Institut für Mathematik, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria (Ulrich.Oberst@uibk.ac.at).

coefficients, where the signals are taken from various function spaces—in particular from those of polynomial-exponential functions. Polderman and Willems in [15, sec. 10.8] and Rocha [19] suggest renouncing the IO structure and output feedback in favor of more general behavior interconnections. The first stabilization results for multidimensional systems in this generality are due to Shankar [21], [22].

In contrast to our approach, most multidimensional stabilization papers use a commutative integral domain S of "SISO-stable plants" and describe the admissible systems, often called plants, by a transfer operator or IO map, which is a matrix with coefficients in the quotient field of S; cf., for instance, [26], [6], [23], [25], [31], [32], [7], [8], [16], [17], [18]. This approach to stabilization theory is originally due to Desoer, Kucera, Vidyasagar, Youla, and their coworkers. These systems are called structurally [6] or internally [17] stable if their transfer matrix has entries in S. In their recent paper [29], Wood, Sule, and Rogers treat stability and causality, but not stabilization of continuous multidimensional IO systems.

An IO behavior $\mathcal{B}$ gives rise to its autonomous part $\mathcal{B}^0$, its transfer matrix H, and the largest controllable subbehavior $\mathcal{B}_{\text{cont}}$ which, in turn, has the autonomous part $\mathcal{B}^0_{\text{cont}}$. The entries of the transfer matrix are complex rational functions in r indeterminates s_ρ, i.e., contained in the quotient field $\mathbb{C}(s)$ of the polynomial algebra $A := \mathbb{C}[s] = \mathbb{C}[s_1, \ldots, s_r]$. In general, the transfer matrix does not act on arbitrary inputs as an operator or IO map, and it is an important task to identify those inputs on which it does. Associated with these behaviors is the complex variety $\text{sing}(\mathcal{B})$ of rank singularities and the characteristic varieties $\text{char}(\mathcal{B}^0) \supseteq \text{char}(\mathcal{B}^0_{\text{cont}})$ of the autonomous subbehaviors, the latter coinciding with the variety of poles of H. In the one-dimensional theory, the elements of $\text{char}(\mathcal{B}^0)$ are called the poles, modes, characteristic values, or natural frequencies of the system.

Stability and stabilization of an IO system are defined with respect to a disjoint decomposition $\mathbb{C}^r = \Lambda_1 \uplus \Lambda_2$ of the complex space into a stable region Λ_1 and an unstable region Λ_2, with the standard continuous (resp., discrete) cases being

$$\begin{aligned} &\Lambda_2 = \overline{\mathbb{C}_+}^r,\ \overline{\mathbb{C}_+} := \{z \in \mathbb{C};\ \Re(z) \geq 0\} \text{ or } \Lambda_2 = \overline{\mathbb{C}_+} \times i\mathbb{R}^{r-1}\ ([29]), \\ &\text{resp.,}\ \Lambda_2 = \{z \in \mathbb{C};\ |z| \geq 1\}^r \text{ or } \Lambda_2 = \{z \in \mathbb{C};\ |z| \geq 1\} \times (S^1)^{r-1}. \end{aligned} \tag{1}$$

A rational function is called stable if it has no poles in the unstable region. The ring of all stable rational functions or SISO-stable plants is the quotient ring $A_T \subseteq \mathbb{C}(s)$ with $T := \{t \in A;\ \forall \lambda \in \Lambda_2 : t(\lambda) \neq 0\}$ [25]. The discrete IO maps in the literature are usually assumed to be causal or proper and are considered rational functions in the indeterminates s_ρ^{-1}; then the set Λ_2 from (1) is replaced with the closed unit polydisc [6], [7]. Properness is not assumed in the present paper because it is rather restrictive for partial differential equations, but properness and the ensuing BIBO stability will be discussed in the paper [20].

An IO system $\mathcal{B}$ is called stable if the characteristic variety of its autonomous part is contained in the stable region or, equivalently, if all polynomial-exponential trajectories in $\mathcal{B}^0$ are stable, i.e., involve exponents in the stable region only. In [29] this is called a characteristic variety (CV) condition and used to define stable *autonomous* systems. Stability of $\mathcal{B}$ in this sense is equivalent to the stability of the transfer matrix and an additional condition (see Theorem/Definition 5.1 and Remark 5.2). In particular, a one-dimensional IO system is stable if and only if its autonomous part is asymptotically stable or, equivalently, if its transfer matrix is stable and its singular variety is contained in the stable region. One of the reviewers points out that these two equivalent descriptions of the stability of one-dimensional IO systems should be

called time domain (resp., frequency domain) stability. Multidimensional stable IO systems are externally or IO stable in the sense that the transfer matrix acts as an operator on interesting classes of inputs and generates outputs of the same type as shown in Theorem/Definition 5.3 and Theorem 5.4; the latter applies results on partial differential equations from [4] as do Theorems 6.4 and 7.2 in [29] which, however, hold in special cases only. Theorem 7.6.2 of [15, p. 265], for instance, treats the connection between asymptotic stability and bounded input/bounded output (BIBO) stability of one-dimensional IO systems as done already by Kalman in his fundamental work. The paper [29] contains the interesting idea that stable systems should generate stable outputs from stable inputs *and initial conditions*, with a necessary requirement that the initial value problem be defined and uniquely solvable. For discrete multidimensional systems this is the case for which we give a partial answer and pose Open Problem 5.13, whose study is also worthwhile for continuous systems and certain function spaces.

An IO system $\mathcal{B}$ is called stabilizable if there is an IO system $\mathcal{B}'$ such that the feedback system (in (20), (21)) of $\mathcal{B}$ and $\mathcal{B}'$ is well-posed [26] and stable; then $\mathcal{B}'$ is called a stabilizing compensator. In Theorems 4.4 and 5.8 we characterize stabilizability of $\mathcal{B}$ and construct one stabilizing compensator, whereas Theorems 2.14 and 4.6 describe the parametrization or construction of all stabilizing compensators of a stabilizable IO behavior. The famous prototype of such a parametrization is that of Kucera, Youla, Bongiorno, and Jabr and is detailed by Vidyasagar in [26, Chap. 5]. The theorems on the stabilization of general multidimensional IO systems and their proofs are new but essentially influenced and inspired by and related to the results on the stabilization of IO maps in the references given above.

The proofs employ localization, after the work of Gabriel, as a new mathematical tool in systems theory which is described in Stenström's book [24] and in section 3 of this paper. At no time do the results and proofs need or employ the so-called fractional representation approach (i.e., matrix fraction descriptions of the transfer matrix of various kinds and the, sometimes long [26], [17], [18], ensuing matrix computations); thus they seem simpler and of interest even in the one-dimensional case (at least to the author). The localization technique also avoids the difficulties in [29] with the lack of ideal-convexity [23] of the unstable regions Λ_2. Such a region is called ideal-convex if

$$V(\mathfrak{a}) \cap \Lambda_2 = \emptyset \;\Rightarrow\; \mathfrak{a} \cap T \neq \emptyset \;\forall \text{ ideals } \mathfrak{a} \subset \mathbb{C}[s], \text{ where} \qquad V(\mathfrak{a}) := \{\lambda \in \mathbb{C}^r;\; \forall t \in \mathfrak{a} :\; t(\lambda) = 0\} \tag{2}$$

denotes the algebraic variety of $\mathfrak{a}$. Ideal-convexity is characterized by the coincidence of Gabriel localization with the standard localization functor $M \mapsto M_T$ on A-modules M (see Theorem/Definition 5.6).

Algorithmic problems are addressed in Remark/Open Problem 5.10. The algorithmic test of stability, stabilizability, and ideal-convexity and the algorithmic construction of one or all stabilizing compensators still encounter many difficulties; see in particular the open problems in [7] and [30] which, however, address these difficulties only for the closed unit polydisc of arbitrary dimension as region of instability. Solutions for the closed unit polydisc are known in interesting special cases [6], [7], [8].

Sections 2, 3, and 4 of this paper consider abstract IO systems whose signal spaces are injective cogenerators over a factorial Noetherian integral domain as in [11, Chap. 7] and use, in particular, Matlis' theory of injective modules over Noetherian

rings as in [12]. This generality makes the proofs more transparent and may possibly be used for other types of systems too. As preparation for the stable and stabilizable systems in section 4, section 2 discusses trivial and trivializable systems and the construction of all trivializing compensators. An IO behavior is called trivial if its autonomous part is zero or, equivalently by Theorem/Definition 2.5, if it is controllable and its transfer matrix is polynomial. In section 5 the results of sections 2, 3, and 4 are specialized to multidimensional systems proper as described above.

Due to the large number of papers on stabilization theory, we list only those references which are actually used. But the author is fully conscious of many other important contributions and contributors, such as, for instance, Bisiacco, Fornasini, Marchesini, and Valcher of the Padovian school.

2. Triviality and trivialization by feedback. Let A denote a commutative Noetherian integral domain with its quotient field $K = \mathrm{quot}(A)$, and let $\mathcal{F}$ denote an injective cogenerator which is used as a signal space with its scalar multiplication $\circ$. In this section we consider $\mathcal{F}$-systems or $\mathcal{F}$-behaviors as introduced and studied in [11], in particular [11, Chap. 7, p. 139]. Refer to the first pages of [28] or to [27] for a newer, more elegant introduction to multidimensional behavioral systems theory. It was shown that many cases of interest for systems theory can be developed in this abstract setting. For instance, the continuous case of systems governed by linear systems of partial differential equations with constant coefficients uses the data

$$\begin{gathered} A := \mathbb{C}[s] = \mathbb{C}[s_1, \dots, s_r],\ \mathcal{F} \ni y = y(z),\ s_\rho \circ y = \partial y / \partial z_\rho, \\ z := (z_1, \dots, z_r) \in \mathbb{R}^r,\ \lambda = (\lambda_1, \dots, \lambda_r) \in \mathbb{C}^r,\ \lambda \bullet z := \lambda_1 z_1 + \cdots + \lambda_r z_r, \\ \mathcal{F} := C^\infty(\mathbb{R}^r, \mathbb{C}) \text{ or } \mathcal{F} := \mathcal{D}'(\mathbb{R}^r) \text{ or} \\ \mathcal{F} := C^\infty(\mathbb{R}^r, \mathbb{C})_{\mathrm{lf}} := \{y \in C^\infty(\mathbb{R}^r, \mathbb{C});\ [\mathbb{C}[s] \circ y : \mathbb{C}] < \infty\} \\ = \mathcal{D}'(\mathbb{R}^r)_{\mathrm{lf}} = \oplus_{\lambda \in \mathbb{C}^r} \mathbb{C}[z] \exp(\lambda \bullet z), \end{gathered} \tag{3}$$

whereas multidimensional discrete or r-dimensional systems theory applies the polynomial algebra A as in (3) and the signal space

$$\begin{gathered} \mathcal{F} := \mathbb{C}^{\mathbb{N}^r} = \mathbb{C}[[z]] = \mathbb{C}[[z_1, \dots, z_r]] \ni y = (y_\mu)_{\mu \in \mathbb{N}^r} = \sum_{\mu \in \mathbb{N}^r} y_\mu z^\mu \\ \text{with } (s^\nu \circ y)_\mu := y_{\mu+\nu},\ \mu, \nu \in \mathbb{N}^r. \end{gathered} \tag{4}$$

The corresponding objects over the real field $\mathbb{R}$ instead of the complex field $\mathbb{C}$ are likewise admissible. Recall that an A-module $\mathcal{F}$ is an injective cogenerator if the contravariant *duality functor*,

$$D := \mathrm{Hom}_A(-, \mathcal{F}) : (\mathrm{Mod}_A)^{op} \to \mathrm{Mod}_A,\ M \mapsto \mathrm{Hom}_A(M, \mathcal{F}), \tag{5}$$

preserves and reflects exact sequences, where Mod_A is the category of A-modules. In [11] we used a *large* injective cogenerator $\mathcal{F}$, but this is unnecessary, as was observed in [12]. The column vectors in $\mathcal{F}^l$ are suggestively called *trajectories* also in the abstract module situation.

A matrix $R \in A^{k \times l}$ gives rise to the row submodule

$$U := A^{1 \times k} R = \sum_{i=1}^{k} A R_{i-} \subset F := A^{1 \times l}, \tag{6}$$

the factor module

$$M := A^{1\times l}/U = \sum_{j=1}^{l} A\overline{\delta_j},\ \delta_j := (0,\dots,0,\overset{j}{1},0,\dots,0) \in A^{1\times l},$$

and the behavior

$$\begin{aligned}\mathcal{B} &:= U^{\perp}\\ &:= \{w=(w_j)_{j=1,\dots,l}\in\mathcal{F}^l;\ \forall f=(f_1,\dots,f_l)\in U : f\circ w = f_1\circ w_1+\cdots+f_l\circ w_l = 0\}\\ &= \{w=(w_j)_{j=1,\dots,l}\in\mathcal{F}^l;\ R\circ w = 0\} \underset{\text{ident.}}{=} \operatorname{Hom}_A(M,\mathcal{F}),\ w=(\overline{\delta_j}\mapsto w_j).\end{aligned}$$

Since A is Noetherian, every submodule $U\subset A^{1\times l}$ arises in this fashion. Like $U^{\perp}$ we define, for every submodule $\mathcal{B}$ of $\mathcal{F}^l$, the orthogonal submodule

$$\mathcal{B}^{\perp} := \{f=(f_1,\dots,f_l)\in A^{1\times l};\ \forall w\in\mathcal{B}:\ f\circ w = 0\}\subset A^{1\times l} \tag{7}$$

of all linear equations which are satisfied by all trajectories in $\mathcal{B}$. If $\mathcal{B}:=U^{\perp}$ is a behavior, the relation $\mathcal{B}^{\perp\perp}=U^{\perp\perp\perp}=U^{\perp}=\mathcal{B}$ holds since $(-)^{\perp}$ is a Galois correspondence, whereas the identity $U=\mathcal{B}^{\perp}=U^{\perp\perp}$ is a consequence of the cogenerator property of $\mathcal{F}$ [11, Cor. 2.47].

The matrix R with M and $\mathcal{B}$ as in (6) also gives rise to its *transfer space* (= *signal flow space* in [11, Thm./Def. 2.91])

$$\widetilde{\mathcal{B}} := \{\widetilde{w}\in K^l;\ R\widetilde{w}=0\} \underset{\text{ident.}}{=} \operatorname{Hom}_A(M,K) \underset{\text{ident.}}{=} \operatorname{Hom}_K(K\otimes_A M,K)\subset K^l \tag{8}$$

and, more precisely, to the contravariant exact functor $\operatorname{Hom}_A(-,K)$ on finitely generated A-modules or to the covariant exact functor $\mathcal{B}\mapsto\widetilde{\mathcal{B}}$ on behaviors. Standard linear algebra over the field K can be applied to the K-space $\widetilde{\mathcal{B}}$. In particular, it determines the number

$$\operatorname{rank}(\mathcal{B}) := \operatorname{rank}(M) := [K\otimes_A M : K] = [\widetilde{\mathcal{B}} : K] = l - \operatorname{rank}(R), \tag{9}$$

where $[F:A]$ denotes the dimension of a *free* A-module F.

Let $w=\binom{y}{u}\in\mathcal{F}^{p+m}$, $l=p+m$, be a decomposition of the trajectories w into two components y and u, possibly after a permutation of the components w_j of w, and let $R=(P,-Q)\in A^{k\times(p+m)}$ be the corresponding decomposition of the matrix R such that $\mathcal{B}:=\{w=\binom{y}{u}\in\mathcal{F}^{p+m};\ P\circ y = Q\circ u\}$. The matrix P gives rise to the module $M^0:=A^{1\times p}/A^{1\times k}P$ and the behavior $\mathcal{B}^0:=\{y\in\mathcal{F}^p;\ P\circ y=0\}$.

Result 2.1 (IO structure and transfer matrix; see [11, Thms. 2.69, 2.94]). *The following assertions are equivalent:*

1. $\operatorname{rank}(P)=\operatorname{rank}(R)=p$.
2. *The projection*

$$\operatorname{proj} := (0\,\mathrm{id}_m)\circ : \widetilde{\mathcal{B}} = \left\{\widetilde{w}=\begin{pmatrix}\widetilde{y}\\ \widetilde{u}\end{pmatrix}\in K^{p+m};\ P\widetilde{y}=Q\widetilde{u}\right\}\to K^m,\ \begin{pmatrix}\widetilde{y}\\ \widetilde{u}\end{pmatrix}\mapsto\widetilde{u}, \tag{10}$$

is a K-isomorphism, i.e., for each $\widetilde{u}\in K^m$ the equation $P\widetilde{y}=Q\widetilde{u}$ is uniquely solvable for $\widetilde{y}\in K^p$.

3. *The A-module sequence*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A^{1\times m} & \overset{\text{inj}:=\circ(0\,\mathrm{id}_m)}{\longrightarrow} & M & \overset{\text{proj}:=\circ\binom{\mathrm{id}_p}{0}}{\longrightarrow} & M^0 & \longrightarrow & 0,\\ & & \eta & \mapsto & \overline{(0,\eta)},\ \overline{(\xi,\eta)} & \mapsto & \overline{\xi} & & \end{array} \tag{11}$$

is exact, and the module M^0 *is a torsion module or* $\operatorname{rank}(M^0) = 0$.

4. *The dual sequence of behaviors*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathcal{B}^0 & \xrightarrow{\binom{\mathrm{id}_p}{0}\circ} & \mathcal{B} & \xrightarrow{\mathrm{proj}:=(0\,\mathrm{id}_m)\circ} & \mathcal{F}^m & \longrightarrow 0, \\ & & y & \mapsto & \binom{y}{0},\ \binom{y}{u} & \mapsto & u & \end{array} \tag{12}$$

is exact, and the behavior $\mathcal{B}^0$ *is autonomous or* $\operatorname{rank}(\mathcal{B}^0) = 0$.

If these equivalent conditions are satisfied, the behavior is called an IO behavior *with the* IO structure $\binom{y}{u}$, *the* input u, *and the* output y. *The rank of* $\mathcal{B}$ *is* m. *There is a unique matrix* $H \in K^{p\times m}$, *the* transfer matrix *of* $\mathcal{B}$, *such that* $PH = Q$. *The inverse of the isomorphism* (10) *has the graph form* $\widetilde{u} \mapsto \binom{H\widetilde{u}}{\widetilde{u}}$, *and for every input* $u \in \mathcal{F}^m$ *there is a trajectory* $\binom{y}{u} \in \mathcal{B}$.

If $\mathcal{B}$ is an IO behavior as in the preceding result, the A-sequence

$$A^{1\times k} \xrightarrow{\circ(P,-Q)} A^{1\times(p+m)} \xrightarrow{\circ\binom{H}{\mathrm{id}_m}} K^{1\times m} \tag{13}$$

is a complex, i.e., $(P,-Q)\binom{H}{id_m} = PH - Q = 0$, and induces an A-epimorphism

$$\begin{gathered} M = A^{1\times(p+m)}/A^{1\times k}(P,-Q) \xrightarrow{\binom{H}{\mathrm{id}_m}} \mathcal{M} := \operatorname{im}\left(\circ\begin{pmatrix} H \\ \mathrm{id}_m \end{pmatrix}\right) = A^{1\times p}H + A^{1\times m}, \\ \overline{(\xi,\eta)} \mapsto \xi H + \eta. \end{gathered} \tag{14}$$

The A-submodule $\mathcal{M}$ of $K^{1\times m}$ is a lattice. Lattices play a decisive part in Quadrat's treatment of stabilization [17, Eq. (38) and Thm. 3], [18] (see also Quadrat's earlier papers quoted there).

Theorem and Definition 2.2 (controllable behaviors and realization). *The module* M *is torsion free if and only if the sequence* (13) *is exact or*

$$\mathcal{B}^\perp = A^{1\times k}(P,-Q) = \{(\xi,\eta) \in A^{1\times(p+m)};\ \xi H + \eta = 0\}. \tag{15}$$

If this is the case, the behavior $\mathcal{B}$ *is called* controllable *and is indeed the* unique controllable *IO behavior with transfer matrix* H *or, in other words, the unique controllable realization of* H, *and moreover,*

$$A^{1\times k}P = \{\xi \in A^{1\times p};\ \xi H \in A^{1\times m}\}. \tag{16}$$

Remark 2.3. As shown by Willems and Rocha for discrete two-dimensional systems and by Pillai and Shankar [14] for continuous multidimensional ones, the term *controllable* is justified by the concatenability of trajectories in controllable systems. A module is torsion free if and only if it can be embedded into a free module and hence, by duality, the system $\mathcal{B}$ is controllable if and only if there is a system epimorphism $\phi : \mathcal{F}^m \to \mathcal{B}$, which Pommaret (resp., Willems) calls a *parametrization* (resp., an *image representation*) of $\mathcal{B}$. The first reviewer suggests calling a controllable system $\mathcal{B}$ torsion free or, more generally, to systematically use the attribute of the system module M also for the system itself. We will stick to the term *controllable* since it is used by most researchers. If d is a common denominator of the entries of H, i.e., if $0 \neq d \in A$ and $dH \in A^{p\times m}$, the multiplication with d is an isomorphism on K, and therefore (15) can also be expressed as

$$A^{1\times k}(P,-Q) = \ker\left(A^{1\times(p+m)} \xrightarrow{\circ\binom{dH}{d\,\mathrm{id}_m}} A^{1\times m}\right).$$

Proof. The theorem is a reformulation of [11, Thm. 7.24]. We give a slightly simpler and more direct proof. If (13) is exact, then (14) is an isomorphism, and hence M is torsion free. If, conversely, M is torsion free, we obtain the *monomorphisms*

(17)
$$\begin{array}{ccccc} M = A^{1\times(p+m)}/A^{1\times k}R & \to & K\otimes_A M = K^{1\times(p+m)}/K^{1\times k}R & \overset{\circ\binom{H}{\mathrm{id}_m}}{\cong} & K^{1\times m}, \\ \overline{(\xi,\eta)} & \mapsto & 1\otimes\overline{(\xi,\eta)} = \overline{(\xi,\eta)} & \mapsto & \xi H+\eta, \end{array}$$

which imply $A^{1\times k}R = \{(\xi,\eta)\in A^{1\times(p+m)};\ \xi H+\eta=0\}$ and the exactness of (13). The first map in (17) is a monomorphism since its kernel is the torsion submodule of M and thus zero. The second isomorphism follows from

$$(P,-Q) = P(\mathrm{id}_p,-H) \text{ and } \mathrm{rank}(P)=p;\ \text{hence } K^{1\times k}(P,-Q) = K^{1\times p}(\mathrm{id}_p,-H),$$

and the exactness of

$$0 \longrightarrow K^{1\times p} \overset{\circ(\mathrm{id}_p,-H)}{\longrightarrow} K^{1\times(p+m)} \overset{\circ\binom{H}{\mathrm{id}_m}}{\longrightarrow} K^{1\times m} \longrightarrow 0$$
$$\text{or } K^{1\times(p+m)}/K^{1\times p}(\mathrm{id}_p,-H) \overset{\circ\binom{H}{\mathrm{id}_m}}{\cong} K^{1\times m}.$$

Concerning equality (16), the identity $PH=Q\in A^{k\times m}$ implies

$$A^{1\times k}P \subseteq \{\xi\in A^{1\times p};\ \xi H\in A^{1\times m}\}.$$

If, conversely, $\eta := -\xi H\in A^{1\times m}$ and thus $\xi H+\eta=0$, the monomorphism (17) implies $(\xi,-\eta)=\zeta(P,-Q)$ for some $\zeta\in A^{1\times k}$ and hence $\xi=\zeta P\in A^{1\times k}P$. ∎

Definition 2.4. *A behavior $\mathcal{B}$ is called* free *(resp.,* projective*) if its module M has this property, and is then controllable. Free behaviors were characterized in* [11, Thm. 7.53] *and were called* strongly *controllable by Rocha in her thesis.*

With these preparations we can now draw the following simple conclusion which, however, is basic for the present paper.

Theorem and Definition 2.5 (trivial IO behaviors). *For an IO behavior $\mathcal{B}$ as in Result* 2.1*, the following conditions are equivalent:*

1. *The autonomous behavior $\mathcal{B}^0$ is zero or $M^0=0$ or $A^{1\times k}P=A^{1\times p}$; i.e., the rows of P generate the full free module $A^{1\times p}$.*
2. *The behavior $\mathcal{B}$ is controllable and $H\in A^{p\times m}$.*

If these equivalent conditions are satisfied, the IO behavior $\mathcal{B}$ is called trivial *and the linear maps*

$$(18)\quad \circ(0\,\mathrm{id}_m) : A^{1\times m}\cong M,\ \eta\mapsto\overline{(0,\eta)},\ \textit{and}\ \ \mathrm{proj} := (0\,\mathrm{id}_m)\circ : \mathcal{B}\cong\mathcal{F}^m,\ \begin{pmatrix} y\\ u\end{pmatrix}\mapsto u,$$

are isomorphisms; in particular M and $\mathcal{B}$ are free. In accordance with the literature, we reserve the term stable *to a more general situation in sections* 4 *and* 5*.*

Proof. $\Rightarrow$: The exact sequences (11) and (12) and $M^0=0$ imply the isomorphisms (18) and hence the freeness and controllability of $\mathcal{B}$. The equality

$$A^{1\times p} = A^{1\times k}P = \{\xi\in A^{1\times p};\ \xi H\in A^{1\times m}\} \text{ implies } A^{1\times p}H\subseteq A^{1\times m};$$

hence $H\in A^{p\times m}$.

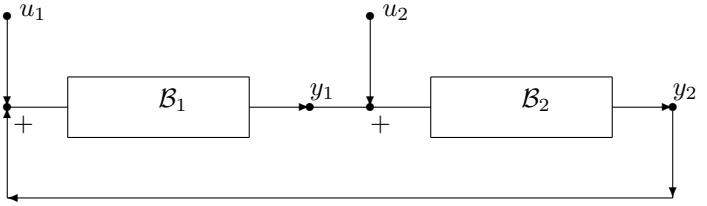


FIG. 2.1.

$\Leftarrow$: The controllability of $\mathcal{B}$ implies equality (16) and then, with $H \in A^{p\times m}$, also $A^{1\times p} = A^{1\times k}P$. ∎

In what follows we consider a feedback system given by the following block diagram [11, Chap. 8] as in Figure 2.1.

The two IO subbehaviors $\mathcal{B}_i$, $i = 1, 2$, of $\mathcal{F}^{p+m}$ are given as

$$\begin{aligned}\mathcal{B}_1 &:= \left\{ \begin{pmatrix} y_1 \\ u_1 \end{pmatrix} \in \mathcal{F}^{p+m};\ P_1 \circ y_1 = Q_1 \circ u_1 \right\},\ P_1H_1 = Q_1,\ P_1 \in A^{k_1\times p},\\ \mathcal{B}_2 &:= \left\{ \begin{pmatrix} u_2 \\ y_2 \end{pmatrix} \in \mathcal{F}^{p+m};\ P_2 \circ y_2 = Q_2 \circ u_2 \right\},\ P_2H_2 = Q_2,\ P_2 \in A^{k_2\times m}.\end{aligned} \tag{19}$$

Let $l := p + m$. The feedback behavior $\mathcal{B} := \text{feedback}(\mathcal{B}_1, \mathcal{B}_2)$ is the behavior

$$\mathcal{B} := \left\{ \begin{pmatrix} y \\ u \end{pmatrix} \in \mathcal{F}^{2l};\ y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, u = \begin{pmatrix} u_2 \\ u_1 \end{pmatrix} \in \mathcal{F}^{p+m} \text{ satisfy (21)} \right\} \tag{20}$$

with

$$P_1 \circ y_1 = Q_1 \circ (u_1 + y_2),\ P_2 \circ y_2 = Q_2 \circ (u_2 + y_1). \tag{21}$$

LEMMA 2.6. *The map*

$$\mathcal{B} = \text{feedback}(\mathcal{B}_1, \mathcal{B}_2) \to \mathcal{B}_1 \times \mathcal{B}_2,\ \begin{pmatrix} y \\ u \end{pmatrix} \mapsto \left(\begin{pmatrix} y_1 \\ u_1 + y_2 \end{pmatrix}, \begin{pmatrix} y_2 \\ u_2 + y_1 \end{pmatrix} \right)$$

is a behavior isomorphism with the inverse map $\left(\binom{y_1}{u_1}, \binom{u_2}{y_2}\right) \mapsto \binom{y}{v}$, *with* $v_1 := u_1 - y_2$ *and* $v_2 := u_2 - y_1$. *In particular,* $\mathcal{B}$ *is controllable (resp., projective) if and only if both* $\mathcal{B}_i$ *have these properties. If* $\mathcal{B}$ *is trivial, then both* $\mathcal{B}_i$ *are projective.*

Proof. It is obvious that the indicated map is indeed the inverse map. Recall that a trivial behavior is free and that projective modules are precisely the direct summands of free ones. ∎

From Result 2.1 we know that the transfer space of $\mathcal{B}_1$ is

$$\widetilde{\mathcal{B}_1} = \left\{ \begin{pmatrix} \widetilde{y_1} \\ \widetilde{u_1} \end{pmatrix} \in K^{p+m};\ \widetilde{y_1} = H_1\widetilde{u_1} \right\}$$

and likewise for $\mathcal{B}_2$. Therefore, the equations of the transfer space of $\mathcal{B}$ according to (8) are

$$\begin{aligned}&\widetilde{y_1} = H_1(\widetilde{u_1} + \widetilde{y_2}),\ \widetilde{y_2} = H_2(\widetilde{u_2} + \widetilde{y_1})\\ \Leftrightarrow\ &(\text{id}_p - H_1H_2)\widetilde{y_1} = H_1\widetilde{u_1} + H_1H_2\widetilde{u_2},\ \widetilde{y_2} = H_2(\widetilde{u_2} + \widetilde{y_1}).\end{aligned} \tag{22}$$

THEOREM 2.7. *The feedback behavior* $\mathcal{B} = \mathrm{feedback}(\mathcal{B}_1, \mathcal{B}_2)$ *is an IO behavior with input* $u = \binom{u_2}{u_1} \in \mathcal{F}^{p+m}$ *if and only if one or, equivalently, all of the matrices*

$$\mathrm{id}_p - H_1 H_2,\ \mathrm{id}_m - H_2 H_1,\ \begin{pmatrix} \mathrm{id}_p & -H_1 \\ -H_2 & \mathrm{id}_m \end{pmatrix}$$

are invertible, i.e., contained in $\mathrm{Gl}_\bullet(K)$. *If this is the case, the autonomous part* $\mathcal{B}^0$ *and the transfer matrix of* $\mathcal{B}$ *are given as*

$$\mathcal{B}^0 = \left\{ y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathcal{F}^{p+m};\ \begin{pmatrix} P_1 & -Q_1 \\ -Q_2 & P_2 \end{pmatrix} \circ y = 0 \right\} = \mathcal{B}_1 \cap \mathcal{B}_2,$$
$$H := \begin{pmatrix} \mathrm{id}_p & -H_1 \\ -H_2 & \mathrm{id}_m \end{pmatrix}^{-1} \begin{pmatrix} 0 & H_1 \\ H_2 & 0 \end{pmatrix}.$$

Moreover,

$$\mathcal{B}_1^\perp \oplus \mathcal{B}_2^\perp = (\mathcal{B}^0)^\perp.$$

Vidyasagar then calls the feedback system well-posed [26, *p.* 100]. *Rocha calls* $\mathcal{B}^0$ *a regular interconnection of* $\mathcal{B}_1$ *and* $\mathcal{B}_2$ [19, *Def.* 1]. *Direct sum decompositions and regular interconnections of systems have also been treated in* [1] *and* [33] *and in papers on the stabilization of IO maps; see, for instance,* [17], [18].

Proof. It is obvious that (22) are uniquely solvable for $\binom{\widetilde{y_1}}{\widetilde{y_2}} \in K^{p+m}$ for given $\binom{\widetilde{u_2}}{\widetilde{u_1}} \in K^{p+m}$ if and only if $\mathrm{id}_p - H_1 H_2 \in \mathrm{Gl}_p(K)$. The assertion then follows from Result 2.1. The simultaneous invertibility of these matrices is standard and follows trivially by elementary row and column operations. According to Result 2.1, the equations of $\mathcal{B}^0$ follow from (21) of $\mathcal{B}$ by setting the input $u = \binom{u_2}{u_1}$ to zero. The expression for H is implied by the equations

$$\begin{pmatrix} P_1 & -Q_1 \\ -Q_2 & P_2 \end{pmatrix} H = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \begin{pmatrix} \mathrm{id}_p & -H_1 \\ -H_2 & \mathrm{id}_m \end{pmatrix} H$$
$$= \begin{pmatrix} 0 & Q_1 \\ Q_2 & 0 \end{pmatrix} = \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} \begin{pmatrix} 0 & H_1 \\ H_2 & 0 \end{pmatrix}$$

by cancelling the matrix $\binom{P_1 \quad 0}{0 \quad P_2}$ of rank $l = p + m$ on the left.

The sum $(\mathcal{B}^0)^\perp = \mathcal{B}_1^\perp + \mathcal{B}_2^\perp$ follows from $\mathcal{B}^0 = \mathcal{B}_1 \cap \mathcal{B}_2$ by duality. Moreover,

$$\mathrm{rank}((\mathcal{B}^0)^\perp) = p + m = \mathrm{rank}(\mathcal{B}_1^\perp) + \mathrm{rank}(\mathcal{B}_2^\perp) \text{ implies } \mathcal{B}_1^\perp \cap \mathcal{B}_2^\perp = 0;$$
$$\text{hence } (\mathcal{B}^0)^\perp = \mathcal{B}_1^\perp \oplus \mathcal{B}_2^\perp. \qquad \square$$

Remark 2.8. If the feedback behavior of the preceding theorem is well-posed, the matrices

$$G := \begin{pmatrix} 0 & H_1 \\ H_2 & 0 \end{pmatrix} \text{ and } (\mathrm{id} - G)^{-1}(\mathrm{id} - G) = \mathrm{id} \text{ in } K^{(p+m)\times(p+m)} \tag{23}$$

imply

$$H = (\mathrm{id} - G)^{-1} G \text{ and } (\mathrm{id} - G)^{-1} = \mathrm{id} + H,$$

and hence that $(\mathrm{id} - G)^{-1}$ and H share many properties, in particular that H belongs to $A^{(p+m)\times(p+m)}$ if and only if $(\mathrm{id} - G)^{-1}$ does [26, Chap. 5, Lem. 9].

COROLLARY AND DEFINITION 2.9. *If the feedback behavior of the preceding theorem is well-posed, then $\mathcal{B}$ is trivial, i.e., $\mathcal{B}^0 = 0$, if and only if*

$$\mathcal{B}_1^\perp \oplus \mathcal{B}_2^\perp = A^{1\times l} \text{ or, by duality, } \mathcal{B}_1 \oplus \mathcal{B}_2 = \mathcal{F}^{p+m}.$$

If this holds, the modules $U_j = \mathcal{B}_j^\perp$, $j = 1, 2$, are complementary direct summands of $A^{1\times l}$ and the modules $M_j = A^{1\times l}/U_j$ and the behaviors $\mathcal{B}_j$ are projective. We say that the IO behavior $\mathcal{B}_2$ trivializes (instead of stabilizes) $\mathcal{B}_1$.

In the following we assume that the IO behavior $\mathcal{B}_1$ is given and construct all trivializing IO behaviors $\mathcal{B}_2$. The set of all these $\mathcal{B}_2$ is parametrized by an open subset (in the Zariski topology) of a finitely generated polynomial module. This parametrization generalizes the important Youla–Kučera parametrization of the one-dimensional stabilization theory. According to the preceding theorem, we make the necessary assumption that the behavior $\mathcal{B}_1$ is *projective*. We use the exact sequence

$$\begin{gathered} F_1 := A^{1\times k_1} \xrightarrow{d_0 := \circ R_1} F_0 := A^{1\times l} \xrightarrow{\text{can}} M_1 := A^{1\times l}/U_1 \to 0 \\ \text{with } l := p+m,\ R_1 = (P_1, -Q_1),\ U_1 := \text{im}(d_0) = A^{1\times k_1} R_1. \end{gathered} \tag{24}$$

The following remark establishes the well-known one-to-one correspondence between idempotent endomorphisms $e = e^2$ of a module and direct sum decompositions. If

$$e = e^2 \in \text{Hom}_A(F_0, F_0)$$

is an idempotent or a projection, then

$$\begin{gathered} F_0 = \text{im}(e) \oplus \ker(e) \ni x = e(x) + (x - e(x)), \\ x = e(x) \ \Leftrightarrow\ x \in \text{im}(e),\ \ker(e) = \text{im}(\text{id} - e). \end{gathered}$$

Conversely, any direct sum decomposition

$$F_0 = V_1 \oplus V_2 \ni x = x_1 + x_2 =: e(x) + (x - e(x))$$

gives rise to the projection $e = e^2 \in \text{Hom}_A(F_0, F_0)$. The map

$$\begin{gathered} \{e \in \text{Hom}_A(F_0, F_0);\ e = e^2\} \to \{F_0 = V_1 \oplus V_2\}, \\ e \mapsto F_0 = \text{im}(e) \oplus \ker(e) \end{gathered}$$

is bijective. By restricting the preceding bijection to decompositions $F_0 = U_1 \oplus U_2$ with the fixed U_1 from above we obtain the following.

COROLLARY 2.10. *For the data from* (24) *the following map is bijective:*

$$\begin{gathered} \{e \in \text{Hom}_A(F_0, F_0);\ e = e^2,\ \text{im}(e) = U_1\} \to \{F_0 = U_1 \oplus U_2\}, \\ e \mapsto F_0 = U_1 \oplus \ker(e),\ \ker(e) = \text{im}(\text{id}_l - e). \end{gathered}$$

The preceding decompositions can also be described by means of homomorphisms $g \in \text{Hom}_A(F_0, F_1) \cong A^{l\times k_1}$, as was already shown more generally in [9, Lem. 6.2, p. 88]. This is important for constructive purposes in particular.

LEMMA 2.11. *The following conditions are equivalent for a submodule U_1 of F_0 and its factor module $M_1 = F_0/U_1 = \text{cok}(d_0)$:*

1. *The module M_1 is projective.*

2. *U_1 is a direct summand of F_0. Then U_1 is also projective and there is an idempotent $e = e^2 \in \text{Hom}_A(F_0, F_0)$ such that $U_1 = \text{im}(e)$.*

3. *There is a homomorphism* $g \in \mathrm{Hom}_A(F_0, F_1)$ *such that* $d_0 g d_0 = d_0$. *Then* $e := d_0 g$ *is an idempotent with* $U_1 = im(d_0) = \mathrm{im}(e)$.

Proof. The equivalence of 1 and 2 is standard since the projectivity of M_1 is equivalent with the splitting of the exact sequence (24).

$2 \Rightarrow 3$: Let e be an idempotent with image $\mathrm{im}(e) = U_1$. Then the map $e : F_0 \to U_1 = \mathrm{im}(e)$ is surjective. Since U_1 is projective and the map $d_0 : F_1 \to U_1$ is surjective, there is a linear section

$$s : U_1 \to F_1 \text{ such that } \mathrm{id}_{U_1} = d_0 s. \text{ Define } g := se : F_0 \xrightarrow{e} U_1 \xrightarrow{s} F_1. \text{ Then}$$
$$e(x) = d_0 s(e(x)) \in U_1 \text{ for } x \in F_0; \text{ hence } e = d_0 se = d_0 g \text{ and}$$
$$d_0 g d_0 = d_0 s e d_0 = e d_0 = d_0 \text{ since } e(d_0(x)) = d_0(x) \text{ for } d_0(x) \in U_1 = \mathrm{im}(e).$$

$3 \Rightarrow 2$: Assume $d_0 g d_0 = d_0$ and define $e := d_0 g$. Then $e \in \mathrm{Hom}_A(F_0, F_0)$ is an idempotent since $e^2 = (d_0 g d_0) g = d_0 g = e$ and, moreover, $\mathrm{im}(e) = \mathrm{im}(d_0 g) \subseteq \mathrm{im}(d_0) = U_1$. From $d_0 = d_0 g d_0 = e d_0$ we infer $U_1 = \mathrm{im}(d_0) \subseteq \mathrm{im}(e)$ in the same fashion; hence $U_1 = \mathrm{im}(e)$. □

THEOREM 2.12. *Assume that* $M_1 = A^{1\times l}/U_1$ *is projective and that* $g_1 \in \mathrm{Hom}_A(F_0, F_1)$ *and* $e_1 := d_0 g_1$ *with* $\mathrm{im}(e_1) = U_1$ *are constructed according to the preceding lemma. Then the map*

$$\varphi : \{h \in \mathrm{Hom}_A(F_0, F_1);\ d_0 h d_0 = 0\}/\{h \in \mathrm{Hom}_A(F_0, F_1);\ d_0 h = 0\}$$
$$\to \{e \in \mathrm{Hom}_A(F_0, F_0);\ e = e^2,\ U_1 = \mathrm{im}(e)\},\ \overline{h} \mapsto \varphi(\overline{h}) := e_1 + d_0 h,$$

is bijective. It furnishes a parametrization of the set of direct complements $U_2 = \ker(e)$ *with* $F_0 = U_1 \oplus U_2$ *by the finitely generated* A*-module on the left. If* A *is a field, then the right side is an affine open subset of the projective Grassmann variety of all* m*-dimensional subspaces of* $A^{1\times(p+m)}$.

Proof. 1. The equation $d_0 g d_0 = d_0$ is an inhomogeneous linear equation for g in the A-module $\mathrm{Hom}_A(F_0, F_1)$ and g_1 is one solution. Hence

$$\{g \in \mathrm{Hom}_A(F_0, F_1);\ d_0 g d_0 = d_0\} = g_1 + \{h \in \mathrm{Hom}_A(F_0, F_1);\ d_0 h d_0 = 0\}.$$

2. The map is well defined: If

$$d_0 h d_0 = 0, \text{ then } g_2 := g_1 + h \text{ satisfies } d_0 g_2 d_0 = d_0, \text{ and hence}$$
$$e_2 := d_0 g_2 = d_0 g_1 + d_0 h = e_1 + d_0 h$$

is an idempotent with $\mathrm{im}(e_2) = U_1$ according to the preceding lemma. If h_2 and h_3 are homogeneous solutions such that

$$\overline{h_2} = \overline{h_3} \text{ or } d_0(h_2 - h_3) = 0, \text{ then } e_1 + d_0 h_2 = e_1 + d_0 h_3.$$

3. It is obvious that the map φ is injective, and it is surjective by the preceding lemma. □

We reformulate the preceding theorem in matrix terms. We identify

$$A^{m\times n} = \mathrm{Hom}_A(A^{1\times m}, A^{1\times n}),\ X = \circ X = (\xi \mapsto \xi X),$$

and emphasize that $(\circ X)(\circ Y) = \circ(YX)$.

THEOREM 2.13. *Let* $\mathcal{B}_1 \subseteq \mathcal{F}^{p+m}$ *be an IO behavior with the transfer matrix* H_1 *and the data*

$$U_1 = \mathcal{B}_1^{\perp} = A^{1\times k_1} R_1,\ M_1 = A^{1\times l}/\mathcal{B}_1^{\perp},$$
$$R_1 = (P_1, -Q_1) \in A^{k_1\times(p+m)},\ \mathrm{rank}(P_1) = p,\ P_1 H_1 = Q_1,\ H_1 \in K^{p\times m}.$$

1. *The module M_1 or the behavior $\mathcal{B}_1$ is projective if and only if there is a matrix*

$$G_1 \in A^{l\times k_1} \text{ with } R_1G_1R_1 = R_1 \text{ or, equivalently, with } (\mathrm{id}_p, -H_1)G_1P_1 = \mathrm{id}_p. \tag{25}$$

Then $E_1 := G_1R_1 = E_1^2 \in A^{l\times l}$ is an idempotent matrix with $\mathcal{B}_1^\perp = A^{1\times l}E_1$.

This algorithm is related to the algorithms of [33]*; its idea goes back at least to MacLane* [9] *as explained above.*

2. *Assume that $\mathcal{B}_1$ is projective with the data from* 1. *Then the map*

$$\begin{aligned}\varphi : \{X \in A^{l\times k_1};\ R_1XR_1 = 0\}/\{X \in A^{l\times k_1};\ XR_1 = 0\}\\ \to \{E \in A^{l\times l};\ E = E^2,\ \mathcal{B}_1^\perp = A^{1\times l}E\},\ \overline{X} \mapsto \varphi(\overline{X}) := E_1 + XR_1,\end{aligned} \tag{26}$$

is bijective. Moreover,

(27)
$$R_1XR_1 = 0 \text{ if and only if } (\mathrm{id}_p, -H_1)XP_1 = 0 \text{ if and only if } (t\,\mathrm{id}_p, -tH_1)XP_1 = 0 \text{ and } XR_1 = 0 \text{ if and only if } XP_1 = 0,$$

where $0 \neq t \in A$ is a common denominator of the entries of H_1, i.e., $tH_1 \in A^{p\times m}$. The direct complement behavior $\mathcal{B}_2$ of $\mathcal{B}_1$ in $\mathcal{F}^{p+m}$, which is constructed by means of the idempotent $E = E_1 + XR_1$, is

$$\begin{aligned}\mathcal{B}_2 := \{w \in \mathcal{F}^{p+m};\ (\mathrm{id}_l - E_1 - XR_1)\circ w = 0\} = U_2^\perp,\ \mathcal{B}_1 \oplus \mathcal{B}_2 = \mathcal{F}^{p+m},\\ \text{where } U_2 := \ker(\circ(E_1 + XR_1)) = A^{1\times l}(\mathrm{id}_l - E_1 - XR_1).\end{aligned} \tag{28}$$

3. *Since $\mathcal{B}_1^\perp = A^{1\times l}E_1$, the bijection* (26) *holds if R_1 is replaced with E_1.*

4. *If U_1 is even free, and if the rows of R_1 are a basis of U_1, i.e., if $R_1 = (P_1, -Q_1) \in A^{p\times(p+m)}$ and $\det(P_1) \neq 0$, then the situation of* (26) *simplifies to*

$$\begin{aligned}G_1 \in A^{l\times p},\ R_1G_1 = \mathrm{id}_p,\ E_1 = G_1R_1,\ \text{and}\\ \varphi : \{X \in A^{l\times k};\ R_1X = 0\} \cong \{E \in A^{l\times l};\ E = E^2,\ \mathcal{B}_1^\perp = A^{1\times l}E\},\\ X \mapsto E_1 + XR_1.\end{aligned} \tag{29}$$

Proof. The equivalence of (25) and (27) follows from

$$R_1 = (P_1, -Q_1) = P_1(\mathrm{id}_p, -H_1) \text{ and } \mathrm{rank}(P_1) = \mathrm{rank}(\mathrm{id}_p, -H_1) = p,$$

which imply that P_1 (resp., $(\mathrm{id}_p, -H_1)$) can be cancelled as a left (resp., right) factor. □

Of course, each direct complement behavior

$$\mathcal{B}_2 = \left\{ w_2 = \begin{pmatrix} u_2 \\ y_2 \end{pmatrix} \in \mathcal{F}^{p+m};\ (\mathrm{id}_l - E)\circ w_2 = 0 \right\} \text{ with}$$
$$\mathrm{rank}(\mathrm{id}_l - E) = l - \mathrm{rank}(E) = l - p = m$$

of the preceding theorem admits an IO structure, but not necessarily that with u_2 as inputs, as needed for the feedback construction. To enforce this additional property we proceed as follows. Let $X_1, \ldots, X_n$ be a system of A-generators of the submodule

$$\{X \in A^{l\times k_1};\ R_1XR_1 = 0(\ \Leftrightarrow\ R_1XP_1 = 0)\}. \tag{30}$$

The same matrices generate $\{X \in K^{l\times k_1};\ (\mathrm{id}_p, -H_1)XP_1 = 0\}$ as a K-vector space. In the relevant examples the X_i can be computed by means of Gröbner bases. The linear map $\theta = (\theta_i)_{i=1,\dots,n} \mapsto \sum_{i=1}^n \theta_i X_i$ induces the A-isomorphism

$$\begin{aligned} &\varphi_1 : A^{1\times n}/V \cong \{X \in A^{l\times k_1};\ R_1XP_1 = 0\}/\{X \in A^{l\times k_1};\ XP_1 = 0\},\\ &\text{where } \varphi_1(\overline{\theta}) := \sum_{i=1}^n \theta_i \overline{X_i} \text{ and } V := \left\{\theta \in A^{1\times n};\ \sum_{i=1}^n \theta_i X_i P_1 = 0\right\}. \end{aligned} \tag{31}$$

Again, in the relevant examples the submodule V and its factor module $A^{1\times n}/V$ can be computed by means of Gröbner bases. In addition we consider the polynomial map

$$d : K^{l\times m} \to K^{\binom{l}{m}},\ Y \mapsto d(Y) = (d_\alpha(Y))_\alpha, \tag{32}$$

where the $d_\alpha(Y)$ are the $\binom{l}{m}$ $m \times m$ minors of the matrix Y. The function d is polynomial in the entries of Y and $d(Y) \neq 0$ if and only if $\mathrm{rank}(Y) = m$. With the data from above, we obtain the induced map

$$\{E \in A^{l\times l};\ E = E^2,\ \mathcal{B}_1^\perp = A^{1\times l}E\} \to A^{\binom{l}{m}},\ E \mapsto d\left((\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix}\right).$$

If

$$\mathrm{id}_l - E =: (-Q_2, P_2) \in A^{l\times(p+m)},\ \text{hence } (\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix} = P_2,$$

the behavior

$$\mathcal{B}_2 := \left\{w_2 = \begin{pmatrix}u_2\\ y_2\end{pmatrix} \in \mathcal{F}^{p+m};\ (\mathrm{id}_l - E)\circ w_2 = 0 \text{ or } P_2 \circ y_2 = Q_2 \circ u_2\right\}$$

is an IO behavior with input u_2 if and only if $\mathrm{rank}(P_2) = m$. Composing the maps just constructed, we obtain the map

$$\begin{aligned} &\Phi : K^{1\times n} \to K^{\binom{l}{m}},\ \theta \mapsto d\left((\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix}\right),\\ &\text{with } E := E_1 + XR_1,\ X := \sum_{i=1}^n \theta_i X_i, \end{aligned} \tag{33}$$

which is polynomial in the components of θ; i.e., its components are contained in $K[\theta]$. Summing up we obtain the final parametrization theorem.

THEOREM 2.14 (parametrization of trivializing behaviors). *Let $\mathcal{B}_1 = \{w_1 = \binom{y_1}{u_1} \in \mathcal{F}^{p+m}\}$ be a projective behavior as in Theorem 2.13, and assume that A is infinite. Then, with the data introduced above, the map*

$$\begin{aligned} &\overline{\theta} \mapsto \mathcal{B}_2 := \left\{w_2 = \begin{pmatrix}u_2\\ y_2\end{pmatrix} \in \mathcal{F}^{p+m};\ (\mathrm{id}_l - E)\circ w_2 = 0\right\},\\ &\textit{where } E := E_1 + XR_1,\ X := \sum_{i=1}^n \theta_i X_i, \end{aligned}$$

is a bijection from the nonempty set

$$\left\{\overline{\theta} \in A^{1\times n}/V;\ \Phi(\theta) = d\left((\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix}\right) \neq 0\right\}$$

onto the set of all IO behaviors $\mathcal{B}_2 \subseteq \mathcal{F}^{p+m}$*, which trivialize* $\mathcal{B}_1$*. This signifies that the feedback system* $\mathcal{B} := \text{feedback}(\mathcal{B}_1, \mathcal{B}_2)$ *is well-posed and trivial. In particular, such behaviors exist.*

Proof. It remains only to show that such behaviors exist. The matrix P_1 of rank p has a left inverse

$$S_1 \in K^{p\times k_1} \text{ with } S_1P_1 = \mathrm{id}_p. \tag{34}$$

Define

$$E := \begin{pmatrix}\mathrm{id}_p\\ 0\end{pmatrix}(\mathrm{id}_p, -H_1) = \begin{pmatrix}\mathrm{id}_p & -H_1\\ 0 & 0\end{pmatrix} \in K^{l\times l} \text{ and } G := \begin{pmatrix}\mathrm{id}_p\\ 0\end{pmatrix}S_1 \in K^{l\times k_1};$$
$$\text{hence } \mathrm{id}_l - E = \begin{pmatrix}0 & H_1\\ 0 & \mathrm{id}_m\end{pmatrix},\ (\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix} = \begin{pmatrix}H_1\\ \mathrm{id}_m\end{pmatrix},$$
$$\text{rank}\left((\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix}\right) = m,\ (\mathrm{id}_p, -H_1)GP_1 = \mathrm{id}_p, \text{ and } E := GR_1.$$

Thus $X := G - G_1 \in K^{l\times k_1}$ satisfies

$$(\mathrm{id}_p, -H_1)XP_1 = 0,\ E = E_1 + XR_1 \text{ and } \text{rank}\left((\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix}\right) = m;$$

in particular, X is of the form

$$X = \sum_{i=1}^{n} \theta_i X_i \text{ with } \theta \in K^{1\times n} \text{ and } \Phi(\theta) \neq 0.$$

Since the components of Φ are polynomials in $K[\theta]$ and since A is an infinite subset of K, there is also a parameter $\theta' \in A^{1\times n}$ with nonzero $\Phi(\theta')$, which induces an IO behavior $\mathcal{B}_2$ such that the feedback system $\text{feedback}(\mathcal{B}_1, \mathcal{B}_2)$ is well-posed and trivial. □

Remark 2.15. We remark that $\{\theta \in K^{1\times n};\ \Phi(\theta) \neq 0\}$ is a nonempty open subset of $K^{1\times n}$ with respect to the Zariski topology and is therefore dense. This signifies that, generically, the behaviors $\mathcal{B}_2$ with $\mathcal{F}^{p+m} = \mathcal{B}_1 \oplus \mathcal{B}_2$ are IO behaviors with the desired input $u_2 \in \mathcal{F}^p$.

3. Localization. The assumptions of the preceding section remain in force. Moreover, we assume that the ring A is *factorial*. We use Matlis' structure theory of injective modules over Noetherian rings and refer to [10, pp. 145–150] and [12, sec. 2], where this theory was used in the system-theoretic context. Moreover, we use Gabriel's theory of localization as detailed in the book [24, Chap. IX].

Let $\text{Spec}(A)$ (resp., $\text{Max}(A)$) denote the set of prime (resp., maximal) ideals of A. A prime ideal $\mathfrak{p}$ is *associated* with an A-module M if and only if there is an $x \in M$ such that $\mathfrak{p} = \text{ann}(x) = \{a \in A;\ ax = 0\}$ or, in other words, that $A/\mathfrak{p}$ is a submodule of M up to isomorphism. Let $\text{Ass}(M) \subset \text{Spec}(A)$ denote the set of prime ideals associated with M. A module M is $\mathfrak{p}$-*coprimary* if $\text{Ass}(M)$ consists exactly of one prime ideal $\mathfrak{p}$; then

$$a : M \to M,\ x \mapsto ax \text{ is } \begin{cases}\text{injective}\\ \text{locally nilpotent}\end{cases} \Leftrightarrow \begin{cases}a \notin \mathfrak{p}\\ a \in \mathfrak{p}\end{cases}. \tag{35}$$

Local or almost nilpotency of a on M means that for all $x \in M$ there is an index m such that $a^m x = 0$.

The injective module $\mathcal{F}$ is a cogenerator if and only if it contains all simple modules $A/\mathfrak{m}$, $\mathfrak{m} \in \operatorname{Max}(A)$, up to isomorphism, i.e., if $\operatorname{Max}(A) \subset \operatorname{Ass}(\mathcal{F})$. Each module M has an *injective envelope* $E(M) \supset M$ which is unique up to noncanonical isomorphism. This signifies that $E(M)$ is injective and that $V \cap M \neq 0$ for each nonzero submodule V of $E(M)$. Since A is Noetherian, a direct sum or coproduct of modules is injective if and only if each direct summand has this property. Each injective module E admits a direct sum decomposition into directly indecomposable injective modules, and this decomposition is unique up to an automorphism of E. An indecomposable injective module E is coprimary, and if $\mathfrak{p}$ is its unique associated prime ideal, then $E = E(A/\mathfrak{p})$. The map

$$\begin{gathered}\operatorname{Spec}(A) \to \{E;\ E \text{ indecomposable injective}\}/\text{isomorphism},\\ \mathfrak{p} \mapsto E(A/\mathfrak{p}),\ \operatorname{Ass}(E(A/\mathfrak{p})) = \{\mathfrak{p}\}\end{gathered} \tag{36}$$

is a bijection of the prime spectrum $\operatorname{Spec}(A)$ onto the set of indecomposable injectives up to isomorphism. Since $E(A/\mathfrak{p})$ is injective and indecomposable, the injective map

$$s : E(A/\mathfrak{p}) \to E(A/\mathfrak{p}),\ x \mapsto sx,\ s \in A \setminus \mathfrak{p},$$

is even bijective, and therefore $E(A/\mathfrak{p})$ is an $A_{\mathfrak{p}}$-module, where

$$A_{\mathfrak{p}} := \left\{ \frac{a}{s};\ a \in A,\ s \in A \setminus \mathfrak{p} \right\} \subset K = \operatorname{quot}(A)$$

is the local ring of the prime ideal $\mathfrak{p}$ with its unique maximal ideal $\mathfrak{p}_{\mathfrak{p}} = A_{\mathfrak{p}}\mathfrak{p}$. For $M \in \operatorname{Mod}_A$, the adjointness isomorphism

$$\begin{gathered}\operatorname{Hom}_A(M, E(A/\mathfrak{p})) \cong \operatorname{Hom}_{A_{\mathfrak{p}}}(M_{\mathfrak{p}}, E(A/\mathfrak{p})) = \operatorname{Hom}_A(M_{\mathfrak{p}}, E(A/\mathfrak{p})),\\ M_{\mathfrak{p}} := A_{\mathfrak{p}} \otimes_A M = \left\{ \frac{x}{s};\ x \in M,\ s \in A \setminus \mathfrak{p} \right\} \in \operatorname{Mod}_{A_{\mathfrak{p}}}\end{gathered} \tag{37}$$

holds. Together with $N = N_{\mathfrak{p}}$ for $N \in \operatorname{Mod}_{A_{\mathfrak{p}}}$ this shows that the duality functor $\operatorname{Hom}_{A_{\mathfrak{p}}}(-, E(A/\mathfrak{p}))$ on $\operatorname{Mod}_{A_{\mathfrak{p}}}$ is exact, and hence $E(A/\mathfrak{p})$ is an injective $A_{\mathfrak{p}}$-module. Since it contains the unique simple $A_{\mathfrak{p}}$-module $(A/\mathfrak{p})_{\mathfrak{p}} = A_{\mathfrak{p}}/\mathfrak{p}_{\mathfrak{p}}$, the module $E(A/\mathfrak{p})$ is the unique minimal injective cogenerator of the category $\operatorname{Mod}_{A_{\mathfrak{p}}}$ to which the considerations of section 2 are applicable. In particular,

$$M_{\mathfrak{p}} = 0 \quad \Leftrightarrow \quad \operatorname{Hom}_A(M, E(A/\mathfrak{p})) = 0. \tag{38}$$

Let

$$\mathcal{F} := \oplus_{i \in I} \mathcal{F}_i,\ \operatorname{Ass}(\mathcal{F}_i) = \{\mathfrak{p}_i\},\ \mathcal{P} := \operatorname{Ass}(\mathcal{F}) = \{\mathfrak{p}_i;\ i \in I\} \tag{39}$$

denote a direct sum decomposition of the injective cogenerator $\mathcal{F}$ into indecomposable injectives $\mathcal{F}_i$ and let

$$\mathcal{P} := \operatorname{Ass}(\mathcal{F}) = \mathcal{P}_1 \uplus \mathcal{P}_2,\ \mathcal{P}_1, \mathcal{P}_2 \neq \emptyset, \tag{40}$$

be a disjoint decomposition of $\operatorname{Ass}(\mathcal{F})$. *All subsequent objects depend on the choice of* $\mathcal{F}$ *and of the decompositions* (39) *and* (40).

In particular, the latter imply direct sum decompositions

$$\begin{gathered}\mathcal{F}=\oplus_{\mathfrak{p}\in\mathcal{P}}\mathcal{F}(\mathfrak{p})=\mathcal{F}_1\oplus\mathcal{F}_2 \text{ with } \mathcal{F}(\mathfrak{p}):=\oplus_{i\in I,\,\mathfrak{p}_i=\mathfrak{p}}\mathcal{F}_i\cong E(A/\mathfrak{p})^{(\mu(\mathfrak{p}))},\\ \mu(\mathfrak{p}):=|\{i\in I;\ \mathfrak{p}_i=\mathfrak{p}\}|,\ \mathcal{F}_j:=\oplus_{\mathfrak{p}\in\mathcal{P}_j}\mathcal{F}(\mathfrak{p}),\ j=1,2,\end{gathered}\tag{41}$$

where $E(A/\mathfrak{p})^{(\mu(\mathfrak{p}))}$ denotes the direct sum of $\mu(\mathfrak{p})$ copies of $E(A/\mathfrak{p})$. All modules in these direct sum decompositions are injective. Like $E(A/\mathfrak{p})$, the module $\mathcal{F}(\mathfrak{p})\cong E(A/\mathfrak{p})^{(\mu(\mathfrak{p}))}$ is an injective cogenerator of the category of $A_{\mathfrak{p}}$-modules. The decompositions (41) induce corresponding decompositions of $\mathcal{F}$ behaviors. Let

$$\begin{gathered}R\in A^{k\times l},\ U:=A^{1\times k}R,\ M:=A^{1\times l}/U,\\ \mathcal{B}:=U^{\perp}\underset{\text{ident.}}{=}\operatorname{Hom}_A(M,\mathcal{F})=\{w\in\mathcal{F}^l;\ R\circ w=0\}.\end{gathered}\tag{42}$$

Since M is finitely generated, the functor $\operatorname{Hom}_A(M,-)$ preserves direct sums, and therefore (41) induces direct decompositions

$$\begin{gathered}\mathcal{B}=\oplus_{i\in I}\mathcal{B}_i\subset\mathcal{F}^l=\oplus_{i\in I}\mathcal{F}_i^l,\\ \mathcal{B}_i:=\mathcal{B}\cap\mathcal{F}_i^l\underset{\text{ident.}}{=}\operatorname{Hom}_A(M,\mathcal{F}_i)\underset{\text{ident.}}{=}\{w\in\mathcal{F}_i^l;\ R\circ w=0\},\\ \mathcal{B}=\oplus_{\mathfrak{p}\in\mathcal{P}}\mathcal{B}(\mathfrak{p})=\mathcal{B}_1\oplus\mathcal{B}_2,\\ \mathcal{B}(\mathfrak{p}):=\mathcal{B}\cap\mathcal{F}(\mathfrak{p})^l=\operatorname{Hom}_A(M,\mathcal{F}(\mathfrak{p}))=\{w\in\mathcal{F}(\mathfrak{p})^l;\ R\circ w=0\}\\ \cong\operatorname{Hom}_{A_{\mathfrak{p}}}(M_{\mathfrak{p}},E(A/\mathfrak{p}))^{(\mu(\mathfrak{p}))},\\ \mathcal{B}_j:=\mathcal{B}\cap\mathcal{F}_j^l=\operatorname{Hom}_A(M,\mathcal{F}_j)=\{w\in\mathcal{F}_j^l;\ R\circ w=0\},\ j=1,2.\end{gathered}\tag{43}$$

We use the following suggestive system-theoretic terminology. The elements of $\mathcal{F}$ (resp., $\mathcal{F}_1$) are called *signals* (resp., *stable signals*). If $y=y_1+y_2\in\mathcal{F}=\mathcal{F}_1\oplus\mathcal{F}_2$ is a signal, then y_1 is called its *stable part* and y_2 its *steady state*.

Example 3.1. We indicate the data for the important complex continuous case. The algebra A is the complex polynomial algebra $\mathbb{C}[s]=\mathbb{C}[s_1,\dots,s_r]$. The map

$$\begin{gathered}\mathbb{C}^r\to\operatorname{Max}(\mathbb{C}[s]),\\ \lambda=(\lambda_1,\dots,\lambda_r)\mapsto\mathfrak{m}_\lambda:=\sum_{\rho=1}^{r}\mathbb{C}[s](s_\rho-\lambda_\rho)=\{t\in\mathbb{C}[s];\ t(\lambda)=0\}\end{gathered}$$

is bijective. As an injective cogenerator we take the module

$$\begin{gathered}\mathcal{F}:=\mathcal{D}'(\mathbb{R}^r)_{lf}:=\{y\in\mathcal{D}'(\mathbb{R}^r);\ [\mathbb{C}[s]\circ y:\mathbb{C}]<\infty\}=\oplus_{\lambda\in\mathbb{C}^r}\mathbb{C}[z]\exp(\lambda\bullet z),\\ z=(z_1,\dots,z_r)\in\mathbb{R}^r,\ \lambda\bullet z:=\lambda_1z_1+\cdots+\lambda_rz_r\end{gathered}\tag{44}$$

of *locally finite* distributions or *polynomial-exponential* functions [12, Thm. 6.6]. Then

$$\begin{gathered}\operatorname{Ass}(\mathcal{F})=\operatorname{Max}(\mathbb{C}[s])=\{\mathfrak{m}_\lambda;\ \lambda\in\mathbb{C}^r\},\\ \mathcal{F}(\lambda):=\mathcal{F}(\mathfrak{m}_\lambda)=\mathbb{C}[z]\exp(\lambda\bullet z)\cong E(\mathbb{C}[s]/\mathfrak{m}_\lambda);\end{gathered}\tag{45}$$

i.e., $\mathcal{F}$ is the unique minimal injective cogenerator of the category of $\mathbb{C}[s]$-modules. We choose a disjoint decomposition

$$\mathbb{C}^r=\Lambda_1\uplus\Lambda_2 \text{ and } \operatorname{Ass}(\mathcal{F})=\operatorname{Max}(\mathbb{C}[s])=\mathcal{P}_1\uplus\mathcal{P}_2,\ \mathcal{P}_j:=\{\mathfrak{m}_\lambda;\ \lambda\in\Lambda_j\}.\tag{46}$$

Standard choices for continuous stability theory are

$$\Lambda_2 := \overline{\mathbb{C}_+}^r, \text{ where } \mathbb{C}_+ := \{w \in \mathbb{C};\ \Re(w) > 0\},\ \overline{\mathbb{C}_+} := \{w \in \mathbb{C};\ \Re(w) \geq 0\}, \tag{47}$$
$$\text{or [29, Def. 3.1] } \Lambda_2 := \overline{\mathbb{C}_+} \times i\mathbb{R}^{r-1}.$$

A discrete analogue is $\Lambda_2 := \{w \in \mathbb{C};\ |w| \geq 1\}^r$; see the end of section 5.

The modules $\mathcal{F}_j$, $j = 1, 2$, are $\mathcal{F}_j = \oplus_{\lambda \in \Lambda_j} \mathbb{C}[z] \exp(\lambda \bullet z)$. For $r = 1$ and time $z = z_1$ the two sets Λ_2 from (47) coincide and yield

$$\mathcal{F} = \oplus_{\lambda \in \mathbb{C}} \mathbb{C}[z] \exp(\lambda z) = \mathcal{F}_1 \oplus \mathcal{F}_2 = \oplus_{\Re(\lambda) < 0} \mathbb{C}[z] \exp(\lambda z) \oplus \oplus_{\Re(\lambda) \geq 0} \mathbb{C}[z] \exp(\lambda z).$$

Then $\mathcal{F}_1 = \{y \in \mathcal{F};\ \lim_{z \to \infty} y(z) = 0\}$. Hence

$$\lim_{z \to \infty} y_1(z) = 0 \text{for} y = y_1 + y_2 \in \mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2.$$

For practical engineering applications the stable part y_1 of y is negligible and the signal y coincides with its steady state y_2. This justifies the suggestive terminology introduced above.

We go on with the general situation. The set

$$T := \cap_{\mathfrak{p}_2 \in \mathcal{P}_2} (A \setminus \mathfrak{p}_2) \subset A \tag{48}$$

is a multiplicative submonoid of A and saturated; i.e., each divisor of an element in T belongs to T. The set T gives rise to its quotient ring

$$A_T := \left\{ \frac{a}{t};\ a \in A,\ t \in T \right\} \subset A_{\mathfrak{p}} \subset \text{quot}(A),\ \mathfrak{p} \in \mathcal{P}_2, \text{ and } (A_T)_{\mathfrak{p}_T} = A_{\mathfrak{p}}. \tag{49}$$

LEMMA 3.2. $A_T = \cap_{\mathfrak{p} \in \mathcal{P}_2} A_{\mathfrak{p}} \subset \text{quot}(A)$.

Proof. The inclusion $\subseteq$ follows from (49). For the reverse inclusion we use that A is factorial. Let $x = \frac{a}{s} \in \text{quot}(A)$ with relatively prime a, s be contained in all

$$A_{\mathfrak{p}},\ \mathfrak{p} \in \mathcal{P}_2, \text{ hence } \frac{a}{s} = \frac{a_{\mathfrak{p}}}{s_{\mathfrak{p}}},\ s_{\mathfrak{p}} \in A \setminus \mathfrak{p},\ \mathfrak{p} \in \mathcal{P}_2.$$

Then

$$a_{\mathfrak{p}} s = s_{\mathfrak{p}} a,\ \gcd(a, s) = 1 \ \Rightarrow\ \forall \mathfrak{p} \in \mathcal{P}_2 :\ s \mid s_{\mathfrak{p}}$$
$$\Rightarrow\ \forall \mathfrak{p} \in \mathcal{P}_2 :\ s \in A \setminus \mathfrak{p} \ \Rightarrow\ s \in \cap_{\mathfrak{p} \in \mathcal{P}_2} (A \setminus \mathfrak{p}) = T \text{ and } x = \frac{a}{s} \in A_T. \qquad \square$$

In the context of Gabriel's localization theory the set $\mathcal{P}_2 \subset \text{Ass}(\mathcal{F})$ and its associated injective module $\mathcal{F}_2$ give rise to a full *localizing or Serre subcategory* or *hereditary torsion class* $\mathfrak{C}$ of Mod_A and a *Gabriel topology* $\mathfrak{T}$ [24, Thm. VI.5.1], where

$$\begin{aligned} \mathfrak{C} &:= \{C \in \text{Mod}_A;\ \text{Hom}_A(C, \mathcal{F}_2) = 0\} \\ &= \{C \in \text{Mod}_A;\ \forall \mathfrak{p} \in \mathcal{P}_2 :\ \text{Hom}_A(C, E(A/\mathfrak{p})) = 0\} \\ &= \{C \in \text{Mod}_A;\ \forall \mathfrak{p} \in \mathcal{P}_2 :\ C_{\mathfrak{p}} = 0\} \text{ and} \\ \mathfrak{T} &:= \{\mathfrak{a} \subseteq A;\ A/\mathfrak{a} \in \mathfrak{C}\}. \end{aligned} \tag{50}$$

Since $\mathcal{F}_2$ is injective, the class $\mathfrak{C}$ of $\mathfrak{T}$-*torsion modules* is obviously closed under taking submodules, factor modules, extensions, and direct sums; in particular

$$\mathfrak{C} = \{C \in \text{Mod}_A;\ \forall x \in C :\ \text{ann}(x) = \{a \in A;\ ax = 0\} \in \mathfrak{T}\} \tag{51}$$

and $\mathfrak{T}$ is directed downward. Moreover, $\mathfrak{C}$ is *stable*, i.e., closed under taking injective envelopes [24, Props. VI.7.1 and VII.4.5]. If M is a finitely generated A-module and $\mathfrak{a}$ an ideal of A, then

$$\begin{gathered} M \in \mathfrak{C},\ \text{resp.},\ \mathfrak{a} \in \mathfrak{T} \\ \Leftrightarrow\ \forall \mathfrak{p} \in \mathcal{P}_2\ \exists s_{\mathfrak{p}} \in A \setminus \mathfrak{p} \text{ such that } s_{\mathfrak{p}} M = 0,\ \text{resp.},\ s_{\mathfrak{p}} \in \mathfrak{a} \cap (A \setminus \mathfrak{p}). \end{gathered} \tag{52}$$

The largest submodule of M in $\mathfrak{C}$ is called the $\mathfrak{T}$-*torsion radical* of M and denoted by $\mathrm{t}_{\mathfrak{T}}(M)$. If $\mathrm{t}_{\mathfrak{T}}(M) = 0$, the module is called $\mathfrak{T}$-*torsion free*. Due to (52) a $\mathfrak{T}$-torsion module is a torsion module in the usual sense, and hence a torsion free module is $\mathfrak{T}$-torsion free. Every subcategory $\mathfrak{C}$ with the indicated properties arises from a suitable injective module $\mathcal{F}_2$; in other words, there are no other localization functors than those described below.

An A-module M also induces its quotient module

$$M_T = \left\{ \frac{x}{t};\ x \in M,\ t \in T \right\} = A_T \otimes_A M.$$

Since $M_{\mathfrak{p}} = (M_T)_{\mathfrak{p}_T}\ \forall \mathfrak{p} \in \mathcal{P}_2$, we see that any module with $M_T = 0$ is a $\mathfrak{T}$-torsion module.

An A-module N is called $\mathfrak{T}$-*closed* if

$$N \cong \mathrm{Hom}_A(\mathfrak{a}, N),\ x \mapsto (a \mapsto ax)\ \ \forall \mathfrak{a} \in \mathfrak{T}. \tag{53}$$

The full additive subcategory of Mod_A of all $\mathfrak{T}$-closed submodules is denoted by $\mathrm{Mod}_{A,\mathfrak{T}}$. According to [24, pp. 195–200; 213–216] the inclusion functor $\mathrm{inj} : \mathrm{Mod}_{A,\mathfrak{T}} \subset \mathrm{Mod}_A$ has an exact left adjoint functor

$$\begin{gathered} \mathrm{Mod}_A \to \mathrm{Mod}_{A,\mathfrak{T}},\ M \mapsto M_{\mathfrak{T}},\ \text{with the functorial adjunction morphism} \\ \eta_M : M \to M_{\mathfrak{T}},\ \text{i.e.,}\ \mathrm{Hom}_A(M_{\mathfrak{T}}, N) \cong \mathrm{Hom}_A(M, N),\ g \mapsto g\eta_M, \end{gathered} \tag{54}$$

for $M \in \mathrm{Mod}_A$ and $N \in \mathrm{Mod}_{A,\mathfrak{T}}$, and is given by the directed colimit [24, Prop. IX.1.7]

$$\begin{gathered} M_{\mathfrak{T}} = \operatorname*{colim}_{\mathfrak{a} \in \mathfrak{T}} \mathrm{Hom}_A(\mathfrak{a}, M) = \{\alpha : \mathfrak{a} \to M;\ \mathfrak{a} \in \mathfrak{T}\}/\sim \ni [\alpha : \mathfrak{a} \to M],\ \text{where} \\ (\alpha : \mathfrak{a} \to M) \sim (\beta : \mathfrak{b} \to M)\ \ \Leftrightarrow\ \ \exists \mathfrak{c} \in \mathfrak{T} \text{ with } \mathfrak{c} \subseteq \mathfrak{a} \cap \mathfrak{b},\ \alpha \mid \mathfrak{c} = \beta \mid \mathfrak{c},\ \text{and} \\ \eta_M : M \to M_{\mathfrak{T}},\ x \mapsto [A \to M,\ a \mapsto ax]. \end{gathered} \tag{55}$$

Moreover [24, IX.1.2,1.3],

$$\mathrm{t}_{\mathfrak{T}}(M) = \ker(\eta_M) \text{ and } M \in \mathfrak{C}\ \ \Leftrightarrow\ \ M = \mathrm{t}_{\mathfrak{T}}(M)\ \ \Leftrightarrow\ \ M_{\mathfrak{T}} = 0. \tag{56}$$

The functor $M \mapsto M_{\mathfrak{T}}$ is called *localization* with respect to $\mathcal{P}_2$, $\mathfrak{C}$, or $\mathfrak{T}$. A module N is $\mathfrak{T}$-closed if and only if η_N is an isomorphism [24, Prop. IX.1.8]; in particular $M_{\mathfrak{T}} = (M_{\mathfrak{T}})_{\mathfrak{T}}$. The category $\mathrm{Mod}_{A,\mathfrak{T}}$ is itself abelian with exact directed colimits; the localization functor is exact, but the inclusion functor inj is left exact only [24, p. 214]. This means that the kernels of an A-linear map between $\mathfrak{T}$-closed modules in Mod_A (resp., $\mathrm{Mod}_{A,\mathfrak{T}}$) are the same, but that an epimorphism in $\mathrm{Mod}_{A,\mathfrak{T}}$ is not surjective in general. Indeed, if $f : N_1 \to N_2$ is an A-linear map between $\mathfrak{T}$-closed modules and $C := \mathrm{cok}_A(f) := N_1/f(N_2)$ is its cokernel in Mod_A, the exact sequence

$$\begin{gathered} N_1 \xrightarrow{f} N_2 \xrightarrow{\mathrm{can}} C \to 0 \text{ in } \mathrm{Mod}_A \text{ induces the exact sequence} \\ N_1 = (N_1)_{\mathfrak{T}} \xrightarrow{f} N_2 = (N_2)_{\mathfrak{T}} \xrightarrow{\mathrm{can}} C_{\mathfrak{T}} \to 0 \text{ in } \mathrm{Mod}_{A,\mathfrak{T}}, \end{gathered} \tag{57}$$

and hence f is an epimorphism in the category $\mathrm{Mod}_{A,\mathfrak{T}}$ if and only if C is a $\mathfrak{T}$-torsion module.

LEMMA 3.3.

1. *The modules* $\mathcal{F}_2 = \oplus_{\mathfrak{p}\in\mathcal{P}_2}\mathcal{F}(\mathfrak{p})$ *and* $\oplus_{\mathfrak{p}\in\mathcal{P}_2}E(A/\mathfrak{p})$ *are injective cogenerators of the category* $\mathrm{Mod}_{A,\mathfrak{T}}$*; in particular, as already indicated,*

$$\mathrm{Hom}_A(M,\mathcal{F}_2)\cong\mathrm{Hom}_A(M_{\mathfrak{T}},\mathcal{F}_2)=0\ \Leftrightarrow\ M_{\mathfrak{T}}=0\ \Leftrightarrow\ M\in\mathfrak{C}. \tag{58}$$

2. *These modules are also injective* A_T*-modules. Only in important special cases are they* A_T*-cogenerators; see Theorem/Definition* 5.6 *of section* 5.
3. *The module* $\mathcal{F}_T\cong\oplus_i\{\mathcal{F}_i;\ \mathfrak{p}_i\cap T=\emptyset\}$ *is* A*- and* A_T*-injective.*

Proof.

1. This is shown in [24, Prop. X.1.9].
2. Since these A_T-modules are injective A-modules, the injectivity as A_T-modules follows from the identity

$$\mathrm{Hom}_A(N,\mathcal{F}_2)=\mathrm{Hom}_{A_T}(N,\mathcal{F}_2),\ N=N_T\in\mathrm{Mod}_{A_T}.$$

3. This follows from

$$E(A/\mathfrak{p})_T=\begin{cases}E(A/\mathfrak{p}) & \text{if } \mathfrak{p}\cap T=\emptyset\\ 0 & \text{if } \mathfrak{p}\cap T\neq\emptyset\end{cases}. \qquad \square$$

Since A is Noetherian and hence the functor $\mathrm{Hom}_A(\mathfrak{a},-)$ in (53) preserves direct sums, a direct sum of A-modules is $\mathfrak{T}$-closed if and only if all summands have this property. This implies in particular that the injection functor from $\mathrm{Mod}_{A,\mathfrak{T}}$ to Mod_A also preserves direct sums and that every projective A_T-module is contained in $\mathrm{Mod}_{A,\mathfrak{T}}$ since $A_T=A_{\mathfrak{T}}$ is $\mathfrak{T}$-closed (see Lemma 3.4).

The module $M_{\mathfrak{T}}$ is always an A_T-module, i.e., the multiplication with $t\in T$ is bijective [24, p. 196], but the functors $M\mapsto M_T$ and $M\mapsto M_{\mathfrak{T}}$ coincide only in exceptional cases [24, Prop. XI.3.4], for instance, for the principal ideal ring $\mathbb{C}[s_1]$ in one-dimensional systems theory. Again in Theorem/Definition 5.6 we characterize this exceptional situation in the most important case and show its relation to *ideal-convexity* in the sense of [29]. The difficulties of [29] with the possible absence of ideal-convexity are avoided in our theory by using $M_{\mathfrak{T}}$ instead of the standard M_T. But see the following lemma.

LEMMA 3.4. $A_T=A_{\mathfrak{T}}\subset K:=\mathrm{quot}(A)$.

Proof. For $\mathfrak{a}\in\mathfrak{T}$, the map

$$\mathrm{Hom}_A(\mathfrak{a},A)\to K,\ \alpha\mapsto x_\alpha:=\alpha(a)/a,\ \text{with}\ 0\neq a\in\mathfrak{a}, \tag{59}$$

is a well-defined A-monomorphism. Indeed, if

$$\begin{gathered}0\neq a_1,a_2\in\mathfrak{a},\ \text{then}\ \alpha(a_1a_2)=a_1\alpha(a_2)=a_2\alpha(a_1);\ \text{hence}\\ \alpha(a_2)/a_2=\alpha(a_1)/a_1\ \text{and}\ \alpha(a_2)=a_2\alpha(a_1)/a_1.\end{gathered}$$

Since directed colimits are exact, the injections (59) induce an injection

$$A_{\mathfrak{T}}=\operatorname*{colim}_{\mathfrak{a}\in\mathfrak{T}}\mathrm{Hom}_A(\mathfrak{a},A)\underset{\text{ident.}}{=}\{x_\alpha=\alpha(a)/a;\ 0\neq a\in\mathfrak{a}\xrightarrow{\alpha}A\}\subset K.$$

But according to (52) and Lemma 3.2,

$$\forall\mathfrak{p}\in\mathcal{P}_2\ \exists a_{\mathfrak{p}}\in\mathfrak{a}\cap(A\setminus\mathfrak{p})\neq\emptyset;\ \text{hence}\ x_\alpha=\alpha(a_{\mathfrak{p}})/a_{\mathfrak{p}}\in\cap_{\mathfrak{p}\in\mathcal{P}_2}A_{\mathfrak{p}}=A_T,$$

and thus $A_{\mathfrak{T}} \subseteq A_T$. If, conversely, $x = a/t \in A_T$, then

$$\forall \mathfrak{p} \in \mathcal{P}_2 : \ t \in At \cap (A \setminus \mathfrak{p}); \text{ hence}$$
$$At \in \mathfrak{T}, \ \alpha : At \to A, \ t \mapsto a, \text{ and } x = a/t = \alpha(t)/t = x_\alpha \in A_{\mathfrak{T}}. \qquad \square$$

As in the preceding lemma, we define for every A-module M and prime ideal $\mathfrak{p} \in \mathcal{P}_2$ the canonical A-linear map

$$\text{can}_{\mathfrak{p}} : M_{\mathfrak{T}} \to M_{\mathfrak{p}}, \ [\alpha : \mathfrak{a} \to M] \mapsto \alpha(s_{\mathfrak{p}})/s_{\mathfrak{p}}, \text{ with } s_{\mathfrak{p}} \in \mathfrak{a} \cap (A \setminus \mathfrak{p}), \tag{60}$$

independent of the choice of $s_{\mathfrak{p}}$.

LEMMA 3.5. *The map*

$$\text{can} : M_{\mathfrak{T}} \to \prod_{\mathfrak{p} \in \mathcal{P}_2} M_{\mathfrak{p}}, \ [\alpha] \mapsto (\alpha(s_{\mathfrak{p}})/s_{\mathfrak{p}})_{\mathfrak{p} \in \mathcal{P}_2},$$

is a functorial monomorphism.

Proof. Only the injectivity has to be shown. Assume therefore that

$$\forall \mathfrak{p} \in \mathcal{P}_2 : \alpha(s_{\mathfrak{p}})/s_{\mathfrak{p}} = 0, \text{ i.e., } \exists t_{\mathfrak{p}} \in A \setminus \mathfrak{p} \text{ such that}$$
$$t_{\mathfrak{p}} \alpha(s_{\mathfrak{p}}) = \alpha(t_{\mathfrak{p}} s_{\mathfrak{p}}) = 0.$$

The ideal $\mathfrak{a} := \sum_{\mathfrak{p} \in \mathcal{P}_2} At_{\mathfrak{p}} s_{\mathfrak{p}}$ belongs to $\mathfrak{T}$ since it contains the $t_{\mathfrak{p}} s_{\mathfrak{p}}$ and, by construction, $\alpha \mid \mathfrak{a} = 0$, and hence $[\alpha] = 0$ in $M_{\mathfrak{T}}$. $\square$

A further functorial homorphism is the map

$$\text{can} : M_T \to M_{\mathfrak{T}}, \ \frac{x}{t} \mapsto [\alpha : At \to M, \ t \mapsto x], \tag{61}$$

which, in general, is neither injective nor surjective. The kernel and cokernel of the canonical map $M \to M_T$ are modules C with $C_T = 0$ and hence $C_{\mathfrak{T}} = 0$, and are thus $\mathfrak{T}$-torsion modules. The same holds for the canonical map $\eta_M : M \to M_{\mathfrak{T}}$ [24, Lemmas IX.1.2, IX.1.5]. Application of the exact functors $(-)_{\mathfrak{T}}$ and $(-)_{\mathfrak{p}}$ thus yields canonical isomorphisms

$$M_{\mathfrak{T}} \cong (M_T)_{\mathfrak{T}} \text{ and } M_{\mathfrak{p}} \cong (M_T)_{\mathfrak{p}} \cong (M_{\mathfrak{T}})_{\mathfrak{p}} \text{ for } \mathfrak{p} \in \mathcal{P}_2,$$
$$K \otimes_A M \cong K \otimes_{A_{\mathfrak{p}}} M_{\mathfrak{p}} \cong K \otimes_A M_T \cong K \otimes_A M_{\mathfrak{T}}; \tag{62}$$
$$\text{hence } \operatorname{rank}(M) = \operatorname{rank}(M_T) = \operatorname{rank}(M_{\mathfrak{T}}) = \operatorname{rank}(M_{\mathfrak{p}}) = [K \otimes_A M : K].$$

For the data from (42), the inclusion $U \subset A^{1\times l}$ induces the inclusions

$$U_T = A_T U \subset U_{\mathfrak{T}} \subset A_T^{1\times l} = A_{\mathfrak{T}}^{1\times l}. \tag{63}$$

LEMMA AND DEFINITION 3.6. *For the data from* (42)*, the submodule* $U_{\text{st}} := U_{\mathfrak{T}} \cap A^{1\times l}$ *of* $A^{1\times l}$ *is the largest one such that*

$$(U_{\text{st}})_{\mathfrak{T}} = U_{\mathfrak{T}} \text{ or } (A^{1\times l}/U_{\text{st}})_{\mathfrak{T}} = M_{\mathfrak{T}}.$$

Moreover,

$$U_{\text{st}} = \cap_{\mathfrak{p} \in \mathcal{P}_2} U_{\mathfrak{p}} \cap A^{1\times l} = \{x \in A^{1\times l}; \ \forall \mathfrak{p} \in \mathcal{P}_2 \ \exists s_{\mathfrak{p}} \in A \setminus \mathfrak{p} \text{ with } s_{\mathfrak{p}} x \in U\},$$
$$U_{\mathfrak{T}} = (U_{\text{st}})_T = A_T U_{\text{st}} \text{ and } K \otimes_A U = K \otimes_A U_{\text{st}} \subset K^{1\times l}.$$

The last equality shows that the behavior $\mathcal{B}_{\rm st} := U_{\rm st}^{\perp}$ *is in the same transfer class, or has the same controllable subbehavior, as* $\mathcal{B}$.

Proof. The exactness of $(-)_{\mathfrak{T}}$ implies

$$(U_{\rm st})_{\mathfrak{T}} = (U_{\mathfrak{T}} \cap A^{1\times l})_{\mathfrak{T}} = (U_{\mathfrak{T}})_{\mathfrak{T}} \cap A_{\mathfrak{T}}^{1\times l} = U_{\mathfrak{T}} \cap A_T^{1\times l} = U_{\mathfrak{T}}.$$
$$\text{If } U_{\mathfrak{T}} = V_{\mathfrak{T}}, \text{ then } V \subset V_{\mathfrak{T}} \cap A^{1\times l} = U_{\mathfrak{T}} \cap A^{1\times l} = U_{\rm st}.$$

Finally,

$$U_{\mathfrak{T}} = A_T(U_{\mathfrak{T}} \cap A^{1\times l}) = A_T U_{\rm st} = (U_{\rm st})_T.$$

The last equations follow from the fact that $U_{\mathfrak{T}}$ is an A_T-submodule of $A_T^{1\times l}$ and that $V = A_T(V \cap A^{1\times l})$ for each A_T-submodule of $A_T^{1\times l}$.

The injection from Lemma 3.5 and the exact sequence

$$0 \to U \xrightarrow{\rm inj} A^{1\times l} \xrightarrow{\rm can} M \to 0$$

induce the exact sequence in Mod_A,

$$0 \to U_{\mathfrak{T}} \xrightarrow{\rm inj} A_T^{1\times l} \xrightarrow{\rm can} \prod_{\mathfrak{p}\in\mathcal{P}_2} A_{\mathfrak{p}}^{1\times l}/U_{\mathfrak{p}},$$

and this implies

$$U_{\rm st} = U_{\mathfrak{T}} \cap A^{1\times l} = \cap_{\mathfrak{p}\in\mathcal{P}_2} U_{\mathfrak{p}} \cap A^{1\times l}. \qquad \square$$

Remark 3.7. If the matrix R with $U = A^{1\times k}R$ is given it is, in general, a difficult problem to compute a matrix

$$R_{\rm st} \in A^{k_1\times l} \text{ such that } U_{\rm st} = A^{1\times k_1}R_{\rm st} \text{ and hence } U_{\mathfrak{T}} = A_T^{1\times k_1}R_{\rm st}.$$

4. Abstract stability and stabilization. The assumptions and notation of the preceding section remain in force. We apply the results of the preceding sections to stability questions, first in the abstract situation and then, in the next section, to multidimensional systems governed by partial differential or difference equations.

THEOREM AND DEFINITION 4.1 (equal steady state). 1. *Let*

$$\mathcal{B} := U^{\perp} = \{w \in \mathcal{F}^l;\ R \circ w = 0\} \text{ and } \mathcal{B}' := U'^{\perp} = \{w \in \mathcal{F}^l;\ R' \circ w = 0\}$$

be two behaviors as in (42) *with their corresponding decompositions* (43) *and the submodules* $U_{\rm st}$ *(resp.,* $U'_{\rm st}$*) according to Lemma/Definition* 3.6. *Then*

$$\mathcal{B}_2 := \mathcal{B} \cap \mathcal{F}_2^l = \mathcal{B}'_2 \ \Leftrightarrow\ U_{\mathfrak{T}} = U'_{\mathfrak{T}} \ \Leftrightarrow\ U_{\rm st} = U'_{\rm st} \ \Leftrightarrow\ M_{\mathfrak{T}} = M'_{\mathfrak{T}}.$$

Then we say that the behaviors $\mathcal{B}$ *and* $\mathcal{B}'$ *have the* same steady state.

2. *The behavior*

$$\mathcal{B}_{\rm st} := U_{\rm st}^{\perp} = \{w \in \mathcal{F}^l;\ R_{\rm st} \circ w = 0\} \text{ with } U_{\rm st} = A^{1\times k_1}R_{\rm st}$$

is the smallest one with the same steady state as $\mathcal{B}$.

Proof. 1. Replacing $\mathcal{B}$ with $\mathcal{B} + \mathcal{B}'$, we assume without loss of generality that $\mathcal{B}' \subseteq \mathcal{B}$ or $U \subseteq U'$. Let $f : U \subset U'$ (resp., $g : M \to M'$) be the canonical injection

(resp., surjection). Application of the exact functor $(-)_{\mathfrak{T}}$ furnishes the commutative diagram with exact rows in $\mathrm{Mod}_{A,\mathfrak{T}}$

$$\begin{array}{ccccccc} 0 & \to & U_{\mathfrak{T}} & \subset A_T^{1\times l} & \xrightarrow{\mathrm{can}_{\mathfrak{T}}} & M_{\mathfrak{T}} & \to & 0 \\ & & \downarrow f_{\mathfrak{T}} & \downarrow \mathrm{id} & & \downarrow g_{\mathfrak{T}} & & \\ 0 & \to & U'_{\mathfrak{T}} & \subset A_T^{1\times l} & \xrightarrow{\mathrm{can}'_{\mathfrak{T}}} & M'_{\mathfrak{T}} & \to & 0, \end{array}$$

where $f_{\mathfrak{T}} : U_{\mathfrak{T}} \subset U'_{\mathfrak{T}}$ and $g_{\mathfrak{T}}$ is the canonical epimorphism. Hence $U_{\mathfrak{T}} = U'_{\mathfrak{T}}$ if and only if $g_{\mathfrak{T}} : M_{\mathfrak{T}} \to M'_{\mathfrak{T}}$ is an isomorphism. Moreover,

$$\begin{gathered} \mathrm{Hom}(g, \mathcal{F}_2) \underset{\text{ident.}}{=} \mathrm{Hom}(g_{\mathfrak{T}}, \mathcal{F}_2) \underset{\text{ident.}}{=} \mathrm{inj} : \\ \mathcal{B}'_2 \underset{\text{ident.}}{=} \mathrm{Hom}_A(M', \mathcal{F}_2) \underset{\text{ident.}}{=} \mathrm{Hom}_A(M'_{\mathfrak{T}}, \mathcal{F}_2) \\ \to \mathcal{B}_2 \underset{\text{ident.}}{=} \mathrm{Hom}_A(M, \mathcal{F}_2) \underset{\text{ident.}}{=} \mathrm{Hom}_A(M_{\mathfrak{T}}, \mathcal{F}_2). \end{gathered}$$

By Lemma 3.3 $\mathcal{F}_2$ is an injective cogenerator in $\mathrm{Mod}_{A,\mathfrak{T}}$ and hence

$$\begin{gathered} \mathcal{B}'_2 = \mathcal{B}_2 \;\Leftrightarrow\; \mathrm{inj} = \mathrm{Hom}(g_{\mathfrak{T}}, \mathcal{F}_2) \text{ is an isomorphism} \\ \Leftrightarrow\; g_{\mathfrak{T}} : M_{\mathfrak{T}} \to M'_{\mathfrak{T}} \text{ is an isomorphism} \;\Leftrightarrow\; M_{\mathfrak{T}} = M'_{\mathfrak{T}}. \end{gathered}$$

2. This is a reformulation of Lemma/Definition 3.6. □

In what follows we assume that an IO system $\mathcal{B} \subset \mathcal{F}^l$ is given with the data

$$\begin{gathered} l = p + m,\ R = (P, -Q) \in A^{k\times(p+m)},\ \mathrm{rank}(P) = \mathrm{rank}(R) = p,\ PH = Q, \\ U := A^{1\times k}R \subset A^{1\times l},\ U^0 := A^{1\times k}P \subset A^{1\times p}, \\ M := A^{1\times l}/U,\ M^0 = A^{1\times p}/U^0, \\ \mathcal{B} := U^{\perp} = \left\{ w = \begin{pmatrix} y \\ u \end{pmatrix} \in \mathcal{F}^{p+m};\ P \circ y = Q \circ u \right\}, \\ \mathcal{B}^0 := \{ y \in \mathcal{F}^p;\ P \circ y = 0 \}. \end{gathered} \tag{64}$$

Theorem and Definition 4.2 (stable systems). *For the data from* (64) *the following assertions are equivalent:*

1. $\mathcal{B}^0 \subseteq \mathcal{F}_1^p$, *i.e.,* $\mathcal{B}_2^0 = 0$ *or* $\mathcal{B}^0(\mathfrak{p}) = 0\ \forall \mathfrak{p} \in \mathcal{P}_2$. *In other words, all trajectories of the autonomous part* $\mathcal{B}^0$ *of* $\mathcal{B}$ *are stable. Recall that the signals* $y \in \mathcal{F}_1$ *are suggestively called stable.*
2. $M_{\mathfrak{T}}^0 = 0$, *i.e., for all* $\mathfrak{p} \in \mathcal{P}_2$ $M_{\mathfrak{p}}^0 = 0$ *or* $A_{\mathfrak{p}}^{1\times k}P = A_{\mathfrak{p}}^{1\times p}$ *or, in the terminology of Theorem/Definition* 2.5, *the* $A_{\mathfrak{p}}$ *IO behavior*

$$\left\{ \begin{pmatrix} y \\ u \end{pmatrix} \in \mathcal{F}(\mathfrak{p})^{p+m};\ P \circ y = Q \circ u \right\} \text{ with the } A_{\mathfrak{p}}\text{-cogenerator } \mathcal{F}(\mathfrak{p})$$

   is trivial.
3. *For all* $\mathfrak{p} \in \mathcal{P}_2$ *the module* $M_{\mathfrak{p}}$ *is torsion free, and* $H \in A_{\mathfrak{p}}^{p\times m}$.
4. *The module* $M_{\mathfrak{T}}$ *is torsion free, and* $H \in A_T^{p\times m}$.

If these conditions are satisfied, the IO system $\mathcal{B}$ *is said to be* stable with respect to

the decompositions (39) and (40), *and then the A-linear map*

$$A^{1\times m} \to M,\ \eta \mapsto \overline{(0,\eta)},\ \textit{induces the isomorphisms}$$
$$A_T^{1\times m} \cong M_{\mathfrak{T}}\ \textit{and}$$
$$\text{(65)}\quad \operatorname{proj} = (0\,\mathrm{id}_m)\circ : \mathcal{B}_2 \to \mathcal{F}_2^m,\ \begin{pmatrix} y_2 \\ u_2 \end{pmatrix} \mapsto u_2,\ \textit{with the inverse}\ u_2 \mapsto \begin{pmatrix} H\circ u_2 \\ u_2 \end{pmatrix};$$
$$\textit{hence}\ \mathcal{B} := \mathcal{B}_1 \oplus \left\{ \begin{pmatrix} H\circ u_2 \\ u_2 \end{pmatrix};\ u_2 \in \mathcal{F}_2^m \right\}.$$

Here $\circ$ denotes the scalar multiplication of $\mathcal{F}_2$ as an A_T-module. In other words, the module $M_{\mathfrak{T}}$ is A_T-free of dimension m. For every input $u_2 \in \mathcal{F}_2^m$, $\binom{H\circ u_2}{u_2}$ is the unique trajectory in $\mathcal{B}\cap\mathcal{F}_2^l$ with input u_2. We remark that H acts on $\mathcal{F}_2^m$ as an operator although no analogue of the customary properness of H is assumed.

Proof. $1 \Leftrightarrow 2$: This follows from

$$\mathcal{B}^0(\mathfrak{p}) \cong \operatorname{Hom}_A(M^0, \mathcal{F}(\mathfrak{p})) \cong \operatorname{Hom}_{A_{\mathfrak{p}}}(M^0_{\mathfrak{p}}, \mathcal{F}(\mathfrak{p}))$$

and the cogenerator property of $\mathcal{F}(\mathfrak{p})$ as an $A_{\mathfrak{p}}$-module.

$2 \Leftrightarrow 3$: This is a special case of Theorem/Definition 2.5.

$3 \Leftrightarrow 4$: The property of H follows from $A_T = \cap_{\mathfrak{p}\in\mathcal{P}_2} A_{\mathfrak{p}}$ according to Lemma 3.2. If all $M_{\mathfrak{p}}$, $\mathfrak{p} \in \mathcal{P}_2$, are torsion free, Lemma 3.5 implies the same property for $M_{\mathfrak{T}}$. Conversely, if $M_{\mathfrak{T}}$ is torsion free, then so are all $M_{\mathfrak{p}} \cong (M_{\mathfrak{T}})_{\mathfrak{p}}$, $\mathfrak{p} \in \mathcal{P}_2$, according to (62).

The condition $M^0_{\mathfrak{T}} = 0$ and the application of the exact functors

$$(-)_{\mathfrak{T}} \text{ and } \operatorname{Hom}_A(-, \mathcal{F}_2) \cong \operatorname{Hom}_A((-)_{\mathfrak{T}}, \mathcal{F}_2)$$

to the exact IO sequence

$$0 \to A^{1\times m} \overset{\circ(0\,\mathrm{id}_m)}{\longrightarrow} M \overset{\text{can}}{\longrightarrow} M^0 \to 0$$

furnishes the isomorphisms (65).

Recall that $\mathcal{F}_2$ is $\mathfrak{T}$-closed and an A_T-module in particular; hence $a\circ y$ is defined for $a \in A_T$ and $y \in \mathcal{F}_2$. The equation $PH = Q$ of matrices with coefficients in A_T hence implies $P\circ(H\circ u_2) = Q\circ u_2$ for $u_2 \in \mathcal{F}_2^m$ and therefore the asserted inverse map of the projection $\operatorname{proj} = (0\,\mathrm{id}_m)\circ : \mathcal{B}_2 \cong \mathcal{F}_2^m$. ∎

THEOREM AND DEFINITION 4.3 (stabilizable systems). *An IO system as in* (64) *is called* stabilizable with respect to the decompositions (39) and (40) *if there is another IO system $\mathcal{B}' = \left\{\binom{u'}{y'} \in \mathcal{F}^{p+m};\ P'\circ y' = Q'\circ u'\right\}$ such that the feedback system $\mathcal{B}'' := \operatorname{feedback}(\mathcal{B},\mathcal{B}')$ according to* (20) *and* 21 *is well-posed and stable in the sense of Theorems* 2.7 *and* 4.2. *If this is the case, then*

$$\text{(66)}\qquad A_T^{1\times l} = U_{\mathfrak{T}} \oplus U'_{\mathfrak{T}}\ \textit{and}\ \mathcal{F}_2^{p+m} = \mathcal{B}_2 \oplus \mathcal{B}'_2,$$

and $\mathcal{B}'$ is called a stabilizing compensator *of $\mathcal{B}$.*

Proof. According to Theorem 2.7 the direct sum decomposition

$$(\mathcal{B}''^0)^\perp = \mathcal{B}^\perp \oplus \mathcal{B}'^\perp \text{ holds and implies } (\mathcal{B}''^0)^\perp_{\mathfrak{T}} = \mathcal{B}^\perp_{\mathfrak{T}} \oplus \mathcal{B}'^\perp_{\mathfrak{T}}.$$

The stability of $\mathcal{B}''$ signifies that $\mathcal{B}''^0_2 = 0$ or $(\mathcal{B}''^0)^\perp_{\mathfrak{T}} = A_T^{1\times l}$, and hence we have the assertion. ∎

Direct decompositions as in (66) appear in almost all stabilization theories, often disguised as Bezout identities for matrices. The next theorem contains a test for stabilizability of the IO system $\mathcal{B}$ and the construction of a stabilizing compensator. It is just a reformulation of Theorem 2.13.

Let

$$U_{\rm st} = A^{1\times k_1} R_{\rm st},\ R_{\rm st} = (P_{\rm st}, -Q_{\rm st}) \in A^{k_1\times(p+m)} \tag{67}$$

be the largest submodule of

$$\begin{gathered} A^{1\times l} \text{ with } U_{\mathfrak{T}} = (U_{\rm st})_{\mathfrak{T}} = A_T^{1\times k_1} R_{\rm st};\ \text{hence} \\ A_{\mathfrak{p}}^{1\times k} R = U_{\mathfrak{p}} = (U_{\rm st})_{\mathfrak{p}} = A_{\mathfrak{p}}^{1\times k_1} R_{\rm st}\ \forall \mathfrak{p} \in \mathcal{P}_2 \end{gathered} \tag{68}$$

according to (62) and Lemma/Definition 3.6. The use of $R_{\rm st}$ is necessary since epimorphisms in $\mathrm{Mod}_{A,\mathfrak{T}}$ are not surjective, and hence $U_{\mathfrak{T}}$ is not generated over A_T by the rows of R in general. Since $U \subset U_{\rm st}$, there is a matrix X such that

$$\begin{gathered} R = XR_{\rm st};\ \text{hence } P = XP_{\rm st} \text{ and } p = \mathrm{rank}(P) \le \mathrm{rank}(P_{\rm st}) \le \mathrm{rank}(R_{\rm st}). \\ \text{Moreover, } p = \mathrm{rank}(R) = \mathrm{rank}(U) = \mathrm{rank}(U_{\rm st}) = \mathrm{rank}(R_{\rm st}), \text{ and therefore} \\ p = \mathrm{rank}(P_{\rm st}) = \mathrm{rank}(R_{\rm st}). \end{gathered}$$

Thus the behavior

$$\mathcal{B}_{\rm st} := U_{\rm st}^{\perp} = \left\{ \begin{pmatrix} y \\ u \end{pmatrix};\ P_{\rm st} \circ y = Q_{\rm st} \circ u \right\} \subset \mathcal{B} \tag{69}$$

is an IO behavior with the same IO structure and steady state as $\mathcal{B}$. Due to $K \otimes_A U = K \otimes_A U_{\rm st}$ from Lemma/Definition 3.6, $\mathcal{B}$ and $\mathcal{B}_{\rm st}$ have also the same transfer matrix H.

THEOREM 4.4 (characterization of stabilizable behaviors). *The following assertions are equivalent for the IO behavior* (64)*:*

1. *The behavior $\mathcal{B}$ is stabilizable.*
2. *$U_{\mathfrak{T}} = A_T^{1\times k_1} R_{\rm st}$ is a direct summand of $A_T^{1\times l}$.*
3. *There is a matrix*

$$\begin{gathered} G_1 \in A_T^{l\times k_1} \text{ such that } R_{\rm st} = R_{\rm st} G_1 R_{\rm st}, \text{ and then} \\ E_1 := G_1 R_{\rm st} = E_1^2 \in A_T^{l\times l} \text{ with } U_{\mathfrak{T}} = A_T^{1\times l} E_1 = U_{\mathfrak{T}}. \end{gathered}$$

4. *With G_1 and E_1 from item 3 there is a matrix*

$$\begin{gathered} X \in A_T^{l\times k_1} \text{ with } R_{\rm st} X R_{\rm st} = 0 \text{ such that } E := E_1 + XR_{\rm st} = E^2 \in A^{l\times l}, \\ U_{\mathfrak{T}} = A_T^{1\times l} E \text{ and } \mathrm{rank}\left((\mathrm{id}_l - E) \begin{pmatrix} 0 \\ \mathrm{id}_m \end{pmatrix} \right) = m. \end{gathered}$$

If these conditions are satisfied, choose $E \in A_T^{l\times l}$ according to item 4. Let $t \in T$ be 1 or the greatest common divisor of the denominators of E. Define

$$\begin{gathered} R' := (-Q'P') := t(\mathrm{id}_l - E) \in A^{l\times(p+m)};\ \text{hence} \\ \mathrm{rank}(R') = l - \mathrm{rank}(E) = l - p = m = \mathrm{rank}(P') \text{ and} \\ \mathcal{B}' := \left\{ \begin{pmatrix} u' \\ y' \end{pmatrix} \in \mathcal{F}^{p+m};\ P' \circ y' = Q' \circ u' \right\}. \end{gathered}$$

Then $\mathcal{B}'$ is a stabilizing compensator *of $\mathcal{B}$, i.e., $\mathcal{B}'$ is an IO behavior with input u', and the feedback system $\mathcal{B}'' = \mathrm{feedback}(\mathcal{B}, \mathcal{B}')$ is well-posed and stable.*

Proof. $1 \Rightarrow 2$: This follows from Theorem 4.3.

$2 \Leftrightarrow 3$: This is an application of Theorem 2.13(1) to the A_T-module $A_T^{1\times l}/U_{\mathfrak{T}}$.

$3 \Rightarrow 4$: This implication follows from Theorem 2.14.

$4 \Rightarrow 1$: Construct R' and $\mathcal{B}'$ as indicated. The equalities $\mathrm{rank}(R') = \mathrm{rank}(P') = m$ show that $\mathcal{B}'$ is an IO system with the desired IO structure. By construction,

$$U_{\mathfrak{T}} = A_T^{1\times k_1} R_{\mathrm{st}} = A_T^{1\times l} E,\ p = \mathrm{rank}(U) = \mathrm{rank}(E),$$
$$U' := \mathcal{B}'^{\perp} = A^{1\times l} t(\mathrm{id}_l - E) \ \Rightarrow\ U'_{\mathfrak{T}} = A_T^{1\times l} t(\mathrm{id}_l - E) = A_T^{1\times l}(\mathrm{id}_l - E),$$
$$U \cap U' \subseteq U_{\mathfrak{T}} \cap U'_{\mathfrak{T}} = 0 \ \Rightarrow\ A^{1\times(k+l)} \begin{pmatrix} P & -Q \\ -Q' & P' \end{pmatrix} = U \oplus U'.$$

The last equation shows that $\mathcal{B}''$ is well-posed with

$$(\mathcal{B}''^0)^{\perp} = U \oplus U' \text{ and } (\mathcal{B}''^0)^{\perp}_{\mathfrak{T}} = U_{\mathfrak{T}} \oplus U'_{\mathfrak{T}} = A_T^{1\times l} E \oplus A_T^{1\times l}(\mathrm{id}_l - E) = A_T^{1\times l}.$$

The last equation signifies the stability of $\mathcal{B}''$. ∎

Remark 4.5. Condition 2 of the preceding theorem is equivalent to the A_T-projectivity of $A_T^{1\times l}/U_{\mathfrak{T}}$, and then this module coincides with $M_{\mathfrak{T}}$. This does not imply that U_T is a direct summand of $A_T^{1\times l}$ or that M_T is A_T-projective. If, however, this is the case, then condition 2 follows and $U_T = U_{\mathfrak{T}}$, $M_T = M_{\mathfrak{T}}$. We remark that, in general, A_T is not projective in $\mathrm{Mod}_{A,\mathfrak{T}}$ since epimorphisms are not surjective in this category. Therefore condition 2 has been expressed without using the notion of projectivity.

THEOREM 4.6 (parametrization of stabilizing controllers). *Assume that $\mathcal{B}$ from* (64) *is stabilizable and that E_1 is the matrix from item* 3 *of the preceding theorem. The steady state parts $\mathcal{B}'_2 \subset \mathcal{F}_2^l$ of the stabilizing compensators $\mathcal{B}'$ of $\mathcal{B}$ are uniquely determined by the idempotent matrices*

$$E = E^2 \in A_T^{l\times l} \ \textit{with}\ \mathrm{rank}\left((\mathrm{id}_l - E)\begin{pmatrix} 0 \\ \mathrm{id}_m \end{pmatrix}\right) = m,$$
$$\mathcal{B}^{\perp}_{\mathfrak{T}} = A_T^{1\times l} E,\ \mathcal{B}'^{\perp}_{\mathfrak{T}} = A_T^{1\times l}(\mathrm{id}_l - E),$$

and these are parametrized by the bijection of Theorem 2.14 *applied to the direct summand $\mathcal{B}^{\perp}_{\mathfrak{T}}$ of $A_T^{1\times l}$ or the projective A_T-module $A_T^{1\times l}/\mathcal{B}^{\perp}_{\mathfrak{T}}$.*

5. Continuous and discrete stability and stabilization. We specialize the results of the preceding section to complex partial differential equations on $\mathbb{R}^r \ni z$; i.e., we assume $A := \mathbb{C}[s] = \mathbb{C}[s_1, \dots, s_r]$, the data from (64), and Example 3.1 with an arbitrary decomposition

$$\mathbb{C}^r = \Lambda_1 \uplus \Lambda_2,\ \mathcal{F} := \mathcal{D}'(\mathbb{R}^r)_{\mathrm{lf}},\ \mathcal{F}(\lambda) := \mathbb{C}[z]\exp(\lambda \bullet z) \cong E(\mathbb{C}[s]/\mathfrak{m}_\lambda),$$
$$\mathrm{Ass}(\mathcal{F}) = \mathrm{Max}(\mathbb{C}[s]) = \mathcal{P}_1 \uplus \mathcal{P}_2,\ \mathcal{P}_j := \{\mathfrak{m}_\lambda;\ \lambda \in \Lambda_j\}, \tag{70}$$
$$\mathcal{F} = \oplus_{\lambda\in\mathbb{C}^r} \mathcal{F}(\lambda) = \mathcal{F}_1 \oplus \mathcal{F}_2,\ \mathcal{F}_1 := \oplus_{\lambda\in\Lambda_1} \mathcal{F}(\lambda),\ \mathcal{F}_2 := \oplus_{\lambda\in\Lambda_2} \mathcal{F}(\lambda).$$

Standard examples for such decompositions are those from (47).

The main goal of this section is to show that for these data the stable systems according to Theorem 4.2 have various stability properties known from the one-dimensional theory, and that therefore the stability terminology and the abstract

theorems of the preceding section are justified. At the end of this section we explain the necessary modifications for the discrete case.

For any polynomial ideal $\mathfrak{a} \subset \mathbb{C}[s]$ we define its *algebraic variety* as usual as

$$\begin{aligned} V(\mathfrak{a}) &:= \{\lambda \in \mathbb{C}^r;\ \forall t \in \mathfrak{a}:\ t(\lambda) = 0\}, \\ \text{in particular } V(t) &:= V(\mathbb{C}[s]t) = \{\lambda \in \mathbb{C}^r;\ t(\lambda) = 0\}. \end{aligned} \tag{71}$$

The maximal ideals of $A = \mathbb{C}[s]$ are the

$$\begin{gathered} \mathfrak{m}_\lambda := \{t \in \mathbb{C}[s];\ t(\lambda) = 0 \text{ or } \lambda \in V(t)\},\ \lambda \in \mathbb{C}^r;\ \text{hence} \\ \lambda \in V(\mathfrak{a}) \ \Leftrightarrow\ \mathfrak{a} \subseteq \mathfrak{m}_\lambda, \end{gathered}$$

and the multiplicative monoid T from (48) is

$$\begin{aligned} T &:= \cap_{\lambda \in \Lambda_2}(A \setminus \mathfrak{m}_\lambda) = \{t \in A;\ \forall \lambda \in \Lambda_2:\ t(\lambda) \neq 0\} \\ &= \{t \in A;\ V(t) \subset \Lambda_1\};\ \text{hence}\ \lambda \in \Lambda_2\ \Rightarrow\ \mathfrak{m}_\lambda \cap T = \emptyset. \end{aligned} \tag{72}$$

The quotient rings $A_\mathfrak{p}$, $\mathfrak{p} \in \mathcal{P}_2$, and A_T are

$$\begin{aligned} A(\lambda) &:= A_{\mathfrak{m}_\lambda} = \left\{\frac{a}{t} \in \mathbb{C}(s);\ t(\lambda) \neq 0\right\}, \\ A_T &= \left\{\frac{a}{t} \in \mathbb{C}(s);\ \forall \lambda \in \Lambda_2:\ t(\lambda) \neq 0\right\}. \end{aligned} \tag{73}$$

Quotient rings A_T of this type are customarily used in stabilization theory as rings of SISO-stable plants [26], [25]. From (70) and (43) we further obtain direct decompositions of $\mathcal{B}$ and likewise of $\mathcal{B}^0$:

$$\begin{gathered} \mathcal{B} := \oplus_{\lambda \in \mathbb{C}^r} \mathcal{B}(\lambda) = \mathcal{B}_1 \oplus \mathcal{B}_2,\ \text{where} \\ \mathcal{B}_1 := \mathcal{B} \cap \mathcal{F}_1^l,\ \mathcal{B}_2 := \mathcal{B} \cap \mathcal{F}_2^l,\ \mathcal{B}(\lambda) := \mathcal{B} \cap \mathcal{F}(\lambda)^l \\ \underset{\text{ident.}}{=} \operatorname{Hom}_{\mathbb{C}[s]}(M, \mathcal{F}(\lambda)) \underset{\text{ident.}}{=} \operatorname{Hom}_{A(\lambda)}(M_{\mathfrak{m}_\lambda}, \mathcal{F}(\lambda)). \end{gathered} \tag{74}$$

For the system $\mathcal{B}$ from (64) its variety $\operatorname{sing}(\mathcal{B})$ of *rank singularities* (resp., the *characteristic variety* $\operatorname{char}(M^0) = \operatorname{char}(\mathcal{B}^0)$) of its autonomous part are [11, Thm. 7.69, Cor. 7.71, Rem. 7.72, Cor. 7.78], [28, Thm. 4.4]

(75)

$$\begin{aligned} \operatorname{sing}(\mathcal{B}) &:= \{\lambda \in \mathbb{C}^r;\ M_{\mathfrak{m}_\lambda} \text{ is not free}\} = \{\lambda \in \mathbb{C}^r;\ \operatorname{rank}(R(\lambda)) < \operatorname{rank}(R) = p\}, \\ \operatorname{char}(\mathcal{B}^0) &:= \operatorname{char}(M^0) := \operatorname{sing}(\mathcal{B}^0) = \{\lambda \in \mathbb{C}^r;\ M^0_{\mathfrak{m}_\lambda} \neq 0\} \\ &= \{\lambda \in \mathbb{C}^r;\ \operatorname{rank}(P(\lambda)) < \operatorname{rank}(P) = p\} = V(\operatorname{ann}(M^0)), \\ &\text{where } \operatorname{ann}(M^0) := \{t \in A;\ tM^0 = 0\}. \end{aligned}$$

We remark that the characteristic variety of a nonautonomous system $\mathcal{B}$ coincides with $\mathbb{C}^r$ since $\operatorname{ann}(M) = 0$. The elements of $\operatorname{char}(\mathcal{B}^0)$ are called the *modes* or *poles* or *characteristic values* of $\mathcal{B}$. The decomposition (74) implies

$$\mathcal{B}^0 = \oplus_{\lambda \in \operatorname{char}(\mathcal{B}^0)} \mathcal{B}(\lambda). \tag{76}$$

We further need the controllable subbehavior $\mathcal{B}_{\text{cont}}$ of $\mathcal{B}$, its autonomous part $\mathcal{B}^0_{\text{cont}}$, and their dual modules M_{cont} (resp., M^0_{cont}). There are the exact sequences [28, diagram (5.1)]

$$\begin{gathered} 0 \to t(M) \to M^0 \xrightarrow{\text{can}} M^0_{\text{cont}} \to 0\ \text{ and hence, } \forall \lambda \in \Lambda_2, \\ 0 \to t(M)_{\mathfrak{m}_\lambda} = t(M_{\mathfrak{m}_\lambda}) \to M^0_{\mathfrak{m}_\lambda} \xrightarrow{\text{can}} M^0_{\text{cont},\mathfrak{m}_\lambda} \to 0, \end{gathered} \tag{77}$$

where $t(M)$ is the torsion submodule of M.

These exact sequences and (16) directly imply [28, Thm. 5.3, Cor. 5.4]

$$\begin{aligned}\operatorname{char}(\mathcal{B}^0_{\mathrm{cont}}) &= \text{set of poles of } H \subset \operatorname{char}(\mathcal{B}^0) = \operatorname{char}(\mathcal{B}^0_{\mathrm{cont}}) \cup \operatorname{char}(t(M)),\\ &\text{where } \operatorname{char}(t(M)) = \{\lambda \in \mathbb{C};\ t(M)_{\mathfrak{m}_\lambda} \neq 0\}.\end{aligned} \tag{78}$$

The modes or poles of $\mathcal{B}$ in $\operatorname{char}(\mathcal{B}^0_{\mathrm{cont}})$ are called *controllable* and coincide with the poles of H. In the trivial one-dimensional example $s_1^2 \circ y = s_1 \circ u$ we have

$$t(M) \cong A/As_1 \neq 0 \text{ and } \operatorname{char}(\mathcal{B}^0) = \operatorname{char}(\mathcal{B}^0_{\mathrm{cont}}) = \operatorname{char}(t(M)) = \{0\}.$$

The following simple theorem characterizes stable behaviors and is related to the principal Theorem 7.2 from [29] which has, however, been shown only for the cases $p = k = 1$ or $r \leq 2$ or for regions Λ_2 which are ideal-convex (see Theorem/Definition 5.6 and Example 5.7 below).

THEOREM AND DEFINITION 5.1 (Λ_2-stability). *For the behavior $\mathcal{B}$ from* (64) *and the decomposition* (70) *the following statements are equivalent:*

1. *$\mathcal{B}$ is stable in the sense of Theorem/Definition* 4.2*, i.e., $\mathcal{B}^0 \subset \mathcal{F}_1^p$ or $M^0_{\mathfrak{m}_\lambda} = 0$ $\forall \lambda \in \Lambda_2$.*
2. *$\operatorname{char}(\mathcal{B}^0) \subset \Lambda_1$ or $\operatorname{char}(\mathcal{B}^0) \cap \Lambda_2 = \emptyset$.*
3. *$M^0_{\mathrm{cont},\mathfrak{m}_\lambda} = t(M_{\mathfrak{m}_\lambda}) = 0$ $\forall \lambda \in \Lambda_2$ or, in other words,*
   (i) *$\operatorname{char}(\mathcal{B}^0_{\mathrm{cont}}) \cap \Lambda_2 = \emptyset$.*
   (ii) *For all $\lambda \in \Lambda_2$ the module $M_{\mathfrak{m}_\lambda}$ is torsion free.*
4. (i) *The transfer matrix H is stable; i.e., by definition, $H \in A_T^{p\times m}$.*
   (ii) *For all $\lambda \in \Lambda_2$ the module $M_{\mathfrak{m}_\lambda}$ is torsion free or the $\mathcal{F}(\lambda)$ behavior $\mathcal{B}(\lambda) := \left\{ \binom{y}{u} \in \mathcal{F}(\lambda)^{p+m};\ P \circ y = Q \circ u \right\}$ is controllable.*

Then $\mathcal{B}$ is called Λ_2-stable.

Proof. $1 \Leftrightarrow 2$ according to definition (75).

$1, 2 \Leftrightarrow 3$ follows from the exactness of the sequences in (77) and the equivalence $1 \Leftrightarrow 2$ applied to $\mathcal{B}^0_{\mathrm{cont}}$ instead of $\mathcal{B}^0$.

$1 \Leftrightarrow 4$ is a special case of Theorem/Definition 4.2, $1 \Leftrightarrow 4$. ∎

Remarks 5.2.

1. Once the data from (70) have been chosen we call the region Λ_1, the polynomials in T, the rational functions in A_T, and the polynomial-exponential functions in $\mathcal{F}_1$ *stable* and Λ_2 the *unstable* region or domain of instability [25, p. 1692]. Stable rational functions are those without poles in the unstable region. Of course, not all unstable regions Λ_2 are of practical importance.
2. The preceding theorem shows that stability of $\mathcal{B}$ in the sense of this paper is characterized both by properties 1 and 2 of its autonomous part, and by the stability of the transfer matrix H together with 4(ii), interpreted as *external* stability due to the following theorems.
3. Trivial systems satisfying $\mathcal{B}^0 = 0$ are stable for each Λ_2.
4. A one-dimensional IO system ($r = 1$) is stable with respect to $\Lambda_1 := \{z \in \mathbb{C};\ \Re(z) < 0\}$ if and only if its autonomous part is asymptotically stable or, equivalently, if its transfer matrix is stable and its singular variety is also contained in Λ_1. In this case, condition 4(ii) is equivalent to

$$\operatorname{sing}(\mathcal{B}) \cap \Lambda_2 = \emptyset \text{ or } \operatorname{rank}(R(\lambda)) = \operatorname{rank}(R)\ \forall \lambda \in \Lambda_2$$

since torsion freeness and freeness of $M_{\mathfrak{m}_\lambda}$ coincide in dimension 1. Recall that $\mathcal{B}$ is controllable if and only if $\operatorname{sing}(\mathcal{B}) = \emptyset$. Lemma 17 from [26, p. 103] is a special case of the implication $4 \Rightarrow 1$ of Theorem/Definition 5.1.

5. In most papers on multidimensional stabilization the systems are discrete and described by a transfer operator or IO map H, and stability signifies just the stability of H [26], [6], [23], [25], [31], [32], [7], [8], [17], [18], so our stability notion is stronger. We remark, however, that for multidimensional systems and nonproper H this transfer matrix cannot, in general, be considered as an operator on the usual function spaces, and therefore properness of H is usually required in the literature. We do not assume this since properness is rather restrictive for multidimensional systems in contrast to one-dimensional systems. For instance, none of the standard equations

$$P \circ y = u,\ P = s_1 + s_2, s_1 - s_2^2, s_1^2 + / - s_2^2$$

has a proper transfer function.
6. If the first independent variable z_1 is distinguished as time and if $\Lambda_2 = \mathbb{C}_+ \times i\mathbb{R}^{r-1}$, Wood, Sule, and Rogers [29, Def. 3.1] call the condition

$$\mathrm{char}(\mathcal{B}^0) \cap \Lambda_2 = \emptyset \ \ (\text{resp.},\ \mathrm{char}(\mathcal{B}^0_{\mathrm{cont}}) \cap \Lambda_2 = \emptyset)$$

the *CV condition* for the autonomous behavior $\mathcal{B}^0$ (resp., $\mathcal{B}^0_{\mathrm{cont}}$) and consider it as the stability condition for autonomous behaviors [29, p. 1500].
7. For the unstable regions Λ_2 from (47) and $r > 1$ the system $\mathcal{B}^0 \subset \mathcal{F}_1^p$ contains, in general, many polynomial-exponential functions which are not stable in a naive sense. This is unavoidable due to the interplay of conditions 2 and 4(i) of the theorem. Indeed, with increasing Λ_1 the sets of stable polynomial-exponential functions, polynomials, and rational functions also grow; hence the generally desired existence of sufficiently many stable rational functions implies the same for the stable polynomial-exponential functions. For instance, for $\Lambda_1 = \emptyset$ a stable system is trivial, whereas for $\Lambda_1 = \mathbb{C}^r$ each system is stable.
8. Consider $\Lambda_2 := \{(\lambda_1, \lambda_2) \in \mathbb{C}^2;\ \Re(\lambda_1), \Re(\lambda_2) \geq 0\}$ and the simple system

$$(s_1 + 1) \circ y = u$$

with the characteristic variety

$$\mathrm{char}(\mathcal{B}^0) = \{-1\} \times \mathbb{C} \text{ and } \mathcal{B}^0 = \oplus_{\lambda_2 \in \mathbb{C}} \mathbb{C}[z_2] \exp(-z_1) \exp(\lambda_2 z_2).$$

This is obviously Λ_2-stable and $\mathcal{B}^0$ contains the functions $f(z) = \exp(-z_1)g(z_2)$ with $g(z_2) := \exp(\lambda_2 z_2)$, $\Re(\lambda_2) > 0$, which grow exponentially with increasing z_2 and are not stable in a naive sense. However, they share this property with their *initial part* $f(0, z_2) = g(z_2)$ and such initial conditions are not permitted according to [29, pp. 1499–1500] and the idea that a *stable system should generate stable outputs for stable inputs and initial conditions.* This requires that the initial value problem can be formulated and uniquely solved. Open Problem 5.13 below addresses this problem for discrete systems.

We are going to show next that a stable system is also *IO stable* in the sense that inputs of various types generate outputs of the same type. Since the used signals are not necessarily bounded, we do not use the acronym BIBO.

The modules $\mathcal{F}(\lambda) = \mathbb{C}[z] \exp(\lambda \bullet z) = E(A/\mathfrak{m}_\lambda)$ (resp., $\mathcal{F}_2$) are $A(\lambda)$-modules (resp., A_T-modules). The module $\mathcal{F}(\lambda)$ has the canonical increasing filtration of

finite-dimensional $A(\lambda)$-submodules [12, Eq. 1.17, Thm. 1.25, Thm. 6.6]

$$\begin{gathered}\mathcal{F}(\lambda) = \cup_{k=0}^{\infty} \mathcal{F}(\lambda)_k \text{ with} \\ \mathcal{F}(\lambda)_k := \{y \in \mathcal{D}'(\mathbb{R}^r);\ \mathfrak{m}_\lambda^{k+1} \circ y = 0\} = \mathbb{C}[z]_{\le k} \exp(\lambda \bullet z), \text{ where} \\ \mathbb{C}[z]_{\le k} := \{f \in \mathbb{C}[z];\ \deg(f) \le k\}\end{gathered} \tag{79}$$

and deg denotes the total degree of a polynomial in $\mathbb{C}[z_1, \dots, z_r]$.

Theorem and Definition 5.3 (IO stability).

1. *Let $t \in A \setminus \mathfrak{m}_\lambda$ or $t(\lambda) \neq 0$ and $y \in \mathcal{F}(\lambda)_k$. Then $\bar{t}$ is invertible in the $\mathbb{C}$ finite-dimensional local ring $A/\mathfrak{m}_\lambda^{k+1}$ or $At + \mathfrak{m}_\lambda^{k+1} = A$. Via Gröbner bases one constructs $t_1 \in A$ such that*

$$t_1 t \equiv 1 (\mathfrak{m}_\lambda^{k+1}) \text{ or } \overline{t_1 t} = 1 \in A/\mathfrak{m}_\lambda^{k+1}.$$

Then the differential operator $t\circ : \mathcal{F}(\lambda)_k \to \mathcal{F}(\lambda)_k$ is bijective and its inverse is the differential operator $t_1\circ : \mathcal{F}(\lambda)_k \to \mathcal{F}(\lambda)_k$. Therefore the scalar multiplication with $\frac{a}{t} \in A(\lambda) = \mathbb{C}[s]_{\mathfrak{m}_\lambda}$ on $\mathcal{F}(\lambda)_k$ is given by the differential operator

$$\frac{a}{t}\circ : \mathcal{F}(\lambda)_k \to \mathcal{F}(\lambda)_k,\ y \mapsto (at_1) \circ y.$$

2. *If $H \in A_T^{p\times m}$ is a stable rational matrix and $u \in \mathcal{F}_2^m = \oplus_{\lambda\in\Lambda_2} \mathbb{C}[z]^m \exp(\lambda \bullet z)$ is an input of the form*

$$u = \sum_{\lambda\in\Lambda_2} f_\lambda(z) \exp(\lambda \bullet z) \text{ with } f_\lambda \in \mathbb{C}[z]^m_{\le k}\ \forall \lambda \in \Lambda_2$$

and $f_\lambda = 0$ for almost all λ, then the corresponding output $y := H \circ u$ has the same form

$$y = \sum_{\lambda\in\Lambda_2} g_\lambda(z) \exp(\lambda \bullet z) \text{ with } g_\lambda \in \mathbb{C}[z]^p_{\le k}\ \forall \lambda \in \Lambda_2$$

and can be computed with the algorithm from part 1. *This property of the operator $H\circ : \mathcal{F}_2^m \to \mathcal{F}_2^p$ is suggestively called its* IO stability *with respect to the decomposition* (70).

3. *Item* 2 *is applicable to any stable system $\mathcal{B}$ according to Theorem/Definition* 5.1 *with its stable transfer matrix H; in particular $\mathcal{B}_2 := \mathcal{B} \cap \mathcal{F}_2^l = \left\{\binom{H\circ u}{u};\ u \in \mathcal{F}_2^m\right\}$ is its steady state part.*

Proof. Item 1 follows from

$$tt_1 + a = 1,\ a \in \mathfrak{m}_\lambda^{k+1},\ \mathfrak{m}_\lambda^{k+1} \circ y = 0 \ \Rightarrow\ t_1 \circ t \circ y = y.$$

Items 2 and 3 are direct consequences of 1 and Theorem/Definition 4.2. ∎

The next IO stability theorems are just reformulations of the results [4, Thm. on p. 16 and Thm. 2 on p. 24]. These results also play an essential part in the stability paper [29, subsections 5, 6, 7]. As in [4, pp. 10–12] let $\mathcal{S}$ ($\supset C_0^\infty(\mathbb{R}^r)$) denote the space of rapidly decreasing C^∞ functions and $\mathcal{S}'$ ($\subset \mathcal{D}'(\mathbb{R}^r)$) its topological dual space of temperate distributions. Further we consider the space $\mathcal{O}$ ($\supset \mathcal{S}$) of slowly increasing C^∞ functions f for which there is an index $m > 0$ and $C > 0$ such that

$$|f^{(\mu)}(z)| \le C(1 + |z|)^m\ \forall\ z \in \mathbb{R}^r,\ \mu \in \mathbb{N}^r,$$

and finally the space $\mathcal{O}'$ ($\subset \mathcal{S}'$) of rapidly decreasing distributions. In the paper [29] the first variable $t := z_1$ is distinguished as time. In this case the spaces

$$\mathbb{R}^r_+ := \{z \in \mathbb{R}^r;\ t = z_1 \geq 0\} \text{ and}$$
$$\mathcal{F}_{3,+} := \{y \in \mathcal{F}_3;\ \operatorname{supp}(y) \subset \mathbb{R}^r_+\},\ \mathcal{F}_3 = \mathcal{S}, \mathcal{S}', \mathcal{O}, \mathcal{O}',$$

play an important part in [4] and [29]. In the next theorem we use the data from (70)–(79).

THEOREM 5.4 (IO stability).

1. *Let $\mathcal{F}_3$ be one of the spaces $\mathcal{S}, \mathcal{S}', \mathcal{O}, \mathcal{O}'$. If $\Lambda_2 \supset i\mathbb{R}^r$, then $\mathcal{F}_3$ is an A_T-module. In particular, if $H \in A_T^{p\times m}$ is a rational matrix and $u \in \mathcal{F}_3^m$ is an input, then $y := H \circ u \in \mathcal{F}_3^p$ is an output of the same type. If $\mathcal{B}$ is stable, then according to Theorem/Definition* 5.1 *this holds.*

2. *If the variable $t := z_1$ is distinguished and if*

$$\Lambda_2 \supset \overline{\mathbb{C}_+} \times i\mathbb{R}^{r-1},\ \overline{\mathbb{C}_+} := \{\lambda \in \mathbb{C};\ \Re(\lambda) \geq 0\}$$

as in [29]*, then the statement of item* 1 *applies to the spaces $\mathcal{S}_+$ and $(\mathcal{O}')_+$.*

If, in addition, $\mathcal{B}^0$ is time-autonomous in the sense of [29, *Def.* 2.1, *Thm.* 2.2]*, then*

$$\mathcal{F}_3^m \cong \mathcal{B} \cap ((\mathcal{D}')^p_+ \times \mathcal{F}_3^m),\ u \mapsto \begin{pmatrix} H \circ u \\ u \end{pmatrix},\ \mathcal{F}_3 = \mathcal{S}_+, (\mathcal{O}')_+. \tag{80}$$

3. *If in the situation of item* 2

$$\Lambda_2 \supset \mathbb{C}_+ \times i\mathbb{R}^{r-1},\ \mathbb{C}_+ := \{\lambda \in \mathbb{C};\ \Re(\lambda) > 0\},$$

then the statement of item 1 *applies to the space $(\mathcal{S}')_+$.*

Proof. 1. Since $i\mathbb{R}^r \subset \Lambda_2$, each $t \in T$ has no zero in $i\mathbb{R}^r$. According to [4, Thm. on p. 16] this is necessary and sufficient for $t\circ : \mathcal{F}_3 \to \mathcal{F}_3$ to be bijective, and this in turn implies that $\mathcal{F}_3$ is an A_T-module with the scalar multiplication

$$\frac{a}{t} \circ u =: y \text{ with } t \circ y = a \circ u.$$

2, 3. These assertions follow from [4, Thm. 2, p. 22] in the same fashion. It remains to show the surjectivity of (80). Assume that

$$u \in \mathcal{F}_3^m,\ y := H \circ u,\ y_1 \in (\mathcal{D}')^p_+, \text{ and } \begin{pmatrix} y_1 \\ u \end{pmatrix} \in \mathcal{B}, \text{ i.e., } P \circ y_1 = Q \circ u. \text{ Then}$$
$$y_1 - y \in (\mathcal{D}')^p_+ \text{ and } P \circ (y_1 - y) = 0.$$

The time-autonomy and [29, Thm. 2.2] imply $y_1 = y = H \circ u$ and hence the assertion. ∎

Remarks 5.5.

1. The preceding theorem suggests to require $i\mathbb{R}^r \subset \Lambda_2$ for the decompositions (70) in context with stability questions of partial differential equations. The standard set

$$\Lambda_2 := \{\lambda \in \mathbb{C}^r;\ \forall \rho = 1, \ldots, r:\ \Re(\lambda_\rho) \geq 0\} = \overline{\mathbb{C}_+}^r$$

satisfies the assumptions of items 1–3.

2. The condition $\binom{H\circ u}{u} \in \mathcal{B}\cap\mathcal{F}_3^{p+m}$ implies in particular that the input u can be freely chosen in $\mathcal{F}_3^m$ and therefore condition 1 of Theorem 7.2 of [29] which, however, was shown only for $p = m = 1$ or $r \le 2$. The isomorphism (80) is the IO stability of $\mathcal{B}$ according to [29, Def. 7.1]. Theorem/Definition 5.1 and Theorem 5.4 show that stability in the sense of the present paper for $\Lambda_2 = \mathbb{C}_+ \times i\mathbb{R}^{r-1}$ implies the various stability notions of that paper.

The next theorem characterizes those sets $\Lambda_2 \subset \mathbb{C}^r$ for which the localization functor $M \mapsto M_{\mathfrak{T}}$ from section 3 is *perfect*, i.e., isomorphic to $M \mapsto M_T$. It turns out that this coincides with *ideal-convexity* of Λ_2 in the sense of [23, Def. on p. 25] or [29, Def. 5.4]. Recall from (62) that $M_T = 0$ implies $M_{\mathfrak{T}} = (M_T)_{\mathfrak{T}} = 0$, in particular $\{\mathfrak{a};\ \mathfrak{a}\cap T \neq \emptyset\} \subseteq \mathfrak{T}$.

THEOREM AND DEFINITION 5.6 (perfect localization and ideal-convexity). *For the decomposition* (70) *and* $A := \mathbb{C}[s]$ *the following assertions are equivalent:*

1. *An A-module is T-torsion if and only if it is $\mathfrak{T}$-torsion, i.e.,*

$$M_T = 0 \ \Leftrightarrow\ M_{\mathfrak{T}} = 0 \underset{(50)}{\Leftrightarrow} \forall\lambda\in\Lambda_2:\ M_{\mathfrak{m}_\lambda} = 0.$$

2. $\mathfrak{T} = \{\mathfrak{a};\ \mathfrak{a}$ *ideal of* A, $T\cap\mathfrak{a} \neq 0\}$.
3. *For each* $M \in \mathrm{Mod}_A$ *the canonical map* (61) $M_T = A_T \otimes_A M \to M_{\mathfrak{T}}$ *is an isomorphism.*
4. $\mathrm{Mod}_{A_T} = \mathrm{Mod}_{A,\mathfrak{T}}$.
5. *The set* Λ_2 *is* ideal-convex *according to* [29, *Def.* 5.4]*; i.e., for each ideal* $\mathfrak{a}$ *of* A *the implication*

$$V(\mathfrak{a})\cap\Lambda_2 = \emptyset \ \Rightarrow\ \mathfrak{a}\cap T \neq \emptyset \text{ or } \exists t\in\mathfrak{a} \text{ with } V(t)\cap\Lambda_2 = \emptyset$$

   holds.
6. *The ideals* $(\mathfrak{m}_\lambda)_T \subset A_T$, $\lambda\in\Lambda_2$, *are the only maximal ideals of* A_T.
7. *The module* $\mathcal{F}_2$ *from* (70) *is a cogenerator in* Mod_{A_T}. *Recall from Lemma* 3.3 *that* $\mathcal{F}_2$ *is an injective A_T-module and an injective cogenerator in* $\mathrm{Mod}_{A,\mathfrak{T}}$.

Stenström [24, *Chap.* XI, *Prop.* 3.4] *talks about* perfect localization *in this context.*

Proof. The equivalence of items 1–4 follows from [24, Chap. XI, Prop. 3.4].

$2 \Leftrightarrow 5$ follows from the equivalences

$$V(\mathfrak{a})\cap\Lambda_2 = \emptyset \ \Leftrightarrow\ \forall\lambda\in\Lambda_2:\ \mathfrak{a}\not\subset\mathfrak{m}_\lambda \text{ or } (A/\mathfrak{a})_{\mathfrak{m}_\lambda} = 0 \underset{(50)}{\Leftrightarrow} \mathfrak{a}\in\mathfrak{T}.$$

The equivalence of items 5 and 6 is a consequence of [23, Prop. 3.1.19], but we give the short proof for completeness.

$5 \Rightarrow 6$: The maximal ideals of A_T are exactly the ideals $\mathfrak{n} = \mathfrak{m}_T$, where $\mathfrak{m}$ is an ideal of A maximal with respect to $\mathfrak{m}\cap T = \emptyset$. Such an $\mathfrak{m}$ is always prime, but not necessarily a maximal ideal of A [10, Thm. 4.1]. Recall that $\mathfrak{m}_\lambda\cap T = \emptyset$ for $\lambda\in\Lambda_2$. Consider such an $\mathfrak{n} = \mathfrak{m}_T$. If $V(\mathfrak{m})\cap\Lambda_2 = \emptyset$, condition 5 implies $\mathfrak{m}\cap T \neq \emptyset$, a contradiction. Hence $V(\mathfrak{m})\cap\Lambda_2 \neq \emptyset$ and therefore there is a $\lambda\in\Lambda_2$ such that $\mathfrak{m}\subseteq\mathfrak{m}_\lambda$ and $\mathfrak{m}_\lambda\cap T = \emptyset$; hence $\mathfrak{m} = \mathfrak{m}_\lambda$ by the maximality of $\mathfrak{m}$.

$6 \Rightarrow 5$: Assume indirectly that $\mathfrak{a}\cap T = \emptyset$ and let $\mathfrak{m}$ be maximal with $\mathfrak{a}\subset\mathfrak{m}$ and $\mathfrak{m}\cap T = \emptyset$. Then $\mathfrak{m}_T$ is a maximal ideal of A_T and therefore by 6 of the form

$$\mathfrak{m}_T = (\mathfrak{m}_\lambda)_T,\ \lambda\in\Lambda_2 \ \Rightarrow\ \mathfrak{a}\subset\mathfrak{m} = \mathfrak{m}_\lambda \ \Rightarrow\ \lambda\in V(\mathfrak{a})\cap\Lambda_2.$$

$4 \Rightarrow 7$: The module $\mathcal{F}_2 = \oplus_{\lambda\in\Lambda_2} E(A/\mathfrak{m}_\lambda)$ is an injective cogenerator in $\mathrm{Mod}_{A,\mathfrak{T}}$ and hence by 4 in Mod_{A_T}.

$7 \Rightarrow 6$: Let $\mathfrak{n} = \mathfrak{m}_T$ be a maximal ideal of A_T as before. The primeness of $\mathfrak{m}$ implies

$$A/\mathfrak{m} \subset (A/\mathfrak{m})_T = A_T/\mathfrak{m}_T = A_T/\mathfrak{n}.$$

Since $\mathcal{F}_2$ is an injective A_T-cogenerator it contains the simple A_T-module $A_T/\mathfrak{n}$ and thus $A/\mathfrak{m}$; hence

$$\mathfrak{m} \in \mathrm{Ass}_A(\mathcal{F}_2) = \{\mathfrak{m}_\lambda;\ \lambda \in \Lambda_2\} \text{ and } \mathfrak{m} = \mathfrak{m}_\lambda \text{ for some } \lambda \in \Lambda_2. \qquad \square$$

Examples 5.7.

1. If $\Lambda_2 \subset \mathbb{C}^r$ is ideal-convex, then

$$\Lambda_2 := \{\lambda \in \mathbb{C}^r;\ \forall t \in T :\ t(\lambda) \neq 0\} \text{ or } \Lambda_1 = \mathbb{C}^r \setminus \Lambda_2 = \cup_{t\in T} V(t), \tag{81}$$

   i.e., Λ_1 is a union of hypersurfaces. Equation (81) is false for

$$\Lambda_2 := \{\lambda \in \mathbb{C}^2;\ \Re(\lambda_1) \geq 0 \text{ or } \Re(\lambda_2) \geq 0\},$$
$$\Lambda_1 = \{\lambda \in \mathbb{C}^r;\ \Re(\lambda_1) < 0, \Re(\lambda_2) < 0\},\ T = \mathbb{C} \setminus \{0\},\ \cup_{t\in T} V(t) = \emptyset.$$

2. In dimension 1 each Λ_2 is ideal-convex.
3. If $r = 2$, (81) characterizes ideal-convexity. In particular, the noncompact regions $\overline{\mathbb{C}^2_+}$ and $\overline{\mathbb{C}_+} \times i\mathbb{R}$ from Theorem 5.4 are ideal-convex, whereas for $r > 2$ this is an open question [29, Thm. 5.5].
4. Each polynomially convex compact subset Λ_2 of $\mathbb{C}^r$ is ideal-convex [23, Prop. 3.1.20, Rem. on p. 28]. Polynomial convexity signifies that

$$\Lambda_2 = \left\{\lambda \in \mathbb{C}^r;\ \forall\ f \in \mathbb{C}[s] :\ |f(\lambda)| \leq \sup_{z\in\Lambda_2} |f(z)|\right\}.$$

   In particular, the closed unit polydisc $\overline{U}^r$, $\overline{U} := \{z \in \mathbb{C};\ |z| \leq 1\}$, is ideal-convex [23, Prop. 3.1.20, Rem. on p. 28].
5. Each compact cuboid Λ_2 of the form

$$\Lambda_2 := \{\lambda \in \mathbb{C}^r;\ \forall \rho = 1, \ldots, r :\ a_\rho \leq \Re(\lambda_\rho) \leq b_\rho,\ c_\rho \leq \Im(\lambda_\rho) \leq d_\rho\},$$
$$\text{where } a_\rho \leq b_\rho,\ c_\rho \leq d_\rho \text{ in } \mathbb{R},\ \rho = 1, \ldots, r, \tag{82}$$

   is ideal-convex. The proof is similar to that of the preceding item in [23] and uses the cohomological theorems A and B concerning coherent modules on Stein spaces for these cuboids and the approximation of holomorphic functions by polynomials on such Λ_2. The proof is omitted.

The proofs of the last two examples use the compactness of K in different places essentially, in particular for the approximation of analytic functions on Λ_2 by polynomials. It seems difficult to derive constructive algorithms from these proofs; see Remark/Open Problem 5.10 below. The theorems A and B, however, are valid in much more generality and can possibly be used to prove ideal-convexity also for certain noncompact subsets Λ_2 of $\mathbb{C}^r$.

THEOREM 5.8 (stabilizability and ideal-convexity).

1. *If $\mathcal{B}$ is stabilizable with respect to* (70)*, then*

$$\mathrm{sing}(\mathcal{B}) \cap \Lambda_2 = \emptyset \text{ or } \mathrm{rank}(R(\lambda)) = \mathrm{rank}(R)\ \forall \lambda \in \Lambda_2.$$

2. *If Λ_2 is ideal-convex, the following assertions are equivalent:*
   (a) *$\mathcal{B}$ is stabilizable.*
   (b) *M_T is a projective A_T-module.*
   (c) i. *The A_T-lattice $\mathcal{M} := A_T^{1\times(p+m)}\binom{H}{\mathrm{id}_m} = A_T^{1\times p}H + A_T^{1\times m} \subset \mathbb{C}(s)^{1\times m}$ is projective.*
   ii. *M_T is torsion free; i.e., $\mathcal{B}_2 := \mathcal{B} \cap \mathcal{F}_2^{p+m}$ is a controllable ${}_{A_T}\mathcal{F}_2$ behavior.*
   (d) $\mathrm{sing}(\mathcal{B}) \cap \Lambda_2 = \emptyset$ *or* $\mathrm{rank}(R(\lambda)) = \mathrm{rank}(R)\ \forall \lambda \in \Lambda_2$.

Proof. 1. According to Theorem 4.4, $U_{\mathfrak{T}}$ is a direct summand of $A_T^{1\times l}$. Hence for all $\lambda_2 \in \Lambda_2$ also $U_{\mathfrak{m}_\lambda} = (U_{\mathfrak{T}})_{\mathfrak{m}_{\lambda,T}}$ is a direct summand of $(A_T)^{1\times l}_{\mathfrak{m}_{\lambda,T}} = A^{1\times l}_{\mathfrak{m}_\lambda}$. Therefore $M_{\mathfrak{m}_\lambda} = A^{1\times l}_{\mathfrak{m}_\lambda}/U_{\mathfrak{m}_\lambda}$ is projective and thus free, and λ_2 is not contained in $\mathrm{sing}(\mathcal{B})$ according (75).

2. (a) $\Leftrightarrow$ (b): Due to the assumed ideal-convexity, we have

$$U_T = U_{\mathfrak{T}},\ M_T = M_{\mathfrak{T}} = A_T^{1\times l}/U_T. \text{ Hence}$$
$$\mathcal{B} \text{ is stabilizable } \underset{\text{Thm. 4.4}}{\Leftrightarrow} U_T \text{ is a direct summand } \Leftrightarrow M_T \text{ is projective.}$$

(b) $\Leftrightarrow$ (c): The assumed torsion freeness of M_T implies the exact sequence (13)

$$A_T^{1\times k} \overset{\circ(P,-Q)}{\longrightarrow} A_T^{1\times(p+m)} \overset{\circ\binom{H}{\mathrm{id}_m}}{\longrightarrow} \mathbb{C}(s)^{1\times m},\ \text{hence}$$
$$M_T = A_T^{1\times(p+m)}/A_T^{1\times k}(P,-Q) \cong A_T^{1\times(p+m)}\begin{pmatrix} H \\ \mathrm{id}_m \end{pmatrix} = \mathcal{M}.$$

Hence M_T is A_T-projective if and only if the lattice $\mathcal{M}$ has this property.

(a) $\Rightarrow$ (d) follows from 1.

(d) $\Rightarrow$ (b): Condition (d) implies that for all $\lambda \in \Lambda_2$ $M_{\mathfrak{m}_\lambda} = (M_T)_{\mathfrak{m}_{\lambda,T}}$ is projective. Since, by Theorem/Definition 5.6(6), the $\mathfrak{m}_{\lambda,T}$, $\lambda \in \Lambda_2$, are all maximal ideals of A_T, the projectivity of M_T follows from [10, Thm. 7.12]. □

Remarks 5.9.

1. For general, not ideal-convex unstable regions, stabilizability is characterized in Theorem/Definition 4.3 and Theorem 4.4.
2. For $r = 1$ and $\Lambda_2 := \{\lambda \in \mathbb{C};\ \Re(\lambda) \geq 0\}$ condition 2(d) coincides with condition 4(ii) from Theorem/Definition 5.1; see item 4 of Remark 5.2. The equivalence (a) $\Leftrightarrow$ (d) is Theorem 5.2.30 of [15], where the stabilizability of $\mathcal{B}$ is, however, defined by asymptotic controllability to zero of its trajectories.
3. The condition 2(c)i of the preceding theorem characterizes internally stabilizable systems according to Quadrat [17, Thm. 3, Eq. (38)] with A_T as a ring of SISO-stable plants.

Remarks and open problems 5.10 (algorithmic questions). The assumptions are those from (64) and (70)–(75). It turns out that *basic algorithmic problems cannot currently be solved.* Refer to [7] and to [30] where the history, the state of the art, and open problems concerning algorithms in multidimensional stability theory are described.

1. According to Theorem/Definition 5.1, the stability of $\mathcal{B}$ is checked via

$$\mathrm{char}(\mathcal{B}^0) \subset \Lambda_1 \text{ or } \mathrm{char}(\mathcal{B}^0) \cap \Lambda_2 = \emptyset, \text{ where}$$
$$\mathrm{char}(\mathcal{B}^0) = V(\mathfrak{a}) \text{ and } \mathfrak{a} := \mathrm{ann}(M^0) = \{f \in \mathbb{C}[s];\ fM^0 = 0\},$$

and $\mathfrak{a}$ can be computed from the matrix $P \in \mathbb{C}[s]^{k\times p}$ via Gröbner bases. Recall from Theorem/Definition 5.6 that

$$V(\mathfrak{a}) \cap \Lambda_2 = \emptyset \;\Leftrightarrow\; \mathfrak{a} \cap T \neq \emptyset \;\forall \text{ ideals } \mathfrak{a}$$

if and only if Λ_2 is ideal-convex. If this is not the case, it is the analogue of the open problem 3 from [7, p. 73] for Λ_2 instead of the closed unit polydisc to decide whether $V(\mathfrak{a}) \cap \Lambda_2 = \emptyset$. No algorithm is presently known. If Λ_2 is ideal-convex, then to decide $\mathfrak{a} \cap T \neq \emptyset$ and to actually construct a $t \in \mathfrak{a} \cap T$ is the analogue of the open problem 4 from [7, p. 73] or problem 1 from [30]. For compact, polynomially convex sets Λ_2, such as the closed unit polydisc, this requires a constructive version of the proof of Example 5.7(4), and even this seems hard to obtain. If

$$[M^0 : \mathbb{C}] < \infty \;\Leftrightarrow\; \mathfrak{a} \text{ is Krull–zero-dimensional } \Leftrightarrow\; V(\mathfrak{a}) = \operatorname{char}(\mathcal{B}^0) \text{ is finite,}$$

then the finite variety $V(\mathfrak{a})$ can be computed via Gröbner bases, and $V(\mathfrak{a}) \cap \Lambda_2 = \emptyset$ can be decided (compare [7, p. 73, Rem.]).

2. Theorem 4.4 is used to test stabilizability and to construct a stabilizing compensator. Presently the computation of a matrix $R_{\mathrm{st}} \in A^{k_{\mathrm{st}}\times l}$ such that

$$\begin{aligned} U_{\mathrm{st}} &= \{x \in A^{1\times l};\ \forall \lambda \in \Lambda_2\ \exists s_\lambda \in A \text{ with} \\ &\qquad s_\lambda(\lambda) \neq 0 \text{ and } s_\lambda x \in U = A^{1\times k}R\} \\ &= A^{1\times k_{\mathrm{st}}}R_{\mathrm{st}} \text{ and hence } U_{\mathfrak{T}} = A_T^{1\times k_{\mathrm{st}}}R_{\mathrm{st}} \end{aligned}$$

is unsolved. If, however, Λ_2 is ideal-convex and $M_{\mathfrak{T}} = M_T$ for all A-modules M, then $U_{\mathfrak{T}} = U_T = A_T^{1\times k}R$ and the construction of R_{st} is superfluous. See Theorem 5.14 for a partial result on the construction of R_{st}. If, in general, a matrix $R_{\mathrm{st}} \in A^{k_{\mathrm{st}}\times l}$ with $U_{\mathfrak{T}} = A_T^{1\times k_{\mathrm{st}}}R_{\mathrm{st}}$ is known, one computes the module L of all solutions $(G_1', t_1) \in A^{l\times k_{\mathrm{st}}} \times A$ of the polynomial linear system

$$R_{\mathrm{st}}G_1'R_{\mathrm{st}} - t_1R_{\mathrm{st}} = 0 \;\Leftrightarrow\; \text{if } t_1 \neq 0 : R_{\mathrm{st}}\frac{G_1'}{t_1}R_{\mathrm{st}} = R_{\mathrm{st}} \tag{83}$$

via Gröbner bases and obtains a system of generators of the ideal

$$\mathfrak{a} := \{t_1 \in A;\ \exists (G_1', t_1) \in L\}. \tag{84}$$

According to Theorem 4.4(3), the system is stabilizable if and only if $\mathfrak{a} \cap T \neq \emptyset$. This is again the open problem quoted in item 1. If

$$t_1 \in \mathfrak{a} \cap T \text{ and } (G_1', t_1) \in L,\ \text{then}\ G_1 := \frac{G_1'}{t_1} \in A_T^{l\times k_{\mathrm{st}}}$$

is a matrix according to Theorem 4.4(3), and

$$E_1 := G_1R_{\mathrm{st}} = E_1^2 \in A_T^{l\times l},\ U_{\mathfrak{T}} = A_T^{1\times l}E_1.$$

Generically, but not always, this matrix already satisfies the rank condition $\operatorname{rank}((\mathrm{id}_l - E_1)\binom{0}{\mathrm{id}_m}) = m$ and can be used to construct a stabilizing compensator of $\mathcal{B}$ according to Theorem 4.4.

3. All stabilizing compensators or idempotent matrices

$$E = E^2 \in A_T^{l\times l} \text{ with } U_{\mathfrak{T}} = A_T^{1\times l}E \text{ and } \operatorname{rank}\left((\mathrm{id}_l - E)\begin{pmatrix}0\\ \mathrm{id}_m\end{pmatrix}\right) = m$$

are constructed according to Theorems 2.14 and 4.6 in the form

$$E = E_1 + XR_{\mathrm{st}},\ X = \sum_{i=1}^{n} \Theta_i X_i,\ \Theta_i \in A_T,$$

where the X_i generate the solution module of the polynomial linear system $R_{\mathrm{st}}X'R_{\mathrm{st}} = 0,\ X' \in A^{l\times k_{\mathrm{st}}}$. Generically, these E satisfy the rank condition $\operatorname{rank}((\mathrm{id}_l - E)\binom{0}{\mathrm{id}_m}) = m$ and can be used to construct all stabilizing compensators of $\mathcal{B}$.

In the remainder of this section we indicate the necessary modifications of the preceding theory for the discrete case of complex *partial difference equations* and the locally finite elements of the A-module $\mathbb{C}^{\mathbb{N}^r}$ [12, Thm. 1.25, Cor. 1.26] with the left shift action

$$(s^\mu \circ y)(\nu) := y(\mu + \nu),\ y \in \mathbb{C}^{\mathbb{N}^r},\ \mu, \nu \in \mathbb{N}^r.$$

For

$$\begin{gathered}\alpha \in \mathbb{C},\ k, i \in \mathbb{N},\ \lambda \in \mathbb{C}^r,\ \mu, \nu \in \mathbb{N}^r \text{ define } e_{\alpha,k} \in \mathbb{C}^{\mathbb{N}},\ e_{\lambda,\mu} \in \mathbb{C}^{\mathbb{N}^r} \text{ by}\\ e_{\alpha,k}(i) := \begin{cases}\binom{i}{k}\alpha^{i-k} & \text{if } \alpha \neq 0\\ \delta_{i,k} & \text{if } \alpha = 0\end{cases},\ e_{\lambda,\mu}(\nu) := \prod_{\rho=1}^{r} e_{\lambda_\rho,\mu_\rho}(\nu_\rho). \text{ Then}\\ (s-\lambda)^\nu \circ e_{\lambda,\mu} = \begin{cases}e_{\lambda,\mu-\nu} & \text{if } \mu \in \nu + \mathbb{N}^r\\ 0 & \text{otherwise}\end{cases} \text{ and}\\ \mathcal{F} := \mathbb{C}_{\mathrm{lf}}^{\mathbb{N}^r} := \{y \in \mathbb{C}^{\mathbb{N}^r};\ [\mathbb{C}[s] \circ y : \mathbb{C}] < \infty\} = \oplus_{\lambda\in\mathbb{C}^r,\,\mu\in\mathbb{N}^r} \mathbb{C}e_{\lambda,\mu}.\end{gathered} \tag{85}$$

Again $\mathcal{F}$ is the minimal injective cogenerator in Mod_A; indeed

$$\mathcal{F}(\lambda) := \oplus_{\mu\in\mathbb{N}^r} \mathbb{C}e_{\lambda,\mu} = E(A/\mathfrak{m}_\lambda)\ \forall \lambda \in \mathbb{C}^r. \tag{86}$$

The Borel isomorphism

$$\mathbb{C}^{\mathbb{N}^r} = \mathbb{C}[[z]] \cong \mathbb{C}[[z]],\ y = \sum_{\mu\in\mathbb{N}^r} y(\mu)z^\mu \mapsto \sum_{\mu\in\mathbb{N}^r} y(\mu)\frac{z^\mu}{\mu!}$$

induces the $\mathbb{C}[s]$ isomorphism

$$\mathcal{F}(\lambda) = \oplus_{\mu\in\mathbb{N}^r} \mathbb{C}e_{\lambda,\mu} \cong \mathbb{C}[z]\exp(\lambda \bullet z),\ e_{\lambda,\mu} \mapsto \frac{z^\mu}{\mu!}\exp(\lambda \bullet z). \tag{87}$$

The decomposition $\mathbb{C}^r = \Lambda_1 \uplus \Lambda_2$ with its implied data from (70) is again arbitrary, but, of course, interesting choices in the discrete case are different from those in the continuous case. Equations (70) to (77) remain valid, and so do Theorems/Definitions 5.1 and 5.3 if $\mathcal{F}(\lambda)_k$ from (79) is replaced with

$$\mathcal{F}(\lambda)_k := \{y \in \mathbb{C}^{\mathbb{N}^r};\ \mathfrak{m}_\lambda^{k+1} \circ y = 0\} = \oplus_{\mu\in\mathbb{N}^r,\,|\mu|\le k} \mathbb{C}e_{\lambda,\mu},\ |\mu| := \mu_1 + \cdots + \mu_r. \tag{88}$$

A standard example for Λ_2 in the discrete case, in particular for $r = 1$, is

$$\Lambda_2 := \{w \in \mathbb{C};\ |w| \geq 1\}^r, \text{ but not } \overline{U}^r \text{ with } \overline{U} := \{w \in \mathbb{C};\ |w| \leq 1\}, \tag{89}$$

where $\overline{U}^r$ is the closed unit polydisc [7, p. 60], and for simplicity we assume this Λ_2 in what follows. This seeming contrast to [7] and most other papers in this field comes from the fact that these papers do not consider IO systems, but systems defined by a *proper* or *causal* IO map H [6], [32, sec. 3.2]. To explain this connection we identify

$$\begin{gathered} z_\rho := s_\rho^{-1},\ A := \mathbb{C}[s] \subset \mathbb{C}(s) = \mathbb{C}(z) = \operatorname{quot}(A),\ \mathbb{C}[z] \subset \mathbb{C}^{\mathbb{N}^r} = \mathbb{C}[[z]]. \text{ Let} \\ \mathfrak{n} :=_{\mathbb{C}[z]} \langle z_1, \ldots, z_r \rangle := \{g \in \mathbb{C}[z];\ g(0) = 0\} \end{gathered} \tag{90}$$

be the distinguished maximal ideal of $\mathbb{C}[z]$. Then [6, Def. 3.16], [11, sec. 6, Thm. 60]

$$\mathbb{C}[z]_{\mathfrak{n}} = \{n(z)/d(z);\ n, d \in \mathbb{C}[z],\ d(0) \neq 0\} = \operatorname{quot}(A) \cap \mathbb{C}[[z]] \tag{91}$$

is the ring of *proper or causal* rational functions where the term "causal" is due to the following property. The power series ring $\mathbb{C}^{\mathbb{N}^r} = \mathbb{C}[[z]]$ is a module over itself via convolution, also denoted by $\circ$, and therefore also a module over the ring of proper rational functions. In particular [11, sec. 6, Cor. 45], if the transfer matrix H of $\mathcal{B}$ is proper, any input $u \in (\mathbb{C}^{\mathbb{N}^r})^m = \mathbb{C}[[z]]^m$ gives rise to an output

$$y := H \circ u \in \mathbb{C}[[z]]^p \text{ with } P \circ y = P \circ (H \circ u) = (PH) \circ u = Q \circ u \text{ or } \begin{pmatrix} y \\ u \end{pmatrix} \in \mathcal{B}.$$

As noted before, properness is a rather restrictive property in the multidimensional situation and therefore we do not assume it in the present paper. For instance, the transfer functions $(s_1 + s_2)^{-1} = \frac{z_1 z_2}{z_1 + z_2}$ of $(s_1 + s_2) \circ y = u$ and of other standard partial difference or differential equations are not proper.

With the data from (89) and the terminology from [6, Def. 3.47], [7, p. 60], we see that

$$S := \left\{ h = \frac{n(z)}{d(z)} \in \mathbb{C}(z) = \mathbb{C}(s);\ \forall w \in \overline{U}^r :\ d(w) \neq 0 \right\} \subset \mathbb{C}[z]_{\mathfrak{n}} \tag{92}$$

is the ring of *structurally stable* rational functions, whereas the denominators $d(z)$ are called structurally stable polynomials.

LEMMA 5.11. $S \subseteq A_T \cap \mathbb{C}[z]_{\mathfrak{n}} = A_T \cap \mathbb{C}[[z]]$.

Proof. Let $h = \frac{n(z)}{d(z)} \in S$ be structurally stable with relatively prime n and d. Choose $\nu \in \mathbb{N}^r$ such $s^\nu n, s^\nu d \in \mathbb{C}[s]$; hence

$$h = \frac{n(z)}{d(z)} = \frac{s^\nu n(z)}{s^\nu d(z)} = \frac{p(s)}{q(s)} \text{ with relatively prime } p, q \in \mathbb{C}[s].$$

The factoriality of $\mathbb{C}[s]$ implies the existence of $f(s) \in \mathbb{C}[s]$ such that

$$\begin{gathered} q(s)f(s) = s^\nu d(z) \in \mathbb{C}[s], \text{ hence } q(\lambda)f(\lambda) = \lambda^\nu d(w) \neq 0\ \forall \\ \lambda \in \Lambda_2 \text{ or } |\lambda_\rho| \geq 1,\ \rho = 1, \ldots, r, \text{ and } w := (\lambda_1^{-1}, \ldots, \lambda_r^{-1}) \in \overline{U}^r, \end{gathered}$$

where $h \in S$ implies $d(w) \neq 0$. Hence $q \in T$ and $h \in A_T$ as asserted. ∎

Remarks 5.12.

1. Thus structural stability of h implies properness and stability, i.e., $h \in A_T$, in the sense of this paper. Equality holds in the preceding lemma if and only if an irreducible polynomial $d(z) \in \mathbb{C}[z]$ is already structurally stable if $d(0) \neq 0$ and if it has no zero w in $\overline{U}^r$ with $w_\rho \neq 0\ \forall \rho = 1, \ldots, r$. For $r = 1$ this is the case; in general it is unknown (to the author). This is another example of the difficulty of characterizing and manipulating the multiplicatively closed sets T of polynomials which have no zero in a given subset Λ_2 of $\mathbb{C}^r$.
2. The standard one-dimensional continuous (resp., discrete) stabilization theory for systems with a given proper transfer operator uses the Euclidean ring $S = A_T \cap \mathbb{C}[[z]]$ [26, Chap. 5] instead of A_T here with

$$\Lambda_2 := \{\lambda \in \mathbb{C};\ \Re(\lambda) \geq 0\},\ \text{resp.},\ \Lambda_2 := \{\lambda \in \mathbb{C};\ |\lambda| \geq 1\}.$$

A counterpart to Theorem 5.4 is the *BIBO stability* of structurally stable transfer matrices [6, sec. 3.4], i.e., the result that the space

$$l^\infty := \left\{ u = \sum_{\mu \in \mathbb{N}^r} u_\mu z^\mu \in \mathbb{C}^{\mathbb{N}^r} = \mathbb{C}[[z]];\ \exists M > 0\ \forall \mu \in \mathbb{N}^r :\ |u_\mu| \leq M \right\}$$

of bounded multisequences is an S-submodule of $\mathbb{C}^{\mathbb{N}^r}$. This is "well-known" and follows from the fact that a structurally stable rational function is a convergent power series in an open polydisc containing $\overline{U}^r$.

For general stable discrete systems, we pose the following open problem. Consider Λ_2 from (89), an arbitrary discrete IO system (64) with signals from $\mathbb{C}[[z]] = \mathbb{C}^{\mathbb{N}^r}$. Let $\Gamma \subset \{1, \ldots, p\} \times \mathbb{N}^r$ denote the associated *canonical initial region* [13, Eq. (14)] with respect to a chosen term order on $\{1, \ldots, p\} \times \mathbb{N}^r$. Then the *canonical Cauchy problem* [13, Thms. 5 and 8]

$$\begin{aligned} P \circ y = Q \circ u,\ u \in \mathbb{C}[[z]]^m,\ y = \left(\sum_{\mu \in \mathbb{N}^r} y_{i,\mu} z^\mu \right)_{i \leq p} \in \mathbb{C}[[z]]^p, \\ y \mid \Gamma = \left(\sum_{\mu} \{ y_{i,\mu} z^\mu;\ (i,\mu) \in \Gamma \} \right)_{i \leq p} = x \in \mathbb{C}^\Gamma \subset \mathbb{C}[[z]]^p \end{aligned} \tag{93}$$

has a unique solution y for given input u and *initial data* x, and this can be computed using Gröbner bases. The space $\mathbb{C}\langle z \rangle$ of *locally convergent* power series is [13, Eq. (54)]

$$\begin{aligned} \mathbb{C}\langle z \rangle = \Bigg\{ u = \sum_{\mu \in \mathbb{N}^r} u_\mu z^\mu \in \mathbb{C}[[z]];\ \exists M > 0\ \exists R \in \mathbb{R}^r_{>0}\ \forall \mu \in \mathbb{N}^r : \\ |u_\mu| \leq M R^\mu \Bigg\} = \left\{ u = \sum_{\mu \in \mathbb{N}^r} u_\mu z^\mu \in \mathbb{C}[[z]];\ \exists \widetilde{R} \in \mathbb{R}^r_{>0} :\ \sum_{\mu \in \mathbb{N}^r} |u_\mu| \widetilde{R}^\mu < \infty \right\} \\ \text{and } l^\infty,\ l^1 := \left\{ u = \sum_{\mu \in \mathbb{N}^r} u_\mu z^\mu \in \mathbb{C}[[z]];\ \sum_{\mu \in \mathbb{N}^r} |u_\mu| < \infty \right\} \subset \mathbb{C}\langle z \rangle \end{aligned}$$

are $\mathbb{C}[s]$-submodules of $\mathbb{C}\langle z\rangle$. The unique solvability of the Cauchy problem (93) also holds for $\mathbb{C}\langle z\rangle$ instead of $\mathbb{C}[[z]]$ [13, Cor. 25], especially

$$\mathcal{B}^0 = \{y \in \mathbb{C}\langle z\rangle^p;\ P \circ y = 0\}$$
$$\cong \left\{x \in \mathbb{C}^\Gamma;\ \forall i = 1,\dots,p:\ \sum_\mu \{x_{i,\mu} z^\mu;\ (i,\mu) \in \Gamma\} \in \mathbb{C}\langle z\rangle\right\},\ y \mapsto y \mid \Gamma.$$

In particular, if u and x have components in l^∞ or l^1, the unique power series solution y satisfies

$$|y_{i,\mu}| \le MR^\mu,\ \text{resp.},\ \sum_{\mu\in\mathbb{N}^r} |y_{i,\mu}|R^\mu < \infty\ \text{for } \textit{some}\ R \in \mathbb{R}^r_{>0}\ \text{and}\ M > 0. \tag{94}$$

If all components R_ρ are smaller (resp., greater) or equal to 1, then $y \in (l^\infty)^p$ (resp., $y \in (l^1)^p$). Now assume that the IO system $\mathcal{B}$ is Λ_2-stable for Λ_2 from (89), i.e., that conditions 2–4 of Theorem/Definition 5.1 are satisfied (the function space differs!!). With the idea from [29, pp. 1499–1500] that stability of a system should imply that *stable inputs and initial conditions generate stable outputs*, we pose the following.

Open problem 5.13. Consider $\Lambda_2 := \{\lambda \in \mathbb{C}^r;\ \forall \rho = 1,\dots,r : |\lambda_\rho| \ge 1\}$, $i = \infty, 1$, and a Λ_2-stable IO system $\mathcal{B}$ as above with input $u \in (l^i)^m$, initial data $x \in (l^i)^p$, and unique output $y \in \mathbb{C}\langle z\rangle^p$. When is $y \in (l^i)^p$, i.e., $R = (1,\dots,1)$ in (94) or, in other words, when is the stable system BIBO- or l^∞-IO-stable (resp., l^1-IO-stable)?

This is always true for one-dimensional discrete IO systems—the properness of the transfer matrix is not required. Compare [15, Thm. 7.6.2] for a continuous analogue under the assumption that the transfer matrix is proper.

The question can also be asked for other natural unstable regions and is also interesting and reasonable in the continuous case of partial differential equations for those spaces of analytic functions for which the Cauchy problem is uniquely solvable. In particular this holds for the space of entire functions of exponential type [13, Thm. 26] which, by means of the Borel isomorphism, is isomorphic to $\mathbb{C}\langle z\rangle$ with the shift action, and for the space $\mathbb{C}\langle z\rangle$, but now with the action by partial differentiation and under the assumption that the term order is graded [13, Thm. 29].

Consider, however, the trivial (in the sense of Theorem/Definition 2.5) and therefore stable one-dimensional system $y = s_1 \circ u$ with an empty initial condition and polynomial transfer function $H = s_1$. The analytic function $u := \exp(iz_1^2)$ is bounded on $\mathbb{R}$, but the output $y = u' = 2iz_1 u$ is not, so BIBO stability does not hold. Likewise, $u(z_1) := (1 + z_1^2)^{-1}\exp(iz_1^2)$ is in $L^1(\mathbb{R})$, but its derivative is not. So additional assumptions, for instance properness, have to be made in the analogue of the open problem for the general continuous case.

The following theorem without a detailed proof improves Theorem 5.8 and item 2 of Remarks/Open Problems 5.10 and was added during the revision process. The data are those from (64) and (70)–(78). According to Theorem/Definition 2.2 we construct the unique controllable realization $\mathcal{B}_{\text{cont}}$ of H via

$$U_{\text{cont}} := \mathcal{B}^\perp_{\text{cont}} := \ker\left(A^{1\times(p+m)} \xrightarrow{\circ\binom{H}{\mathrm{id}_m}} K^{1\times m}\right),\ \text{hence}$$
$$M/t(M) \cong M_{\text{cont}} := A^{1\times(p+m)}/U_{\text{cont}}, \tag{95}$$
$$t(M) = U_{\text{cont}}/U \subset M = A^{1\times(p+m)}/U,$$
$$\mathfrak{a} := \operatorname{ann}_A(t(M)).$$

The submodule U_{cont} and the annihilator $\mathfrak{a}$ of the torsion module $t(M) = U_{\mathrm{cont}}/U$ are computed via Buchberger's algorithm. For the Gabriel localization $U_{\mathfrak{T}}$, Lemma/Definition 3.6 implies

$$U_T \subset U_{\mathfrak{T}} = A_T^{1\times(p+m)} \cap \cap_{\lambda\in\Lambda_2} U_{\mathfrak{m}_\lambda} \subset A_T^{1\times(p+m)}. \tag{96}$$

Generators of $U_{\mathfrak{T}}$ are needed for the application of the algorithm in Remarks/Open Problems 5.10(2), and in general are hard to obtain. The next theorem gives a partial solution.

THEOREM 5.14. *We use data as just introduced.*

(1) *All localized modules $M_{\mathfrak{m}_\lambda}$, $\lambda \in \Lambda_2$, are torsion free if and only if $V(\mathfrak{a})\cap\Lambda_2 = \emptyset$. If this is the case, then $U_{\mathfrak{T}} = U_{\mathrm{cont},T}$, and $\mathcal{B}$ is stabilizable if and only if $M_{\mathrm{cont},T}$ is a projective A_T-module.*

(2) *The module M_T is torsion free if and only if $\mathfrak{a}\cap T \neq \emptyset$. If this is the case, then*

$$U_{\mathfrak{T}} = U_{\mathrm{cont},T} = U_T \text{ and } M_T \cong \mathcal{M} := A_T^{1\times p}H + A_T^{1\times m} \subset \mathbb{C}(s)^{1\times m},$$

and $\mathcal{B}$ is stabilizable if and only if the lattice $\mathcal{M}$ is a projective A_T-module. According to Quadrat [17, *Cor.* 3] *this characterizes the stabilizability of the transfer matrix H in the usual sense.*

6. Conclusion. In contrast to the literature on the stabilization of *discrete* multidimensional IO *maps*, the present paper developed the stabilization theory for *continuous or discrete* multidimensional IO *behaviors*. An important technical tool in this context was the generalized localization theory due to Gabriel. Algorithmic problems were addressed, but only partially solved. For a complete solution one needs constructive solutions for the following problems from algebraic and analytic geometry. Let Λ_2 be an arbitrary subset of the r-dimensional complex space and consider

$$T := \{t \in \mathbb{C}[s_1,\dots,s_r];\ \forall\lambda\in\Lambda_2 :\ t(\lambda)\neq 0\} =: \{\text{stable polynomials}\},$$
$$\text{nonzero polynomials } f \text{ and ideals } \mathfrak{a} \text{ in } \mathbb{C}[s].$$

OPEN ALGORITHMIC PROBLEMS 6.1 (compare [7], [30]).

1. *Decide $f \in T$.*
2. *Decide $V(\mathfrak{a}) \cap \Lambda_2 = \emptyset$.*
3. *Decide $\mathfrak{a}\cap T \neq \emptyset$ and construct $t \in \mathfrak{a}\cap T$.*
4. *The implication $\mathfrak{a}\cap T \neq \emptyset \ \Rightarrow\ V(\mathfrak{a})\cap\Lambda_2 = \emptyset$ is obvious. The reverse implication is true for all ideals $\mathfrak{a}$ if and only if $\mathfrak{a}$ is ideal-convex. Decide ideal-convexity constructively and construct $t \in \mathfrak{a}\cap T$ if $V(\mathfrak{a})\cap\Lambda_2 = \emptyset$.*
5. *Of course, the preceding tasks have to be solved only for suitable Λ_2, for instance for the typical continuous (resp., discrete) cases*

$$\Lambda_2 := \{z\in\mathbb{C};\ \Re(z)\geq 0\}^r,\ \textit{resp.},\ \Lambda_2 = \{z\in\mathbb{C};\ |z|\geq 1\}^r.$$

   As shown in Remarks/Open Problems 5.10 *and in Theorem* 5.14 *a special stabilization problem gives rise only to few ideals $\mathfrak{a}$ for which the preceding tasks have to be solved. The paper* [8] *treats these problems for the closed unit polydisc and a special class of ideals.*
6. *If $\mathfrak{a}$ is Krull–zero-dimensional or $V(\mathfrak{a})$ is finite, then this variety can be computed via Buchberger's algorithm and the constructive tasks can in general be solved.*

At the recent workshops D2 and D3 of the Gröbner semester in Linz, Austria (May 8–19, 2006), S. Tsarev (TU Berlin) pointed out the very interesting and highly useful fact that some of the preceding open problems can actually be constructively solved by means of the Tarski–Seidenberg theorem for real semi-algebraic geometry and cylindrical algebraic decomposition.

Acknowledgments. The author thanks the two referees for their reviewing work and critical comments which have been incorporated into the revised version of the paper.

REFERENCES

[1] M. Bisiacco and M. E. Valcher, *Two-dimensional behavior decompositions with finite-dimensional intersection: A complete characterization*, Multidimens. Systems Signal Process., 16 (2005), pp. 335–354.
[2] V. D. Blondel and A. Megretski, eds., *Unsolved Problems in Mathematical Systems and Control Theory*, Princeton University Press, Princeton, NJ, 2004.
[3] N. K. Bose, ed., *Multidimensional Systems Theory*, D. Reidel, Dordrecht, 1985.
[4] Y. V. Egorov and M. A. Shubin, eds., *Partial Differential Equations* III, Springer-Verlag, Berlin, 1991.
[5] K. Galkowski and J. Wood, eds., *Multidimensional Signals, Circuits, and Systems*, Taylor and Francis, London, 2001.
[6] J. P. Guiver and N. K. Bose, *Causal and weakly causal* 2*-D filters with application in stabilization*, in Multidimensional Systems Theory, N. K. Bose, ed., D. Reidel, Dordrecht, 1985, pp. 52–100.
[7] Z. Lin, *Output feedback stabilizability and stabilization of linear n-D systems*, in Multidimensional Signals, Circuits, and Systems, K. Galkowski and J. Wood, eds., Taylor and Francis, London, 2001, pp. 59–76.
[8] Z. Lin, J. Lam, K. Galkowski, and S. Xu, *A constructive approach to stabilizability and stabilization of a class of nD systems*, Multidimens. Systems Signal Process., 12 (2001), pp. 329–343.
[9] S. MacLane, *Homology*, Springer-Verlag, Berlin, 1967.
[10] H. Matsumura, *Commutative Ring Theory*, Cambridge University Press, Cambridge, UK, 1986.
[11] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.
[12] U. Oberst, *Variations on the fundamental principle for linear systems of partial differential and difference equations with constant coefficients*, Appl. Algebra Engrg. Comm. Comput., 6 (1995), pp. 211–243.
[13] U. Oberst and F. Pauer, *The constructive solution of linear systems of partial difference and differential equations with constant coefficients*, Multidimens. Systems Signal Process., 12 (2001), pp. 253–308.
[14] H. K. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1998), pp. 388–408.
[15] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory*, Springer-Verlag, New York, 1998.
[16] A. Quadrat, *An algebraic interpretation to the operator-theoretic approach to stabilizability. Part* I: *SISO systems*, Acta Appl. Math., 88 (2005), pp. 1–45.
[17] A. Quadrat, *A lattice approach to analysis and synthesis problems*, Math. Control Signals Systems, 18 (2006), pp. 147–186.
[18] A. Quadrat, *On a generalization of the Youla–Kučera parametrization. Part* II: *The lattice approach to MIMO systems*, Math. Control Signals Systems, to appear.
[19] P. Rocha, *Feedback control of multidimensional behaviors*, Systems Control Lett., 45 (2002), pp. 207–215.
[20] M. Scheicher and U. Oberst, *Proper Stabilization of Multidimensional Input/Output Systems*, in preparation.
[21] S. Shankar, *Can one control the vibrations of a drum?*, Multidimens. Systems Signal Process., 11 (2000), pp. 67–81.
[22] S. Shankar, *The lattice structure of behaviors*, SIAM J. Control Optim., 39 (2001), pp. 1817–1832.
[23] S. Shankar and V. R. Sule, *Algebraic geometric aspects of feedback stabilization*, SIAM J. Control Optim., 30 (1992), pp. 11–30.

[24] B. Stenström, *Rings of Quotients*, Springer-Verlag, Berlin, 1975.
[25] V. R. Sule, *Feedback stabilization over commutative rings: The matrix case*, SIAM J. Control Optim., 32 (1994), pp. 1675–1695.
[26] M. Vidyasagar, *Control Systems Synthesis: A Factorization Approach*, The MIT Press, Cambridge, MA, 1985.
[27] J. Wood, *Modules and behaviors in nD systems theory*, Multidimens. Systems Signal Process., 11 (2000), pp. 11–48.
[28] J. Wood, U. Oberst, E. Rogers, and D. H. Owens, *A behavioral approach to the pole structure of one-dimensional and multidimensional linear systems*, SIAM J. Control Optim., 38 (2000), pp. 627–661.
[29] J. Wood, V. R. Sule, and E. Rogers, *Causal and stable input/output structures on multidimensional behaviors*, SIAM J. Control Optim., 43 (2005), pp. 1493–1520.
[30] L. Xu, Z. Lin, J.-Q. Ying, O. Saito, and Y. Anazawa, *Open problems in control of linear discrete multidimensional systems*, in Unsolved Problems in Mathematical Systems and Control Theory, V. D. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, pp. 221–228.
[31] J. Q. Ying, *On the strong stabilizability of MIMO n-dimensional linear systems*, SIAM J. Control Optim., 38 (1999), pp. 313–335.
[32] E. Zerz, *Topics in Multidimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 256, Springer-Verlag, London, 2000.
[33] E. Zerz and V. Lomadze, *A constructive solution to interconnection and decomposition problems with multidimensional behaviors*, SIAM J. Control Optim., 40 (2001), pp. 1072–1086.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 4, pp. 1508–1539

A TWO-PERSON GAME FOR PRICING CONVERTIBLE BONDS*

MIHAI SÎRBU† AND STEVEN E. SHREVE‡

Abstract. A firm issues a convertible bond. At each subsequent time, the bondholder must decide whether to continue to hold the bond, thereby collecting coupons, or to convert it to stock. The bondholder wishes to choose a conversion strategy to maximize the bond value. Subject to some restrictions, the bond can be called by the issuing firm, which presumably acts to maximize the equity value of the firm by minimizing the bond value. This creates a two-person game. We show that if the coupon rate is below the interest rate times the call price, then conversion should precede call. On the other hand, if the dividend rate times the call price is below the coupon rate, call should precede conversion. In either case, the game reduces to a problem of optimal stopping.

Key words. convertible bonds, optimal stopping, two-person game, viscosity solutions, free boundary

AMS subject classifications. 90A09, 60H30, 60G44

DOI. 10.1137/050630222

1. Introduction.

Firms raise capital by issuing debt (bonds) and equity (shares of stock). The *convertible bond* is intermediate between these two instruments. A convertible bond is a bond in that it entitles its owner to receive coupons plus the return of principal at maturity. However, prior to maturity the holder may *convert* the bond, surrendering it for a preset number of shares of stock. The price of the bond is thus dependent on the price of the firm's stock. Finally, prior to maturity, the firm may *call* the bond, forcing the bondholder to either surrender it to the firm for a previously agreed price or convert it to stock as above.

After issuing a convertible bond, the firm's objective is to exercise its call option in order to maximize the value of shareholder equity. The bondholder's objective is to exercise the conversion option in order to maximize the value of the bond. Because the firm must pay coupons to the bondholder, it may call the bond if it can subsequently reissue a bond with a lower coupon rate. This happens as the firm's fortunes improve, for then the risk of default has diminished and investors will accept a lower coupon rate on the firm's bonds. In the case of a convertible bond, the firm has a second incentive to call: as the firm's fortunes improve, the investor may convert, becoming a shareholder of a profitable firm and diluting the value of the stock owned by the original shareholders. The firm can prevent this by calling the bond. The bondholder has an incentive to convert the bond to stock before maturity under exactly the same scenario; the bondholder may want to become a shareholder of a profitable firm, for the promise of future dividends may be more valuable than the promise of future coupons.

If stock and convertible bonds are the only assets issued by a firm, then the value of the firm is the aggregate value of these two types of assets. In idealized

*Received by the editors April 28, 2005; accepted for publication (in revised form) May 24, 2006; published electronically October 20, 2006. This material is based upon work supported by the National Science Foundation under grants DMS-0139911 and DMS-0404682. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

http://www.siam.org/journals/sicon/45-4/63022.html

†Department of Mathematics, Columbia University, New York, NY 10027 (sirbu@math.columbia.edu).

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (shreve@andrew.cmu.edu).

markets, where the Miller–Modigliani [32], [33] assumptions hold (see Hennessy and Tserlukevich [20] for a model in which they do not), changes in corporate capital structure do not affect firm value. In particular, the value of the firm does not change at the time of conversion, and the only change in the value of the firm at the time of call is a reduction by the call price paid to the bondholder if the bondholder surrenders rather than converting the bond. By acting to maximize the value of equity, the firm is in fact minimizing the value of the convertible bond. By acting to maximize the value of the bond, the bondholder is in fact minimizing the value of equity. This creates a *two-person, zero-sum game*. The game is complicated by the fact that one can expect the dividend payment policy of the firm to depend on the bond price, a feature explicitly modeled in this paper. This feature causes the bond price to be governed by a *nonlinear* second-order partial differential equation, a novel feature of this paper.

This is a companion paper to Sîrbu, Pikovsky, and Shreve [36]. In [36], the bond did not mature and hence time was not a variable, whereas in the present paper, the bond has finite maturity and the bond price depends on the time to maturity.

Brennan and Schwartz [8] and Ingersoll [22] address the convertible bond pricing problem via the arbitrage pricing theory developed by Merton [30], [31] and underlying the option pricing formula of Black and Scholes [7]. In the Brennan–Schwartz [8] model, dividends and coupons are paid at discrete dates. Between these dates, the value of the firm is a geometric Brownian motion and the price of the convertible bond is governed by the *linear* partial differential equation developed by Black and Scholes [7]. This sets up a backward recursion over payment dates, which permits a numerical solution of the bond pricing problem but is not readily amenable to qualitative analysis. In Ingersoll [22], coupons are paid out continuously. For most of the results obtained in [22], dividends are zero, and because of this the bond price is again governed by a *linear* partial differential equation.

The present paper differs from the classical literature in a second respect. In [8], the bond should not be converted except possibly immediately prior to a dividend payment; in [22], the bond should not be converted except possibly at maturity. Therefore, neither of these papers needs to address the *free boundary problem* that arises if early conversion (other than at discrete dates) is optimal.

Ingersoll [22] provides a heuristic argument that the firm should call as soon as the conversion value of the bond (the value the bondholder would receive if he converts the bond to stock) rises to the call price. It is observed that firms tend to call later than this, and several reasons have been advanced to explain this departure from the model; see, e.g., [2], [3], [16], [19], [23]. We show here by a rigorous analysis of the model that, although the Ingersoll conclusion is often valid, it is also possible that the firm should call *before* the conversion value of the bond rises to the call price. In these cases, explanation of observed firm behavior is more difficult than previously believed.

The present paper assumes that a firm's value comprises equity and convertible bonds. To simplify the discussion, we assume that equity is in the form of a single share of stock, and that there is a single convertible bond. We assume that the value of the issuing firm has constant volatility, the bond continuously pays coupons at a fixed rate, and the firm continuously pays dividends at a rate that is a fixed fraction of equity. Default occurs if the coupon payments cause the firm value to fall to zero, in which case the bond has zero recovery. In this model, both the bond price and the stock price are functions of the underlying firm value. Because the stock price is the difference between firm value and bond price, and dividends are paid proportionally

to the stock price, the differential equation characterizing the bond price as a function of the firm value is *nonlinear*.

In section 2 we provide a no-arbitrage argument, which states that once the firm and the bondholder choose their call and conversion strategies, the price of the bond is the expected value under the *risk-neutral measure* of the cash flows that accrue from ownership of the bond. The determination of the optimal call and conversion strategies then becomes a Dynkin game between the firm and the bondholder, and the bond is almost a game option in the sense of Kifer [26]. In contrast to [26], here the evolution of the underlying process, the firm value, depends on the solution to the game. Kallsen and Kühn [24] consider a game option setting that includes this possibility.

Recognizing that convertible bond pricing is a game is implicit in previous work. For example, [8] observes that the pricing problem " ... results in a pair of conversion-call strategies which are in equilibrium in the sense that neither party could improve his position by adopting any other strategy." Here we make the game explicit and obtain a good qualitative description of its value. In particular, if the dividend rate is below the interest rate, then the game reduces to one of two possible optimal stopping problems, either the problem of optimal call or the problem of optimal conversion, and we are able to determine in advance from the model parameters which of these two problems is relevant.

Convertible bonds can have several features that must be captured by any model intended for practical application; see [29]. These include periods of call protection, time-dependent conversion factors, and exposure to interest rate and default risk. The model of this paper captures only the default risk, and that via a simple structural model in which default occurs at the time the firm value falls to zero. Loschak [27] allows nonconvertible senior debt and uses a more sophisticated structural model for default. Brennan and Schwartz [9] also allow senior debt. Another interesting issue is the process of conversion when bonds are held by competing investors; see Constantinides [11] and Constantinides and Rosenthal [12].

Practical models have been built around the idea that the cash flow from a convertible bond can be separated into an "equity" part, which should be discounted at the interest rate, and a "bond" part, which should be discounted at the interest rate plus a credit spread. Papers taking this approach are McConnell and Schwartz [28], Cheung and Nelken [10], Ho and Pteffer [21], Tsiveriotis and Fernandes [39], and Yigitbasioglu [40]. Ayache, Forsyth, and Vetzal [4] analyze some of this work and conclude that its failure to account for the effect of default on equity introduces significant pricing errors. This deficiency is corrected in Davis and Lischka [14], Takahashi, Kobayashi, and Nakagawa [38], and Andersen and Buffum [1], who build intensity-based models for default affecting equity value.

We describe our model in section 2 and report our main results in section 3. In particular, the Dynkin game that describes the bond price reduces to one of two optimal stopping problems and a fixed point problem. Section 4 provides a probabilistic justification for the reduction of the game to optimal stopping. Viscosity solution results concerning the Hamilton–Jacobi–Bellman equations governing the optimal stopping problems are provided in section 5. This permits the proof in section 6 of the existence and uniqueness of the solution to the fixed point problem, and this solution is the bond pricing function. Section 7 relates this paper to perpetual convertible bonds. In section 8 we provide some results on the nature of the stopping and continuation regions of the optimal stopping problems of this paper.

2. The model. We denote the value of the firm at each time t by X_t. We assume the value of the firm consists of equity and debt. The debt D_t is due to a single outstanding convertible bond. This assumption of a single bond means that all debt is called and/or converted simultaneously. We denote by S_t the total value of equity, which, following the standard finance model (e.g., Merton [31, bottom of p. 453]) is given by

$$S_t = X_t - D_t. \tag{2.1}$$

Equity owners receive dividends paid continuously over time at a rate δS_t, and the bondholder receives coupons paid continuously over time at a rate c. We assume that $\delta \geq 0$ and $c > 0$ are both constant. If there is no call or conversion prior to *maturity* T, then at maturity the bondholder receives the *par value* L from the firm, provided $X_T \geq L$. Otherwise, the bondholder receives X_T. However, at any time $t \in [0, T]$, the bondholder may *convert* the bond to stock, thereby immediately receiving stock valued at the *conversion factor* $\gamma \in (0, 1)$ times the firm value X_t. The firm value is not affected by this conversion. On the other hand, at any time t when $X_t \geq K$, the firm may *call* the bond, forcing the bondholder to either immediately surrender the bond in exchange for the *call price* K or else immediately convert the bond as described above. We assume $K \geq L > 0$; it is common to have $L = K$. If K were less than L, then L would be irrelevant since the firm could always call at maturity to avoid paying L.

In order to model the firm value process, which is the primitive in our analysis, we note that, prior to maturity, as long as the bond has not been called or converted and the firm value has not fallen to zero, there are three financial instruments in the market: the stock, the convertible bond(s), and a money market account with risk-free rate of interest $r > 0$. The wealth V_t of an investor holding Δ_t shares of stock and Γ_t convertible bonds at each time t, investing or borrowing in the money market account as necessary in order to finance this, evolves according to the stochastic differential equation

$$dV_t = \Delta_t[dS_t + \delta S_t\, dt] + \Gamma_t[dD_t + c\, dt] + r[V(t) - \Delta_t S_t - \Gamma_t D_t]\, dt, \tag{2.2}$$

and the discounted wealth thus satisfies

$$d\big(e^{-rt}V_t\big) = \Delta_t\big[d\big(e^{-rt}S_t\big) + e^{-rt}\delta S_t\, dt\big] + \Gamma_t\big[d\big(e^{-rt}D_t\big) + e^{-rt}c\, dt\big]. \tag{2.3}$$

Such an investor should not be able to produce arbitrage. To ensure this, according to the *first fundamental theorem of asset pricing*, both the discounted stock price plus the cumulative discounted dividend payments,

$$M_t = e^{-rt}S_t + \int_0^t e^{-ru}\delta S_u\, du,$$

and the discounted convertible bond price plus the cumulative discounted coupons,

$$N_t = e^{-rt}D_t + \int_0^t e^{-ru}c\, du,$$

must be local martingales under some *risk-neutral* probability measure $\mathbb{P}$ (see the argument due to [18] and developed in great generality by [15]). Adding the above equations, using the relation $X_t = S_t + D_t$, we obtain

$$dX_t = rX_t\, dt - c\, dt - \delta S_t\, dt + e^{rt}d(M_t + N_t). \tag{2.4}$$

Assuming constant volatility $\sigma > 0$ for X_t means that the local martingale term $e^{rt}d(M_t + N_t)$ is in fact equal to $\sigma X_t\, dW_t$, where W is a Brownian motion. To summarize, in our model the value of the firm evolves according to the equation

$$dX_t = rX_t\, dt - c\, dt - \delta S_t\, dt + \sigma X_t\, dW_t, \tag{2.5}$$

where W_t, $0 \le t \le T$, is a Brownian motion under the risk-neutral probability measure $\mathbb{P}$. This is the starting point of our model, and it is a common starting point for treatments of the convertible bond pricing problem; see, e.g., [8], [22]. Equation (2.5) says that under the risk-neutral measure the mean rate of growth of X_t is the interest rate r adjusted by the payouts being made.

Finally, we observe from (2.5) that $X_t \le X_0 e^{(r-\frac{1}{2}\sigma^2)t+\sigma W_t}$. Therefore, since $0 \le S_t \le X_t$ and $0 \le D_t \le X_t$, we conclude that the local martingales M_t and N_t are martingales under $\mathbb{P}$.

We adopt throughout the *standing assumption*

$$0 \le \delta < r, \tag{2.6}$$

but see Remark 3.5 below.

We generalize slightly the previous discussion by permitting the initial time to be $s \in [0,T]$ rather than requiring it to be 0. We shall price the bond at time s under the assumption that $X_s = x$. Given these initial conditions, we denote by $X_t^{s,x}$ the solution to (2.5) at time $t \in [s,T]$ and set

$$\theta_y^{s,x} \triangleq \min\{t \in [s,T] : X_t^{s,x} = y\}, \quad y \ge 0,$$

where we adopt the convention that $\min \emptyset = \infty$. The firm defaults on the bond at time $\theta_0^{s,x}$ if $\theta_0^{s,x} \le T$, and $\theta_0^{s,x} = \infty$ corresponds to no default.

The firm adopts a *call strategy* ρ and the bondholder adopts a *conversion strategy* τ. Both of these are stopping times for the filtration generated by $W_u - W_s$, $u \in [s,T]$ (augmented by $\mathbb{P}$-null sets), and they must satisfy $\rho, \tau \in [s, T \wedge \theta_0^{s,x}] \cup \{\theta_0^{s,x}\}$. We denote the set of all such stopping times by $\mathcal{S}^{s,x}$. We interpret ρ and τ to be the times of call and conversion, respectively, except on the set $\{\rho = \theta_0^{s,x}\}$, where there is no call. Similarly, on the set $\{\tau = \theta_0^{s,x}\}$, there is no conversion. On the set $\{\rho = \tau < \theta_0^{s,x}\}$, there is simultaneous call and conversion, and the conversion takes priority. This is the standard contractual specification for convertible bonds. There is no requirement that call or conversion must take place, and we capture the absence of call (respectively, conversion) by permitting $\rho = \infty$ (respectively, $\tau = \infty$) if $\theta_0^{s,x} = \infty$. The firm can call at time $\rho < \theta_0^{s,x}$ only if $X_\rho^{s,x} \ge K$. We denote by $\mathcal{S}_K^{s,x}$ the set of stopping times in $\mathcal{S}^{s,x}$ satisfying the additional condition that $X_\rho^{s,x} \ge K$ on $\rho < \theta_0^{s,x}$, and we require that $\rho \in \mathcal{S}_K^{s,x}$. Once the call and conversion strategies $\rho \in \mathcal{S}_K^{s,x}$ and $\tau \in \mathcal{S}^{s,x}$ are chosen, we use the fact that N_t stopped at $\rho \wedge \tau \wedge T$ is a martingale under $\mathbb{P}$ to write the value of the bond D_s at time s as

$$\begin{aligned} J(s,x;\rho,\tau) &\triangleq e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau\wedge T} e^{-ru}c\, du + e^{-r(\rho\wedge\tau\wedge T)}D_{\rho\wedge\tau\wedge T}\right] \\ &= e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau\wedge T} e^{-ru}c\, du + e^{-r(\rho\wedge\tau\wedge T)}\big(\mathbb{I}_{\{\tau\le\rho\wedge T\}}\gamma X_\tau^{s,x}\right. \\ &\qquad \left. + \mathbb{I}_{\{\rho<\tau\}}K + \mathbb{I}_{\{\rho\wedge\tau=\infty\}}(X_T^{s,x}\wedge L)\big)\right]. \end{aligned} \tag{2.7}$$

3. The method and principal results. We must deal with the fact that the process S_t in section 2 is endogenous. In fact, the bond price, the firm value, and the equity value S_t are related by (2.1). Just as in [8], [22], [31], and even [7], for the case of options rather than convertible bonds, we make the ansatz that there is a function $g(t,x)$ such that prior to call and conversion, $D_t = g(t, X_t)$ and hence $S_t = X_t - g(t, X_t)$. This is a reasonable step because the only source of uncertainty in the model is the uncertainty in the firm value (equivalently, the uncertainty in the Brownian motion driving the firm value), and thus all asset prices should depend on only this and the time variable.

We eventually see (Lemma 4.1 below) that if $\gamma X_t \geq K$, then it is optimal to convert, and hence $D_t = \gamma X_t$. Hence, the function $g(t,x)$ should satisfy

$$g(t,x) = \gamma x \text{ for } 0 \leq t \leq T \text{ and } x \geq \frac{K}{\gamma}. \tag{3.1}$$

Also, we expect both the value of the bond and the value of the equity to increase with increasing firm value, which is equivalent to

$$0 \leq g(t,y) - g(t,x) \leq y - x \text{ for } 0 \leq t \leq T \text{ and } 0 \leq x \leq y. \tag{3.2}$$

The bond is never worth less than its conversion value and never worth more than the firm value. Since the firm can always call when $\gamma x \leq K$, in which case the call does not result in conversion, the bond is not worth more than the call price. In other words,

$$\gamma x \leq g(t,x) \leq x \wedge K \text{ for } 0 \leq t \leq T \text{ and } 0 \leq x \leq \frac{K}{\gamma}. \tag{3.3}$$

We shall show that within the collection of functions

$$\mathcal{G} = \{g\colon [0,T] \times [0,\infty) \to [0,\infty) : g \text{ is continuous and (3.1)–(3.3) hold}\},$$

there exists a unique function g^* such that $g^*(t, X_t)$ gives a bond price consistent with our modeling assumptions.

To get started, we simply choose an arbitrary $g \in \mathcal{G}$ and define

$$S_t = X_t - g(t, X_t). \tag{3.4}$$

We substitute this value of S_t into (2.5), thereby obtaining a stochastic differential equation for X. The Lipschitz continuity (3.2) guarantees that corresponding to every initial condition $(s,x) \in [0,T] \times [0,\infty)$ this equation has a strong solution, and we thus obtain $X^{s,x}$. We proceed as in section 2 and conclude with the function J of (2.7), which we now denote J_g.

For each fixed $g \in \mathcal{G}$, we can construct a Dynkin game, where now the evolution of the underlying process is specified by (2.5) and (3.4). Kallsen and Kühn [24] show that the value of this game will be the no-arbitrage price of the bond, provided the function g has been chosen "correctly" (see the next paragraph). This game has lower and upper values

$$\underline{v}_g(s,x) \triangleq \sup_{\tau \in \mathcal{S}^{s,x}} \inf_{\rho \in \mathcal{S}_K^{s,x}} J_g(s,x;\rho,\tau), \quad \overline{v}_g(s,x) \triangleq \inf_{\rho \in \mathcal{S}_K^{s,x}} \sup_{\tau \in \mathcal{S}^{s,x}} J_g(s,x;\rho,\tau),$$

respectively. Clearly, $\underline{v}_g \leq \overline{v}_g$. In fact,

$$\underline{v}_g(s,x) = \overline{v}_g(s,x) \text{ for } 0 \leq s \leq T \text{ and } x \geq 0. \tag{3.5}$$

This is a consequence of the theory of Dynkin games, but rather than appeal to that theory, we obtain (3.5) as a by-product of our characterization of the solution of the game; see Lemma 4.1 and Propositions 4.5 and 4.6 below.

The function $\underline{v}_g = \overline{v}_g$ provides the price of the convertible bond if we choose g to be the pricing function of the convertible bond. That is to say, we want to find a function $g^* \in \mathcal{G}$ such that $\underline{v}_{g^*} = \overline{v}_{g^*} = g^*$. Let us define the operator $\mathcal{T}$ on $\mathcal{G}$ by $\mathcal{T}g \triangleq \underline{v}_g = \overline{v}_g$. We shall prove the following.

THEOREM 3.1. *$\mathcal{T}$ maps $\mathcal{G}$ into $\mathcal{G}$ and has a unique fixed point g^*.*

When $x = 0$, the only stopping time in $\mathcal{S}^{T,x}$ is $\tau \equiv T$, whereas when $0 < x \le \frac{K}{\gamma}$, the set $\mathcal{S}^{T,x}$ also contains the stopping time $\tau \equiv \infty$. When $x = 0$, the only stopping in $\mathcal{S}_K^{T,x}$ is $\rho = T$, whereas when $0 < x < K$, the only stopping time in $\mathcal{S}_K^{T,x}$ is $\rho \equiv \infty$. When $K \le x \le \frac{K}{\gamma}$, the set $\mathcal{S}_K^{T,x}$ consists of $\rho \equiv T$ and $\rho \equiv \infty$. We further have

$$J_g(T,x;\rho,\tau) = \mathbb{I}_{\{\tau \le \rho \wedge T\}} \gamma x + \mathbb{I}_{\{\rho < \tau\}} K + \mathbb{I}_{\{\rho \wedge \tau = \infty\}} (x \wedge L).$$

It is now straightforward to compute $\underline{v}_g(T,x)$ and $\overline{v}_g(T,x)$ for $0 \le x \le \frac{K}{\gamma}$, and irrespective of the choice of g, this results in the terminal convertible bond pricing function $g^*(T,\cdot)$ given by

$$g^*(T,x) = (x \wedge L) \vee (\gamma x) \text{ for } 0 \le x \le \frac{K}{\gamma}. \tag{3.6}$$

Because g^* also satisfies (3.1), we need only describe this function on $[0,T) \times \left[0, \frac{K}{\gamma}\right)$. From (3.3) we have the boundary conditions

$$g^*(t,0) = 0, \quad g^*\left(t, \frac{K}{\gamma}\right) = K \text{ for } 0 \le t \le T. \tag{3.7}$$

THEOREM 3.2 (Case I). *If $c \le rK$, the time of optimal call is the first time the conversion value γX_t rises to the call price K. The bond pricing function g^* is determined by solving the problem of optimal conversion in $[0,T] \times \left[0, \frac{K}{\gamma}\right]$. In particular, g^* is the unique continuous viscosity solution of the variational inequality*

$$\min\left\{-v_t + rv - (rx - c)v_x + \delta(x - v)v_x - \frac{1}{2}\sigma^2 x^2 v_{xx} - c, v - \gamma x\right\} = 0 \tag{3.8}$$

on $[0,T] \times [0, \frac{K}{\gamma}]$ satisfying (3.6) and (3.7).

(Case II). *If $\delta K \le c$, the time of optimal conversion is at the first time the conversion value γX_t rises to the call price K, or at maturity if the conversion value exceeds the par value. The bond pricing function g^* is determined by solving the problem of optimal call in $[0,T] \times \left[0, \frac{K}{\gamma}\right]$. In particular, g^* is the unique continuous viscosity solution of the variational inequality*

$$\max\left\{-v_t + rv - (rx - c)v_x + \delta(x - v)v_x - \frac{1}{2}\sigma^2 x^2 v_{xx} - c, v - K\right\} = 0 \tag{3.9}$$

on $[0,T] \times [0, \frac{K}{\gamma}]$ satisfying (3.6) and (3.7).

Remark 3.3. By a "continuous viscosity solution on $[0,T] \times [0, \frac{K}{\gamma}]$" we mean a continuous function on $[0,T] \times [0, \frac{K}{\gamma}]$ that is a solution in the interior of the domain, namely, $(0,T) \times (0, \frac{K}{\gamma})$.

Remark 3.4. Because of standing assumption (2.6), Cases I and II overlap. In other words, we can have $\delta K \le c \le rK$, and optimal call and conversion both occur the first time γX_t rises to K. In this case, g^* is the unique continuous viscosity solution on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$ of the partial differential equation

$$-v_t + rv - (rv - c)v_x + \delta(x - v)v_x - \frac{1}{2}\sigma^2 x^2 v_{xx} = c$$

satisfying (3.6) and (3.7).

Remark 3.5. The proofs in this paper do not actually require standing assumption (2.6), but rather that either $c \le rK$ or $\delta K \le c$. Under either of these conditions, Theorems 3.1 and 3.2 hold. However, Theorem 3.6 below requires (2.6) for the pricing of the perpetual convertible bond; see [36].

THEOREM 3.6. *As the time to maturity approaches ∞, the price of the finite-maturity convertible bond approaches the price of the perpetual convertible bond of* [36]*, which is characterized in Theorem* 7.2 *below, and this convergence is uniform in the firm value.*

4. Construction and properties of v_g.

4.1. Reduction to $[0,T] \times \left[0, \frac{K}{\gamma}\right]$.

LEMMA 4.1. *Assume that g defined on $[0,T] \times [0,\infty)$ satisfies* (3.2)*, so that* (3.4) *and* (2.5) *uniquely determine a process $X^{s,x}$ for $(s,x) \in [0,T] \times [0,\infty)$. Then*

$$\underline{v}_g(s,x) = \overline{v}_g(s,x) = \gamma x \text{ for } 0 \le s \le T \text{ and } x \ge \frac{K}{\gamma}. \tag{4.1}$$

Proof. With $\tau \equiv s$, (2.7) implies $J_g(s,x;\rho,s) = \gamma x$ for $\rho \in \mathcal{S}_K^{s,x}$, and thus

$$\underline{v}_g(s,x) \ge \inf_{\rho \in \mathcal{S}_K^{s,x}} J_g(s,x;\rho,s) = \gamma x. \tag{4.2}$$

For $x \ge \frac{K}{\gamma}$, we may set $\rho \equiv s$ and then have for every $\tau \in \mathcal{S}^{s,x}$ that $J_g(s,x;s,\tau) = \gamma x \mathbb{I}_{\{\tau = s\}} + K\mathbb{I}_{\{s<\tau\}} \le \gamma x$, and hence

$$\overline{v}_g(s,x) \le \sup_{\tau \in \mathcal{S}_K^{s,x}} J_g(s,x;s,\tau) \le \gamma x. \tag{4.3}$$

But directly from their definitions, we know that $\underline{v}_g \le \overline{v}_g$. ∎

We fix a function $g \in \mathcal{G}$ for the remainder of section 4.

4.2. Modification of payoffs. The contractual features of convertible bonds require that we define the value of the bond by (2.7) once the call and conversion strategies ρ and τ are specified. This formulation is not readily amenable to analysis, since the stopping times are allowed to take the value ∞, the different players have different sets of stopping times at their disposal, and the payoff in the event of call is not always less than or equal to the payoff in the event of conversion. In this section we create an auxiliary problem that has all these desirable features, and we subsequently show in Propositions 4.5 and 4.6 that both problems have the same value.

We restrict our attention to stopping times in $\mathcal{S}_T^{s,x} \triangleq \{\theta \in \mathcal{S}^{s,x} : \theta \le T\}$. In particular, we do not allow stopping times to take the value ∞ and we do not require the call strategy ρ to satisfy $X_\rho^{s,x} \ge K$ on $\{\rho < \theta_0^{s,x}\}$. We also change the payoffs

appearing in (2.7). We define

$$\psi(t,x) \triangleq \begin{cases} \gamma x & \text{for } 0 \le t < T,\ x \ge 0, \\ (x \wedge L) \vee (\gamma x) & \text{for } t = T,\ x \ge 0, \end{cases}$$

$$\varphi(t,x) \triangleq \begin{cases} (x \wedge K) \vee (\gamma x) & \text{for } 0 \le t < T,\ x \ge 0, \\ (x \wedge L) \vee (\gamma x) & \text{for } t = T,\ x \ge 0. \end{cases}$$

Then $\psi < \varphi$ on $[0,T) \times \big(0, \frac{K}{\gamma}\big)$ and $\psi = \varphi$ on the parabolic boundary

$$\partial_p D_0 \triangleq \left([0,T) \times \left\{0, \frac{K}{\gamma}\right\} \right) \cup \left(\{T\} \times \left[0, \frac{K}{\gamma}\right] \right). \tag{4.4}$$

For $\rho, \tau \in \mathcal{S}_T^{s,x}$, we set

$$\begin{aligned} &\widetilde{J}_g(s,x;\rho,\tau) \\ &\triangleq e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau} e^{-ru} c\, du + e^{-r(\rho\wedge\tau)} \big(\mathbb{I}_{\{\tau<\rho\}} \psi(\tau, X_\tau^{s,x}) + \mathbb{I}_{\{\rho\le\tau\}} \varphi(\rho, X_\rho^{s,x}) \big) \right]. \end{aligned}$$

The interpretation of $\widetilde{J}_g$ is that if the firm value is insufficient to pay the call price at the time of the call, then the bondholder receives the firm value. Also, call takes priority over conversion, but the bondholder receives the conversion value if that is greater than the call price at the time of the call. We show in Propositions 4.5 and 4.6 that changing the payoffs in this way does not change the value of the convertible bond pricing problem. We begin with the following straightforward modification of Lemma 4.1.

LEMMA 4.2. *For $0 \le s \le T$ and $x \ge \frac{K}{\gamma}$, we have*

$$\inf_{\rho \in \mathcal{S}_T^{s,x}} \sup_{\tau \in \mathcal{S}_T^{s,x}} \widetilde{J}_g(s,x;\rho,\tau) = \sup_{\tau \in \mathcal{S}_T^{s,x}} \inf_{\rho \in \mathcal{S}_T^{s,x}} \widetilde{J}_g(s,x;\rho,\tau) = \gamma x.$$

4.3. Technical preparations. Itô's formula implies that if h is a continuous function on $[0,T] \times \big[0, \frac{K}{\gamma}\big]$, h is $C^{1,2}$ on the interior of its domain, and the derivatives of h have limits at the boundary of its domain, then for $(s,x) \in [0,T] \times \big[0, \frac{K}{\gamma}\big]$,

$$\begin{aligned} &d\left(\int_s^t e^{-ru} c\, du + e^{-rt} h(t, X_t^{s,x}) \right) \\ &\qquad = e^{-rt}\big[-\mathcal{L}_g h(t, X_t^{s,x}) + c \big]\, dt + e^{-rt} \sigma X_t^{s,x} h_x(t, X_t^{s,x})\, dW_t \end{aligned} \tag{4.5}$$

for $t \in [s, \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T]$, where

$$\begin{aligned} \mathcal{L}_g h(t,x) \triangleq\ & -h_t(t,x) + rh(t,x) - (rx - c) h_x(t,x) \\ & + \delta\big(x - g(t,x)\big) h_x(t,x) - \frac{1}{2}\sigma^2 x^2 h_{xx}(t,x). \end{aligned} \tag{4.6}$$

LEMMA 4.3. *Let $\widetilde{c} > 0$ be given, and for $0 \le s \le T$ and $0 \le x \le \frac{K}{\gamma}$, define*

$$\begin{aligned} &k(s,x) \\ &\triangleq e^{rs}\mathbb{E}\left[\int_s^{\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T} e^{-ru} \widetilde{c}\, du + e^{-r(\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T)} \psi\left(\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T, X_{\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T}^{s,x} \right) \right] \\ &= e^{rs}\mathbb{E}\left[\int_s^{\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T} e^{-ru} \widetilde{c}\, du + e^{-r(\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T)} \varphi\left(\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T, X_{\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T}^{s,x} \right) \right], \end{aligned} \tag{4.7}$$

where we have used the fact that ψ and φ agree on the parabolic boundary $\partial_p D_0$. Then k is continuous and satisfies $k = \varphi = \psi$ on $\partial_p D_0$.

Remark 4.4. We would expect the function k to satisfy the partial differential equation $\mathcal{L}_g k = \widetilde{c}$, but since g is only continuous and not Hölder continuous with respect to time, we do not know that this equation has a classical solution. Hence, we give a probabilistic proof of Lemma 4.3.

Proof of Lemma 4.3. It is apparent that $k = \psi$ on $\partial_p D_0$. It remains to prove the continuity.

We extend g to a jointly continuous function, globally Lipschitz in its second variable, defined on $[0,T] \times \mathbb{R}$, so that for every $(s,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$, $X_t^{s,x}$ can be defined by (2.5) and (3.4) for all $t \in [s,T]$. We define $X_t^{s,x} = x$ for $t \in [0,s)$. All the processes $X^{s,x}$ are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in $C[0,T]$.

For $(s,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$, denote by $\mathbb{P}^{s,x}$ the distribution of $X^{s,x}$ on $C[0,T]$. According to [37, p. 152], $\mathbb{P}^{s,x}$ is continuous in (s,x). For $(s,x) \in [0,T] \times \mathbb{R}$, we define the measure $\mathbb{Q}^{s,x} \triangleq \delta_s \times \mathbb{P}^{s,x}$ on $[0,T] \times C[0,T]$, where δ_s denotes the unit point mass at s. Then $\mathbb{Q}^{s,x}$ is also continuous in (s,x) [6, Thm. 2.8, p. 23], which means that

$$\int_{C[0,T]} f(s_n, y)\, d\mathbb{P}^{s_n,x_n}(y) \to \int_{C[0,T]} f(s,y)\, d\mathbb{P}^{s,x}(y) \tag{4.8}$$

whenever $s_n \to s$, $x_n \to x$ and f defined on $[0,T] \times C[0,T]$ is a bounded function that is continuous except on a $\mathbb{Q}^{s,x}$-null set.

We define $\tau \colon [0,T] \times C[0,T] \to [0,T]$ by

$$\tau(s,y) \triangleq T \wedge \min\left\{ t \in [s,T] : y(t) \notin \left(0, \frac{K}{\gamma}\right) \right\},$$

so that $\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T = \tau\big(s, X^{s,x}\big)$. For $\mathbb{P}$-almost every ω, if $\tau(s, X^{s,x}(\omega)) < T$, then there is a sequence $\epsilon_n \downarrow 0$, depending on s, x, and ω, such that $X^{s,x}_{\tau(s,X^{s,x}(\omega))+\epsilon_n} \notin \left[0, \frac{K}{\gamma}\right]$ for every n. Indeed, if $X^{s,x}(\omega)$ exits $\left(0, \frac{K}{\gamma}\right)$ at 0, this follows from the fact that $c > 0$ and all other terms on the right-hand side of (2.5) are zero at the time of exit. On the other hand, if $X^{s,x}(\omega)$ exits $\left(0, \frac{K}{\gamma}\right)$ at $\frac{K}{\gamma}$, this follows from the nondegeneracy of the diffusion term in (2.5). Because of the fact that, almost surely, if $X^{s,x}$ exits $(0, \frac{K}{\gamma})$ strictly prior to time T, then it also exits $[0, \frac{K}{\gamma}]$ immediately thereafter, a small perturbation of the path of $X^{s,x}$ results in a small perturbation of the time of exit. More precisely, for every $(s,x) \in [0,T] \times \mathbb{R}$, τ is continuous except on a $\mathbb{Q}^{s,x}$-null set. We conclude by rewriting (4.7) as

$$\begin{aligned} &k(s_n, x_n) \\ &= e^{rs_n} \int_{C[0,T]} \left[\int_{s_n}^{\tau(s_n,y)} e^{-ru} \widetilde{c}\, du + e^{-r\tau(s_n,y)} \psi\big(\tau(s_n,y), y(\tau(s_n,y))\big) \right] d\mathbb{P}^{s_n,x_n}(y) \end{aligned}$$

and observing that because the argument of ψ is in $\partial_p D_0$, where ψ is bounded and continuous, (4.8) implies $k(s_n,x_n) \to k(s,x)$ as $s_n \to s$, $x_n \to x$. ∎

4.4. Characterization of game value.

Proposition 4.5 (Case I). *Assume $c \le rK$. In this case, we define*

$$v_g(s,x) \triangleq \sup_{\tau \in \mathcal{S}_T^{s,x}, \tau \le \theta_{\frac{K}{\gamma}}^{s,x}} e^{rs} \mathbb{E}\left[\int_s^\tau e^{-ru} c\, du + e^{-r\tau} \psi(\tau, X_\tau^{s,x}) \right] \tag{4.9}$$

for $(s,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$, *and we define* $v_g(s,x) \triangleq \gamma x$ *for* $(s,x) \in [0,T] \times (\frac{K}{\gamma}, \infty)$. *Then* $v_g = \underline{v}_g = \overline{v}_g$ *on* $[0,T] \times \left[0,\infty\right)$. *Furthermore,*

$$v_g(s,x) = \inf_{\rho \in \mathcal{S}_T^{s,x}} \sup_{\tau \in \mathcal{S}_T^{s,x}} \widetilde{J}_g(s,x;\rho,\tau) = \sup_{\tau \in \mathcal{S}_T^{s,x}} \inf_{\rho \in \mathcal{S}_T^{s,x}} \widetilde{J}_g(s,x;\rho,\tau) \tag{4.10}$$

for $0 \le s \le T$ *and* $x \ge 0$.

Proof. The claims about v_g for $x > \frac{K}{\gamma}$ follow immediately from Lemmas 4.1 and 4.2. We thus restrict our attention to $0 \le x \le \frac{K}{\gamma}$.

Step 1: *Construction of an upper bound on* v_g. Define $h_1(t,x) \triangleq K$ and $h_2(t,x) \triangleq x$ for $0 \le t \le T$ and $0 \le x \le \frac{K}{\gamma}$. Both h_1 and h_2 dominate ψ on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$. Because $c \le rK$, we have $-\mathcal{L}_g h_1 + c \le 0$. Therefore, for any stopping time $\tau \in \mathcal{S}_T^{s,x}$ satisfying $\tau \le \theta^{s,x}_{\frac{K}{\gamma}}$, (4.5) implies

$$\begin{aligned} h_1(s,x) &\ge e^{rs}\mathbb{E}\left[\int_s^\tau e^{-ru} c\, du + e^{-r\tau} h_1(\tau, X_\tau^{s,x})\right] \\ &\ge e^{rs}\mathbb{E}\left[\int_s^\tau e^{-ru} c\, du + e^{-r\tau} \psi(\tau, X_\tau^{s,x})\right]. \end{aligned} \tag{4.11}$$

It follows that $v_g \le h_1$ on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$. On the other hand,

$$\mathcal{L}_g h_2 = c + \delta(x - g(x)) \ge c \tag{4.12}$$

because of (3.3), and the above argument applied with h_2 in place of h_1 yields $v_g \le h_2$ on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$. We conclude that

$$v_g(s,x) \le x \wedge K \text{ for } 0 \le s \le T,\ 0 \le x \le \frac{K}{\gamma}. \tag{4.13}$$

By definition, $v_g(T,\cdot) = \psi(T,\cdot) = \varphi(T,\cdot)$. Because of this and (4.13),

$$v_g(s,x) \le \varphi(s,x) \text{ for } 0 \le s \le T,\ 0 \le x \le \frac{K}{\gamma}. \tag{4.14}$$

Step 2: *Optimal stopping time.* The theory of optimal stopping we use here requires that we replace ψ on the right-hand side of (4.9) with a continuous function. Let $\widetilde{c} \in (0,c)$ be given, and let k be the continuous function defined by (4.7). For $0 \le s < T$ and $0 < x < \frac{K}{\gamma}$, we have

$$\begin{aligned} k(s,x) &< e^{rs}\mathbb{E}\left[\int_s^{\theta_0^{s,x} \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge T} e^{-ru} c\, du \right. \\ &\qquad \left. + e^{-r(\theta_0^{s,x} \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge T)} \psi\left(\theta_0^{s,x} \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge T, X^{s,x}_{\theta_0^{s,x} \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge T}\right)\right] \\ &\le v_g(s,x). \end{aligned} \tag{4.15}$$

Set $\widetilde{\psi} = \psi \vee k$. Since $k(T,x) = \psi(T,x) \ge \gamma x$ for $0 \le x \le \frac{K}{\gamma}$ and $\psi(s,x) = \gamma x$ for $0 \le s < T$, $0 \le x \le \frac{K}{\gamma}$, we have $\widetilde{\psi}(s,x) = \max\{\gamma x, k(s,x)\}$ for $0 \le s \le T$, $0 \le x \le \frac{K}{\gamma}$. Being the maximum of two continuous functions, $\widetilde{\psi}$ is continuous. Also, $\psi \le \widetilde{\psi} \le v_g$.

According to the principle of dynamic programming,

$$v_g(s,x) = \sup_{\tau \in \mathcal{S}_T^{s,x}, \tau \le \theta^{s,x}_{\frac{K}{\gamma}}} e^{rs}\mathbb{E}\left[\int_s^\tau e^{-ru}c\,du + e^{-r\tau}v_g(\tau, X_\tau^{s,x})\right]. \tag{4.16}$$

Comparing (4.9) and (4.16), we see that

$$v_g(s,x) = \sup_{\tau \in \mathcal{S}_T^{s,x}, \tau \le \theta^{s,x}_{\frac{K}{\gamma}}} e^{rs}\mathbb{E}\left[\int_s^\tau e^{-ru}c\,du + e^{-r\tau}\widetilde{\psi}(\tau, X_\tau^{s,x})\right]. \tag{4.17}$$

We fix $(s,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$ and define

$$Y_t^{s,x} \triangleq e^{rs}\left[\int_s^t e^{-ru}c\,du + e^{-rt}v_g(t, X_t^{s,x})\right] \text{ for } s \le t \le \theta_0^{s,x} \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge T.$$

We set $\overline{\tau} \triangleq \min\{t \in [s,T] : v_g(t, X_t^{s,x}) = \widetilde{\psi}(t, X_t^{s,x})\}$. Since $v_g = \widetilde{\psi}$ on $\partial_p D_0$, we have $\overline{\tau} \le \theta_0^{s,x} \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge T$. According to the theory of optimal stopping applied to (4.17) (see, e.g., [25, Thms. D.12, D.13] or [35, pp. 124–127]), $Y^{s,x}_{t\wedge\theta_0^{s,x}\wedge\theta^{s,x}_{\frac{K}{\gamma}}}$ is a supermartingale and the stopped process $Y^{s,x}_{t\wedge\overline{\tau}}$ is a martingale.

Step 3: *Optimal strategies for the game.* From (4.15) and the fact that $\widetilde{\psi} = \psi \vee k$, we see that $\overline{\tau} = \min\{t \in [s,T] : v_g(t, X_t^{s,x}) = \psi(t, X_t^{s,x})\}$ and

$$v_g(\overline{\tau}, X_{\overline{\tau}}^{s,x}) = \psi(\overline{\tau}, X_{\overline{\tau}}^{s,x}) = \begin{cases} \gamma X_{\overline{\tau}}^{s,x} & \text{if } \overline{\tau} < T, \\ (X_T^{s,x} \wedge L) \vee (\gamma X_T^{s,x}) & \text{if } \overline{\tau} = T. \end{cases} \tag{4.18}$$

Define $\tau^* = \infty$ if $\overline{\tau} = T$ and $0 < X_T^{s,x} < \frac{L}{\gamma}$, and define $\tau^* = \overline{\tau}$ otherwise so that $\tau^* \in \mathcal{S}^{s,x}$ and $\overline{\tau} = \tau^* \wedge T$. We have

$$v_g(\tau^*, X_{\tau^*}^{s,x}) = \gamma X_{\tau^*}^{s,x} \text{ if } \tau^* \le T. \tag{4.19}$$

For every $\rho \in \mathcal{S}_K^{s,x}$, (4.19), (4.13), and the fact that $v_g(T,x) = x \wedge L$ when $0 \le x \le \frac{L}{\gamma}$ imply

$$\begin{aligned} &J_g(s,x;\rho,\tau^*) \\ &= e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau^*\wedge T} e^{-ru}c\,du + e^{-r(\rho\wedge\tau^*\wedge T)}\big(\mathbb{I}_{\{\tau^*\le\rho\wedge T\}}\gamma X_{\tau^*}^{s,x} + \mathbb{I}_{\{\rho<\tau^*\}}K \right. \\ &\qquad \left. + \mathbb{I}_{\{\rho\wedge\tau^*=\infty\}}(X_T^{s,x} \wedge L)\big)\right] \\ &\ge e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau^*\wedge T} e^{-ru}c\,du + e^{-r(\rho\wedge\tau^*\wedge T)}v_g(\rho\wedge\tau^*\wedge T, X^{s,x}_{\rho\wedge\tau^*\wedge T})\right] \\ &= \mathbb{E}Y^{s,x}_{\rho\wedge\overline{\tau}} = \mathbb{E}Y_s^{s,x} = v_g(s,x). \end{aligned} \tag{4.20}$$

This implies $\underline{v}_g(s,x) \ge v_g(s,x)$.

To show that $v_g(s,x) \geq \overline{v}_g(s,x)$, we set $\rho^* \triangleq \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x}$, which is in $\mathcal{S}_K^{s,x}$. For every $\tau \in \mathcal{S}^{s,x}$, we have $\rho^* \wedge \tau \wedge T \in \mathcal{S}_T^{s,x}$, and thus

$$\begin{aligned}
&J_g(s,x;\rho^*,\tau)\\
&= e^{rs}\mathbb{E}\left[\int_s^{\rho^*\wedge\tau\wedge T} e^{-ru}c\,du + e^{-r(\rho^*\wedge\tau\wedge T)}\big(\mathbb{I}_{\{\tau\leq\rho^*\wedge T\}}\gamma X_\tau^{s,x} + \mathbb{I}_{\{\rho^*<\tau\}}K\right.\\
&\qquad\left. + \mathbb{I}_{\{\rho^*\wedge\tau=\infty\}}(X_T^{s,x}\wedge L)\big)\right]\\
&\leq e^{rs}\mathbb{E}\left[\int_s^{\rho^*\wedge\tau\wedge T} e^{-ru}c\,du + e^{-r(\rho^*\wedge\tau\wedge T)}\psi(\rho^*\wedge\tau\wedge T, X_{\rho^*\wedge\tau\wedge T}^{s,x})\right]\\
&\leq v_g(s,x). \qquad (4.21)
\end{aligned}$$

This implies $\overline{v}_g(s,x) \leq v_g(s,x)$. We conclude that $v_g = \underline{v}_g = \overline{v}_g$.

Step 4: *Proof of* (4.10). With $\overline{\tau} \in \mathcal{S}_T^{s,x}$ as defined in Step 2, we have from (4.14) and (4.18) that for every $\rho \in \mathcal{S}_T^{s,x}$,

$$\begin{aligned}
&\widetilde{J}_g(s,x;\rho,\overline{\tau})\\
&= e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\overline{\tau}} e^{-ru}c\,du + e^{-r(\rho\wedge\overline{\tau})}\big(\mathbb{I}_{\{\overline{\tau}<\rho\}}\gamma X_{\overline{\tau}}^{s,x} + \mathbb{I}_{\{\rho\leq\overline{\tau}\}}\varphi(\rho, X_\rho^{s,x})\big)\right]\\
&\geq e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\overline{\tau}} e^{-ru}c\,du + e^{-r(\rho\wedge\overline{\tau})}v_g(\rho\wedge\overline{\tau}, X_{\rho\wedge\overline{\tau}}^{s,x})\right]\\
&= \mathbb{E}Y_{\rho\wedge\overline{\tau}}^{s,x} \;=\; \mathbb{E}Y_s^{s,x} \;=\; v_g(s,x).
\end{aligned}$$

On the other hand, with ρ^* defined as in Step 3 and

$$\overline{\rho} \triangleq \rho^* \wedge T = \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T \in \mathcal{S}_T^{s,x}, \qquad (4.22)$$

we have $\varphi(\overline{\rho}, X_{\overline{\rho}}^{s,x}) = \psi(\overline{\rho}, X_{\overline{\rho}}^{s,x})$. Thus, for every $\tau \in \mathcal{S}_T^{s,x}$,

$$\begin{aligned}
&\widetilde{J}_g(s,x;\overline{\rho},\tau)\\
&= e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}\wedge\tau} e^{-ru}c\,du + e^{-r(\overline{\rho}\wedge\tau)}\big(\mathbb{I}_{\{\tau<\overline{\rho}\}}\gamma X_\tau^{s,x} + \mathbb{I}_{\{\overline{\rho}\leq\tau\}}\varphi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\big)\right]\\
&= e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}\wedge\tau} e^{-ru}c\,du + e^{-r(\overline{\rho}\wedge\tau)}\psi(\overline{\rho}\wedge\tau, X_{\overline{\rho}\wedge\tau}^{s,x})\right]\\
&\leq v_g(s,x). \qquad (4.23)
\end{aligned}$$

We complete the argument as in Step 3. □

PROPOSITION 4.6 (Case II). *Assume $\delta K \leq c$. In this case, we define*

$$v_g(s,x) \triangleq \inf_{\rho\in\mathcal{S}_T^{s,x},\rho\leq\theta_{\frac{K}{\gamma}}^{s,x}} e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}\varphi(\rho, X_\rho^{s,x})\right] \qquad (4.24)$$

for $(s,x) \in [0,T]\times\big[0,\frac{K}{\gamma}\big]$, and we define $v_g(s,x) \triangleq \gamma x$ for $(s,x) \in [0,T]\times(\frac{K}{\gamma},\infty)$. Then $v_g = \underline{v}_g = \overline{v}_g$ on $[0,T]\times\big[0,\infty\big)$. Furthermore,

$$v_g(s,x) = \inf_{\rho\in\mathcal{S}_T^{s,x}}\sup_{\tau\in\mathcal{S}_T^{s,x}} \widetilde{J}_g(s,x;\rho,\tau) = \sup_{\tau\in\mathcal{S}_T^{s,x}}\inf_{\rho\in\mathcal{S}_T^{s,x}} \widetilde{J}_g(s,x;\rho,\tau) \qquad (4.25)$$

for $0 \leq s \leq T$ and $x \geq 0$.

Proof. The claims about v_g for $x > \frac{K}{\gamma}$ follow immediately from Lemmas 4.1 and 4.2. We thus restrict our attention to $0 \le x \le \frac{K}{\gamma}$.

Step 1: *Construction of bounds on* v_g. Define $h_3(t,x) \triangleq \gamma x$ and $h_2(t,x) \triangleq x$ for $0 \le t \le T$ and $0 \le x \le \frac{K}{\gamma}$, so that $h_3 \le \varphi \le h_2$. For $(t,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$, we have (4.12) and

$$\mathcal{L}_g h_3(t,x) = c\gamma + \delta\gamma\big(x - g(t,x)\big) \le c + (1-\gamma)(\delta\gamma x - c) \le c. \tag{4.26}$$

Let $\rho \in \mathcal{S}_T^{s,x}$ satisfy $\rho \le \theta_{\frac{K}{\gamma}}^{s,x}$, and apply (4.5) and (4.26) to conclude

$$\begin{aligned} h_3(s,x) &\le e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}h_3(\rho, X_\rho^{s,x})\right] \\ &\le e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}\varphi(\rho, X_\rho^{s,x})\right]. \end{aligned} \tag{4.27}$$

Taking the infimum over ρ, we obtain

$$\gamma x \le v_g(s,x) \text{ for } 0 \le s \le T,\ 0 \le x \le \frac{K}{\gamma}. \tag{4.28}$$

We repeat the above argument with h_2 and $\rho = \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}} \wedge T$, using (4.12) to reverse the first inequality and $\varphi \le h_2$ to reverse the second, to obtain

$$\begin{aligned} h_2(s,x) &\ge e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}h_2(\rho, X_\rho^{s,x})\right] \\ &\ge e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}\varphi(\rho, X_\rho^{s,x})\right] \\ &\ge v_g(s,x) \text{ for } 0 \le s \le T,\ 0 \le x \le \frac{K}{\gamma}. \end{aligned} \tag{4.29}$$

In fact, since for $(s,x) \in [0,T) \times \left(0, \frac{K}{\gamma}\right)$, with positive probability $X^{s,x}$ exits $[0,T] \times \left(0, \frac{K}{\gamma}\right)$ through the set $\{T\} \times \left(L, \frac{K}{\gamma}\right]$, where h_2 is strictly greater than φ, the second inequality in (4.29) is strict for such (s,x). This implies

$$v_g(s,x) < x \text{ for } 0 \le s < T,\ 0 < x < \frac{K}{\gamma}. \tag{4.30}$$

Step 2: *Optimal stopping time.* Let $\widetilde{c} \in (c, \infty)$ be given and let k be defined by (4.7). For $0 \le s < T$ and $0 < x < \frac{K}{\gamma}$, using the second part of (4.7), we have

$$\begin{aligned} k(s,x) &> e^{rs}\mathbb{E}\left[\int_s^{\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T} e^{-ru}c\,du \right. \\ &\qquad \left. + e^{-r(\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T)} \varphi\left(\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T, X_{\theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T}^{s,x}\right)\right] \\ &\ge v_g(s,x). \end{aligned} \tag{4.31}$$

We set $\widetilde{\varphi} = \varphi \wedge k$. Because $\varphi(T,x) = k(T,x) \le x \wedge K$ for $0 \le x \le \frac{K}{\gamma}$, and $\varphi(t,x) = x \wedge K$ for $0 \le t < T$, $0 \le x \le \frac{K}{\gamma}$, we have

$$\widetilde{\varphi}(t,x) = (x \wedge K) \wedge k(t,x) \text{ for } 0 \le t \le T,\ 0 \le x \le \frac{K}{\gamma}.$$

This shows that $\widetilde{\varphi}$ is continuous. From (4.24) we have $v_g \le \varphi$, and hence $v_g \le \widetilde{\varphi} \le \varphi$. According to the principle of dynamic programming,

$$v_g(s,x) \triangleq \inf_{\rho \in \mathcal{S}_T^{s,x}, \rho \le \theta_{\frac{K}{\gamma}}^{s,x}} e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}v_g(\rho, X_\rho^{s,x})\right]. \tag{4.32}$$

Comparing (4.24) and (4.32), we see that

$$v_g(s,x) \triangleq \inf_{\rho \in \mathcal{S}_T^{s,x}, \rho \le \theta_{\frac{K}{\gamma}}^{s,x}} e^{rs}\mathbb{E}\left[\int_s^\rho e^{-ru}c\,du + e^{-r\rho}\widetilde{\varphi}(\rho, X_\rho^{s,x})\right]. \tag{4.33}$$

We fix $(s,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$ and define

$$Z_t^{s,x} = e^{rs}\left[\int_s^t e^{-ru}c\,du + e^{-rt}v_g(t, X_t^{s,x})\right] \text{ for } s \le t \le \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T.$$

We set

$$\overline{\rho} \triangleq \min\{t \in [s,T] : v_g(t, X_t^{s,x}) = \widetilde{\varphi}(t, X_t^{s,x})\}. \tag{4.34}$$

Since $v_g = \widetilde{\varphi}$ on $\partial_p D_0$, we have $\overline{\rho} \le \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T$. According to the theory of optimal stopping, $Z_{t \wedge \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x}}^{s,x}$ is a submartingale and the stopped process $Z_{t \wedge \overline{\rho}}^{s,x}$ is a martingale.

Step 3: *Optimal strategies for the game.* Because of (4.31), we have that $\overline{\rho} = \min\{t \in [s,T] : v_g(t, X_t^{s,x}) = \varphi(t, X_t^{s,x})\}$. In particular, $v_g(\overline{\rho}, X_{\overline{\rho}}^{s,x}) = X_{\overline{\rho}}^{s,x} \wedge K$ on $\{\overline{\rho} < T\}$. Inequality (4.30) then implies

$$v_g(\overline{\rho}, X_{\overline{\rho}}^{s,x}) = K < X_{\overline{\rho}}^{s,x} \text{ on } \{\overline{\rho} < \theta_0^{s,x} \wedge T\}. \tag{4.35}$$

Define

$$\rho^* \triangleq \begin{cases} \infty & \text{if } \overline{\rho} = T \text{ and } X_{\overline{\rho}}^{s,x} > 0, \\ \overline{\rho} & \text{otherwise,} \end{cases}$$

so that $\rho^* \in \mathcal{S}_K^{s,x}$ and $\overline{\rho} = \rho^* \wedge T$. For every $\tau \in \mathcal{S}^{s,x}$, (4.28), (4.35), and the fact that $v_g(T,x) = \varphi(T,x) \ge x \wedge L$ when $0 \le x \le \frac{K}{\gamma}$ imply

$$\begin{aligned} &J_g(s,x;\rho^*,\tau) \\ &= e^{rs}\mathbb{E}\left[\int_s^{\rho^* \wedge \tau \wedge T} ce^{-ru}\,du + e^{-r(\rho^* \wedge \tau \wedge T)}\big(\mathbb{I}_{\{\tau \le \rho^* \wedge T\}}\gamma X_\tau^{s,x} + \mathbb{I}_{\{\rho^* < \tau\}}K \right. \\ &\qquad \left. + \mathbb{I}_{\{\rho^* \wedge \tau = \infty\}}(X_T^{s,x} \wedge L)\big)\right] \\ &\le e^{rs}\mathbb{E}\left[\int_s^{\rho^* \wedge \tau \wedge T} ce^{-ru}\,du + e^{-r(\rho^* \wedge \tau \wedge T)}v_g(\rho^* \wedge \tau \wedge T, X_{\rho^* \wedge \tau \wedge T}^{s,x})\right] \\ &= \mathbb{E}Z_{\overline{\rho} \wedge \tau}^{s,x} = Z_s^{s,x} = v_g(s,x). \end{aligned} \tag{4.36}$$

This implies $\overline{v}_g(s,x) \le v_g(s,x)$.

We set

$$\overline{\tau} \triangleq \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge T, \tag{4.37}$$

$$\tau^* \triangleq \begin{cases} \overline{\tau} & \text{if } \overline{\tau} < T, \\ T & \text{if } \overline{\tau} = T,\ X_T^{s,x} \geq \frac{L}{\gamma} \text{ or if } \overline{\tau} = T,\ X_T^{s,x} = 0, \\ \infty & \text{if } \overline{\tau} = T, 0 < X_T^{s,x} < \frac{L}{\gamma}, \end{cases} \tag{4.38}$$

so that $\tau^* \in \mathcal{S}^{s,x}$. For every $\rho \in \mathcal{S}_K^{s,x}$, we have

$$\begin{aligned} &J_g(s,x;\rho,\tau^*) \\ &= e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau^*\wedge T} ce^{-ru}\,du + e^{-r(\rho\wedge\tau^*\wedge T)}\big(\mathbb{I}_{\{\tau^*\leq\rho\wedge T\}}\gamma X_\tau^{s,x} + \mathbb{I}_{\{\rho<\tau^*\}}K \right. \\ &\qquad \left. + \mathbb{I}_{\{\rho\wedge\tau^*=\infty\}}(X_T^{s,x}\wedge L)\big)\right] \\ &\geq e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\tau^*\wedge T} ce^{-ru}\,du + e^{-r(\rho\wedge\tau^*\wedge T)}\varphi(\rho^*\wedge\tau\wedge T, X_{\rho\wedge\tau^*\wedge T}^{s,x})\right] \\ &\geq v_g(s,x). \end{aligned} \tag{4.39}$$

This implies $\underline{v}_g(s,x) \geq v_g(s,x)$. We conclude that $v_g = \underline{v}_g = \overline{v}_g$.

Step 4. *Proof of* (4.25). With $\overline{\rho} \in \mathcal{S}^{s,x}$ given by (4.34), we have $v_g(\overline{\rho}, X_{\overline{\rho}}^{s,x}) = \varphi(\overline{\rho}\wedge\tau, X_{\overline{\rho}\wedge\tau}^{s,x})$, and (4.28) implies that for $\tau \in \mathcal{S}_T^{s,x}$,

$$\begin{aligned} &\widetilde{J}_g(s,x;\overline{\rho},\tau) \\ &= e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}\wedge\tau} e^{-ru}c\,du + e^{-r(\overline{\rho}\wedge\tau)}\big(\mathbb{I}_{\{\tau<\overline{\rho}\}}\gamma X_\tau^{s,x} + \mathbb{I}_{\{\overline{\rho}\leq\tau\}}\varphi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\big)\right] \\ &\leq e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}\wedge\tau} e^{-ru}c\,du + e^{-r(\overline{\rho}\wedge\tau)}v_g(\overline{\rho}\wedge\tau, X_{\overline{\rho}\wedge\tau}^{s,x})\right] \\ &= \mathbb{E}Z_{\overline{\rho}\wedge\tau}^{s,x} \;=\; Z_s^{s,x} \;=\; v_g(s,x). \end{aligned}$$

With $\overline{\tau} \in \mathcal{S}_T^{s,x}$ defined by (4.37), we have for every $\rho \in \mathcal{S}_T^{s,x}$,

$$\begin{aligned} &\widetilde{J}_g(s,x;\rho,\overline{\tau}) \\ &= e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\overline{\tau}} e^{-rs}c\,du + e^{-r(\rho\wedge\overline{\tau})}\big(\mathbb{I}_{\{\overline{\tau}<\rho\}}\gamma X_{\overline{\tau}}^{s,x} + \mathbb{I}_{\{\rho\leq\overline{\tau}\}}\varphi(\rho, X_\rho^{s,x})\big)\right] \\ &= e^{rs}\mathbb{E}\left[\int_s^{\rho\wedge\overline{\tau}} e^{-rs}c\,du + e^{-r(\rho\wedge\overline{\tau})}\varphi(\rho\wedge\overline{\tau}, X_{\rho\wedge\overline{\tau}}^{s,x})\right] \;\geq\; v_g(s,x). \end{aligned} \tag{4.40}$$

We complete the argument as in Step 3. □

PROPOSITION 4.7 (overlapping case). *Assume* $\delta K \leq c \leq rK$. *In this case,* v_g *defined by* (4.9) *agrees with* v_g *defined by* (4.24), *and for* $0 \leq s \leq T$ *and* $0 \leq x \leq \frac{K}{\gamma}$,

$$\begin{aligned} &v_g(s,x) \\ &= e^{rs}\mathbb{E}\left[\int_s^{\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T} e^{-ru}c\,du + e^{-r(\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T)}\psi\left(\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T, X_{\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T}^{s,x}\right)\right] \\ &= e^{rs}\mathbb{E}\left[\int_s^{\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T} e^{-ru}c\,du + e^{-r(\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T)}\varphi\left(\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T, X_{\theta_0^{s,x}\wedge\theta_{\frac{K}{\gamma}}^{s,x}\wedge T}^{s,x}\right)\right]. \end{aligned} \tag{4.41}$$

Furthermore,

$$\gamma x < v_g(s,x) < x \wedge K \text{ for } 0 \le s < T \text{ and } 0 < x < \frac{K}{\gamma}. \tag{4.42}$$

Proof. The function v_g defined by (4.9) satisfies (4.10), the function v_g defined by (4.24) satisfies (4.25), and so these definitions of v_g coincide. With $\overline{\rho} \triangleq \overline{\tau}$ given by (4.22) and (4.37), inequalities (4.23) and (4.40) imply

$$\begin{aligned}
&e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}} e^{-ru}c\,du + e^{-r\overline{\rho}}\varphi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\right] \\
&= e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}} e^{-ru}c\,du + e^{-r\overline{\rho}}\psi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\right] \\
&\le v_g(s,x) \\
&\le e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}} e^{-ru}c\,du + e^{-r\overline{\rho}}\varphi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\right],
\end{aligned}$$

which gives us (4.41).

We return to (4.11), replacing τ with $\overline{\rho}$ and using the fact that when $0 \le s < T$ and $0 < x < \frac{K}{\gamma}$, there is a positive probability that $X^{s,x}$ exits $[0,T] \times \left[0, \frac{K}{\gamma}\right]$ through the set $\left\{(t,x) : t = T, 0 < x < \frac{K}{\gamma}\right\}$, where $h_1 = K$ is strictly larger than ψ. This implies

$$K > e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}} e^{-ru}c\,du + e^{-r\overline{\rho}}\psi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\right] = v_g(s,x).$$

The second inequality in (4.42) follows from this and (4.30). For the first inequality in (4.42), we replace ρ in (4.27) with $\overline{\rho}$ and use the fact that when $0 \le s < T$ and $0 < x < \frac{K}{\gamma}$, there is a positive probability that $X^{s,x}$ exits $[0,T] \times \left[0, \frac{K}{\gamma}\right]$ through the set $\left\{(t,x) : t = T, 0 < x < \frac{L}{\gamma}\right\}$, where $h_3 = \gamma x$ is strictly smaller than φ, to obtain

$$\gamma x < e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho}} e^{-ru}c\,du + e^{-r\overline{\rho}}\varphi(\overline{\rho}, X_{\overline{\rho}}^{s,x})\right] = v_g(s,x). \qquad \square$$

4.5. Membership of v_g in $\mathcal{G}$. To show that $v_g \in \mathcal{G}$ whenever $g \in \mathcal{G}$, we must verify that v_g is continuous and satisfies (3.1)–(3.3). Property (3.1) is provided by Propositions 4.5 and 4.6. When $c \le rK$, we obtain the lower bound in (3.3) directly from (4.9) and the fact that $\psi \ge \gamma x$, and (4.13) provides the upper bound. When $\delta K \le C$, the upper bound in (3.3) comes from (4.24), the fact that $\varphi \le K$ on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$, and (4.29). The lower bound comes from (4.28). It remains to verify that v_g is continuous and satisfies (3.2), which is the subject of this section.

LEMMA 4.8. *We have*

$$0 \le v_g(s,y) - v_g(s,x) \le y - x \text{ for } 0 \le s \le T \text{ and } 0 \le x \le y. \tag{4.43}$$

Proof. In Step 4 of the proofs of Propositions 4.5 and 4.6, we produced stopping times $\overline{\rho}, \overline{\tau} \in \mathcal{S}_T^{s,x}$ such that

$$\widetilde{J}_g(s,x;\overline{\rho},\tau) \le v_g(s,x) \le \widetilde{J}_g(s,x;\rho,\overline{\tau}) \text{ for all } \rho, \tau \in \mathcal{S}_T^{s,x}. \tag{4.44}$$

It follows from this that $v_g(s,x) = \widetilde{J}_g(s,x;\overline{\rho},\overline{\tau})$. Relation (4.44) was developed for $(s,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right]$, but in light of Lemma 4.2, it holds as well for $(s,x) \in [0,T] \times \left[\frac{K}{\gamma}, \infty\right)$ if we define $\overline{\rho} = \overline{\tau} = s$ in this case.

We note that ψ and φ satisfy (3.2), and we use the representations (4.10), (4.25) to show that v_g does as well. Without loss of generality, we consider only the case $s = 0$. We let $0 \le x \le y < \infty$ be given. Then $X_t^{0,x} \le X_t^{0,y}$ for $0 \le t \le T$, almost surely, and $\mathcal{S}_T^{0,x} \subset \mathcal{S}_T^{0,y}$.

Consider the nonnegative martingale $Z_t = e^{-\sigma W_t - \frac{1}{2}\sigma^2 t}$. We compute

$$\begin{aligned} d\big((X_t^{0,y} - X_t^{0,x})Z_t\big) &= (r-\sigma^2)(X_t^{0,y} - X_t^{0,x})Z_t\,dt \\ &\quad -\delta\big[(X_t^{0,y} - X_t^{0,x}) - \big(g(t,X_t^{0,y}) - g(t,X_t^{0,x})\big)\big]Z_t\,dt \\ &\le (r-\sigma^2)(X_t^{0,y} - X_t^{0,x})Z_t\,dt. \end{aligned}$$

Gronwall's inequality implies $(X_t^{0,y} - X_t^{0,x})Z_t \le (y-x)e^{(r-\sigma^2)t}$, or equivalently,

$$e^{-rt}\big(X_t^{0,y} - X_t^{0,x}\big) \le (y-x)e^{\sigma W_t - \frac{1}{2}\sigma^2 t},\ 0 \le t \le \tau_0^{0,x}.$$

Let $\overline{\tau}, \overline{\rho} \in \mathcal{S}_T^{0,x}$ be the stopping times appearing in (4.44) corresponding to the initial condition $(0,x)$. For every $\tau \in \mathcal{S}^{0,x}$, we have

$$\begin{aligned} &\widetilde{J}_g(0,x;\overline{\rho},\tau) \\ &= \mathbb{E}\left[\int_0^{\overline{\rho}\wedge\tau} ce^{-ru}\,du + e^{-r(\overline{\rho}\wedge\tau)}\big(\mathbb{I}_{\{\tau<\overline{\rho}\}}\psi(\tau, X_\tau^{0,x}) + \mathbb{I}_{\{\overline{\rho}\le\tau\}}\varphi(\overline{\rho}, X_{\overline{\rho}}^{0,x})\big)\right] \\ &= \widetilde{J}_g(0,y;\overline{\rho},\tau) - \mathbb{E}\left[e^{-r(\overline{\rho}\wedge\tau)}\Big(\mathbb{I}_{\{\tau<\overline{\rho}\}}\big(\psi(X_\tau^{0,y}) - \psi(X_\tau^{0,x})\big)\right. \\ &\qquad\qquad \left. +\mathbb{I}_{\{\overline{\rho}\le\tau\}}\big(\varphi(\overline{\rho}, X_{\overline{\rho}}^{0,y}) - \varphi(\overline{\rho}, X_{\overline{\rho}}^{0,x})\big)\Big)\right] \\ &\ge \widetilde{J}_g(0,y;\overline{\rho},\tau) - \mathbb{E}\left[e^{-r(\overline{\rho}\wedge\tau)}(X_{\overline{\rho}\wedge\tau}^{0,y} - X_{\overline{\rho}\wedge\tau}^{0,x})\right] \\ &\ge \widetilde{J}_g(0,y;\overline{\rho},\tau) - (y-x)\mathbb{E}e^{\sigma W(\overline{\rho}\wedge\tau) - \frac{1}{2}\sigma^2(\overline{\rho}\wedge\tau)} \\ &= \widetilde{J}_g(0,y;\overline{\rho},\tau) - (y-x). \end{aligned}$$

Furthermore, $\overline{\rho}\wedge\tau \in \mathcal{S}_T^{0,x}$ whenever $\tau \in \mathcal{S}_T^{0,y}$, and for $z = x$ and $z = y$, we have $\widetilde{J}_g(0,z;\overline{\rho},\tau) = \widetilde{J}_g(0,z;\overline{\rho},\overline{\rho}\wedge\tau)$. Therefore,

$$\begin{aligned} v_g(0,x) + y - x &= \widetilde{J}_g(0,x;\overline{\rho},\overline{\tau}) + y - x \\ &= \sup_{\tau\in\mathcal{S}_T^{0,x}} \widetilde{J}_g(0,x;\overline{\rho},\overline{\rho}\wedge\tau) + y - x \\ &\ge \sup_{\tau\in\mathcal{S}_T^{0,x}} \widetilde{J}_g(0,y;\overline{\rho},\overline{\rho}\wedge\tau) \\ &= \sup_{\tau\in\mathcal{S}_T^{0,y}} \widetilde{J}_g(0,y;\overline{\rho},\overline{\rho}\wedge\tau) \\ &= \sup_{\tau\in\mathcal{S}_T^{0,y}} \widetilde{J}_g(0,y;\overline{\rho},\tau) \\ &\ge \inf_{\rho\in\mathcal{S}_T^{0,y}} \sup_{\tau\in\mathcal{S}_T^{0,y}} \widetilde{J}_g(0,y;\rho,\tau) \\ &= v_g(0,y). \end{aligned}$$

This establishes the second inequality in (4.43).

The set of stopping times $\mathcal{S}_T^{0,x}$ is the set of all stopping times of the form $\tau \wedge \theta_0^{0,x}$, where τ is any stopping time in the set $\mathcal{S}_T$ of all stopping times satisfying $\tau \leq T$ almost surely. Therefore,

$$v_g(0,x) = \sup_{\tau \in \mathcal{S}_T} \inf_{\rho \in \mathcal{S}_T} \widetilde{J}_g(0,x;\rho \wedge \theta_0^{0,x}, \tau \wedge \theta_0^{0,x}),$$
$$v_g(0,y) = \sup_{\tau \in \mathcal{S}_T} \inf_{\rho \in \mathcal{S}_T} \widetilde{J}_g(0,y;\rho \wedge \theta_0^{0,y}, \tau \wedge \theta_0^{0,y}).$$

Thus, to prove the first inequality in (4.43), it suffices to show that

$$\widetilde{J}_g(0,x;\rho \wedge \theta_0^{0,x}, \tau \wedge \theta_0^{0,x}) \leq \widetilde{J}_g(0,y;\rho \wedge \theta_0^{0,y}, \tau \wedge \theta_0^{0,y})$$

for all $\rho, \tau \in \mathcal{S}_T$. This follows from the definition of $\widetilde{J}_g$ and $\theta_0^{0,x} \leq \theta_0^{0,y}$. □

The value functions of optimal stopping problems with continuous payoff functions are continuous (see [5]), and thus the representations (4.17) and (4.33) of v_g imply continuity of v_g. In this model, however, continuity can be proved without invoking the general theory. We have already shown in Lemma 4.8 that $v_g(s,x)$ is Lipschitz in $x \in [0,\infty)$, uniformly in $s \in [0,T]$. Given this, it is not difficult to show that $v_g(s,x)$ is jointly continuous in (s,x), and we do that here.

LEMMA 4.9. *The function v_g is continuous on $[0,T] \times [0,\infty)$.*

Proof. Because of Lemmas 4.1 and 4.8, we need only show for each fixed $x \in (0, \frac{K}{\gamma})$ that the function $s \mapsto v_g(s,x)$ is continuous. With x fixed, $s \in [0,T]$, $\epsilon > 0$, and $\delta > 0$, we define

$$A_{\epsilon,\delta}^{s,x} \triangleq \left\{ \max_{u \in [s, \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x} \wedge (s+\delta) \wedge T]} \left| X_u^{s,x} - x \right| \leq \epsilon \right\}.$$

Because g is bounded on $[0,T] \times [0, \frac{K}{\gamma}]$, (3.4) and (2.5) imply

$$\lim_{\delta \downarrow 0} \min_{s \in [0,T]} \mathbb{P}(A_{\epsilon,\delta}^{s,x}) = 1 \text{ for every } \epsilon > 0. \tag{4.45}$$

We proceed under the Case II assumption $\delta K \leq c$; the argument in Case I is similar. In Case II, the submartingale $Z_{t \wedge \theta_0^{s,x} \wedge \theta_{\frac{K}{\gamma}}^{s,x}}^{s,x}$ of Step 2 of the proof of Proposition 4.6 is a martingale when stopped at $\overline{\rho}$ given by (4.34). Let s and t satisfy $0 \leq s < t \leq (s+\delta) \wedge T$. Then

$$\begin{aligned} & v_g(s,x) \\ & = e^{rs}\mathbb{E}\left[\int_s^{\overline{\rho} \wedge t} e^{-rs} c \, du + e^{-r(\overline{\rho} \wedge t)} v_g(\overline{\rho} \wedge t, X_{\overline{\rho} \wedge t}^{s,x}) \right] \\ & \leq e^{rs}\mathbb{E}\left[\int_s^{\rho \wedge t} e^{-ru} c \, du + e^{-r\rho} v_g(\rho \wedge t, X_{\rho \wedge t}^{s,x}) \right] \text{ for } \rho \in \mathcal{S}_T^{s,x}, \rho \leq \theta_{\frac{K}{\gamma}}^{s,x}. \end{aligned} \tag{4.46}$$

If $\overline{\rho} < t$, then $v_g(\overline{\rho} \wedge t, X_{\overline{\rho} \wedge t}^{s,x}) = \widetilde{\varphi}(\overline{\rho}, X_{\overline{\rho}}^{s,x})$. But $\widetilde{\varphi} = \varphi \wedge k$, and (4.31) shows that for $(u,y) \in [0,T) \times \left(0, \frac{K}{\gamma}\right)$, we have $v_g(u,y) = \widetilde{\varphi}(u,y)$ if and only if $v_g(u,y) = \varphi(u,y) = y \wedge K$. This observation combined with (4.30) yields

$$v_g(u,y) = \widetilde{\varphi}(u,y) \;\Leftrightarrow\; v_g(u,y) = K \text{ for } u \in [0,T),\ y \in \left(0, \frac{K}{\gamma}\right). \tag{4.47}$$

We now choose $\epsilon > 0$ so that $0 < x - \epsilon < x + \epsilon < \frac{K}{\gamma}$. On the set $\{\overline{\rho} < t\} \cap A^{s,x}_{\epsilon,\delta}$ we have $0 < X^{s,x}_{\overline{\rho}\wedge t} = X^{s,x}_{\overline{\rho}} < \frac{K}{\gamma}$, and thus

$$v_g(\overline{\rho} \wedge t, X^{s,x}_{\overline{\rho}\wedge t}) = v_g(\overline{\rho}, X^{s,x}_{\overline{\rho}}) = K \geq v_g(t,x) - \epsilon.$$

On the set $\{\overline{\rho} \geq t\} \cap A^{s,x}_{\epsilon,\delta}$, we also have $v_g(\overline{\rho}\wedge t, X^{s,x}_{\overline{\rho}\wedge t}) \geq v_g(t,x) - \epsilon$, this time because of the Lipschitz continuity (4.43). The equality in (4.46) implies

$$v_g(s,x) \geq e^{rs}\mathbb{E}\big[e^{-r(\overline{\rho}\wedge t)} v_g(\overline{\rho}\wedge t, X^{s,x}_{\overline{\rho}\wedge t})\big] \geq e^{-r\delta}\mathbb{P}(A^{s,x}_{\epsilon,\delta})\big[v_g(t,x) - \epsilon\big]. \tag{4.48}$$

On the other hand, on the set $A^{s,x}_{\epsilon,\delta}$, we have $\theta^{s,x}_0 \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge t = t$, and the inequality in (4.46) implies

$$\begin{aligned}
&v_g(s,x)\\
&\leq e^{rs}\mathbb{E}\left[\int_s^{\theta^{s,x}_0 \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge t} e^{-ru}c\,du + e^{-r(\theta^{s,x}_0 \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge t)} v_g\left(\theta^{s,x}_0 \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge t, X^{s,x}_{\theta^{s,x}_0 \wedge \theta^{s,x}_{\frac{K}{\gamma}} \wedge t}\right)\right]\\
&\leq e^{rs}\int_s^{s+\delta} e^{-ru}c\,du + \big[1 - \mathbb{P}(A^{s,x}_{\epsilon,\delta})\big]K + \mathbb{E}\left[\mathbb{I}_{A^{s,x}_{\epsilon,\delta}} v_g(t, X^{s,x}_t)\right]\\
&\leq \frac{c}{r}(1 - e^{-r\delta}) + [1 - \mathbb{P}(A^{s,x}_{\epsilon,\delta})]K + \mathbb{P}(A^{s,x}_{\epsilon,\delta})\big(v_g(t,x) + \epsilon\big).
\end{aligned} \tag{4.49}$$

From (4.48) and (4.49), using the fact that $0 \leq v_g(t,x) \leq K$, we obtain

$$-[1 - e^{-r\delta}\mathbb{P}(A^{s,x}_{\epsilon,\delta})]K - \epsilon \leq v_g(s,x) - v_g(t,x) \leq \frac{c}{r}(1 - e^{-r\delta}) + [1 - \mathbb{P}(A^{s,x}_{\epsilon,\delta})]K + \epsilon.$$

Continuity of $s \mapsto v_g(s,x)$ follows from this and (4.45). □

5. Viscosity solution characterization of v_g. Propositions 4.5 and 4.6 establish (3.5). Except for the fact that we have fixed a function $g \in \mathcal{G}$ that may not satisfy the fixed point condition $v_g = g$, Proposition 4.5 says further that when $c \leq rK$, the convertible bond pricing problem reduces to the problem of optimal conversion in the region $[0,T] \times \big[0, \frac{K}{\gamma}\big]$. In particular, (4.20) and (4.21) show that the firm should use the call strategy $\rho^* = \theta^{s,x}_0 \wedge \theta^{s,x}_{\frac{K}{\gamma}}$. Proposition 4.6 shows that when $\delta K \leq c$, the convertible bond pricing problem reduces to the problem of optimal call. In particular, (4.36) and (4.39) show that the bondholder should use the conversion strategy τ^* of (4.38). Note that at maturity, τ^* mandates conversion if and only if the conversion value $\gamma X^{s,x}_T$ exceeds the par value L. These are the main assertions of Theorem 3.2.

In this section, we examine the versions of (3.8) and (3.9) appropriate for the situation with $g \in \mathcal{G}$ chosen a priori. These equations are

$$\min\{\mathcal{L}_g v - c, v - \gamma x\} = 0, \tag{5.1}$$

$$\max\{\mathcal{L}_g v - c, v - K\} = 0, \tag{5.2}$$

where $\mathcal{L}_g$ is given by (4.6). The proofs that the value function of the optimal stopping problem (4.9) satisfies (5.1) and that the value function of problem (4.24) satisfies (5.2), both in the viscosity sense on $(0,T) \times \big(0, \frac{K}{\gamma}\big)$ (see Definition 5.1 below), are standard and are omitted. Uniqueness of the continuous viscosity solutions of (5.1)

and (5.2) subject to the boundary conditions (3.6) and (3.7) follows from Lemma 6.1 below; see Remark 6.2.

We refer the reader to [13] and [17] for a detailed development of the theory of second-order viscosity solutions for Hamilton–Jacobi–Bellman equations and to [34] for an application of this theory to optimal stopping.

DEFINITION 5.1. *Let v be a continuous function defined on $(0,T)\times\left(0,\frac{K}{\gamma}\right)$.*

(a) *The function v is a* viscosity subsolution *of* (5.1) (*respectively,* (5.2)) *if, for every point $(t_0,x_0)\in(0,T)\times\left(0,\frac{K}{\gamma}\right)$ and for every "test function" $h\in C^{1,2}\left((0,T)\times\left(0,\frac{K}{\gamma}\right)\right)$ satisfying $v\le h$ on $(0,T)\times\left(0,\frac{K}{\gamma}\right)$ and $v(t_0,x_0)=h(t_0,x_0)$, we have $\min\{\mathcal{L}_g h(t_0,x_0)-c, h(t_0,x_0)-\gamma x_0\}\le 0$ (respectively, $\max\{\mathcal{L}_g h(t_0,x_0)-c, h(t_0,x_0)-K\}\le 0$).*

(b) *The function v is a* viscosity supersolution *of* (5.1) (*respectively,* (5.2)) *if, for every point $(t_0,x_0)\in(0,T)\times\left(0,\frac{K}{\gamma}\right)$ and for every "test function" $h\in C^{1,2}\left((0,T)\times\left(0,\frac{K}{\gamma}\right)\right)$ satisfying $v\ge h$ on $(0,T)\times\left(0,\frac{K}{\gamma}\right)$ and $v(t_0,x_0)=h(t_0,x_0)$, we have $\min\{\mathcal{L}_g h(t_0,x_0)-c, h(t_0,x_0)-\gamma x_0\}\ge 0$ (respectively, $\max\{\mathcal{L}_g h(t_0,x_0)-c, h(t_0,x_0)-K\}\ge 0$).*

A function v is a viscosity solution *of one of these equations if it is both a viscosity subsolution and a viscosity supersolution.*

The pricing function for the convertible bond satisfies a variational inequality, and the solution of a variational inequality is often the solution to a *free boundary problem*, where the "free boundary" divides the region in which an action (conversion or call) should take place from a region in which no action should occur. We show in section 8 that the bond pricing function does indeed satisfy a free boundary problem, and we derive properties of the free boundary. To prepare for that analysis, we introduce the following sets.

In Case I ($c\le rK$) of Proposition 4.5, we define the *continuation set*

$$\begin{aligned}\mathcal{C}_T^I &\triangleq \left\{(t,x)\in(0,T)\times\left(0,\frac{K}{\gamma}\right) : v_g(t,x)>\gamma x\right\}\\ &= \left\{(t,x)\in(0,T)\times\left(0,\frac{K}{\gamma}\right) : v_g(t,x)>\widetilde{\psi}(t,x)\right\},\end{aligned}\tag{5.3}$$

where $\widetilde{\psi}(t,x)=\max\{\gamma x, k(t,x)\}$ is defined in Step 2 of the proof of Proposition 4.5. Because k satisfies (4.15), $v_g(t,x)>\gamma x$ if and only if $v_g(t,x)>\widetilde{\psi}(t,x)$. Because v_g and $\widetilde{\psi}$ are continuous, $\mathcal{C}_T^I$ is open. Define the *stopping set*

$$\begin{aligned}\mathcal{S}_T^I &\triangleq \left\{(t,x)\in[0,T]\times\left[0,\frac{K}{\gamma}\right] : v_g(t,x)=\psi(t,x)\right\}\\ &= \left\{(t,x)\in[0,T]\times\left[0,\frac{K}{\gamma}\right] : v_g(t,x)=\widetilde{\psi}(t,x)\right\}.\end{aligned}$$

The equality is justified by the same argument that justified the equality in (5.3) and the additional observation that $\psi(T,\cdot)=\widetilde{\psi}(T,\cdot)$. The set $\mathcal{S}_T^I$ is closed. Under the Case I assumption, v_g is a viscosity solution of (5.1), which is equivalent to the following three conditions:

(i) $v_g\ge\gamma x$ on $[0,T]\times\left[0,\frac{K}{\gamma}\right]$,

(ii) v_g is a viscosity supersolution of $\mathcal{L}_g v-c=0$ on $(0,T)\times\left(0,\frac{K}{\gamma}\right)$, and

(iii) v_g is a viscosity solution of $\mathcal{L}_g v - c = 0$ on $\mathcal{C}_T^I$.

In Case II ($\delta K \le c$) of Proposition 4.6, we define the *continuation set*

$$\begin{aligned}\mathcal{C}_T^{II} &\triangleq \left\{(t,x) \in (0,T) \times \left(0, \frac{K}{\gamma}\right) : v_g(t,x) < K\right\} \\ &= \left\{(t,x) \in (0,T) \times \left(0, \frac{K}{\gamma}\right) : v_g(t,x) < \widetilde{\varphi}(t,x)\right\}, \end{aligned} \tag{5.4}$$

where $\widetilde{\varphi} = \varphi \wedge k$ is defined in Step 2 of the proof of Proposition 4.6, and the equality in (5.4) is justified by (4.47). The set $\mathcal{C}_T^{II}$ is open. Define the *stopping set*

$$\begin{aligned}\mathcal{S}_T^{II} &\triangleq \left\{(t,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right] : v_g(t,x) = \varphi(t,x)\right\} \\ &= \left\{(t,x) \in [0,T] \times \left[0, \frac{K}{\gamma}\right] : v_g(t,x) = \widetilde{\varphi}(t,x)\right\}. \end{aligned}$$

The equality is justified by the argument that justified (5.4) and the additional observations that $\varphi(T,\cdot) = \widetilde{\varphi}(T,\cdot)$. The set $\mathcal{S}_T^{II}$ is closed. Under the Case II assumption, v_g is a viscosity solution of (5.2), which is equivalent to the following conditions:

(iv) $v_g \le K$ on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$,

(v) v_g is a viscosity subsolution of $\mathcal{L}_g v - c = 0$ on $(0,T) \times \left(0, \frac{K}{\gamma}\right)$, and

(vi) v_g is a viscosity solution of $\mathcal{L}_g v - c = 0$ on $\mathcal{C}_T^{II}$.

Remark 5.2. In the overlapping case, $\delta K \le c \le rK$, we have from Proposition 4.7 that $\mathcal{C}_T^I = \mathcal{C}_T^{II} = (0,T) \times (0, \frac{K}{\gamma})$ and v_g is a viscosity solution of $\mathcal{L}_g u - c = 0$ on this set. Remark 4.4 applies in the overlapping case, which is why we require $\mathcal{L}_g v_g - c = 0$ to hold only in the viscosity sense.

6. Proof of Theorem 3.1. In this section we prove Theorem 3.1 and also prove that the continuous viscosity solutions of (5.1) and (5.2) with boundary conditions (3.6) and (3.7) are unique. In light of Propositions 4.5 and 4.6 and the discussion of section 5, this provides the final step in the proof of Theorem 3.2.

For $\epsilon \in \left[0, \frac{K}{\gamma}\right)$, we define the sets

$$D_\epsilon \triangleq [0,T] \times \left[\epsilon, \frac{K}{\gamma}\right], \quad \widetilde{D}_\epsilon \triangleq [0,T] \times \left[\log \epsilon, \log \frac{K}{\gamma}\right],$$

their parabolic boundaries

$$\begin{aligned}\partial_p D_\epsilon &\triangleq \left([0,T] \times \left\{\epsilon, \frac{K}{\gamma}\right\}\right) \cup \left(\{T\} \times \left(\epsilon, \frac{K}{\gamma}\right)\right) \\ \partial_p \widetilde{D}_\epsilon &\triangleq \left([0,T] \times \left\{\log \epsilon, \log \frac{K}{\gamma}\right\}\right) \cup \left(\{T\} \times \left(\log \epsilon, \log \frac{K}{\gamma}\right)\right), \end{aligned}$$

and their topological boundaries

$$\partial D_\epsilon \triangleq \partial_p D_\epsilon \cup \left(\{0\} \times \left(\epsilon, \frac{K}{\gamma}\right)\right), \quad \partial \widetilde{D}_\epsilon \triangleq \partial_p \widetilde{D}_\epsilon \cup \left(\{0\} \times \left(\log \epsilon, \log \frac{K}{\gamma}\right)\right).$$

In the above definitions, we use the convention $\log 0 = -\infty$, so $\widetilde{D}_0$, $\partial_p \widetilde{D}_0$, and $\partial \widetilde{D}_0$ are subsets of the extended real numbers. The following comparison lemma is a

modification of Theorem 8.2 of [13], differing by the fact that the functions u and v satisfy different equations rather than the same equation.

LEMMA 6.1 (comparison). *Let f, g in $C(D_0)$ be given. Let $u, v \in C(D_0)$ be respective viscosity sub- and supersolutions on $D_0 \setminus \partial D_0$ of the equations*

$$\min\{\mathcal{L}_f u - c, u - \gamma x\} = 0, \tag{6.1}$$

$$\min\{\mathcal{L}_g v - c, v - \gamma x\} = 0. \tag{6.2}$$

Alternatively, let u, v be respective viscosity sub- and supersolutions of the equations

$$\max\{\mathcal{L}_f u - c, u - K\} = 0, \tag{6.3}$$

$$\max\{\mathcal{L}_g v - c, v - K\} = 0. \tag{6.4}$$

Assume further that one of the functions u or v (let us say u) satisfies

$$0 \le u(t,y) - u(t,x) \le y - x \text{ for } (t,x), (t,y) \in D_0. \tag{6.5}$$

Then for every $\lambda \ge 0$, we have

$$\begin{aligned} &\max_{(t,x)\in D_0} e^{\lambda t}\big(u(t,x) - v(t,x)\big)^+ \\ &\le \max\left\{\frac{\delta}{r+\lambda} \max_{(t,x)\in D_0} e^{\lambda t}\big(f(t,x) - g(t,x)\big)^+, \max_{(t,x)\in \partial_p D_0} e^{\lambda t}\big(u(t,x) - v(t,x)\big)^+\right\}. \end{aligned} \tag{6.6}$$

Proof. We provide the proof under the assumptions that u is a subsolution of (6.1) and v is a supersolution of (6.2). Because f, g, u, and v are continuous, it suffices to prove

$$\begin{aligned} &\max_{(t,x)\in D_\epsilon} e^{\lambda t}\big(u(t,x) - v(t,x)\big)^+ \\ &\le \max\left\{\frac{\delta}{r+\lambda} \max_{(t,x)\in D_\epsilon} e^{\lambda t}\big(f(t,x) - g(t,x)\big)^+, \max_{(t,x)\in \partial_p D_\epsilon} e^{\lambda t}\big(u(t,x) - v(t,x)\big)^+\right\} \end{aligned}$$

for every $\epsilon \in \big(0, \frac{K}{\gamma}\big)$. To do this, we define $\widetilde{u}(t,\xi) \triangleq e^{\lambda t} u\big(t, e^\xi\big)$, $\widetilde{v}(t,\xi) \triangleq e^{\lambda t} v\big(t, e^\xi\big)$, $\widetilde{f}(t,\xi) \triangleq e^{\lambda t} f\big(t, e^\xi\big)$, and $\widetilde{g}(t,\xi) \triangleq e^{\lambda t} g\big(t, e^\xi\big)$. In terms of these functions, we need to prove that for every $\epsilon \in \big(0, \frac{K}{\gamma}\big)$,

$$\begin{aligned} &\max_{(t,\xi)\in \widetilde{D}_\epsilon} \big(\widetilde{u}(t,\xi) - \widetilde{v}(t,\xi)\big)^+ \\ &\le \max\left\{\frac{\delta}{r+\lambda} \max_{(t,\xi)\in \widetilde{D}_\epsilon} \big(\widetilde{f}(t,\xi) - \widetilde{g}(t,\xi)\big)^+, \max_{(t,\xi)\in \partial_p \widetilde{D}_\epsilon} \big(\widetilde{u}(t,\xi) - v(t,\xi)\big)^+\right\}. \end{aligned} \tag{6.7}$$

For $\eta > 0$, we define $\widetilde{u}^\eta(t,\xi) \triangleq \widetilde{u}(t,\xi) - \frac{\eta}{t}$, so that $\lim_{t\downarrow 0} \widetilde{u}^\eta(t,\xi) = -\infty$ uniformly in ξ. We will show for all $\epsilon \in \big(0, \frac{K}{\gamma}\big)$ that

$$\begin{aligned} &\max_{(t,\xi)\in \widetilde{D}_\epsilon} \big(\widetilde{u}^\eta(t,\xi) - \widetilde{v}(t,\xi)\big)^+ \\ &\le \max\left\{\frac{\delta}{r+\lambda} \max_{(t,\xi)\in \widetilde{D}_\epsilon} \big(\widetilde{f}(t,\xi) - \widetilde{g}(t,\xi)\big)^+, \max_{(t,\xi)\in \partial_p \widetilde{D}_\epsilon} \big(\widetilde{u}^\eta(t,\xi) - v(t,\xi)\big)^+\right\}. \end{aligned} \tag{6.8}$$

We can then let $\eta \downarrow 0$ in (6.8) to obtain (6.7) and conclude the proof.

The change of variable transforms (6.1) and (6.2) into

$$\min\Big\{ -\widetilde{u}_t + (r+\lambda)\widetilde{u} - \Big(r-\delta-\frac{1}{2}\sigma^2\Big)\widetilde{u}_\xi - \delta e^{-\lambda t-\xi}\widetilde{f}(t,\xi)\widetilde{u}_\xi + ce^{-\xi}\widetilde{u}_\xi \\ -\frac{1}{2}\sigma^2\widetilde{u}_{\xi\xi} - e^{\lambda t}c, \widetilde{u} - \gamma e^{\lambda t+\xi}\Big\} = 0,$$

$$\min\Big\{ -\widetilde{v}_t + (r+\lambda)\widetilde{v} - \Big(r-\delta-\frac{1}{2}\sigma^2\Big)\widetilde{v}_\xi - \delta e^{-\lambda t-\xi}\widetilde{g}(t,\xi)\widetilde{v}_\xi + ce^{-\xi}\widetilde{v}_\xi \\ -\frac{1}{2}\sigma^2\widetilde{v}_{\xi\xi} - e^{\lambda t}c, \widetilde{v} - \gamma e^{\lambda t+\xi}\Big\} = 0.$$

On the set

$$\widetilde{\mathcal{C}}_{\widetilde{u}} \triangleq \big\{(t,\xi) \in \widetilde{D}_0 \setminus \partial\widetilde{D}_0 : \widetilde{u}(t,\xi) > \gamma e^{\lambda t+\xi}\big\},$$

the function $\widetilde{u}$ is a viscosity subsolution of

(6.9)
$$-\widetilde{u}_t + (r+\lambda)\widetilde{u} - \Big(r-\delta-\frac{1}{2}\sigma^2\Big)\widetilde{u}_\xi - \frac{1}{2}\sigma^2\widetilde{u}_{\xi\xi} - e^{\lambda t}c = \delta e^{-\lambda t-\xi}\widetilde{f}(t,\xi)\widetilde{u}_\xi - ce^{-\xi}\widetilde{u}_\xi,$$

and so for $\eta > 0$, the function $\widetilde{u}^\eta$ is also a viscosity subsolution of this equation on $\widetilde{\mathcal{C}}_{\widetilde{u}}$. On $\widetilde{D}_0 \setminus \partial\widetilde{D}_0$, $\widetilde{v}(t,\xi) \geq \gamma e^{\lambda t+\xi}$, and $\widetilde{v}$ is a viscosity supersolution of

(6.10)
$$-\widetilde{v}_t + (r+\lambda)\widetilde{v} - \Big(r-\delta-\frac{1}{2}\sigma^2\Big)\widetilde{v}_\xi - \frac{1}{2}\sigma^2\widetilde{v}_{\xi\xi} - e^{\lambda t}c = \delta e^{-\lambda t-\xi}\widetilde{g}(t,\xi)\widetilde{v}_\xi - ce^{-\xi}\widetilde{v}_\xi.$$

Let us assume that (6.8) is violated for some $\eta > 0$ and $\epsilon \in \big(0, \frac{K}{\gamma}\big)$. This means that

$$\max_{(t,x)\in\widetilde{D}_\epsilon} \big(\widetilde{u}^\eta(t,\xi) - \widetilde{v}(t,\xi)\big)^+ > \frac{\delta}{r+\lambda} \max_{(t,\xi)\in\widetilde{D}_\epsilon} \big(\widetilde{f}(t,\xi) - \widetilde{g}(t,\xi)\big)^+. \tag{6.11}$$

Let $\alpha > 0$ be given, and set

$$M_\alpha \triangleq \max_{(t,\xi),(t,\zeta)\in\widetilde{D}_\epsilon} \Big(\widetilde{u}^\eta(t,\xi) - \widetilde{v}(t,\zeta) - \frac{\alpha}{2}|\xi-\zeta|^2\Big).$$

The maximum is attained at some point $(t_\alpha, \xi_\alpha, \zeta_\alpha)$. According to a slight variant of Lemma 3.1 of [13],

$$\lim_{\alpha\to\infty} \alpha|\xi_\alpha - \zeta_\alpha|^2 = 0 \text{ and } \lim_{\alpha\to\infty} M_\alpha = \max_{(t,\xi)\in\widetilde{D}_\epsilon} \big(\widetilde{u}^\eta(t,\xi) - \widetilde{v}(t,\xi)\big). \tag{6.12}$$

Violation of (6.8) implies that for large α, the points (t_α, ξ_α) and (t_α, ζ_α) are bounded away from the parabolic boundary $\partial_p\widetilde{D}_\epsilon$. Furthermore, because $\lim_{t\downarrow 0} \widetilde{u}^\eta(t,\xi) = -\infty$, these points are bounded away from the topological boundary $\partial\widetilde{D}_\epsilon$ as well.

There are two cases to consider. In the first case, $(t_\alpha, \xi_\alpha) \notin \widetilde{\mathcal{C}}_{\widetilde{u}}$, and so $\widetilde{u}^\eta(t_\alpha, \xi_\alpha) = \gamma e^{\lambda t_\alpha+\xi_\alpha} - \frac{\eta}{t_\alpha} \leq \gamma e^{\lambda t_\alpha+\xi_\alpha}$. We have

$$M_\alpha \leq \widetilde{u}^\eta(t_\alpha,\xi_\alpha) - \widetilde{v}(t_\alpha,\zeta_\alpha) \leq \gamma e^{\lambda t_\alpha}\big(e^{\xi_\alpha} - e^{\zeta_\alpha}\big). \tag{6.13}$$

In the other case, (t_α, ξ_α) is in $\widetilde{\mathcal{C}}_{\widetilde{u}}$. Because $\widetilde{u}^\eta$ is a subsolution of (6.9) in a neighborhood of (t_α, ξ_α), $\widetilde{v}$ is a supersolution of (6.10) in a neighborhood of (t_α, ζ_α), and these points are bounded away from $\partial\widetilde{D}_\epsilon$, condition (8.5) of Theorem 8.3 of [13] is satisfied (our time variable is reversed from that of [13]). That theorem with $\epsilon = \frac{1}{\alpha}$ implies the existence of numbers b, X, and Y such that $X \le Y$ and[1]

$$(b, \alpha(\xi_\alpha - \zeta_\alpha), X) \in \overline{\mathcal{P}}^{2,+}\widetilde{u}^\eta(t_\alpha, \xi_\alpha) \text{ and } (b, \alpha(\xi_\alpha - \zeta_\alpha), Y) \in \overline{\mathcal{P}}^{2,-}\widetilde{v}(t_\alpha, \zeta_\alpha)$$

(see the use of Theorem 8.3 on page 50 in [13]; see also page 17). Note that (t_α, ξ_α) and (t_α, ζ_α) are in the open set $\widetilde{D}_\epsilon \setminus \partial\widetilde{D}_\epsilon$. Moreover, they provide terms that can replace the time derivative, the spatial derivative, and the second spatial derivative in the subsolution and supersolution inequalities for (6.9) and (6.10):

$$\begin{aligned} -b + (r+\lambda)\widetilde{u}^\eta(t_\alpha, \xi_\alpha) - \left(r - \delta - \frac{1}{2}\sigma^2\right)\alpha(\xi_\alpha - \zeta_\alpha) - \frac{1}{2}\sigma^2 X - e^{\lambda t_\alpha}c \\ \le \delta e^{-\lambda t_\alpha - \xi_\alpha}\widetilde{f}(t_\alpha, \xi_\alpha)\alpha(\xi_\alpha - \zeta_\alpha) - ce^{-\xi_\alpha}\alpha(\xi_\alpha - \zeta_\alpha), \end{aligned} \tag{6.14}$$

$$\begin{aligned} -b + (r+\lambda)\widetilde{v}(t_\alpha, \zeta_\alpha) - \left(r - \delta - \frac{1}{2}\sigma^2\right)\alpha(\xi_\alpha - \zeta_\alpha) - \frac{1}{2}\sigma^2 Y - e^{\lambda t_\alpha}c \\ \ge \delta e^{-\lambda t_\alpha - \zeta_\alpha}\widetilde{g}(t_\alpha, \zeta_\alpha)\alpha(\xi_\alpha - \zeta_\alpha) - ce^{-\zeta_\alpha}\alpha(\xi_\alpha - \zeta_\alpha). \end{aligned} \tag{6.15}$$

Subtracting (6.15) from (6.14) and using $\sup_{\xi,\zeta \ge \log\epsilon, \xi\ne\zeta}\left|\frac{e^{-\xi} - e^{-\zeta}}{\xi - \zeta}\right| = \frac{1}{\epsilon}$, we obtain

$$\begin{aligned} M_\alpha &\le \widetilde{u}^\eta(t_\alpha, \xi_\alpha) - \widetilde{v}(t_\alpha, \zeta_\alpha) \\ &\le \frac{\delta}{r+\lambda}e^{-\lambda t_\alpha - \xi_\alpha}\big(\widetilde{f}(t_\alpha, \xi_\alpha) - \widetilde{g}(t_\alpha, \zeta_\alpha)\big)\alpha(\xi_\alpha - \zeta_\alpha) \\ &\quad + \frac{\delta}{\epsilon(r+\lambda)}e^{-\lambda t_\alpha}|\widetilde{g}(t_\alpha, \zeta_\alpha)|\alpha(\xi_\alpha - \zeta_\alpha)^2 + \frac{c}{\epsilon(r+\lambda)}\alpha(\xi_\alpha - \zeta_\alpha)^2. \end{aligned} \tag{6.16}$$

But also, (6.5) implies, at least formally, that $0 \le u_x(t,x) \le 1$, or equivalently, $0 \le \widetilde{u}_\xi(t,\xi) \le e^{\lambda t + \xi}$. Of course u_x and $\widetilde{u}_\xi$ may not exist, but (6.5) implies that $\alpha(\xi_\alpha - \zeta_\alpha)$, the surrogate for $\widetilde{u}_\xi(t_\alpha, \xi_\alpha)$, must satisfy $0 \le \alpha(\xi_\alpha - \zeta_\alpha) \le e^{\lambda t_\alpha + \xi_\alpha}$. Using this inequality in (6.16), we obtain

$$\begin{aligned} M_\alpha &\le \frac{\delta}{r+\lambda}\big(\widetilde{f}(t_\alpha, \xi_\alpha) - \widetilde{g}(t_\alpha, \zeta_\alpha)\big) + O\big(\alpha(\xi_\alpha - \zeta_\alpha)^2\big) \\ &= \frac{\delta}{r+\lambda}\big(\widetilde{f}(t_\alpha, \xi_\alpha) - \widetilde{g}(t_\alpha, \xi_\alpha)\big) + \frac{\delta}{r+\lambda}\big(\widetilde{g}(t_\alpha, \xi_\alpha) - \widetilde{g}(t_\alpha, \zeta_\alpha)\big) \\ &\quad + O\big(\alpha(\xi_\alpha - \zeta_\alpha)^2\big) \\ &\le \frac{\delta}{r+\lambda}\max_{(t,\xi)\in\widetilde{D}_\epsilon}\big(\widetilde{f}(t,\xi) - \widetilde{g}(t,\xi)\big)^+ + \frac{\delta}{r+\lambda}\big(\widetilde{g}(t_\alpha, \xi_\alpha) - \widetilde{g}(t_\alpha, \zeta_\alpha)\big) \\ &\quad + O\big((\xi_\alpha - \zeta_\alpha)^2\big). \end{aligned} \tag{6.17}$$

[1]Following [13], for a real-valued function u defined on a domain $D \subset \mathbb{R}^2$ and for (t,ξ) in the interior of D, we set $\mathcal{P}^{2,+}u(t,\xi) \triangleq \{(b,\alpha,X) \in \mathbb{R}^3 : u(s,y) \le u(t,\xi) + b(s-t) + \alpha(y-\xi) + \frac{1}{2}X(y-\xi)^2 + o(s-t) + o((y-x)^2) \text{ as } (s,y) \to (t,\xi)\}$ and define $\mathcal{P}^{2,-}u(t,\xi)$ analogously to $\mathcal{P}^{2,+}$ but with the inequality in the definition reversed. We next define $\overline{\mathcal{P}}^{2,\pm}u(t,\xi) \triangleq \{(b,\alpha,X) \in \mathbb{R}^3 : \text{there exists } (t_n, \xi_n, b_n, \alpha_n, X_n) \in D \times \mathbb{R}^3 \text{ such that } (b_n, \alpha_n, X_n) \in \mathcal{P}^{2,\pm}u(t_n, \xi_n) \text{ and } (t_n, \xi_n, u(t_n, \xi_n), b_n, \alpha_n, X_n) \to (t, \xi, u(t,\xi), b, \alpha, X)\}$.

Letting $\alpha \to \infty$ in (6.13) and (6.17), using (6.12), and using the uniform continuity of $\widetilde{g}$ on $\widetilde{D}_\epsilon$, we contradict (6.11). □

Proof of Theorem 3.1. Set $\lambda \triangleq \delta + 1$ and endow $\mathcal{G}$ with the metric

$$d(f,g) \triangleq \max_{(t,x)\in D_0} e^{\lambda t}|f(t,x) - g(t,x)| \text{ for all } f,g \in \mathcal{G}. \tag{6.18}$$

Under this metric, $\mathcal{G}$ is complete. Let $f, g \in \mathcal{G}$ be given, and define $u = \mathcal{T}f$ and $v = \mathcal{T}g$. According to subsection 4.5, u and v are in $\mathcal{G}$. In particular, (6.5) is satisfied. We apply Lemma 6.1, noting that u and v are viscosity solutions of (6.1) and (6.2), respectively, or viscosity solutions of (6.3) and (6.4), respectively, and they agree on $\partial_p D_0$, to conclude that

$$d(u,v) \le \frac{\delta}{r+\lambda} \max_{(t,x)\in D_0} e^{\lambda t}\big(f(t,x) - g(t,x)\big)^+.$$

Reversing the roles of f and g, we obtain the contraction property for $\mathcal{T}$. □

Remark 6.2. Uniqueness of the continuous viscosity solution of (6.1) or (6.3) with boundary conditions (3.6) and (3.7) follows from Lemma 6.1 with $f = g$.

7. Asymptotic behavior. We relate the problem of this paper to the perpetual convertible bond. To do this, we reverse time, denoting by $u_L(t,x)$ the price of the bond for fixed par value $L \in [0,K]$ when the time to maturity is t and the firm value is x. This section requires standing assumption (2.6). We have the following variation of Lemma 6.1.

LEMMA 7.1. *Fix $T > 0$ and let g_1 and g_2 in $C\big([0,T] \times \big[0, \frac{K}{\gamma}\big]\big)$ be a viscosity subsolution and a viscosity supersolution, respectively, of*

$$\min\{g_t + \mathcal{N}g - c, g - \gamma x\} = 0 \text{ on } (0,T) \times \left(0, \frac{K}{\gamma}\right) \tag{7.1}$$

or a viscosity subsolution and viscosity supersolution, respectively, of

$$\max\{g_t + \mathcal{N}g - c, g - K\} = 0 \text{ on } (0,T) \times \left(0, \frac{K}{\gamma}\right), \tag{7.2}$$

where $\mathcal{N}$ is the nonlinear operator

$$\begin{aligned}\mathcal{N}g(t,x) \triangleq\ & rg(t,x) - (rx - c)g_x(t,x) \\ & + \delta\big(x - g(t,x)\big)g_x(t,x) - \frac{1}{2}\sigma^2 x^2 g_{xx}(t,x).\end{aligned}$$

Assume that either g_1 or g_2 satisfies (3.2). *If $g_1(0,\cdot) \le g_2(0,\cdot)$ and*

$$g_1(t,0) \le g_2(t,0),\ g_1\Big(t, \frac{K}{\gamma}\Big) \le g_2\Big(t, \frac{K}{\gamma}\Big), \quad 0 \le t \le T, \tag{7.3}$$

then $g_1 \le g_2$. In particular, if g_1 and g_2 are viscosity solutions of (7.1) *or* (7.2), *$g_1(0,\cdot) = g_2(0,\cdot)$, and equality holds in both parts of* (7.3), *then $g_1 = g_2$.*

Proof. Apply the time-reversed version of Lemma 6.1 with $\lambda = 0$, $u = f = g_1$, and $v = g = g_2$ to conclude that

$$\max_{(t,x)\in[0,T]\times[0,\frac{K}{\gamma}]} \big(g_1(t,x) - g_2(t,x)\big)^+ \le \frac{\delta}{r} \max_{(t,x)\in[0,T]\times[0,\frac{K}{\gamma}]} \big(g_1(t,x) - g_2(t,x)\big)^+.$$

Since $\delta < r$, we have $g_1 \leq g_2$. □

Regardless of the initial time to maturity, as a function of the firm value and remaining time to maturity, the convertible bond price must satisfy one (or both) of (7.1) and (7.2), depending on whether $c \leq rK$ or $\delta K \leq c$. The uniqueness assertion in Lemma 7.1 guarantees that while the bond price does depend on the time to maturity, it does not depend on the time when the bond was issued, or equivalently, it does not depend on the *initial* time to maturity.

A perpetual convertible bond never matures, and hence the time variable and the par value are irrelevant. Its price $p(x)$ is a function of the underlying firm value alone. The following result is proved in [36].

THEOREM 7.2. *The perpetual convertible bond price function p is continuous on $[0,\infty)$, continuously differentiable on $\big(0, \frac{K}{\gamma}\big)$, and satisfies $0 \leq p'(x) \leq 1$ for $0 < x < \frac{K}{\gamma}$ and $p(x) = \gamma x$ for $x \geq \frac{K}{\gamma}$.*

If $c \leq rK$, then p, regarded as a function of (t,x) with $p_t = 0$, is a continuous viscosity solution of (7.1) *satisfying*

$$p(0) = 0, \quad p\left(\frac{K}{\gamma}\right) = K. \tag{7.4}$$

Furthermore, there exists $C_o^* \in \big(0, \frac{K}{\gamma}\big]$ such that p restricted to $(0, C_o^*)$ is strictly greater than γx and is a classical solution of $\mathcal{N}p = c$, whereas $p(x) = \gamma x$ for $x \geq C_o^*$.

If $\delta K \leq c$, then p is a continuous viscosity solution of (7.2) *satisfying* (7.4). *Furthermore, there exists $C_a^* \in \big(0, \frac{K}{\gamma}\big]$ such that p restricted to $(0, C_a^*)$ is strictly less than K and is a classical solution of $\mathcal{N}p = c$, whereas $p(x) = K$ for $C_a^* \leq x \leq \frac{K}{\gamma}$.*

Uniqueness of p in [36] is proved only in the class of functions that are smooth in the continuation region, not within the class of all continuous functions. We upgrade the uniqueness result to the larger class here.

LEMMA 7.3. *Let p be the perpetual convertible bond price function. If $c \leq rK$, then p is the unique viscosity solution of* (7.1) *on $\big(0, \frac{K}{\gamma}\big)$ that is continuous on $\big[0, \frac{K}{\gamma}\big]$ and satisfies* (7.4). *If $\delta K \leq c$, then p is the unique viscosity solution of* (7.2) *on $\big(0, \frac{K}{\gamma}\big)$ that is continuous on $\big[0, \frac{K}{\gamma}\big]$ and satisfies* (7.4).

Proof. We provide the proof for the case $c \leq rK$. In the second case, $\delta K \leq c$, a similar proof is possible.

Let $q \in C\big[0, \frac{K}{\gamma}\big]$ be a viscosity solution of (7.1) on $\big(0, \frac{K}{\gamma}\big)$ satisfying (7.4). Assume

$$\max_{x \in [0, \frac{K}{\gamma}]} \big(p(x) - q(x)\big) = p(x_0) - q(x_0) > 0. \tag{7.5}$$

Then $x_0 \in \big(0, \frac{K}{\gamma}\big)$ and $p(x_0) > q(x_0) \geq \gamma x_0$, so $x_0 \in (0, C_o^*)$. Because p is twice continuously differentiable in $(0, C_o^*)$, we can use $p + q(x_0) - p(x_0)$ as a test function for the viscosity supersolution q to obtain

$$rq(x_0) - (rx_0 - c)p'(x_0) + \delta\big(x_0 - q(x_0)\big)p'(x_0) - \frac{1}{2}\sigma^2 x_0^2 p''(x_0) \geq c.$$

But p satisfies $\mathcal{N}p(x_0) = c$, so

$$rp(x_0) - (rx_0 - c)p'(x_0) + \delta\big(x_0 - p(x_0)\big)p'(x_0) - \frac{1}{2}\sigma^2 x_0^2 p''(x_0) = c.$$

Subtracting these relations, we obtain

$$r\big(p(x_0) - q(x_0)\big) \leq \delta\big(p(x_0) - q(x_0)\big)p'(x_0).$$

But $0 \le p'(x_0) \le 1$ and $0 \le \delta < r$, so we have a contradiction to (7.5).

Assume on the other hand that

$$\max_{x \in [0, \frac{K}{\gamma}]} \big(q(x) - p(x)\big) = q(x_0) - p(x_0) > 0. \tag{7.6}$$

Then $x_0 \in \big(0, \frac{K}{\gamma}\big)$ and $q(x_0) > p(x_0) \ge \gamma x_0$. We have $q \le p + q(x_0) - p(x_0)$, and if $x_0 \ne C_o^*$, so that p is twice continuously differentiable in a neighborhood of x_0, we can use $p + q(x_0) - p(x_0)$ as a test function for the viscosity subsolution q to obtain

$$rq(x_0) - (rx_0 - c)p'(x_0) + \delta\big(x_0 - q(x_0)\big)p'(x_0) - \frac{1}{2}\sigma^2 x_0^2 p''(x_0) \le c. \tag{7.7}$$

But $\mathcal{N}p(x_0) \ge c$ means that

$$rp(x_0) - (rx_0 - c)p'(x_0) + \delta\big(x_0 - p(x_0)\big)p'(x_0) - \frac{1}{2}\sigma^2 x_0^2 p''(x_0) \ge c. \tag{7.8}$$

Subtracting these relations, we obtain

$$r\big(q(x_0) - p(x_0)\big) \le \delta\big(q(x_0) - p(x_0)\big)p'(x_0), \tag{7.9}$$

and we conclude as before.

The only other possibility is that (7.6) holds and $x_0 = C_o^* \in \big(0, \frac{K}{\gamma}\big)$. According to Theorem 7.2, p' is defined on $\big(0, \frac{K}{\gamma}\big)$, and using the equations $\mathcal{N}p = c$ to the left of C_o^* and $p(x) = \gamma x$ to the right of x_0, we see that the left- and right-hand second derivatives $p''(x_0-)$ and $p''(x_0+) = 0$ exist. Furthermore, $p(x) - \gamma x$ attains its minimum value of 0 at x_0, so $p''(x_0-) \ge 0$. We need only rule out the case $p''(x_0-) > 0$, for in the event $p''(x_0-) = 0$, the function p is twice continuously differentiable at x_0 and we can use $p + q(x_0) - p(x_0)$ as a test function as above.

Suppose $p''(x_0-) > 0 = p''(x_0+)$. Let $\underline{p}$ be the solution in $\big(0, \frac{K}{\gamma}\big]$ of the ordinary differential equation $\mathcal{N}\underline{p} = c$ satisfying $\underline{p}(x_0) = \gamma x_0$ and $\underline{p}'(x_0) = \gamma$. On $(0, x_0]$, p is a solution to this terminal value problem and hence agrees with $\underline{p}$. In particular, $\underline{p}''(x_0) = p''(x_0-) > 0$, and this implies $\underline{p}(x) > \gamma x = p(x)$ for x in some interval $(x_0, x_0+\epsilon)$, where $\epsilon > 0$. The function $q - \underline{p}$ attains a local maximum at x_0 because $q - p$ does, and we can use $\underline{p} + q(x_0) - p(x_0)$ as a test function for the viscosity subsolution q as above. This leads to (7.7), with $p''(x_0-)$ replacing $p''(x_0)$. Inequality (7.8) holds for all $x \in (0, x_0)$, and letting $x \uparrow x_0$, we obtain (7.8), with $p''(x_0-)$ replacing $p''(x_0)$ as well. This implies (7.9), and (7.6) is contradicted. □

Proof of Theorem 3.6. The terminal condition (3.6), with time reversed, states that for $0 \le L \le K$, we have

$$\gamma x = u_0(0, x) \le u_L(0, x) \le u_K(0, x) = x \wedge K, \quad 0 \le x \le \frac{K}{\gamma}.$$

The functions u_0, u_L, and u_K are continuous viscosity solutions of (7.1) or (7.2), depending on whether $c \le rK$ or $\delta K \le c$. Lemma 7.1 and the membership of u_0 and u_K in $\mathcal{G}$ (see, in particular, (3.3)) imply that for $0 \le L \le K$,

$$\gamma x \le u_0(t, x) \le u_L(t, x) \le u_K(t, x) \le x \wedge K, \quad t \ge 0,\ 0 \le x \le \frac{K}{\gamma}. \tag{7.10}$$

For $0 \le t_1 \le t_2$, we have $u_0(0, \cdot) = \gamma x \le u_0(t_2 - t_1, \cdot)$, and we can apply Lemma 7.1 with $g_1(0, \cdot) = u_0(0, \cdot)$ and $g_2(0, \cdot) = u_0(t_2 - t_1, \cdot)$ to conclude that $u_0(t_1, \cdot) \le u_0(t_2, \cdot)$.

In other words, $u_0(t,x)$ is nondecreasing in t for each fixed x. On the other hand, $u_K(0,x) = x \wedge K \geq u_K(t_2 - t_1, x)$, and this leads to the conclusion that $u_K(t,x)$ is nonincreasing in t for each fixed x. Both $u_0(t,\cdot)$ and $u_K(t,\cdot)$ are Lipschitz continuous with constant 1, and the Arzela–Ascoli theorem implies that they converge uniformly on $\left[0, \frac{K}{\gamma}\right]$ to Lipschitz continuous limits $u_-(\cdot)$ and $u_+(\cdot)$, respectively, as $t \to \infty$. Uniform convergence preserves the viscosity solution property (see [13]), and so u_- and u_+ are also continuous viscosity solutions of either (7.1) or (7.2). Lemma 7.3 implies $u_- = p = u_+$. Relation (7.10) then implies $\lim_{t\to\infty} u_L(t,x) = p(x)$ for all $x \in \left[0, \frac{K}{\gamma}\right]$, and the convergence is uniform in x. Of course, for $x \geq \frac{K}{\gamma}$, $u_L(t,x) = p(x) = \gamma x$. □

Remark 7.4. The proof of Theorem 3.6 shows that for all $t \geq 0$,

$$u_0(t,x) \leq \lim_{s\to\infty} u_0(s,x) = p(x) = \lim_{s\to\infty} u_K(s,x) \leq u_K(t,x) \tag{7.11}$$

and the convergence is uniform in $x \in \left(0, \frac{K}{\gamma}\right)$.

8. Continuation and stopping sets. We continue with the time reversal introduced in section 7, denoting by $u_L(t,x)$ the price of the convertible bond when time to maturity is t and the underlying firm value is x. Following section 5, in Case I ($c \leq rK$) of Proposition 4.5, we define

$$\mathcal{C}_L^I \triangleq \left\{(t,x) \in (0,\infty) \times \left(0, \frac{K}{\gamma}\right) : u_L(t,x) > \gamma x\right\}, \tag{8.1}$$

$$\mathcal{S}_L^I \triangleq \left\{(t,x) \in (0,\infty) \times \left(0, \frac{K}{\gamma}\right] : u_L(t,x) = \gamma x\right\}. \tag{8.2}$$

In Case II ($\delta K \leq c$) of Proposition 4.6, we define

$$\mathcal{C}_L^{II} \triangleq \left\{(t,x) \in (0,\infty) \times \left(0, \frac{K}{\gamma}\right) : u_L(t,x) < K\right\}, \tag{8.3}$$

$$\mathcal{S}_L^{II} \triangleq \left\{(t,x) \in (0,\infty) \times \left(0, \frac{K}{\gamma}\right] : u_L(t,x) = K\right\}. \tag{8.4}$$

To relate (8.4) to the definition of $\mathcal{S}_T^{II}$ in section 5, recall (4.47). The present section provides some information about the nature of the sets in (8.1)–(8.4).

LEMMA 8.1. *For all $t \geq 0$, the mapping $x \mapsto \frac{1}{x} u_L(t,x)$ is nonincreasing.*

Proof. We rescale u_L. Let $\ell > 0$ be given and define $\overline{u} : [0,\infty) \times \left[0, \frac{\ell K}{\gamma}\right] \to [0,\infty)$ by $\overline{u}(t,x) \triangleq \ell u\left(t, \frac{x}{\ell}\right)$. Because we have formally that $0 \leq u_x(t,x) \leq 1$, we also have formally that $0 \leq \overline{u}_x(t,x) \leq 1$. Furthermore,

$$\overline{u}_t(t,x) + \mathcal{N}\overline{u}(t,x) = \ell\left[u_t\left(t, \frac{x}{\ell}\right) + \mathcal{N}u\left(t, \frac{x}{\ell}\right)\right] + c(1-\ell)\overline{u}_x(t,x). \tag{8.5}$$

In Case I ($c \leq rK$) we let $0 < a < b < \frac{K}{\gamma}$ be given and set $\ell = \frac{b}{a} > 1$. Because $u_t + \mathcal{N}u \geq c$, (8.5) implies

$$\overline{u}_t(t,x) + \mathcal{N}\overline{u}(t,x) \geq \ell c + c(1-\ell)\overline{u}_x(t,x) \geq c. \tag{8.6}$$

In other words, $\overline{u}(t,x)$ is a viscosity supersolution of $\overline{u}_t + \mathcal{N}\overline{u} \geq c$ on $(0,\infty) \times \left(0, \frac{\ell K}{\gamma}\right)$. But also, $\overline{u}(t,x) \geq \gamma x$ for $0 \leq x \leq \frac{\ell K}{\gamma}$ because $u(t,x) \geq \gamma x$ for $0 \leq x \leq \frac{K}{\gamma}$. It

follows that for every $T > 0$, $\overline{u}$ is defined and continuous on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$ and is a supersolution of (7.1). Furthermore, $\overline{u}(0,\cdot) \geq u(0,\cdot)$, $\overline{u}(t,0) = u(t,0) = 0$, and $\overline{u}\left(t, \frac{K}{\gamma}\right) \geq K = u\left(t, \frac{K}{\gamma}\right)$. Lemma 7.1 implies $\overline{u} \geq u$ on $[0,T] \times \left[0, \frac{K}{\gamma}\right]$ for every $T > 0$. In particular, $\frac{b}{a} u(t,a) = \overline{u}(t,b) \geq u(t,b)$, which yields the desired result.

In Case II ($\delta K \leq c$) we again let $0 < a < b < \frac{K}{\gamma}$ be given, but now set $\ell = \frac{a}{b} < 1$. In this case, $u_t + \mathcal{N}u \leq c$ and both inequalities in (8.6) are reversed. But also, $\overline{u} \leq \ell K \leq K$. It follow that $\overline{u}$ is a subsolution of (7.2), but it is defined only on the set $[0,\infty) \times \left[0, \frac{\ell K}{\gamma}\right] \subset [0,\infty) \times \left[0, \frac{K}{\gamma}\right]$. However, on the upper boundary $[0,\infty) \times \left\{\frac{\ell K}{\gamma}\right\}$ of this set, $\overline{u} = \ell K$ and $u\left(t, \frac{\ell K}{\gamma}\right) \geq \ell K$. The function $u(0,\cdot)$ also dominates $\overline{u}(0,\cdot)$. We fix an arbitrary $T > 0$ and apply Lemma 7.1 on the smaller domain $[0,T] \times \left[0, \frac{\ell K}{\gamma}\right]$ (just take γ in Lemma 7.1 to be $\frac{\gamma}{\ell}$) to conclude that $\overline{u} \leq u$ on this domain. In particular, $u(t,a) \geq \overline{u}(t,a) = \frac{a}{b} u(t,b)$, which yields the desired result. ∎

In Case I, we define the free boundary

$$c_L(t) \triangleq \inf\left\{x \in \left(0, \frac{K}{\gamma}\right] : u_L(t,x) = \gamma x\right\}, \quad t > 0, \tag{8.7}$$

and in Case II, we define the free boundary

$$d_L(t) \triangleq \inf\left\{x \in \left(0, \frac{K}{\gamma}\right] : u_L(t,x) = K\right\}, \quad t > 0. \tag{8.8}$$

In the overlapping case $\delta K \leq c \leq rK$, Remark 5.2 says that $c_L(t) = d_L(t) = \frac{K}{\gamma}$ for all $t > 0$. We see in Remark 8.4 below that $c_L(t)$ and $d_L(t)$ are positive, so inf could be replaced by min in (8.7) and (8.8).

Remark 8.2. Because $u_L(t,x)$ is nondecreasing in L, we have $\mathcal{S}_K^I \subset \mathcal{S}_L^I \subset \mathcal{S}_0^I$ in Case I and $\mathcal{S}_0^{II} \subset \mathcal{S}_L^{II} \subset \mathcal{S}_K^{II}$ in Case II. This implies that $c_L(t)$ is nondecreasing in L in Case I and $d_L(t)$ is nonincreasing in L in Case II. In the proof of Theorem 3.6 at the end of section 7, we saw $u_0(t,x)$ is nondecreasing in t and $u_K(t,x)$ is nonincreasing in t. This implies in Case I that $c_0(t)$ is nondecreasing and $c_K(t)$ is nonincreasing, while in Case II, $d_0(t)$ is nonincreasing and $d_K(t)$ is nondecreasing.

The following theorem asserts that the continuation set and stopping set are divided by the free boundary $c_L(\cdot)$ in Case I and $d_L(\cdot)$ in Case II.

THEOREM 8.3. *In Case* I ($c \leq rK$) *we have*

$$\mathcal{S}_L^I = \left\{(t,x) \in (0,\infty) \times \left(0, \frac{K}{\gamma}\right] : c_L(t) \leq x \leq \frac{K}{\gamma}\right\}.$$

In Case II ($\delta K \leq c$) *we have*

$$\mathcal{S}_L^{II} = \left\{(t,x) \in (0,\infty) \times \left(0, \frac{K}{\gamma}\right] : d_L(t) \leq x \leq \frac{K}{\gamma}\right\}.$$

Proof. In Case I, we must show that if $u_L(t,x) = \gamma x$ for some $x \in \left(0, \frac{K}{\gamma}\right)$, then $u_L(t,y) = \gamma y$ for all $y \in \left[x, \frac{K}{\gamma}\right]$. This follows immediately from Lemma 8.1. In Case II, the result follows from the nondecrease in x of $u_L(t,x)$. ∎

Remark 8.4. Consider Case I. If $(t_0,x_0) \in \mathcal{S}_L^I$, then $u_L(t_0,x_0) = \gamma x_0$. Because u_L is a viscosity solution of $\min\{u_t + \mathcal{N}u - c, u - \gamma x\} = 0$, we may use $h(x) = \gamma x$ as a "test function" at the point (t_0,x_0) for the viscosity supersolution property to obtain

$\mathcal{N}h(x_0) \geq c$, or equivalently, $x_0 \geq \frac{c}{\delta\gamma}$. It follows that $\min\left\{\frac{K}{\gamma}, \frac{c}{\delta\gamma}\right\} \leq c_L(t) \leq \frac{K}{\gamma}$ for every $t > 0$. In Case II, we have $K = u_L(t, d_L(t)) \leq d_L(t) \leq \frac{K}{\gamma}$.

THEOREM 8.5. *Let C_o^* and C_a^* be as in Theorem* 7.2. *In Case* I, *we have* $\lim_{t\to\infty} c_L(t) = C_o^*$. *In Case* II, *we have* $\lim_{t\to\infty} d_L(t) = C_a^*$.

Proof. In light of Remark 8.2, it suffices to prove the theorem for the limiting cases $L = 0$ and $L = \frac{K}{\gamma}$. We treat Case I only.

Since c_0 is nondecreasing and c_K is nonincreasing, these functions have limits $c_0(\infty)$, and $c_K(\infty)$ in $\left(0, \frac{K}{\gamma}\right]$ as $t \to \infty$, and we must show $c_0(\infty) = C_o^* = c_K(\infty)$. Because $C_o^* = \min\left\{x \in \left(0, \frac{K}{\gamma}\right] : p(x) = \gamma x\right\}$, we have from the first inequality in (7.11) that $c_0(t) \leq C_o^*$ and hence $c_0(\infty) \leq C_o^*$. But $u_0(t, c_0(t)) = \gamma c_0(t)$ and $u_0(t, \cdot)$ converges uniformly to $p(\cdot)$, so $c_0(\infty) \geq C_o^*$.

Using the second inequality in (7.11), we obtain $c_K(t) \geq C_o^*$, and hence $c_K(\infty) \geq C_o^*$. Assume $C_o^* < c_K(\infty) \leq \frac{K}{\gamma}$. Because $c_K(\cdot)$ is nonincreasing, u_K is a viscosity solution of $u_t + \mathcal{N}u = c$ on $(0,\infty) \times (0, c_K(\infty))$. Hence the limit $p(\cdot)$ is a viscosity solution of this equation on $(0, c_K(\infty))$. But $p(x) = \gamma x$ for $x \in (C_o^*, c_K(\infty))$, and this does not satisfy $\mathcal{N}p = c$. This contradiction implies $c_K(\infty) = C_0^*$. ∎

Acknowledgments. We benefited from discussions with H. Mete Soner and Luc Tartar about the partial differential equations in this work, and we thank them for their assistance.

REFERENCES

[1] L. ANDERSEN AND D. BUFFUM, *Calibration and implementation of convertible bond models*, J. Comput. Finance, 2 (2003–2004), pp. 1–34.
[2] P. ASQUITH, *Convertible bonds are not called late*, J. Finance, 50 (1995), pp. 1275–1289.
[3] P. ASQUITH AND D. MULLINS, JR., *Convertible debt: Corporate call policy and voluntary conversion*, J. Finance, 46 (1991), pp. 1273–1289.
[4] E. AYACHE, P. A. FORSYTH, AND K. R. VETZAL, *The valuation of convertible bonds with credit risk*, J. Derivatives, 7 (2003), pp. 9–29.
[5] B. BASSAN AND C. CECI, *Optimal stopping problems with discontinuous reward: Regularity of the value function and viscosity solutions*, Stoch. Stoch. Rep., 72 (2002), pp. 55–77.
[6] P. BILLINGSLEY, *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.
[7] F. BLACK AND M. SCHOLES, *The pricing of options and corporate liabilities*, J. Political Economy, 81 (1973), pp. 637–659.
[8] M. BRENNAN AND E. SCHWARTZ, *Convertible bonds: Valuation and optimal strategies for call and conversion*, J. Finance, 32 (1977), pp. 1699–1715.
[9] M. BRENNAN AND E. SCHWARTZ, *Analyzing convertible bonds*, J. Financial Quant. Anal., 15 (1980), pp. 907–929.
[10] W. CHEUNG AND I. NELKEN, *Costing converts*, Risk, 7 (1994), pp. 47–49.
[11] G. CONSTANTINIDES, *Warrant exercise and bond conversion in competitive markets*, J. Financial Econom., 13 (1984), pp. 371–397.
[12] G. CONSTANTINIDES AND R. ROSENTHAL, *Strategic analysis of the competitive exercise of certain financial options*, J. Econom. Theory, 32 (1984), pp. 128–138.
[13] M. CRANDALL, H. ISHII, AND P.-L. LIONS, *User's guide to viscosity solutions of second-order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.
[14] M. H. A. DAVIS AND F. LISCHKA, *Convertible bonds with market risk and credit risk*, in Applied Probability, R. Chan, Y.-K. Kwok, D. Yao, and Q. Zhang, eds., AMS/IP Stud. Adv. Math. 26, Amer. Math. Soc., Providence, RI, 2002, pp. 45–58.
[15] F. DELBAEN AND W. SCHACHERMAYER, *A general version of the fundamental theorem of asset pricing*, Math. Ann., 300 (1994), pp. 463–520.
[16] K. DUNN AND K. EADES, *Voluntary conversion of convertible securities and the optimal call strategy*, J. Financial Econom., 23 (1984), pp. 273–301.
[17] W. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer, New York, 1993.
[18] J. M. HARRISON AND S. PLISKA, *Martingales and stochastic integrals in the theory of continuous*

trading, Stoch. Process. Appl., 11 (1981), pp. 215–260.
[19] M. Harris and A. Raviv, *A sequential model of convertible debt call policy*, J. Finance, 40 (1985), pp. 1263–1282.
[20] C. Hennessy and Y. Tserlukevich, *Analyzing Callable and Convertible Bonds when the Modigliani-Miller Assumptions are Violated*, EFA Moscow Meetings Paper, Haas School, University California at Berkeley, Berkeley, CA, 2005.
[21] T. Ho and M. Pteffer, *Convertible bonds: Model, value, attribution and analytics*, Financial Analysts J., 52 (1996), pp. 35–44.
[22] J. E. Ingersoll, *A contingent-claims valuation of convertible securities*, J. Financial Econom., 4 (1977), pp. 289–322.
[23] J. E. Ingersoll, *An examination of corporate call policies on convertible securities*, J. Finance, 32 (1977), pp. 463–478.
[24] J. Kallsen and C. Kühn, *Convertible bonds: Financial derivatives of game type*, in Exotic Option Pricing and Advanced Lévy Models, A. Kyprianov, W. Schoutems, and P. Willmott, eds., Wiley, New York, 2005, pp. 277–292.
[25] I. Karatzas and S. Shreve, *Methods of Mathematical Finance*, Springer, New York, 1998.
[26] I. Kifer, *Game options*, Finance Stoch., 4 (2002), pp. 443–463.
[27] B. Loschak, *The Valuation of Defaultable Convertible Bonds under Stochastic Interest Rates*, Ph.D. dissertation, Krannert Graduate School of Management, Purdue University, West Lafayette, IN, 1996.
[28] J. J. McConnell and E. S. Schwartz, *LYON taming*, J. Finance, 41 (1986), pp. 561–577.
[29] A. McNee, *The challenge of equity derivatives*, Risk, Technology Supplement, Aug. (1999), pp. S29–S31.
[30] R. C. Merton, *Theory of rational option pricing*, Bell J. Econom. Manag. Sci., 4 (1973), pp. 141–183.
[31] R. C. Merton, *On the pricing of corporate debt: The risk structure of interest rates*, J. Finance, 29 (1974), pp. 449–470.
[32] M. Miller and F. Modigliani, *The cost of capital, corporation finance, and the theory of investment*, Amer. Econ. Rev., 48 (1958), pp. 261–297.
[33] M. Miller and F. Modigliani, *Dividend policy, growth and the valuation of shares*, J. Business, 34 (1961), pp. 411–433.
[34] B. Øksendal and K. Reikvam, *Viscosity solutions of optimal stopping problems*, Stoch. Stoch. Rep., 62 (1998), pp. 285–301.
[35] A. N. Shiryayev, *Optimal Stopping Rules*, Springer, Berlin, 1977.
[36] M. Sîrbu, I. Pikovsky, and S. Shreve, *Perpetual convertible bonds*, SIAM J. Control Optim., 43 (2004), pp. 58–85.
[37] D. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer, New York, 1979.
[38] A. Takahashi, T. Kobayashi, and N. Nakagawa, *Pricing convertible bonds with default risk*, J. Fixed Income, 11 (2001), pp. 20–29.
[39] K. Tsiveriotis and C. Fernandes, *Valuing convertible bonds with credit risk*, J. Fixed Income, 8 (1998), pp. 95–102.
[40] A. B. Yigitbasioglu, *Pricing Convertible Bonds with Interest Rate, Equity, Credit and FX Risk*, EFMA London Meeting Paper, ISMA Center, University of Reading, Reading, UK, 2002.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1541–1560

MODEL REDUCTION OF UNCERTAIN SYSTEMS WITH MULTIPLICATIVE NOISE BASED ON BALANCING*

LIQIAN ZHANG†, BIAO HUANG†, AND TONGWEN CHEN‡

Abstract. This paper investigates the problem of model reduction based on balancing for uncertain discrete-time systems with multiplicative noise. Such systems can be considered as linear systems with both deterministic and stochastic uncertainties. Two linear matrix inequalities (LMIs) are proposed to find the balancing transformation, through which the original uncertain model with multiplicative noise is balanced. The reduced order model with the same structure as that of the original one is obtained by truncating the balanced model. An upper bound of the model reduction error is guaranteed. Based on the derived model reduction error bound, an optimization problem is suggested so that the solutions of the LMIs can be uniquely found and the model reduction error is ensured to be small.

Key words. discrete-time systems, linear matrix inequalities, multiplicative noise, balanced truncation, model reduction, uncertain systems

AMS subject classifications. 37H99, 37M99, 41A29

DOI. 10.1137/S0363012904443063

1. Introduction.

Stochastic systems have received much attention since stochastic modeling has come to play an important role in many branches of science and engineering applications. One particular kind of stochastic model includes linear models whose parameter uncertainties are modeled as white noise processes in a linear setting, or simply, multiplicative noise. This kind of stochastic model uncertainty comes from dynamics driven by noise signals and uncertainties due to fast parameter variations. Linear systems with multiplicative noise have been widely studied in [2, 3, 5, 10, 18, 23], just to name a few, and have applications in many areas such as nuclear fission and heat transfer, population models, and immunology and gain scheduling (see [5] and the references therein). Other commonly studied uncertainties in the literature, which are deterministic in nature, result from unmodelled dynamics, modelling and linearization errors, and slow parameter deviations. Different kinds of descriptions have been introduced to account for deterministic uncertainties, such as norm-bounded uncertainties, polytopic uncertainties [2], and linear fractional transformation uncertainties [1]. As a result of the existence of the aforementioned stochastic uncertainties and deterministic uncertainties, in recent years, systems with both deterministic and stochastic uncertainties have drawn much attention. They were first studied in [9], and later they have been widely investigated, e.g., in [4, 17, 23] for the discrete-time case and [10, 12, 14, 18, 19, 20, 22] for the continuous-time case, with a variety of deterministic uncertainties.

Model reduction has also been an active topic for several decades. It is well known that a fundamental problem in system theory concerns the construction of mathematical models for physical processes. To describe such processes accurately, high order

*Received by the editors April 16, 2004; accepted for publication (in revised form) April 10, 2006; published electronically November 14, 2006. The research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).
http://www.siam.org/journals/sicon/45-5/44306.html
†Department of Chemical and Materials Engineering, University of Alberta, Edmonton T6G 2G6, AB, Canada (liqian@ualberta.ca, biao.huang@ualberta.ca).
‡Corresponding author. Department of Electrical and Computer Engineering, University of Alberta, Edmonton T6G 2V4, AB, Canada (tchen@ece.ualberta.ca).

finite dimensional models may be obtained through standard modeling procedures. These models may be unnecessarily complicated as far as representing the physical systems is concerned. Given a high order model, a logical way to take is therefore to look for possible simplified model descriptions. As far as deterministic linear models are concerned, model reduction has been well studied, and various efficient model reduction methods have been proposed based on different model reduction criteria. One of the most powerful methods for deterministic linear systems is model reduction based on balancing. Such work can be seen in [6, 11, 15, 16] for nominal linear systems and [1, 7, 8] for linear systems with deterministic uncertainties.

In spite of the wide study of control of stochastic systems and model reduction of deterministic systems, there are only a few pieces of work in the literature contributing to the model reduction of stochastic systems. The H_∞ model reduction for linear systems with only stochastic uncertainties (multiplicative noise) was investigated in [21], and the model reduction based on balancing still remains open for linear systems with multiplicative noise. Furthermore, the model reduction for linear systems with both deterministic and stochastic uncertainties has never been studied. In this paper, we consider the model reduction problem based on balancing for uncertain linear systems with multiplicative noise. It is assumed that the systems to be considered consist of norm-bounded time-varying deterministic uncertainties and stochastic uncertainties, namely, multiplicative noise. By solving two linear matrix inequalities (LMIs), two positive matrices are obtained. By balancing these two positive matrices, the uncertain model with multiplicative noise is balanced, and the reduced order model can be derived by truncating the balanced model. It is shown that the model reduction error between the original and truncated models is bounded. Numerical examples show the efficiency of the proposed model reduction method.

2. Preliminaries. Consider the following discrete-time system with norm-bounded time-varying uncertainties and multiplicative noise:

$$S:\begin{cases} x(k+1) = A(k)x(k) + B(k)u(k) + \displaystyle\sum_{i=1}^{L}(A_i x(k) + B_i u(k))p_i(k), \\ y(k) = C(k)x(k) + D(k)u(k) + \displaystyle\sum_{i=1}^{L}(C_i x(k) + D_i u(k))p_i(k), \end{cases}$$

where $x(k) \in R^n$ is the state, $u(k) \in R^{n_u}$ is the control input, $y(k) \in R^{n_y}$ is the output, the initial state is x_0, and $p(0), p(1), \ldots, p(L)$ are independent, identically distributed random variable sequences on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ satisfying

$$\mathcal{E}\{p(k)\} = 0, \quad \mathcal{E}\left\{p(k)p(k)^T\right\} = \Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_L^2) \tag{1}$$

with $\mathcal{E}\{\cdot\}$ denoting the expectation operator and

$$p(k) = \begin{bmatrix} p_1(k) & p_2(k) & \cdots & p_L(k) \end{bmatrix}^T.$$

Denote $\mathbb{N} = \{0, 1, 2, \ldots\}$. Let $\mathcal{F}_k$ be the σ-algebras generated by $(p(k))_{k\in\mathbb{N}}$. We denote by $l_2(\Omega, R^k)$ the space of square-summable R^k-valued vector functions on $(\Omega, \mathcal{F}, \mathcal{P})$ and by $l_{e_2}([0,\infty); R^k)$ the space of k-dimensional nonanticipatory square-summable stochastic processes $f(\cdot) = (f(k))_{k\in\mathbb{N}}$ on $\mathbb{N}$ with respect to $\mathcal{F}_k$ satisfying

$$\|f\|_{e_2}^2 = \mathcal{E}\left\{\sum_{k\in\mathbb{N}} |f(k)|^2\right\} = \sum_{k\in\mathbb{N}} \mathcal{E}\left\{|f(k)|^2\right\} < \infty.$$

We assume that $u(k)$ belongs to $l_{e_2}([0,\infty);R^{n_u})$ and is $\mathcal{F}_{k-1}$ measurable for all $k\in\mathbb{N}$, and x_0 is independent of the random process p. $A_i,\ B_i,\ C_i,\ D_i\ (i=1,\dots,L)$ are constant matrices with appropriate dimensions, and $A(k)$, $B(k)$, $C(k)$, $D(k)$ are time-varying matrices with norm-bounded uncertainties, which are given by

$$\begin{bmatrix} A(k) & B(k) \\ C(k) & D(k) \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} + \begin{bmatrix} U \\ V \end{bmatrix} E(k) \begin{bmatrix} M & N \end{bmatrix}, \tag{2}$$

where $E(k)\in R^{s_1\times s_2}$ is time-varying, and A,B,C,D,U,V,M,N are constant matrices with appropriate dimensions. The uncertainty matrix $E(k)$ is feasible if it satisfies

$$\|E(k)\|_2 \le 1,$$

where the norm used is the Euclidean one.

By rewriting stochastic system S as

$$\begin{cases} x(k+1) = \left(A + UE(k)M + \sum_{i=1}^{L} A_i p_i(k)\right) x(k) + (B + UE(k)N + \sum_{i=1}^{L} B_i p_i(k))u(k), \\ y(k) = \left(C + VE(k)M + \sum_{i=1}^{L} C_i p_i(k)\right) x(k) + \left(D + VE(k)N + \sum_{i=1}^{L} D_i p_i(k)\right) u(k), \end{cases}$$

it can be seen that system S can be considered as a linear system

$$\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right)$$

with deterministic time-varying uncertainty

$$\left(\begin{bmatrix} U \\ V \end{bmatrix} E(k) \begin{bmatrix} M & N \end{bmatrix}\right)$$

and stochastic uncertainty

$$\begin{pmatrix} \sum_{i=1}^{L} A_i p_i(k) & \sum_{i=1}^{L} B_i p_i(k) \\ \sum_{i=1}^{L} C_i p_i(k) & \sum_{i=1}^{L} D_i p_i(k) \end{pmatrix}.$$

We refer to

$$\begin{bmatrix} \begin{bmatrix} A(k) & B(k) \\ C(k) & D(k) \end{bmatrix} \\ \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \\ \cdots \\ \begin{bmatrix} A_L & B_L \\ C_L & D_L \end{bmatrix} \end{bmatrix} \tag{3}$$

as a realization of stochastic system S. Throughout this paper, we use the following definitions.

DEFINITION 2.1.

$$\left[\begin{array}{c} \left[\begin{array}{cc} \bar{A}(k) & \bar{B}(k) \\ \bar{C}(k) & \bar{D}(k) \end{array}\right] \\ \left[\begin{array}{cc} \bar{A}_1 & \bar{B}_1 \\ \bar{C}_1 & \bar{D}_1 \end{array}\right] \\ \cdots \\ \left[\begin{array}{cc} \bar{A}_L & \bar{B}_L \\ \bar{C}_L & \bar{D}_L \end{array}\right] \end{array}\right] \tag{4}$$

is an equivalent realization to (3) *if there exists a nonsingular matrix* T *such that*

$$\left[\begin{array}{cc} \bar{A}(k) & \bar{B}(k) \\ \bar{C}(k) & \bar{D}(k) \end{array}\right] = \left[\begin{array}{cc} T & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{cc} A(k) & B(k) \\ C(k) & D(k) \end{array}\right] \left[\begin{array}{cc} T^{-1} & 0 \\ 0 & I \end{array}\right],$$

$$\left[\begin{array}{cc} \bar{A}_i & \bar{B}_i \\ \bar{C}_i & \bar{D}_i \end{array}\right] = \left[\begin{array}{cc} T & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{cc} A_i & B_i \\ C_i & D_i \end{array}\right] \left[\begin{array}{cc} T^{-1} & 0 \\ 0 & I \end{array}\right], \ i = 1, \ldots, L.$$

In this case, T *is called an equivalent transformation, and* (4) *is one realization of the stochastic system* S.

It can be verified that equivalent realizations induce the same input-output behavior.

DEFINITION 2.2 (see [2, 21]). *The uncertain stochastic system* S *is said to be robustly mean-square stable if for any feasible* $E(k)$, *the following equation holds:*

$$\lim_{k\to\infty} \mathcal{E}\left\{x(k)x(k)^T\right\} = 0$$

when $u(k) = 0$.

DEFINITION 2.3 (see [4, 21]). *The uncertain stochastic system* S *is said to be externally stable if for every* $u(k) \in l_{e_2}([0,\infty); R^{n_u})$,

$$y(k) \in l_{e_2}([0,\infty); R^{n_y})$$

and there exists a scalar $\eta > 0$ *such that*

$$\|y\|_{e_2} \leq \eta \|u\|_{e_2}. \tag{5}$$

DEFINITION 2.4 (see [4, 21]). *Suppose that system* S *is externally stable. Under the zero initial condition, the operator*

$$\mathbb{L} : l_{e_2}([0,\infty); R^{n_u}) \to l_{e_2}([0,\infty); R^{n_y})$$

defined by

$$(\mathbb{L}u)(k) = y(k)$$

is called the perturbation operator of system S. *Its norm* $\|\mathbb{L}\|$ *is defined as the minimum* $\eta > 0$ *such that* (5) *is satisfied.*

For the mean-square stability, we have the following lemma.

LEMMA 2.5 (see [2]). *The uncertain stochastic system* S *is robustly mean-square stable if and only if there exists* $Q > 0$ *satisfying the inequality*

$$-Q + A(k)^T Q A(k) + \sum_{i=1}^{L} \sigma_i^2 A_i^T Q A_i < 0,$$

or dually there exists $P > 0$ *satisfying*

$$-P + A(k)PA(k)^T + \sum_{i=1}^{L} \sigma_i^2 A_i P A_i^T < 0$$

for any feasible $E(k)$.

The following lemmas will be useful to the derivation of our main results.

LEMMA 2.6 (see [21]). *The uncertain stochastic system* S *is robustly mean-square stable and* $\|\mathbb{L}\| < \gamma$ *if and only if there exists* $X > 0$ *satisfying*

$$-\begin{bmatrix} X & 0 \\ 0 & \gamma^2 I \end{bmatrix} + \begin{bmatrix} A(k) & B(k) \\ C(k) & D(k) \end{bmatrix}^T \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A(k) & B(k) \\ C(k) & D(k) \end{bmatrix} + \sum_{i=1}^{L} \sigma_i^2 \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix}^T \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix} < 0$$

for any feasible $E(k)$.

LEMMA 2.7 (see [13]). *Let* $\Sigma_1, \Sigma_2, \Sigma_3$, *and* Γ *be given constant matrices with appropriate dimensions such that* $\Gamma = \Gamma^T < 0$. *Then there exists a matrix* $\Lambda = \Lambda^T > 0$ *such that*

$$(\Sigma_1 + \Sigma_2 E(k) \Sigma_3)^T \Lambda (\Sigma_1 + \Sigma_2 E(k) \Sigma_3) + \Gamma < 0$$

for any feasible $E(k)$ *if and only if there exists a scalar* $\varepsilon > 0$ *such that the following condition holds:*

$$\begin{bmatrix} \Gamma + \varepsilon (\Sigma_3)^T \Sigma_3 & (\Sigma_1)^T \\ \Sigma_1 & -\Lambda^{-1} + \varepsilon^{-1} \Sigma_2 (\Sigma_2)^T \end{bmatrix} < 0.$$

Remark 1. The results to be developed in this paper can be extended to the case where there are also uncertainties in A_i, B_i, C_i, and D_i $(i = 1, \ldots, L)$, provided that the uncertainties have the same structures as those of $A(k), B(k), C(k)$, and $D(k)$, respectively. In this case, Lemma 2.7 should be extended as follows: Let $\Sigma_1^j, \Sigma_2^j, \Sigma_3^j$ $(j = 1, \ldots, L)$, and Γ be given constant matrices with appropriate dimensions such that $\Gamma = \Gamma^T$. Then there exists a matrix $\Lambda = \Lambda^T > 0$ such that

$$\sum_{j=1}^{N} \left(\Sigma_1^j + \Sigma_2^j E(k) \Sigma_3^j \right)^T \Lambda \left(\Sigma_1^j + \Sigma_2^j E(k) \Sigma_3^j \right) + \Gamma < 0 \tag{6}$$

for any $E(k) \in \Omega$ if there exist scalars $\varepsilon_1 > 0, \ldots, \varepsilon_L > 0$ such that the following condition holds:

$$\Gamma + \sum_{j=1}^{N} \left(\varepsilon_j^{-1} (\Sigma_3^j)^T \Sigma_3^j + (\Sigma_1^j)^T (\Lambda^{-1} - \varepsilon_j \Sigma_2^j (\Sigma_2^j)^T)^{-1} \Sigma_1^j \right) < 0. \tag{7}$$

It is noticed that in this case, (7) is only a sufficient condition for (6).

The model reduction problem for the uncertain stochastic system S can be stated as follows: for the robustly mean-square stable stochastic system S, find an uncertain mean-square stable stochastic system

$$S_r : \begin{cases} x_r(k+1) = A_r(k) x_r(k) + B_r(k) u(k) + \sum_{i=1}^{L} \left(A_{ri} x_r(k) + B_{ri} u(k) \right) p_i(k), \\ y_r(k) = C_r(k) x_r(k) + D_r(k) u(k) + \sum_{i=1}^{L} \left(C_{ri} x_r(k) + D_{ri} u(k) \right) p_i(k), \end{cases}$$

where $x_r(k) \in R^r$ and $y_r(k) \in R^{n_y}$ with $r < n$ such that S_r approximates S according to a certain criterion. In order to preserve the structure of S, it is also assumed that $A_r(k)$, $B_r(k)$, $C_r(k)$, and $D_r(k)$ are time-varying and have the structure of

$$\begin{bmatrix} A_r(k) & B_r(k) \\ C_r(k) & D_r(k) \end{bmatrix} = \begin{bmatrix} A_r & B_r \\ C_r & D_r \end{bmatrix} + \begin{bmatrix} U_r \\ V_r \end{bmatrix} E(k) \begin{bmatrix} M_r & N_r \end{bmatrix} \tag{8}$$

with $E(k)$ feasible. This assumption means that the uncertainties in both the original models and reduced order models should come from the same sources but with the different weighting matrices $\{U, V, M, N\}$ and $\{U_r, V_r, M_r, N_r\}$. A similar assumption was made in [1].

3. Balanced realization and truncation. Consider the robustly mean-square stable stochastic system S and the following two inequalities:

$$\begin{bmatrix} -P & 0 & \cdots & 0 \\ 0 & -P & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -P \end{bmatrix} + \begin{bmatrix} A(k) & B(k) \\ \sigma_1 A_1 & \sigma_1 B_1 \\ \vdots & \vdots \\ \sigma_L A_L & \sigma_L B_L \end{bmatrix} \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A(k) & B(k) \\ \sigma_1 A_1 & \sigma_1 B_1 \\ \vdots & \vdots \\ \sigma_L A_L & \sigma_L B_L \end{bmatrix}^T < 0, \tag{9}$$

$$-Q + A(k)^T Q A(k) + C(k)^T C(k) + \sum_{i=1}^{L} \sigma_i^2 (A_i Q A_i^T + C_i^T C_i) < 0. \tag{10}$$

By Lemma 2.5, it can be seen from (9) and (10) that the robust mean-square stability is necessary for the feasibilities of (9) and (10) for $P > 0$ and $Q > 0$.

Remark 2. If no multiplicative noise, or stochastic uncertainty, appears in S, that is, $\sigma_i = 0$, then inequalities (9) and (10) reduce to

$$-P + A(k) P A(k)^T + B(k) B(k)^T < 0,$$
$$-Q + A(k)^T Q A(k) + C(k)^T C(k) < 0,$$

respectively, which coincide with the Lyapunov inequalities given in [1] for deterministic linear systems.

Assume that (9) and (10) are feasible for $P > 0$ and $Q > 0$. Since P and Q are positive definite, there exists a nonsingular matrix T such that P and Q are balanced; that is,

$$T P T^T = T^{-T} Q T^{-1} = \Sigma, \tag{11}$$

where Σ=diag$(\alpha_1, \alpha_2, \ldots, \alpha_n)$ with $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n > 0$. The balanced realization of the stochastic system S is defined as follows.

DEFINITION 3.1. *Assume that* (9) *and* (10) *are feasible for* $P > 0$ *and* $Q > 0$ *and the nonsingular matrix* T *is such that* (11) *holds. Let*

$$\begin{bmatrix} A_b(k) & B_b(k) \\ C_b(k) & D_b(k) \end{bmatrix} := \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A(k) & B(k) \\ C(k) & D(k) \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}, \tag{12}$$

$$\begin{bmatrix} A_{bi} & B_{bi} \\ C_{bi} & D_{bi} \end{bmatrix} := \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}, \ i = 1, \ldots, L. \tag{13}$$

Then

$$\begin{bmatrix} A_b(k) & B_b(k) \\ C_b(k) & D_b(k) \\ A_{b1} & B_{b1} \\ C_{b1} & D_{b1} \\ \vdots & \vdots \\ \vdots & \vdots \\ A_{bL} & B_{bL} \\ C_{bL} & D_{bL} \end{bmatrix} \tag{14}$$

is called a balanced realization of the stochastic system S.

By substituting (8) into (12) and defining

$$\begin{bmatrix} A_b & B_b \\ C_b & D_b \end{bmatrix} := \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}, \tag{15}$$

$$\begin{bmatrix} U_b \\ V_b \end{bmatrix} := \begin{bmatrix} T & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix}, \tag{16}$$

$$\begin{bmatrix} M_b & N_b \end{bmatrix} := \begin{bmatrix} M & N \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & I \end{bmatrix}, \tag{17}$$

we have

$$\begin{bmatrix} A_b(k) & B_b(k) \\ C_b(k) & D_b(k) \end{bmatrix} = \begin{bmatrix} A_b & B_b \\ C_b & D_b \end{bmatrix} + \begin{bmatrix} U_b \\ V_b \end{bmatrix} E(k) \begin{bmatrix} M_b & N_b \end{bmatrix}.$$

To find the balanced realization of the stochastic system S, the key is how to solve $P > 0$ and $Q > 0$ from (9) and (10) if they are feasible. It is noticed that inequalities (9) and (10) are nonlinear and time-varying, and hence they are difficult to solve with the current forms. The following theorem will give the equivalent time-invariant LMIs by the Schur complements and Lemma 2.7.

THEOREM 3.2. *Assume that* (9) *and* (10) *are feasible for* $P > 0$ *and* $Q > 0$. *Then* $P > 0$ *and* $Q > 0$ *can be solved from the following two LMIs:*

$$\begin{bmatrix} -P + \varepsilon_1 UU^T & 0 & \cdots & 0 & AP & B & 0 \\ 0 & -P & \cdots & 0 & \sigma_1 A_1 P & \sigma_1 B_1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -P & \sigma_L A_L P & \sigma_L B_L & 0 \\ PA^T & \sigma_1 PA_1^T & \cdots & \sigma_L PA_L^T & -P & 0 & PM^T \\ B^T & \sigma_1 B_1^T & \cdots & \sigma_L B_L^T & 0 & -I & N^T \\ 0 & 0 & \cdots & 0 & MP & N & -\varepsilon_1 I \end{bmatrix} < 0, \tag{18}$$

$$\begin{bmatrix} -Q + \sum_{i=1}^{L} \sigma_i^2 (A_i^T Q A_i + C_i^T C_i) + \varepsilon_2 M^T M & A^T Q & C^T & 0 \\ QA & -Q & 0 & QU \\ C & 0 & -I & V \\ 0 & U^T Q & V^T & -\varepsilon_2 I \end{bmatrix} < 0, \tag{19}$$

where $\varepsilon_1 > 0$ *and* $\varepsilon_2 > 0$ *are two scalar variables.*

Proof. Notice that (9) and (10) can be written, respectively, as

$$\begin{bmatrix} -P & 0 & \cdots & 0 \\ 0 & -P & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & -P \end{bmatrix} + \left(\begin{bmatrix} A(k) & B(k) \\ \sigma_1 A_1 & \sigma_1 B_1 \\ \vdots & \vdots \\ \sigma_L A_L & \sigma_L B_L \end{bmatrix} + \begin{bmatrix} U \\ 0 \\ \vdots \\ 0 \end{bmatrix} E(k) \begin{bmatrix} M & N \end{bmatrix} \right)$$
$$\times \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} A(k) & B(k) \\ \sigma_1 A_1 & \sigma_1 B_1 \\ \vdots & \vdots \\ \sigma_L A_L & \sigma_L B_L \end{bmatrix} + \begin{bmatrix} U \\ 0 \\ \vdots \\ 0 \end{bmatrix} E(k) \begin{bmatrix} M & N \end{bmatrix} \right)^T < 0,$$

$$-Q + \sum_{i=1}^{L} \sigma_i^2 (A_i^T Q A_i + C_i^2 C_i)$$
$$+ \left(\begin{bmatrix} A \\ C \end{bmatrix} + \begin{bmatrix} U \\ V \end{bmatrix} E(k) M \right)^T \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} A \\ C \end{bmatrix} + \begin{bmatrix} U \\ V \end{bmatrix} E(k) M \right) < 0.$$

By Lemma 2.7, it can be seen that (9) and (10) are equivalent to

$$\begin{bmatrix} \begin{bmatrix} -P & 0 & \cdots & 0 \\ 0 & -P & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & -P \end{bmatrix} + \varepsilon_1 \begin{bmatrix} U \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} U \\ 0 \\ \vdots \\ 0 \end{bmatrix}^T & \begin{bmatrix} A & B \\ \sigma_1 A_1 & \sigma_1 B_1 \\ \vdots & \vdots \\ \sigma_L A_L & \sigma_L B_L \end{bmatrix} \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \\ \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ \sigma_1 A_1 & \sigma_1 B_1 \\ \vdots & \vdots \\ \sigma_L A_L & \sigma_L B_L \end{bmatrix}^T & -\begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} + \varepsilon_1^{-1} \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} M^T \\ N^T \end{bmatrix} \\ & \times \begin{bmatrix} M & N \end{bmatrix} \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \end{bmatrix} < 0,$$

$$\begin{bmatrix} -Q + \sum_{i=1}^{L} \sigma_i^2 (A_i^T Q A_i + C_i^2 C_i) + \varepsilon_2 M^T M & \begin{bmatrix} A^T & C^T \end{bmatrix} \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \\ \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A \\ C \end{bmatrix} & -\begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} + \varepsilon_2^{-1} \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix} \\ & \times \begin{bmatrix} U \\ V \end{bmatrix}^T \begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} \end{bmatrix} < 0,$$

respectively. By the Schur complements again, we obtain (18) and (19). ∎

Remark 3. It is noticed that when S has only stochastic uncertainty (multiplicative noise) and no deterministic uncertainty, that is, $E(k) = 0$, then (18) and (19) reduce to

$$\begin{bmatrix} -P & 0 & \cdots & 0 & AP & B \\ 0 & -P & \cdots & 0 & \sigma_1 A_1 P & \sigma_1 B_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & -P & \sigma_L A_L P & \sigma_L B_L \\ PA^T & \sigma_1 PA_1^T & \cdots & \sigma_L PA_L^T & -P & 0 \\ B^T & \sigma_1 B_1^T & \cdots & \sigma_L B_L^T & 0 & -I \end{bmatrix} < 0,$$

$$\begin{bmatrix} -Q + \sum_{i=1}^{L} \sigma_i^2 (A_i^T Q A_i + C_i^T C_i) & A^T Q & C^T \\ QA & -Q & 0 \\ C & 0 & -I \end{bmatrix} < 0,$$

respectively. On the contrary, when S has only deterministic uncertainty and no stochastic uncertainty, that is, $\sigma_i = 0$, then (18) and (19) reduce to

$$\begin{bmatrix} -P+\varepsilon_1 UU^T & AP & B & 0 \\ PA^T & -P & 0 & PM^T \\ B^T & 0 & -I & N^T \\ 0 & MP & N & -\varepsilon_1 I \end{bmatrix} < 0,$$

$$\begin{bmatrix} -Q+\varepsilon_2 M^T M & A^T Q & C^T & 0 \\ QA & -Q & 0 & QU \\ C & 0 & -I & V \\ 0 & U^T Q & V^T & -\varepsilon_2 I \end{bmatrix} < 0,$$

respectively.

Premultiplying and postmultiplying both sides of (10) by T^T and T, respectively, we have

$$-\Sigma + A_b(k)^T \Sigma A_b(k) + C_b(k)^T C_b(k) + \sum_{i=1}^{L} \sigma_i^2 (A_{bi}^T \Sigma A_{bi} + C_{bi}^T C_{bi}) < 0. \tag{20}$$

Similarly, (9) can be transformed to

$$\begin{bmatrix} -\Sigma & 0 & \cdots & 0 \\ 0 & -\Sigma & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -\Sigma \end{bmatrix} + \begin{bmatrix} A_b(k) & B_b(k) \\ \sigma_1 A_{b1} & \sigma_1 B_{b1} \\ \vdots & \vdots \\ \sigma_L A_{bL} & \sigma_L B_{bL} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_b(k) & B_b(k) \\ \sigma_1 A_{b1} & \sigma_1 B_{b1} \\ \vdots & \vdots \\ \sigma_L A_{bL} & \sigma_L B_{bL} \end{bmatrix}^T < 0. \tag{21}$$

The balanced truncation is a simple but powerful model reduction method. In order to obtain an rth reduced order model S_r for the stochastic system S, partition $A_b(k)$, $B_b(k)$, $C_b(k)$, A_{b1}, B_{b1}, C_{b1} conformably as

$$A_b(k) = \begin{bmatrix} A_{11}(k) & A_{12}(k) \\ A_{21}(k) & A_{22}(k) \end{bmatrix}, \; B_b(k) = \begin{bmatrix} \tilde{B}_1(k) \\ \tilde{B}_2(k) \end{bmatrix}, \; C_b(k) = \begin{bmatrix} \tilde{C}_1(k) & \tilde{C}_2(k) \end{bmatrix}, \tag{22}$$

$$A_{bi} = \begin{bmatrix} A_{11}^i & A_{12}^i \\ A_{21}^i & A_{22}^i \end{bmatrix}, \; B_{bi} = \begin{bmatrix} \tilde{B}_1^i \\ \tilde{B}_2^i \end{bmatrix}, \; C_{bi} = \begin{bmatrix} \tilde{C}_1^i & \tilde{C}_2^i \end{bmatrix}, \; i = 1, \ldots, L, \tag{23}$$

$$A_b = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \; B_b = \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix}, \; \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{24}$$

$$C_b = \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix}, \; U_b = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \; M_b = \begin{bmatrix} M_1 & M_2 \end{bmatrix}, \tag{25}$$

where $A_{11} \in R^{r\times r}$ and $A_{11}^i \in R^{r\times r}$ for $i = 1, \ldots, L$. Then an rth reduced order

stochastic model S_r is derived as

$$\begin{bmatrix} A_r(k) & B_r(k) \\ C_r(k) & D_r(k) \\ A_{r1} & B_{r1} \\ C_{r1} & D_{r1} \\ \vdots & \vdots \\ \vdots & \vdots \\ A_{rL} & B_{rL} \\ C_{rL} & D_{rL} \end{bmatrix}, \tag{26}$$

where

$$\begin{bmatrix} A_r(k) & B_r(k) \\ C_r(k) & D_r(k) \end{bmatrix}$$

has the structure of (8) with

$$\begin{bmatrix} A_r & B_r \\ C_r & D_r \end{bmatrix} = \begin{bmatrix} A_{11} & \tilde{B}_1 \\ \tilde{C}_1 & D_b \end{bmatrix}, \begin{bmatrix} A_{ri} & B_{ri} \\ C_{ri} & D_{ri} \end{bmatrix} = \begin{bmatrix} A_{11}^i & \tilde{B}_1^i \\ \tilde{C}_1^i & D_b^i \end{bmatrix}, \tag{27}$$

$$\begin{bmatrix} M_r & N_r \end{bmatrix} = \begin{bmatrix} M_1 & N_b \end{bmatrix}, \begin{bmatrix} U_r \\ V_r \end{bmatrix} = \begin{bmatrix} U_1 \\ V_b \end{bmatrix}, \ i = 1, \ldots, L. \tag{28}$$

For the obtained reduced order model, the following properties can be derived.

LEMMA 3.3. *For the balanced model with the realization in* (14)*, the truncated model with the realization in* (26) *is robustly mean-square stable.*

Proof. From the (1,1) block of (20), we have

$$\begin{aligned} &-\Sigma_1 + A_{11}(k)^T \Sigma_1 A_{11}(k) + A_{21}(k)^T \Sigma_2 A_{21}(k) + \tilde{C}_1(k)^T \tilde{C}_1(k) \\ &\quad + \sum_{i=1}^{L} \sigma_i^2 \left((A_{11}^i)^T \Sigma_1 A_{11}^i + (A_{21}^i)^T \Sigma_2 A_{21}^i + (\tilde{C}_1^i)^T \tilde{C}_1^i \right) < 0, \end{aligned}$$

which further implies

$$-\Sigma_1 + A_{11}(k)^T \Sigma_1 A_{11}(k) + \tilde{C}_1(k)^T \tilde{C}_1(k) + \sum_{i=1}^{L} \sigma_i^2 \left((A_{11}^i)^T \Sigma_1 A_{11}^i + (\tilde{C}_1^i)^T \tilde{C}_1^i \right) < 0.$$

From Lemma 2.5, it can be seen that the stochastic model in (26) is robustly mean-square stable. ∎

The following theorem shows that the balanced truncation of the uncertain stochastic systems can guarantee an upper bound of the model reduction error in the sense of perturbation operator norm. The proof is quite involved and is therefore given in the appendix.

THEOREM 3.4. *If the truncated model with the realization in* (26) *is denoted as* S_r*, then the* H_∞ *model reduction error between* S *and* S_r *satisfies*

$$\|\mathbb{L}_e\| < 2 \sum_{i=r+1}^{n} \alpha_i,$$

where $\mathbb{L}_e$ *is the perturbation operator of the error system between* S *and* S_r*.*

It is noticed that in this paper we use two Lyapunov inequalities for balanced truncation; the distinctness of α_i in Σ_1 and Σ_2 is not necessary for the proofs of stability and error bound, though it is necessary when two Lyapunov equations are used for the continuous-time linear time-invariant systems (see [16]).

4. Model reduction algorithm. It is noticed that the solutions of the LMIs in (18) and (19) are not unique, and this results in the nonuniqueness of the balanced realization and the truncated model. On the other hand, Theorem 3.4 notes that the model reduction error bound depends on α_i's, and it can be seen that α_i^2's are the eigenvalues of PQ. Thus, in order to achieve a small model reduction error bound, a reasonable way is to solve $P > 0$ and $Q > 0$ from the optimization

$$\min_{P>0,Q>0} \operatorname{trace}(PQ) \tag{29}$$

subject to (18) and (19). By linearization, the following LMI-based iterative algorithm is proposed.

LMI-based Iterative Algorithm. Given $A, B, C, D, A_1, B_1, C_1, D_1, \sigma_i, U, V, M, N$, and $\varepsilon > 0$,

Step 1 Solve $P_0 > 0$ and $Q_0 > 0$ from (18) and (19);

Step 2 Define the linear function

$$f_k(P, Q) \triangleq \operatorname{trace}(QP_k + PQ_k); \tag{30}$$

Step 3 Solve $P_{\min} > 0$ and $Q_{\min} > 0$ from the following convex minimization problem:

$$\min_{P>0,Q>0} f_k \tag{31}$$

$$\text{s.t. (18) and (19);} \tag{32}$$

let $P_{k+1} = P_{\min}$ and $Q_{k+1} = Q_{\min}$, and compute the minimum as

$$f_k^* = \operatorname{trace}(Q_{k+1}P_k + P_{k+1}Q_k);$$

Step 4 If $\left|f_{k+1}^* - f_k^*\right| < \varepsilon$, then $P = X_{k+1}, Q = Y_{k+1}$; otherwise, set $k = k + 1$, and go to Step 2.

It can be seen that Step 1 is a simple LMI problem, and Step 3 is a convex programming with LMI constraints. The proposed algorithm has the following properties:

(i) $0 < f_{k+1}^* \le f_k^*$.

(ii) If the minimization problem (31) is feasible, then it is also feasible for $k + 1$.

Property (i) can be simply obtained from the fact that (P_{k+1}, Q_{k+1}) is such that f_k is minimized and (P_k, Q_k) is such that f_{k-1} is minimized. It shows that $\{f_k^*\}$ converges. Property (ii) holds since (P_k, Q_k) is a feasible solution to (31) when k is set to $k+1$. Consequently, from these two properties, this algorithm can be iteratively implemented if (18) and (19) are feasible.

5. Numerical example.

Example 1. Consider the following stochastic system S without deterministic uncertainty:

$$A = \begin{bmatrix} 0 & 0 & 0.0267 & 0 & 0 \\ 1.1111 & 0 & -0.2889 & 0 & 0 \\ 0 & 1.1111 & 1 & 0 & 0 \\ 0 & 0 & 0.2222 & 0 & -0.0667 \\ 0 & 0 & 0.1667 & 1.1111 & 0.5556 \end{bmatrix},$$

$$B = \begin{bmatrix} 0.8 \\ 0.6 \\ 0.4 \\ 0.2 \\ 0.5 \end{bmatrix},\ B_1 = \begin{bmatrix} 0.1 \\ 0.04 \\ 0.06 \\ 0.2 \\ 0.1 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0.3649 & 0.0081 & 0.2189 & 0.0906 & -0.0820 \\ -0.0722 & 0.2038 & -0.1152 & -0.0059 & 0.0340 \\ -0.1063 & 0.0137 & -0.0233 & -0.0450 & 0.0426 \\ 0.0254 & -0.0285 & 0.0271 & 0.1013 & 0.0073 \\ 0.1412 & 0.0370 & 0.1104 & 0.0642 & 0.1033 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.2 \\ 0.4 \\ 0.6 \\ 0.8 \\ 1 \end{bmatrix}^T,\ C_1 = B_1^T,$$

$$D = 0.5,\ D_1 = 0.2,\ \sigma_1 = 1,\ U = 0,\ V = 0,\ M = 0,\ N = 0.$$

Our purpose is to find the 2nd and 3rd order reduced models without deterministic uncertainty using the proposed balanced truncation. By balancing P and Q, we obtain

$$\Sigma = \text{diag}\{4.2739, 0.6855, 0.0724, 0.0037, 0.0002\}.$$

The 2nd order reduced model is given by

$$A_r = \begin{bmatrix} 0.9097 & 0.1024 \\ -0.4046 & 0.6910 \end{bmatrix},\ A_{r1} = \begin{bmatrix} 0.2335 & 0.0032 \\ 0.0576 & 0.0840 \end{bmatrix},$$

$$B_r = \begin{bmatrix} -2.2336 \\ 0.1314 \end{bmatrix},\ B_{r1} = \begin{bmatrix} -0.3461 \\ 0.3163 \end{bmatrix},$$

$$C_r = \begin{bmatrix} -0.5390 & 0.3859 \end{bmatrix},\ C_{r1} = \begin{bmatrix} -0.1098 & 0.0856 \end{bmatrix},\ D_r = 0.5,\ D_{1r} = 0.2,$$

and the 3rd order reduced model has state matrices as

$$A_r = \begin{bmatrix} 0.9097 & 0.1024 & -0.0103 \\ -0.4046 & 0.6910 & 0.0644 \\ 0.6961 & -0.9390 & -0.2286 \end{bmatrix},\ A_{r1} = \begin{bmatrix} 0.2335 & 0.0032 & -0.0278 \\ 0.0576 & 0.0840 & -0.0917 \\ 0.0373 & 0.0044 & 0.1728 \end{bmatrix},$$

$$B_r = \begin{bmatrix} -2.2336 \\ 0.1314 \\ -0.4830 \end{bmatrix},\ B_{r1} = \begin{bmatrix} -0.3461 \\ 0.3163 \\ -0.0944 \end{bmatrix},$$

$$C_r = \begin{bmatrix} -0.5390 \\ 0.3859 \\ -0.1691 \end{bmatrix}^T,\ C_{r1} = \begin{bmatrix} -0.1098 \\ 0.0856 \\ 0.0334 \end{bmatrix}^T,\ D_r = 0.5,\ D_{r1} = 0.2.$$

The outputs of the original model and the 2nd and 3rd order reduced models are given in Figure 1 when the input is set to e^{-k}. It can be seen that the 3rd order reduced model approximates the original model better than the 2nd order reduced model.

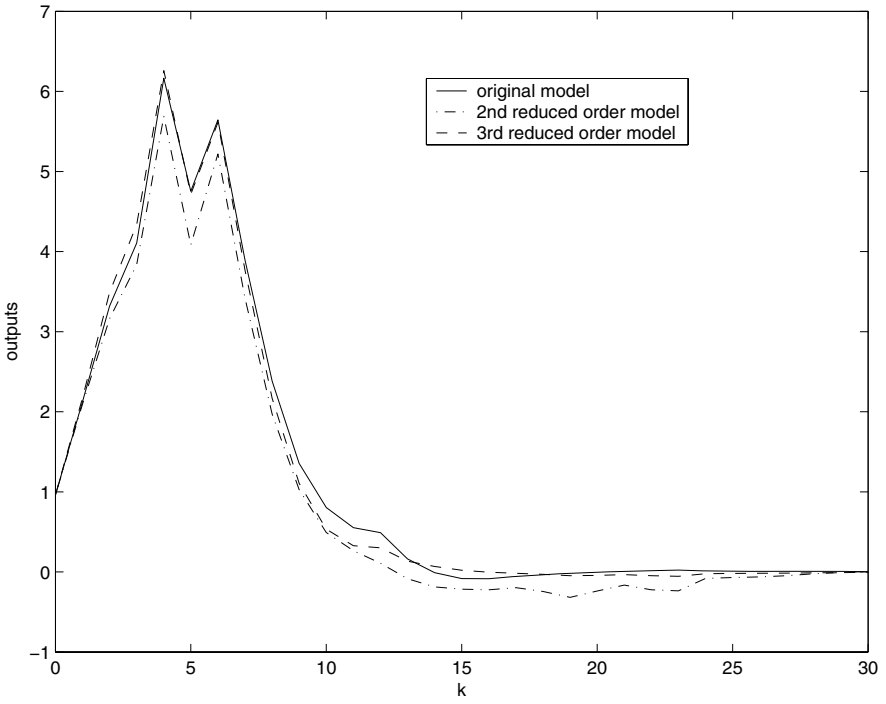


FIG. 1. *Comparison of the outputs: without uncertainty.*

Example 2. Consider the uncertain stochastic system S, in which $A, A_1, B, B_1, C, C_1, D$, and D_1 are given in Example 1, and U, V, M, N are

$$U = \begin{bmatrix} 0.1188 \\ 0.0289 \\ 0.0759 \\ 0.0607 \\ 0.1114 \end{bmatrix},\ V = -0.1618,\ M = \begin{bmatrix} 0.1191 \\ 0.0921 \\ -0.3434 \\ 0.0761 \\ -0.0009 \end{bmatrix},\ N = -0.0871.$$

By balancing P and Q, we obtain

$$\Sigma = \operatorname{diag}\{16.3927, 6.5846, 4.7287, 3.3965, 0.7145\}.$$

The 2nd order reduced model is given by

$$A_r = \begin{bmatrix} 0.7938 & -0.1855 \\ 0.1607 & 0.5968 \end{bmatrix},\ A_{r1} = \begin{bmatrix} 0.1532 & -0.0742 \\ -0.0903 & 0.1062 \end{bmatrix},$$
$$B_r = \begin{bmatrix} -1.3901 \\ 1.1269 \end{bmatrix},\ B_{r1} = \begin{bmatrix} -0.2208 \\ 0.0778 \end{bmatrix},$$
$$C_r = \begin{bmatrix} -1.8819 & -0.8822 \end{bmatrix},\ C_{r1} = \begin{bmatrix} -0.2338 & -0.0904 \end{bmatrix},\ D_r = 0.5,\ D_{1r} = 0.2,$$
$$U_r = \begin{bmatrix} -0.1935 \\ 0.1166 \end{bmatrix},\ V_r = -0.1618,\ M_r = \begin{bmatrix} 0.2850 & 0.5415 \end{bmatrix},\ N_r = -0.0871,$$

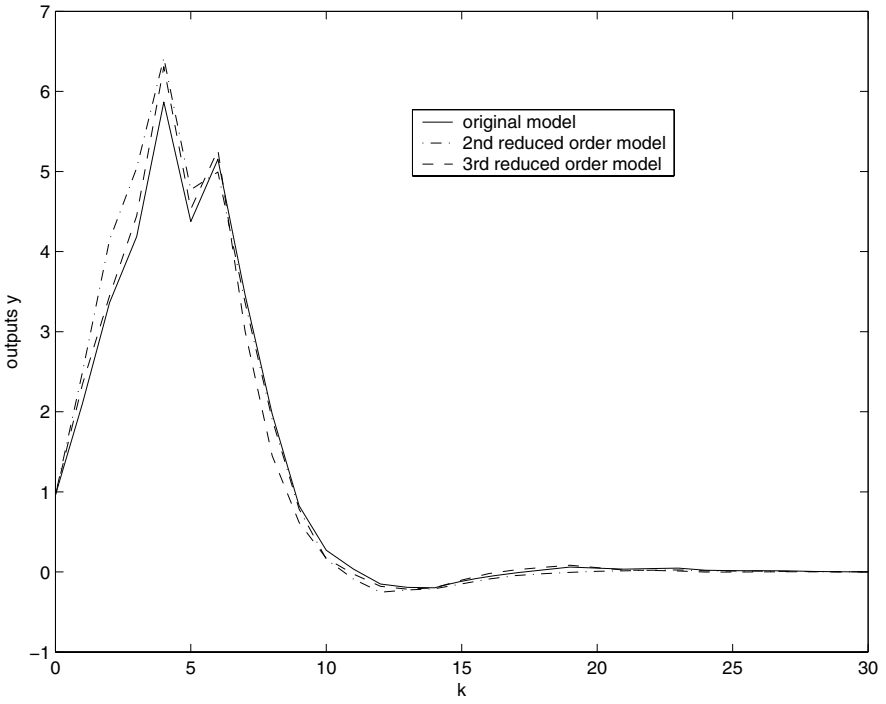


FIG. 2. *Comparison of the outputs: with uncertainty.*

and the 3rd order reduced model has state matrices as

$$A_r = \begin{bmatrix} 0.7938 & -0.1855 & -0.0950 \\ 0.1607 & 0.5968 & -0.4748 \\ -0.0373 & 0.3754 & 0.3300 \end{bmatrix},\ A_{r1} = \begin{bmatrix} 0.1532 & -0.0742 & 0.0178 \\ -0.0903 & 0.1062 & -0.0003 \\ 0.1267 & 0.1577 & 0.1645 \end{bmatrix},$$

$$B_r = \begin{bmatrix} -1.3901 \\ 1.1269 \\ 0.0363 \end{bmatrix},\ B_{r1} = \begin{bmatrix} -0.2208 \\ 0.0778 \\ -0.0452 \end{bmatrix},$$

$$C_r = \begin{bmatrix} -1.8819 \\ -0.8822 \\ -0.9196 \end{bmatrix}^T,\ C_{r1} = \begin{bmatrix} -0.2338 \\ -0.0904 \\ -0.2257 \end{bmatrix}^T,\ D_r = 0.5,\ D_{r1} = 0.2,$$

$$U_r = \begin{bmatrix} -0.1935 \\ 0.1166 \\ -0.0206 \end{bmatrix},\ V_r = -0.1618,\ M_r = \begin{bmatrix} 0.2850 & 0.5415 & -0.2667 \end{bmatrix},$$

$$N_r = -0.0871.$$

The outputs of the original model and the 2nd and 3rd order reduced models are given in Figure 2 when the input is set to e^{-k} and $E(k) = \sin(k)$. It can also be seen that the 3rd order reduced model approximates the original model better than the 2nd order reduced model.

6. Conclusions. In this paper, we have studied the model reduction problem based on balancing for uncertain linear systems with multiplicative noise, or linear systems with both deterministic and stochastic uncertainties. The deterministic un-

certainties are structured time-varying and norm-bounded, and the stochastic uncertainties are expressed as a multiplicative noise. The original model can be balanced through a balanced transformation, which can be found by balancing the solutions of two LMIs. By truncating the balanced model, the reduced model with a desired order is obtained. The upper bound of the model reduction error is guaranteed in the sense of perturbation operator norm. Since the solutions of the LMIs are not unique, an optimization problem that ensures small model reduction error bound has been proposed. Two numerical examples have been given for illustration of the main results.

Appendix: Proof of Theorem 3.4. For deterministic linear discrete-time systems, the model reduction based on balancing and LMIs was considered in [1]. Here, we will prove Theorem 3.4 by generalizing the technique given in [1]. For simplicity, the proof will be given in the case when $L = 1$. The case when $L > 1$ can be obtained similarly. The following definition and lemmas are essential for the proof.

DEFINITION 6.1. *A matrix set* $\{X_1, X_2, \ldots, X_s\}$ *is contractive if* $\sum_{i=1}^{s} X_i^T X_i \le I$ *and strictly contractive if* $\sum_{i=1}^{s} X_i^T X_i < I$.

When $s = 1$, Definition 6.1 reduces to the contractiveness of a matrix defined in [1]. Based on Definition 6.1 and Lemma 2.6, the following lemma follows.

LEMMA 6.2. *The uncertain stochastic system* S *is robustly mean-square stable and* $\|\mathbb{L}\| < \gamma$ *if and only if there exists one realization, denoted as*

$$\begin{bmatrix} \bar{A}(k) & \bar{B}(k) \\ \bar{C}(k) & \bar{D}(k) \\ \bar{A}_1 & \bar{B}_1 \\ \bar{C}_1 & \bar{D}_1 \end{bmatrix} \tag{33}$$

such that the matrix set

$$\left\{ \begin{bmatrix} \bar{A}(k) & \frac{1}{\sqrt{\gamma}}\bar{B}(k) \\ \frac{1}{\sqrt{\gamma}}\bar{C}(k) & \frac{1}{\gamma}\bar{D}(k) \end{bmatrix}, \begin{bmatrix} \sigma_1\bar{A}_1 & \frac{\sigma_1}{\sqrt{\gamma}}\bar{B}_1 \\ \frac{\sigma_1}{\sqrt{\gamma}}\bar{C}_1 & \frac{\sigma_1}{\gamma}\bar{D}_1 \end{bmatrix} \right\} \tag{34}$$

is strictly contractive.

Proof. From Lemma 2.6, $\|\mathbb{L}\| < \gamma$ if and only if there exists $X > 0$ such that

$$\begin{bmatrix} A(k)^T X A(k) + C(k)^T C(k) & A(k)^T X B(k) + C(k)^T D(k) \\ B(k)^T X A(k) + D(k)^T C(k) & B(k)^T X B(k) + D(k)^T D(k) \end{bmatrix} + \sigma_1^2 \begin{bmatrix} A_1^T X A_1 + C_1^T C_1 & A_1^T X B_1 + C_1^T D_1 \\ B_1^T X A_1 + D_1^T C_1 & B_1^T X B_1 + D_1^T D_1 \end{bmatrix} < \begin{bmatrix} X & 0 \\ 0 & \gamma^2 I \end{bmatrix}$$

holds. Premultiplying and postmultiplying the above inequality by

$$\begin{bmatrix} \frac{1}{\sqrt{\gamma}} I & 0 \\ 0 & \frac{1}{\gamma} I \end{bmatrix},$$

respectively, and letting $\tilde{X} = \frac{1}{\gamma} X$, we obtain

$$\begin{bmatrix} A(k)^T \tilde{X} A(k) + \frac{1}{\gamma} C(k)^T C(k) & \frac{1}{\sqrt{\gamma}} A(k)^T \tilde{X} B(k) + \frac{1}{\gamma\sqrt{\gamma}} C(k)^T D(k) \\ \frac{1}{\sqrt{\gamma}} B(k)^T \tilde{X} A(k) + \frac{1}{\gamma\sqrt{\gamma}} D(k)^T C(k) & \frac{1}{\gamma} B(k)^T \tilde{X} B(k) + \frac{1}{\gamma^2} D(k)^T D(k) \end{bmatrix} + \sigma_1^2 \begin{bmatrix} A_1^T \tilde{X} A_1 + \frac{1}{\gamma} C_1^T C_1 & \frac{1}{\sqrt{\gamma}} A_1^T \tilde{X} B_1 + \frac{1}{\gamma\sqrt{\gamma}} C_1^T D_1 \\ \frac{1}{\sqrt{\gamma}} B_1^T \tilde{X} A_1 + \frac{1}{\gamma\sqrt{\gamma}} D_1^T C_1 & \frac{1}{\gamma} B_1^T \tilde{X} B_1 + \frac{1}{\gamma^2} D_1^T D_1 \end{bmatrix} < \begin{bmatrix} \tilde{X} & 0 \\ 0 & I \end{bmatrix},$$

or equivalently,

$$\begin{bmatrix} A(k) & \frac{1}{\sqrt{\gamma}}B(k) \\ \frac{1}{\sqrt{\gamma}}C(k) & \frac{1}{\gamma}D(k) \end{bmatrix}^T \begin{bmatrix} \tilde{X} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A(k) & \frac{1}{\sqrt{\gamma}}B(k) \\ \frac{1}{\sqrt{\gamma}}C(k) & \frac{1}{\gamma}D(k) \end{bmatrix}$$
$$+\sigma_1^2 \begin{bmatrix} A_1 & \frac{1}{\sqrt{\gamma}}B_1 \\ \frac{1}{\sqrt{\gamma}}C_1 & \frac{1}{\gamma}D_1 \end{bmatrix}^T \begin{bmatrix} \tilde{X} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_1 & \frac{1}{\sqrt{\gamma}}B_1 \\ \frac{1}{\sqrt{\gamma}}C_1 & \frac{1}{\gamma}D_1 \end{bmatrix} < \begin{bmatrix} \tilde{X} & 0 \\ 0 & I \end{bmatrix}.$$

Premultiplying and postmultiplying both sides of the above inequality by

$$\begin{bmatrix} \tilde{X}^{-\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix}$$

and letting

$$\begin{bmatrix} \bar{A}(k) & \bar{B}(k) \\ \bar{C}(k) & \bar{D}(k) \end{bmatrix} = \begin{bmatrix} \tilde{X}^{\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A(k) & B(k) \\ C(k) & D(k) \end{bmatrix} \begin{bmatrix} \tilde{X}^{-\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix},$$

we have

$$\begin{bmatrix} \bar{A}_1 & \bar{B}_1 \\ \bar{C}_1 & \bar{D}_1 \end{bmatrix} = \begin{bmatrix} \tilde{X}^{\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} \tilde{X}^{-\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix}.$$

It can be seen that (33) is one realization of S, and (34) is strictly contractive. ∎

LEMMA 6.3. *Given a balanced realization* (14) *with* (22)–(25), *if* $\Sigma_2 = I$, *the matrix set*

$$\left\{ \begin{bmatrix} \Sigma_1^{-\frac{1}{2}}A_{12}(k) & \Sigma_1^{-\frac{1}{2}}A_{11}(k)\Sigma_1^{\frac{1}{2}} & \Sigma_1^{-\frac{1}{2}}\tilde{B}_1(k) \\ A_{22}(k) & A_{21}(k)\Sigma_1^{\frac{1}{2}} & \tilde{B}_2(k) \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} \sigma_1\Sigma_1^{-\frac{1}{2}}A_{12}^1 & \sigma_1\Sigma_1^{-\frac{1}{2}}A_{11}^1\Sigma_1^{\frac{1}{2}} & \sigma_1\Sigma_1^{-\frac{1}{2}}\tilde{B}_1^1 \\ \sigma_1 A_{22}^1 & \sigma_1 A_{21}^1\Sigma_1^{\frac{1}{2}} & \sigma_1\tilde{B}_2^1 \end{bmatrix} \right\} \tag{35}$$

and

$$\left\{ \begin{bmatrix} A_{21}(k)\Sigma_1^{-\frac{1}{2}} & A_{22}(k) \\ \Sigma_1^{\frac{1}{2}}A_{11}(k)\Sigma_1^{-\frac{1}{2}} & \Sigma_1^{\frac{1}{2}}A_{12}(k) \\ \tilde{C}_1(k)\Sigma_1^{-\frac{1}{2}} & \tilde{C}_2(k) \end{bmatrix}, \begin{bmatrix} \sigma_1 A_{21}^1\Sigma_1^{-\frac{1}{2}} & \sigma_1 A_{22}^1 \\ \sigma_1\Sigma_1^{\frac{1}{2}}A_{11}^1\Sigma_1^{-\frac{1}{2}} & \sigma_1\Sigma_1^{\frac{1}{2}}A_{12}^1 \\ \sigma_1\tilde{C}_1^1\Sigma_1^{-\frac{1}{2}} & \sigma_1\tilde{C}_2^1 \end{bmatrix} \right\} \tag{36}$$

are strictly contractive.

Proof. By the Schur complements, (21) is equivalent to

$$\begin{bmatrix} -\Sigma^{-1} & 0 & A_b(k)^T & \sigma_1 A_{b1}^T \\ 0 & -I & B_b(k)^T & \sigma_1 B_{b1}^T \\ A_b(k) & B_b(k) & -\Sigma & 0 \\ \sigma_1 A_{b1} & \sigma_1 B_{b1} & 0 & -\Sigma \end{bmatrix} < 0,$$

or

$$\begin{bmatrix} -\Sigma^{-1} & 0 \\ 0 & -I \end{bmatrix} + \begin{bmatrix} A_b(k)^T & \sigma_1 A_{b1}^T \\ B_b(k)^T & \sigma_1 B_{b1}^T \end{bmatrix} \begin{bmatrix} -\Sigma^{-1} & 0 \\ 0 & -\Sigma^{-1} \end{bmatrix} \begin{bmatrix} A_b(k) & B_b(k) \\ \sigma_1 A_{b1} & \sigma_1 B_{b1} \end{bmatrix} < 0,$$

or

$$\begin{bmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} A_b(k)^T \\ B_b(k)^T \end{bmatrix} \Sigma^{-1} \begin{bmatrix} A_b(k) & B_b(k) \end{bmatrix} + \begin{bmatrix} \sigma_1 A_{b1}^T \\ \sigma_1 B_{b1}^T \end{bmatrix} \Sigma^{-1} \begin{bmatrix} \sigma_1 A_{b1} & \sigma_1 B_{b1} \end{bmatrix} \right) \begin{bmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix} < I,$$

which implies the strict contractiveness of

$$\left\{ \Sigma^{-\frac{1}{2}} \begin{bmatrix} A_b(k) & B_b(k) \end{bmatrix} \begin{bmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix}, \Sigma^{-\frac{1}{2}} \begin{bmatrix} \sigma_1 A_{b1} & \sigma_1 B_{b1} \end{bmatrix} \begin{bmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & I \end{bmatrix} \right\}. \tag{37}$$

On the other hand, (20) gives

$$\Sigma^{-\frac{1}{2}} A_b(k)^T \Sigma A_b(k) \Sigma^{-\frac{1}{2}} + \Sigma^{-\frac{1}{2}} C_b(k)^T C_b(k) \Sigma^{-\frac{1}{2}} + \Sigma^{-\frac{1}{2}} \sigma_1^2 (A_{b1}^T \Sigma A_{b1} + C_{b1}^T C_{b1}) \Sigma^{-\frac{1}{2}} < I$$

or

$$\Sigma^{-\frac{1}{2}} \left(\begin{bmatrix} A_b(k)^T & C_b(k)^T \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_b(k) \\ C_b(k) \end{bmatrix} + \sigma_1^2 \begin{bmatrix} A_{b1}{}^T & C_{b1}{}^T \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{b1} \\ C_{b1} \end{bmatrix} \right) \Sigma^{-\frac{1}{2}} < I.$$

This implies the strict contractiveness of

$$\left\{ \begin{bmatrix} \Sigma^{1/2} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_b(k) \\ C_b(k) \end{bmatrix} \Sigma^{-\frac{1}{2}}, \sigma_1 \begin{bmatrix} \Sigma^{1/2} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{b1} \\ C_{b1} \end{bmatrix} \Sigma^{-\frac{1}{2}} \right\}. \tag{38}$$

Substituting (22)–(25) into (37) and (38) and permuting the rows of the resulting matrices, we can see the strict contractiveness of (35) and (36), respectively. □

LEMMA 6.4. *Suppose that the two matrix set*

$$\left\{ \begin{bmatrix} U_{11} & U_{12} \\ Z & U_{22} \end{bmatrix}, \begin{bmatrix} U_{11}^1 & U_{12}^1 \\ Z^1 & U_{22}^1 \end{bmatrix} \right\}$$

and

$$\left\{ \begin{bmatrix} V_{11} & Z \\ V_{21} & V_{22} \end{bmatrix}, \begin{bmatrix} V_{11}^1 & Z^1 \\ V_{21}^1 & V_{22}^1 \end{bmatrix} \right\}$$

are strictly contractive. Then

$$\{K_1, K_2\} := \left\{ \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} U_{11} & U_{12} \\ \frac{1}{\sqrt{2}} V_{11} & Z & \frac{1}{\sqrt{2}} U_{22} \\ V_{21} & \frac{1}{\sqrt{2}} V_{22} & 0 \end{bmatrix}, \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} U_{11}^1 & U_{12}^1 \\ \frac{1}{\sqrt{2}} V_{11}^1 & Z^1 & \frac{1}{\sqrt{2}} U_{22}^1 \\ V_{21}^1 & \frac{1}{\sqrt{2}} V_{22}^1 & 0 \end{bmatrix} \right\} \tag{39}$$

is strictly contractive.

Proof. Dilate the matrix set in (39) as

$$\left\{ \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} U_{11} & U_{12} & \frac{1}{\sqrt{2}} U_{11} \\ \frac{1}{\sqrt{2}} V_{11} & Z & \frac{1}{\sqrt{2}} U_{22} & 0 \\ V_{21} & \frac{1}{\sqrt{2}} V_{22} & 0 & -\frac{1}{\sqrt{2}} V_{22} \\ \frac{1}{\sqrt{2}} V_{11} & 0 & -\frac{1}{\sqrt{2}} U_{22} & -Z \end{bmatrix}, \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} U_{11}^1 & U_{12}^1 & \frac{1}{\sqrt{2}} U_{11}^1 \\ \frac{1}{\sqrt{2}} V_{11}^1 & Z^1 & \frac{1}{\sqrt{2}} U_{22}^1 & 0 \\ V_{21}^1 & \frac{1}{\sqrt{2}} V_{22}^1 & 0 & -\frac{1}{\sqrt{2}} V_{22}^1 \\ \frac{1}{\sqrt{2}} V_{11}^1 & 0 & -\frac{1}{\sqrt{2}} U_{22}^1 & -Z^1 \end{bmatrix} \right\},$$

and it can be verified that the resulting matrix set is strictly contractive. Thus, (39) is strictly contractive. □

Proof of Theorem 3.4. Without loss of generality, we assume that $\Sigma_2 = 1$; in this case, we need to show that

$$\|\mathbb{L}_e\| < 2. \tag{40}$$

The general case follows from scaling and applying this result recursively. It is noticed that one realization of the error system $S - S_r$ is given by

$$\begin{bmatrix} \begin{bmatrix} \begin{bmatrix} A_{11}(k) & 0 \\ 0 & A_b(k) \end{bmatrix} & \begin{bmatrix} \tilde{B}_1(k) \\ B_b(k) \end{bmatrix} \\ \begin{bmatrix} -\tilde{C}_1(k) & C_b(k) \end{bmatrix} & 0 \end{bmatrix} \\ \begin{bmatrix} \begin{bmatrix} A_{11}^1 & 0 \\ 0 & A_b^1 \end{bmatrix} & \begin{bmatrix} \tilde{B}_1^1 \\ B_b^1 \end{bmatrix} \\ \begin{bmatrix} -\tilde{C}_1^1 & C_b^1 \end{bmatrix} & 0 \end{bmatrix} \end{bmatrix}.$$

Letting T be an equivalent transformation, which is given by

$$T = \begin{bmatrix} -\frac{1}{\sqrt{2}}\Sigma_1^{1/2} & \frac{1}{\sqrt{2}}\Sigma_1^{1/2} & 0 \\ 0 & 0 & I \\ \frac{1}{\sqrt{2}}\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\Sigma_1^{-1/2} & 0 \end{bmatrix},$$

the resulting equivalent realization is

$$\begin{bmatrix} \begin{bmatrix} \begin{bmatrix} \Sigma_1^{1/2}A_{11}(k)\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\Sigma_1^{1/2}A_{12}(k) & 0 \\ \frac{1}{\sqrt{2}}A_{21}(k)\Sigma_1^{-1/2} & A_{22}(k) & \frac{1}{\sqrt{2}}A_{21}(k)\Sigma_1^{1/2} \\ 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-1/2}A_{12}(k) & \Sigma_1^{-1/2}A_{11}(k)\Sigma_1^{1/2} \end{bmatrix} & \begin{bmatrix} 0 \\ \tilde{B}_2(k) \\ \frac{2}{\sqrt{2}}\Sigma_1^{-1/2}\tilde{B}_1(k) \end{bmatrix} \\ \begin{bmatrix} \frac{2}{\sqrt{2}}\tilde{C}_1\Sigma_1^{-1/2} & \tilde{C}_2 & 0 \end{bmatrix} & 0 \end{bmatrix} \\ \begin{bmatrix} \begin{bmatrix} \Sigma_1^{1/2}A_{11}^1\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\Sigma_1^{1/2}A_{12}^1 & 0 \\ \frac{1}{\sqrt{2}}A_{21}^1\Sigma_1^{-1/2} & A_{22}^1 & \frac{1}{\sqrt{2}}A_{21}^1\Sigma_1^{1/2} \\ 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-1/2}A_{12}^1 & \Sigma_1^{-1/2}A_{11}^1\Sigma_1^{1/2} \end{bmatrix} & \begin{bmatrix} 0 \\ \tilde{B}_2^1 \\ \frac{2}{\sqrt{2}}\Sigma_1^{-1/2}\tilde{B}_1^1 \end{bmatrix} \\ \begin{bmatrix} \frac{2}{\sqrt{2}}\tilde{C}_1^1\Sigma_1^{-1/2} & \tilde{C}_2^1 & 0 \end{bmatrix} & 0 \end{bmatrix} \end{bmatrix}.$$

By Lemma 6.2, if we can show that the matrix set

$$\{L_1, L_2\} := \left\{ \begin{bmatrix} \Sigma_1^{1/2}A_{11}(k)\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\Sigma_1^{1/2}A_{12}(k) & 0 & 0 \\ \frac{1}{\sqrt{2}}A_{21}(k)\Sigma_1^{-1/2} & A_{22}(k) & \frac{1}{\sqrt{2}}A_{21}(k)\Sigma_1^{1/2} & \frac{1}{\sqrt{2}}\tilde{B}_2(k) \\ 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-1/2}A_{12}(k) & \Sigma_1^{-1/2}A_{11}(k)\Sigma_1^{1/2} & \Sigma_1^{-1/2}\tilde{B}_1(k) \\ \tilde{C}_1(k)\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\tilde{C}_2(k) & 0 & 0 \end{bmatrix}, \right.$$
$$\left. \sigma_1 \begin{bmatrix} \Sigma_1^{1/2}A_{11}^1\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\Sigma_1^{1/2}A_{12}^1 & 0 & 0 \\ \frac{1}{\sqrt{2}}A_{21}^1\Sigma_1^{-1/2} & A_{22}^1 & \frac{1}{\sqrt{2}}A_{21}^1\Sigma_1^{1/2} & \frac{1}{\sqrt{2}}\tilde{B}_2^1 \\ 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-1/2}A_{12}^1 & \Sigma_1^{-1/2}A_{11}^1\Sigma_1^{1/2} & \Sigma_1^{-1/2}\tilde{B}_1^1 \\ \tilde{C}_1^1\Sigma_1^{-1/2} & \frac{1}{\sqrt{2}}\tilde{C}_2^1 & 0 & 0 \end{bmatrix} \right\}$$

is strictly contractive, then the proof is obtained. By Lemmas 6.3 and 6.4, it can be

seen that the matrix set

$$\{\tilde{L}_1, \tilde{L}_2\} := \left\{ \begin{bmatrix} 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-\frac{1}{2}} A_{12}(k) & \Sigma_1^{-1/2} A_{11}(k) \Sigma_1^{1/2} & \Sigma_1^{-1/2} \tilde{B}_1 \\ \frac{1}{\sqrt{2}} A_{21}(k) \Sigma_1^{-\frac{1}{2}} & A_{22}(k) & \frac{1}{\sqrt{2}} A_{21}(k) \Sigma_1^{1/2} & \frac{1}{\sqrt{2}} \tilde{B}_2(k) \\ \Sigma_1^{1/2} A_{11}(k) \Sigma_1^{-1/2} & \frac{1}{\sqrt{2}} \Sigma_1^{1/2} A_{12}(k) & 0 & 0 \\ \tilde{C}_1(k) \Sigma_1^{-1/2} & \frac{1}{\sqrt{2}} \tilde{C}_2(k) & 0 & 0 \end{bmatrix}, \right.$$
$$\left. \sigma_1 \begin{bmatrix} 0 & \frac{1}{\sqrt{2}}\Sigma_1^{-1/2} A_{12}^1 & \Sigma_1^{-1/2} A_{11}^1 \Sigma_1^{1/2} & \Sigma_1^{-1/2} \tilde{B}_1^1 \\ \frac{1}{\sqrt{2}} A_{21}^1 \Sigma_1^{-1/2} & A_{22}^1 & \frac{1}{\sqrt{2}} A_{21}^1 \Sigma_1^{1/2} & \frac{1}{\sqrt{2}} \tilde{B}_2^1 \\ \Sigma_1^{1/2} A_{11}^1 \Sigma_1^{-1/2} & \frac{1}{\sqrt{2}} \Sigma_1^{1/2} A_{12}^1 & 0 & 0 \\ \tilde{C}_1^1 \Sigma_1^{-1/2} & \frac{1}{\sqrt{2}} \tilde{C}_2^1 & 0 & 0 \end{bmatrix} \right\}$$

is strictly contractive. By permuting the rows of $\tilde{L}_1, \tilde{L}_2$, the strict contractiveness of $\{L_1, L_2\}$ is obtained. □

REFERENCES

[1] C. L. Beck, J. Doyle, and K. Glover, *Model reduction of multidimensional and uncertain systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1466–1477.

[2] S. Boyd, L. E Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.

[3] O. L. V. Costa and C. S. Kubrusly, *State-feedback H_∞-control for discrete-time infinite-dimensional stochastic bilinear systems*, J. Math. Systems Estim. Control, 6 (1996), pp. 1–32.

[4] A. El Bouhtouri, D. Hinrichsen, and A. J. Pritchard, *H^∞-type control for discrete-time stochastic systems*, Internat. J. Robust and Nonlinear Control, 9 (1999), pp. 923–948.

[5] E. Gershon, Shaked U., and I. Yaesh, *H_∞ control and filtering of discrete-time stochastic systems with multiplicative noise*, Automatica J. IFAC, 37 (2001), pp. 409–417.

[6] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

[7] K. Glover, P. J. Goddard, and Y. C. Chu, *Model reduction for classes of uncertain, multi-dimensional, parameter varying and non-linear systems*, in Learning, Control and Hybrid Systems, Springer, London, 1999, pp. 269–280.

[8] P. J. Goddard, *Performance-Preserving Controller Approximation*, Ph.D. thesis, Cambridge University, Cambridge, UK, 1995.

[9] D. Hinrichsen and A. J. Pritchard, *Stability margins for systems with deterministic and stochastic uncertainty*, in Proceedings of the 33rd IEEE Conference on Decision and Control, 1994, pp. 3825–3826.

[10] D. Hinrichsen and A. J. Pritchard, *Stochastic H^∞*, SIAM J. Control Optim., 36 (1998), pp. 1504–1538.

[11] J. Lam and Y. S. Hung, *On the stability of projections of balanced realizations*, Linear Algebra Appl., 257 (1997), pp. 163–182.

[12] C. Lu, J. S. Tsai, G. Jong, and T. Su, *An LMI-based approach for robust stabilization of uncertain stochastic systems with time-varying delays*, IEEE Trans. Automat. Control, 48 (2003), pp. 286–289.

[13] M. S. Mahmoud and L. Xie, *Guaranteed cost control of uncertain discrete systems with delays*, Internat. J. Control, 73 (2000), pp. 105–114.

[14] X. Mao, N. Koroleva, and A. Rodkina, *Robust stability of uncertain stochastic differential delay equations*, Systems Control Lett., 33 (1998), pp. 325–336.

[15] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.

[16] L. Pernebo and L. M. Silverman, *Model reduction via balanced state space representations*, IEEE Trans. Automat. Control, 27 (1982), pp. 382–387.

[17] I. R. Petersen, M. R. James, and P. Dupuis, *Minmax optimal control of stochastic uncertain systems with relative entropy constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 398–412.

[18] V. A. Ugrinovskii and I. R. Petersen, *Absolute stabilization and minimax optimal control of uncertain systems with stochastic uncertainty*, SIAM J. Control Optim., 37 (1999), pp. 1089–1122.

[19] Z. Wang, H. Qiao, and K. J. Burnham, *On stabilization of bilinear uncertain stochastic time-delay stochastic systems with Markovian jumping parameters*, IEEE Trans. Automat. Control, 47 (2002), pp. 640–646.

[20] S. Xu and T. Chen, *Robust H_∞ control for uncertain stochastic systems with state delay*, IEEE Trans. Automat. Control, 47 (2002).

[21] S. Xu and T. Chen, *H_∞ model reduction in the stochastic framework*, SIAM J. Control Optim., 42 (2003), pp. 1293–1309.

[22] S. Xu and T. Chen, *Robust H_∞ filtering for uncertain impulsive stochastic systems under sampled measurements*, Automatica J. IFAC, 39 (2003), pp. 509–516.

[23] F. Yang, Z. Wang, and Y. S. Hung, *Robust Kalman filtering for discrete time-varying uncertain systems with multiplicative noises*, IEEE Trans. Automat. Control, 47 (2002), pp. 1179–1183.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1561–1585

STABILITY AND INSTABILITY RESULTS OF THE WAVE EQUATION WITH A DELAY TERM IN THE BOUNDARY OR INTERNAL FEEDBACKS*

SERGE NICAISE† AND CRISTINA PIGNOTTI‡

Abstract. In this paper we consider, in a bounded and smooth domain, the wave equation with a delay term in the boundary condition. We also consider the wave equation with a delayed velocity term and mixed Dirichlet–Neumann boundary condition. In both cases, under suitable assumptions, we prove exponential stability of the solution. These results are obtained by introducing suitable energies and by using some observability inequalities. If one of the above assumptions is not satisfied, some instability results are also given by constructing some sequences of delays for which the energy of some solutions does not tend to zero.

Key words. wave equation, delay feedbacks, stabilization

AMS subject classifications. 35L05, 93D15

DOI. 10.1137/060648891

1. Introduction.

We investigate the effect of time delay in boundary or internal stabilization of the wave equation in domains of $\mathbb{R}^n$. Such effects arise in many practical problems, and it is well known, at least in one dimension, that they can induce some instabilities; see [4, 5, 6, 17].

To our knowledge, analysis in higher dimensions has not yet been done. In this paper, we give some stability results under a sufficient condition, and we further show that if this condition is not satisfied, then there exist some delays for which the system is destabilized. So, in a certain sense, our sufficient condition is also necessary in order to have a general stability result.

Let $\Omega \subset \mathbb{R}^n$ be an open bounded set with a boundary Γ of class C^2. We assume that Γ is divided into two parts Γ_D and Γ_N, i.e., $\Gamma = \Gamma_D \cup \Gamma_N$, with $\overline{\Gamma}_D \cap \overline{\Gamma}_N = \emptyset$ and $\Gamma_D \neq \emptyset$.

In this domain Ω, we consider the initial boundary value problem

$$u_{tt}(x,t) - \Delta u(x,t) = 0 \quad \text{in} \quad \Omega \times (0,+\infty), \tag{1.1}$$

$$u(x,t) = 0 \quad \text{on} \quad \Gamma_D \times (0,+\infty), \tag{1.2}$$

$$\frac{\partial u}{\partial \nu}(x,t) = -\mu_1 u_t(x,t) - \mu_2 u_t(x,t-\tau) \quad \text{on} \quad \Gamma_N \times (0,+\infty), \tag{1.3}$$

$$u(x,0) = u_0(x) \quad \text{and} \quad u_t(x,0) = u_1(x) \quad \text{in} \quad \Omega, \tag{1.4}$$

$$u_t(x,t-\tau) = f_0(x,t-\tau) \quad \text{in } \Gamma_N \times (0,\tau), \tag{1.5}$$

where $\nu(x)$ denotes the outer unit normal vector to the point $x \in \Gamma$ and $\frac{\partial u}{\partial \nu}$ is the normal derivative. Moreover, $\tau > 0$ is the time delay, μ_1 and μ_2 are positive real numbers, and the initial data (u_0, u_1, f_0) belong to a suitable space.

We are interested in giving an exponential stability result for such a problem.

*Received by the editors January 3, 2006; accepted for publication (in revised form) May 24, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/sicon/45-5/64889.html

†Université de Valenciennes et du Hainaut Cambrésis, MACS, Institut des Sciences et Techniques de Valenciennes, 59313 Valenciennes Cedex 9, France (snicaise@univ-valenciennes.fr).

‡Dipartimento di Matematica Pura e Applicata, Università di L'Aquila, Via Vetoio, Loc. Coppito, 67010 L'Aquila, Italy (pignotti@univaq.it).

Let us denote by $\langle v, w\rangle$ or, equivalently, by $v \cdot w$ the Euclidean inner product between two vectors $v, w \in \mathbb{R}^n$.

We assume that there exists a scalar function $v \in C^2(\overline{\Omega})$ such that

(i) v is strictly convex in $\overline{\Omega}$; that is, there exists $\alpha > 0$ such that

$$\langle D^2(v)(x)\xi, \xi\rangle \geq 2\alpha|\xi|^2 \quad \forall x \in \overline{\Omega},\ \forall \xi \in \mathbb{R}^n, \tag{1.6}$$

where $D^2(v)$ denotes the Hessian matrix of v;

(ii) the vector field $H := \nabla v$ verifies

$$H(x) \cdot \nu(x) \leq 0 \quad \forall x \in \Gamma_D. \tag{1.7}$$

For the above assumptions see [14], where some observability estimates for second-order hyperbolic equations are given.

It is well known that if $\mu_2 = 0$, that is, in absence of delay, the energy of problem (1.1)–(1.5) is exponentially decaying to zero. See for instance Chen [2, 3], Lagnese [11, 12], Lasiecka and Triggiani [13], Komornik and Zuazua [10], and Komornik [8, 9]. On the contrary, if $\mu_1 = 0$, that is, if we have only the delay part in the boundary condition on Γ_N, system (1.1)–(1.5) becomes unstable. See, for instance Datko, Lagnese, and Polis [6].

Although these examples involve only one space dimension, we can expect a similar phenomenon to occur in higher space dimensions. So, it is interesting to seek a stabilization result when both μ_1 and μ_2 are nonzero. In this case, the boundary feedback is composed of two parts and only one of them has a delay.

This problem has been studied in one space dimension by Xu, Yung, and Li [17]. After a spectral analysis these authors proved a stability result for the case when $\mu_2 < \mu_1$. In their paper it is also shown that if $\mu_2 > \mu_1$, the system is unstable and if $\mu_1 = \mu_2$, some instabilities may occur.

Here, in agreement with [17] and assuming that

$$\mu_2 < \mu_1, \tag{1.8}$$

we obtain a stabilization result in a general space dimension by using a suitable observability estimate. This is done by applying inequalities obtained from Carleman estimates for the wave equation by Lasiecka, Triggiani, and Yao in [14] and by using compactness-uniqueness arguments. If $\mu_1 = \mu_2$, we further show that there exists a sequence of arbitrary small (and large) delays such that instabilities occur. In the case $\mu_2 > \mu_1$, we also obtain delays which destabilize the system. More precisely, we show the next results.

Under assumption (1.8) let us define the energy of a solution of problem (1.1)–(1.5) as

$$E(t) := \frac{1}{2}\int_\Omega \{u_t^2(x,t) + |\nabla u(x,t)|^2\}dx + \frac{\xi}{2}\int_{\Gamma_N}\int_0^1 u_t^2(x, t-\tau\rho)d\rho d\Gamma, \tag{1.9}$$

where ξ is a positive constant verifying

$$\tau\mu_2 < \xi < \tau(2\mu_1 - \mu_2). \tag{1.10}$$

Clearly this energy is larger than the standard energy

$$\frac{1}{2}\int_\Omega \{u_t^2(x,t) + |\nabla u(x,t)|^2\}dx$$

and contains an additional term that comes from the delay term.

THEOREM 1.1. *Assume that* (1.8) *holds. There exist positive constants* C_1, C_2 *such that, for any solution of problem* (1.1)–(1.5),

$$E(t) \le C_1 E(0) e^{-C_2 t} \qquad \forall t \ge 0. \tag{1.11}$$

THEOREM 1.2. *If* (1.8) *does not hold, there exist a sequence of delays, and solutions of problem* (1.1)–(1.5) *corresponding to these delays, such that their standard energy is constant.*

In this paper we also study the problem for the wave equation with internal feedback. In particular, we consider the system

$$u_{tt}(x,t) - \Delta u(x,t) + a(x)[\mu_1 u_t(x,t) + \mu_2 u_t(x,t-\tau)] = 0 \ \text{ in } \Omega \times (0,+\infty), \tag{1.12}$$

$$u(x,t) = 0 \quad \text{on} \quad \Gamma_D \times (0,+\infty), \tag{1.13}$$

$$\frac{\partial u}{\partial \nu}(x,t) = 0 \quad \text{on} \quad \Gamma_N \times (0,+\infty), \tag{1.14}$$

$$u(x,0) = u_0(x) \quad \text{and} \quad u_t(x,0) = u_1(x) \quad \text{in} \quad \Omega, \tag{1.15}$$

$$u_t(x,t-\tau) = g_0(x,t-\tau) \quad \text{in } \Omega \times (0,\tau), \tag{1.16}$$

where $a \in L^\infty(\Omega)$ is a function such that

$$a(x) \ge 0 \quad \text{a. e.} \quad \text{in} \quad \Omega, \tag{1.17}$$

and

$$a(x) > a_0 > 0 \quad \text{a. e.} \quad \text{in} \quad \omega, \tag{1.18}$$

where $\omega \subset \Omega$ is an open neighborhood of Γ_N.

Exponential stability results for the above problem in the case of $\mu_2 = 0$, that is, without delay, have been obtained by several authors. See for instance Zuazua [18] and Liu [16].

On the contrary, at least for the one-dimensional case, Datko [4] has shown that the wave equation with a velocity term and mixed Dirichlet–Neumann boundary condition is destabilized by a time delay in the velocity term.

In this paper, in the case $\mu_2 < \mu_1$, we show that the energy is exponentially decaying to zero. This is done, as for the problem with boundary feedback, by using a suitable observability estimate. If $\mu_2 \ge \mu_1$, we obtain an explicit sequence of arbitrary small delays that destabilize the system.

As before, under assumption (1.8) let us define the energy of a solution of (1.12)–(1.16) as

$$\mathcal{F}(t) := \frac{1}{2}\int_\Omega \{u_t^2(x,t) + |\nabla u(x,t)|^2\}dx + \frac{\xi}{2}\int_\Omega a(x)\int_0^1 u_t^2(x,t-\tau\rho)d\rho dx, \tag{1.19}$$

where ξ is a positive constant verifying (1.10). Again $\mathcal{F}$ is larger than the standard energy and contains an extra term due to the delay.

THEOREM 1.3. *Let assumption* (1.8) *be satisfied. Then there exist positive constants* C_1, C_2 *such that, for any solution of problem* (1.12)–(1.16),

$$\mathcal{F}(t) \le C_1 \mathcal{F}(0) e^{-C_2 t} \qquad \forall t \ge 0. \tag{1.20}$$

THEOREM 1.4. *If* (1.8) *does not hold, there exist a sequence of arbitrary small (or large) delays, and solutions of problem* (1.12)–(1.16) *corresponding to these delays, such that their standard energy does not tend to* 0.

Remark 1.5. In [14], in order to deal with variable coefficients, the authors assume that there exists a scalar function v strictly convex with respect to the Riemannian metric induced by the spatial operator. Here, we are principally interested in the effect of the delay term in the boundary or internal feedback. So, in order to avoid technicalities, we consider constant coefficients. Actually, our stability results hold even for variable coefficients under the assumption of [14].

The paper is organized as follows. Well-posedness of the problems is analyzed in section 2 using semigroup theory. In subsection 2.1 we study the well-posedness of problem (1.1)–(1.5), while in subsection 2.2 we concentrate on problem (1.12)–(1.16). In sections 3 and 4 we prove the exponential stability of the problem with boundary and internal feedbacks, respectively. Finally, section 5 is devoted to some instability examples.

2. Well-posedness of the problems. In this section we will give well-posedness results for problem (1.1)–(1.5) and problem (1.12)–(1.16) using semigroup theory.

2.1. Boundary feedback. Let us set

$$z(x,\rho,t) = u_t(x,t-\tau\rho), \quad x \in \Gamma_N,\ \rho \in (0,1),\ t>0. \tag{2.1}$$

Then, problem (1.1)–(1.5) is equivalent to

$$u_{tt}(x,t) - \Delta u(x,t) = 0 \quad \text{in} \quad \Omega \times (0,+\infty), \tag{2.2}$$

$$\tau z_t(x,\rho,t) + z_\rho(x,\rho,t) = 0 \quad \text{in} \quad \Gamma_N \times (0,1) \times (0,+\infty), \tag{2.3}$$

$$u(x,t) = 0 \quad \text{on} \quad \Gamma_D \times (0,+\infty), \tag{2.4}$$

$$\frac{\partial u}{\partial \nu}(x,t) = -\mu_1 u_t(x,t) - \mu_2 z(x,1,t) \quad \text{on} \quad \Gamma_N \times (0,+\infty), \tag{2.5}$$

$$z(x,0,t) = u_t(x,t) \quad \text{on} \quad \Gamma_N \times (0,\infty), \tag{2.6}$$

$$u(x,0) = u_0(x) \quad \text{and} \quad u_t(x,0) = u_1(x) \quad \text{in} \quad \Omega, \tag{2.7}$$

$$z(x,\rho,0) = f_0(x,-\rho\tau) \quad \text{in} \quad \Gamma_N \times (0,1). \tag{2.8}$$

If we denote

$$U := (u, u_t, z)^T,$$

then

$$U' := (u_t, u_{tt}, z_t)^T = \left(u_t, \Delta u, -\tau^{-1} z_\rho\right)^T.$$

Therefore, problem (2.2)–(2.8) can be rewritten as

$$\begin{cases} U' = \mathcal{A}U, \\ U(0) = (u_0, u_1, f_0(\cdot, -\cdot\,\tau))^T, \end{cases} \tag{2.9}$$

where the operator $\mathcal{A}$ is defined by

$$\mathcal{A}\begin{pmatrix} u \\ v \\ z \end{pmatrix} := \begin{pmatrix} v \\ \Delta u \\ -\tau^{-1} z_\rho \end{pmatrix},$$

with domain

(2.10)
$$\mathcal{D}(\mathcal{A}) := \Big\{ (u,v,z)^T \in \big(E(\Delta, L^2(\Omega)) \cap H^1_{\Gamma_D}(\Omega)\big) \times H^1(\Omega) \times L^2(\Gamma_N; H^1(0,1)) \ : \ \frac{\partial u}{\partial \nu} = -\mu_1 v - \mu_2 z(\cdot,1) \text{ on } \Gamma_N;\ v = z(\cdot,0) \text{ on } \Gamma_N \Big\},$$

where, as usual,

$$H^1_{\Gamma_D}(\Omega) = \{ u \in H^1(\Omega) \ : \ u = 0 \quad \text{on} \ \ \Gamma_D \},$$

and

$$E(\Delta, L^2(\Omega)) = \{u \in H^1(\Omega) : \Delta u \in L^2(\Omega)\}.$$

Recall that for a function $u \in E(\Delta, L^2(\Omega))$, $\frac{\partial u}{\partial \nu}$ belongs to $H^{-1/2}(\Gamma_N)$ and the next Green formula is valid (see section 1.5 of [7])

$$\int_\Omega \nabla u \nabla w dx = -\int_\Omega \Delta u w dx + \left\langle \frac{\partial u}{\partial \nu}; w \right\rangle_{\Gamma_N} \quad \forall w \in H^1_{\Gamma_D}(\Omega), \tag{2.11}$$

where $\langle \cdot; \cdot \rangle_{\Gamma_N}$ means the duality pairing between $H^{-1/2}(\Gamma_N)$ and $H^{1/2}(\Gamma_N)$.

Note further that for $(u,v,z)^T \in \mathcal{D}(\mathcal{A})$, $\frac{\partial u}{\partial \nu}$ belongs to $L^2(\Gamma_N)$ since $z(\cdot,1)$ is in $L^2(\Gamma_N)$.

Denote by $\mathcal{H}$ the Hilbert space

$$\mathcal{H} := H^1_{\Gamma_D}(\Omega) \times L^2(\Omega) \times L^2(\Gamma_N \times (0,1)). \tag{2.12}$$

Assuming that

$$\mu_2 \le \mu_1, \tag{2.13}$$

we will show that $\mathcal{A}$ generates a C_0 semigroup on $\mathcal{H}$.

Let ξ be a positive real number such that

$$\tau\mu_2 \le \xi \le \tau(2\mu_1 - \mu_2). \tag{2.14}$$

Note that, from (2.13), such a constant ξ exists.

Let us define on the Hilbert space $\mathcal{H}$ the inner product

(2.15)
$$\left\langle \begin{pmatrix} u \\ v \\ z \end{pmatrix}, \begin{pmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{z} \end{pmatrix} \right\rangle_{\mathcal{H}} := \int_\Omega \{\nabla u(x) \nabla \tilde{u}(x) + v(x)\tilde{v}(x)\} dx + \xi \int_{\Gamma_N} \int_0^1 z(x,\rho)\tilde{z}(x,\rho) d\rho d\Gamma.$$

THEOREM 2.1. *For any initial datum $U_0 \in \mathcal{H}$ there exists a unique solution $U \in C([0,+\infty), \mathcal{H})$ of problem (2.9). Moreover, if $U_0 \in \mathcal{D}(\mathcal{A})$, then*

$$U \in C([0,+\infty), \mathcal{D}(\mathcal{A})) \cap C^1([0,+\infty), \mathcal{H}).$$

Proof. Take $U = (u,v,z)^T \in \mathcal{D}(\mathcal{A})$. Then

$$\begin{aligned}(\mathcal{A}U, U) &= \left\langle \begin{pmatrix} v \\ \Delta u \\ -\tau^{-1} z_\rho \end{pmatrix}, \begin{pmatrix} u \\ v \\ z \end{pmatrix} \right\rangle_{\mathcal{H}} \\ &= \int_\Omega \{\nabla v(x) \nabla u(x) + v(x) \Delta u(x)\} dx - \xi\tau^{-1} \int_{\Gamma_N} \int_0^1 z_\rho(x,\rho) z(x,\rho) d\rho d\Gamma.\end{aligned}$$

So, by Green's formula,

$$(\mathcal{A}U,U)=\int_{\Gamma_N}\frac{\partial u}{\partial\nu}(x)v(x)d\Gamma-\xi\tau^{-1}\int_{\Gamma_N}\int_0^1 z_\rho(x,\rho)z(x,\rho)d\rho d\Gamma. \tag{2.16}$$

Integrating by parts in ρ, we get

$$\int_{\Gamma_N}\int_0^1 z_\rho(x,\rho)z(x,\rho)d\rho d\Gamma=-\int_{\Gamma_N}\int_0^1 z_\rho(x,\rho)z(x,\rho)d\rho d\Gamma+\int_{\Gamma_N}\{z^2(x,1)-z^2(x,0)\}d\Gamma,$$

that is

$$\int_{\Gamma_N}\int_0^1 z_\rho(x,\rho)z(x,\rho)d\rho d\Gamma=\frac{1}{2}\int_{\Gamma_N}\{z^2(x,1)-z^2(x,0)\}d\Gamma. \tag{2.17}$$

Therefore, from (2.16) and (2.17),

$$\begin{aligned}(\mathcal{A}U,U)&=\int_{\Gamma_N}\frac{\partial u}{\partial\nu}(x)v(x)d\Gamma-\frac{\xi\tau^{-1}}{2}\int_{\Gamma_N}\{z^2(x,1)-z^2(x,0)\}d\Gamma\\&=-\int_{\Gamma_N}(\mu_1v(x)+\mu_2z(x,1))v(x)d\Gamma-\frac{\xi\tau^{-1}}{2}\int_{\Gamma_N}\{z^2(x,1)-z^2(x,0)\}d\Gamma\\&=-\mu_1\int_{\Gamma_N}v^2(x)d\Gamma-\mu_2\int_{\Gamma_N}z(x,1)v(x)d\Gamma-\frac{\xi\tau^{-1}}{2}\int_{\Gamma_N}z^2(x,1)d\Gamma+\frac{\xi\tau^{-1}}{2}\int_{\Gamma_N}v^2(x)d\Gamma,\end{aligned}$$

from which follows, using the Cauchy–Schwarz inequality,

$$(\mathcal{A}U,U)\le\left(-\mu_1+\frac{\mu_2}{2}+\frac{\xi\tau^{-1}}{2}\right)\int_{\Gamma_N}v^2(x)d\Gamma+\left(\frac{\mu_2}{2}-\frac{\xi\tau^{-1}}{2}\right)\int_{\Gamma_N}z^2(x,1)d\Gamma. \tag{2.18}$$

Now, observe that from (2.14),

$$-\mu_1+\frac{\mu_2}{2}+\frac{\xi\tau^{-1}}{2}\le 0,\quad \frac{\mu_2}{2}-\frac{\xi\tau^{-1}}{2}\le 0.$$

Then, $(\mathcal{A}U,U)\le 0$, which means that the operator $\mathcal{A}$ is dissipative.

Now, we will show that $\lambda I-\mathcal{A}$ is surjective for a fixed $\lambda>0$. Given $(f,g,h)^T\in\mathcal{H}$, we seek a $U=(u,v,z)^T\in\mathcal{D}(\mathcal{A})$ solution of

$$(\lambda I-\mathcal{A})\begin{pmatrix}u\\v\\z\end{pmatrix}=\begin{pmatrix}f\\g\\h\end{pmatrix},$$

that is, verifying

$$\begin{cases}\lambda u-v=f,\\ \lambda v-\Delta u=g,\\ \lambda z+\tau^{-1}z_\rho=h.\end{cases} \tag{2.19}$$

Suppose that we have found u with the appropriate regularity. Then,

$$v:=\lambda u-f \tag{2.20}$$

and we can determine z. Indeed, by (2.10),

$$z(x,0) = v(x) \quad \text{for} \quad x \in \Gamma_N, \tag{2.21}$$

and, from (2.19),

$$\lambda z(x,\rho) + \tau^{-1} z_\rho(x,\rho) = h(x,\rho) \quad \text{for} \quad x \in \Gamma_N,\ \rho \in (0,1). \tag{2.22}$$

Then, by (2.21) and (2.22), we obtain

$$z(x,\rho) = v(x)e^{-\lambda\rho\tau} + \tau e^{-\lambda\rho\tau} \int_0^\rho h(x,\sigma)e^{\lambda\sigma\tau} d\sigma.$$

So, from (2.20),

$$z(x,\rho) = \lambda u(x)e^{-\lambda\rho\tau} - f(x)e^{-\lambda\rho\tau} + \tau e^{-\lambda\rho\tau} \int_0^\rho h(x,\sigma)e^{\lambda\sigma\tau} d\sigma \quad \text{on } \Gamma_N \times (0,1), \tag{2.23}$$

and, in particular,

$$z(x,1) = \lambda u(x)e^{-\lambda\tau} + z_0(x), \quad x \in \Gamma_N, \tag{2.24}$$

with $z_0 \in L^2(\Gamma_N)$ defined by

$$z_0(x) = -f(x)e^{-\lambda\tau} + \tau e^{-\lambda\tau} \int_0^1 h(x,\sigma)e^{\lambda\sigma\tau} d\sigma, \quad x \in \Gamma_N. \tag{2.25}$$

By (2.20) and (2.19), the function u verifies

$$\lambda(\lambda u - f) - \Delta u = g,$$

that is,

$$\lambda^2 u - \Delta u = g + \lambda f. \tag{2.26}$$

Problem (2.26) can be reformulated as

$$\int_\Omega (\lambda^2 u - \Delta u) w dx = \int_\Omega (g + \lambda f) w dx \quad \forall w \in H^1_{\Gamma_D}(\Omega). \tag{2.27}$$

Integrating by parts,

$$\begin{aligned}
\int_\Omega (\lambda^2 u - \Delta u) w dx &= \int_\Omega (\lambda^2 uw + \nabla u \nabla w) dx - \int_{\Gamma_N} \frac{\partial u}{\partial \nu} w d\Gamma \\
&= \int_\Omega (\lambda^2 uw + \nabla u \nabla w) dx + \int_{\Gamma_N} (\mu_1 vw + \mu_2 z(x,1)) w d\Gamma \\
&= \int_\Omega (\lambda^2 uw + \nabla u \nabla w) dx + \int_{\Gamma_N} \{\mu_1(\lambda u - f)w + \mu_2(\lambda u e^{-\lambda\tau} + z_0)w\} d\Gamma,
\end{aligned}$$

where we have used (2.20) and (2.24). Therefore, (2.27) can be rewritten as

$$\begin{aligned}
&\int_\Omega (\lambda^2 uw + \nabla u \nabla w) dx + \int_{\Gamma_N} (\mu_1 + \mu_2 e^{-\lambda\tau}) \lambda uw d\Gamma \\
&\qquad = \int_\Omega (g + \lambda f) w dx + \mu_1 \int_{\Gamma_N} f w d\Gamma - \mu_2 \int_{\Gamma_N} z_0 w d\Gamma \quad \forall w \in H^1_{\Gamma_D}(\Omega).
\end{aligned} \tag{2.28}$$

As the left-hand side of (2.28) is coercive on $H^1_{\Gamma_D}(\Omega)$, the Lax–Milgram lemma guarantees the existence and uniqueness of a solution $u \in H^1_{\Gamma_D}(\Omega)$ of (2.28).

If we consider $w \in \mathcal{D}(\Omega)$ in (2.28), we have that u solves in $\mathcal{D}'(\Omega)$

$$\lambda^2 u - \Delta u = g + \lambda f, \tag{2.29}$$

and thus $u \in E(\Delta, L^2(\Omega))$.

Using Green's formula (2.11) in (2.28) and using (2.29), we obtain

$$\int_{\Gamma_N} (\mu_1 + \mu_2 e^{-\lambda\tau})\lambda u w d\Gamma + \left\langle \frac{\partial u}{\partial \nu}; w \right\rangle_{\Gamma_N} = \mu_1 \int_{\Gamma_N} f w d\Gamma - \mu_2 \int_{\Gamma_N} z_0 w d\Gamma,$$

from which follows

$$\frac{\partial u}{\partial \nu} + (\mu_1 + \mu_2 e^{-\lambda\tau})\lambda u = \mu_1 f - \mu_2 z_0 \quad \text{on} \quad \Gamma_N. \tag{2.30}$$

Therefore, from (2.30),

$$\frac{\partial u}{\partial \nu} = -\mu_1 v - \mu_2 z(\cdot, 1) \quad \text{on} \quad \Gamma_N,$$

where we have used (2.20) and (2.24). So, we have found $(u, v, z)^T \in \mathcal{D}(\mathcal{A})$, which verifies (2.19).

Now, the well-posedness result follows from the Hille–Yosida theorem. □

2.2. Internal feedback. Setting

$$z(x, \rho, t) = u_t(x, t - \tau\rho), \quad x \in \Omega, \ \rho \in (0, 1), \ t > 0, \tag{2.31}$$

problem (1.12)–(1.16) is equivalent to

$$u_{tt} - \Delta u + a(x)[\mu_1 u_t(x, t) + \mu_2 u_t(x, t - \tau)] = 0 \quad \text{in} \quad \Omega \times (0, +\infty), \tag{2.32}$$

$$\tau z_t(x, \rho, t) + z_\rho(x, \rho, t) = 0 \quad \text{in} \quad \Omega \times (0, 1) \times (0, +\infty), \tag{2.33}$$

$$u(x, t) = 0 \quad \text{on} \quad \Gamma_D \times (0, +\infty), \tag{2.34}$$

$$\frac{\partial u}{\partial \nu}(x, t) = 0 \quad \text{on} \quad \Gamma_N \times (0, +\infty), \tag{2.35}$$

$$z(x, 0, t) = u_t(x, t) \quad \text{on} \quad \Gamma_N \times (0, +\infty), \tag{2.36}$$

$$u(x, 0) = u_0(x) \quad \text{and} \quad u_t(x, 0) = u_1(x) \quad \text{in} \quad \Omega, \tag{2.37}$$

$$z(x, \rho, 0) = g_0(x, -\rho\tau) \quad \text{in} \quad \Omega \times (0, 1). \tag{2.38}$$

If we denote by

$$U := (u, u_t, z)^T,$$

then

$$U' := (u_t, u_{tt}, z_t)^T = \big(u_t, \Delta u - a(\mu_1 u_t + \mu_2 z(\cdot, 1, \cdot)), -\tau^{-1} z_\rho\big)^T.$$

Therefore, problem (2.32)–(2.38) can be rewritten as

$$\begin{cases} U' = \mathcal{A}^0 U, \\ U(0) = (u_0, u_1, g_0(\cdot, -\cdot\,\tau))^T, \end{cases} \tag{2.39}$$

where the operator $\mathcal{A}^0$ is defined by

$$\mathcal{A}^0 \begin{pmatrix} u \\ v \\ z \end{pmatrix} := \begin{pmatrix} v \\ \Delta u - a\mu_1 v - a\mu_2 z(\cdot,1) \\ -\tau^{-1} z_\rho \end{pmatrix},$$

with domain

$$\mathcal{D}(\mathcal{A}^0) := \Big\{ (u,v,z)^T \in \big(H^2(\Omega) \cap H^1_{\Gamma_D}(\Omega)\big) \times H^1(\Omega) \times L^2(\Omega; H^1(0,1)) \ : \ \frac{\partial u}{\partial \nu} = 0 \text{ on } \Gamma_N;\ v = z(\cdot,0) \text{ in } \Omega \Big\}. \tag{2.40}$$

Denote by $\mathcal{H}^0$ the Hilbert space

$$\mathcal{H}^0 := H^1_{\Gamma_D}(\Omega) \times L^2(\Omega) \times L^2(\Omega \times (0,1)), \tag{2.41}$$

equipped with the inner product

$$\left\langle \begin{pmatrix} u \\ v \\ z \end{pmatrix}, \begin{pmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{z} \end{pmatrix} \right\rangle_{\mathcal{H}^0} := \int_\Omega \{\nabla u(x) \nabla \tilde{u}(x) + v(x)\tilde{v}(x)\} dx + \xi \int_\Omega \int_0^1 z(x,\rho)\tilde{z}(x,\rho) d\rho dx, \tag{2.42}$$

where ξ is a fixed positive number satisfying (2.14).

Arguing analogously to the previous case, we can show that the operator $\mathcal{A}^0$ generates a C_0 semigroup on $\mathcal{H}^0$. Consequently we have the following well-posedness result.

THEOREM 2.2. *For any initial datum $U_0 \in \mathcal{H}^0$ there exists a unique solution $U \in C([0,+\infty), \mathcal{H}^0)$ of problem* (2.39). *Moreover, if $U_0 \in \mathcal{D}(\mathcal{A}^0)$, then*

$$U \in C([0,+\infty), \mathcal{D}(\mathcal{A}^0)) \cap C^1([0,+\infty), \mathcal{H}^0).$$

3. Boundary stability result. In this section, in order to prove an exponential stability result for problem (1.1)–(1.5), we assume (1.8).

Let $E(\cdot)$ be the energy defined by (1.9) and (1.10). We have the following result.

PROPOSITION 3.1. *For any regular solution of problem* (1.1)–(1.5)*, the energy is decreasing and there exists a positive constant C such that*

$$E'(t) \le -C \int_{\Gamma_N} \{u_t^2(x,t) + u_t^2(x,t-\tau)\} d\Gamma. \tag{3.1}$$

Proof. Differentiating (1.9), we obtain

$$E'(t) = \int_\Omega \{u_t u_{tt} + \nabla u \nabla u_t\} dx + \xi \int_{\Gamma_N} \int_0^1 u_t(x,t-\tau\rho) u_{tt}(x,t-\tau\rho) d\rho d\Gamma,$$

and then, applying Green's formula,

$$E'(t) = \int_{\Gamma_N} u_t \frac{\partial u}{\partial \nu} d\Gamma + \xi \int_{\Gamma_N} \int_0^1 u_t(x,t-\tau\rho) u_{tt}(x,t-\tau\rho) d\rho d\Gamma. \tag{3.2}$$

Now, observe that

$$u_t(x,t-\tau\rho) = -\tau^{-1} u_\rho(x,t-\tau\rho)$$

and

$$u_{tt}(x,t-\tau\rho)=\tau^{-2}u_{\rho\rho}(x,t-\tau\rho).$$

Therefore,

(3.3)
$$\int_{\Gamma_N}\int_0^1 u_t(x,t-\tau\rho)u_{tt}(x,t-\tau\rho)d\rho d\Gamma=-\tau^{-3}\int_{\Gamma_N}\int_0^1 u_\rho(x,t-\tau\rho)u_{\rho\rho}(x,t-\tau\rho)d\rho d\Gamma.$$

Integrating by parts in ρ, we obtain

$$\begin{aligned}\int_{\Gamma_N}\int_0^1 u_\rho(x,t-\tau\rho)u_{\rho\rho}(x,t-\tau\rho)d\rho d\Gamma=&-\int_{\Gamma_N}\int_0^1 u_\rho(x,t-\tau\rho)u_{\rho\rho}(x,t-\tau\rho)d\rho d\Gamma\\&+\int_{\Gamma_N}\{u_\rho^2(x,t-\tau)-u_\rho^2(x,t)\}d\Gamma,\end{aligned}$$

that is,

$$\int_{\Gamma_N}\int_0^1 u_\rho(x,t-\tau\rho)u_{\rho\rho}(x,t-\tau\rho)d\rho d\Gamma=\frac{1}{2}\int_{\Gamma_N}\{u_\rho^2(x,t-\tau)-u_\rho^2(x,t)\}d\Gamma. \tag{3.4}$$

Then, from (3.3) and (3.4),

(3.5)
$$\begin{aligned}\int_{\Gamma_N}\int_0^1 u_t(x,t-\tau\rho)u_{tt}(x,t-\tau\rho)d\rho d\Gamma&=-\tau^{-3}\int_{\Gamma_N}\int_0^1 u_\rho(x,t-\tau)u_{\rho\rho}(x,t-\tau\rho)d\rho d\Gamma\\&=\frac{\tau^{-3}}{2}\int_{\Gamma_N}\{u_\rho^2(x,t)-u_\rho^2(x,t-\tau)\}d\Gamma=\frac{\tau^{-1}}{2}\int_{\Gamma_N}\{u_t^2(x,t)-u_t^2(x,t-\tau)\}d\Gamma.\end{aligned}$$

Using (3.2), (3.5), and the boundary condition (1.3) on Γ_N, we have

$$\begin{aligned}E'(t)=&-\mu_1\int_{\Gamma_N}u_t^2(x,t)d\Gamma-\mu_2\int_{\Gamma_N}u_t(x,t)u_t(x,t-\tau)d\Gamma\\&+\frac{\xi\tau^{-1}}{2}\int_{\Gamma_N}u_t^2(x,t)d\Gamma-\frac{\xi\tau^{-1}}{2}\int_{\Gamma_N}u_t^2(x,t-\tau)d\Gamma.\end{aligned} \tag{3.6}$$

From (3.6), applying the Cauchy–Schwarz inequality we obtain

$$E'(t)\le\left(-\mu_1+\frac{\mu_2}{2}+\frac{\xi\tau^{-1}}{2}\right)\int_{\Gamma_N}u_t^2(x,t)d\Gamma+\left(\frac{\mu_2}{2}-\frac{\xi\tau^{-1}}{2}\right)\int_{\Gamma_N}u_t^2(x,t-\tau)d\Gamma,$$

which implies

$$E'(t)\le-C\int_{\Gamma_N}\{u_t^2(x,t)+u_t^2(x,t-\tau)\}d\Gamma,$$

with

$$C=\min\left\{\left(\mu_1-\frac{\mu_2}{2}-\frac{\xi\tau^{-1}}{2}\right),\left(-\frac{\mu_2}{2}+\frac{\xi\tau^{-1}}{2}\right)\right\}.$$

Since ξ is chosen satisfying assumption (1.10), the constant C is positive. □

We can write

$$E(t) = \mathcal{E}(t) + E_N(t),$$

where $\mathcal{E}(t)$ is the standard energy for the wave equation

$$\mathcal{E}(t) := \frac{1}{2}\int_\Omega \{u_t^2(x,t) + |\nabla u(x,t)|^2\}dx, \tag{3.7}$$

and

$$E_N(t) := \frac{\xi}{2}\int_{\Gamma_N}\int_0^1 u_t^2(x,t-\tau\rho)d\rho d\Gamma. \tag{3.8}$$

With a change of variable we can rewrite

$$E_N(t) = \frac{\xi}{2\tau}\int_{\Gamma_N}\int_{t-\tau}^t u_t^2(x,s)dsd\Gamma. \tag{3.9}$$

We can now give a boundary observability inequality which we will use to prove the exponential decay of the energy $E(t)$.

PROPOSITION 3.2. *There exists a time $\overline{T} > 0$ such that for all times $T > \overline{T}$, there exists a positive constant C_0 (depending on T) for which*

$$E(0) \le C_0\int_0^T\int_{\Gamma_N}\{u_t^2(x,t) + u_t^2(x,t-\tau)\}d\Gamma dt \tag{3.10}$$

for any regular solution u of problem (1.1)–(1.5).

Proof. From Proposition 6.3 of [14], for T greater than a sufficiently large time T_0, and any $\varepsilon > 0$, we have

$$\mathcal{E}(0) \le c\int_0^T\int_{\Gamma_N}\Big\{\Big(\frac{\partial u}{\partial \nu}\Big)^2 + u_t^2\Big\}d\Gamma dt + c\|u\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))} \tag{3.11}$$

for a suitable constant c (depending on T). Estimate (3.11) is obtained by Carleman estimates under the assumption that there exists a function v of class C^2 satisfying (1.6) and (1.7). The function v is needed to construct a suitable weight function for Carleman estimates (see the proof of Proposition 4.2 below). Then, by (3.11) and the boundary condition (1.3), we have

$$\mathcal{E}(0) \le c\int_0^T\int_{\Gamma_N}\{u_t^2(x,t) + u_t^2(x,t-\tau)\}d\Gamma dt + c\|u\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))} \tag{3.12}$$

for a suitable positive constant c. From (3.9) we have that

$$E_N(0) \le c\int_{\Gamma_N}\int_{-\tau}^0 u_t^2(x,s)dsd\Gamma. \tag{3.13}$$

By a change of variable in (3.13) we obtain, for $T \ge \tau$,

$$E_N(0) \le c\int_0^T\int_{\Gamma_N} u_t^2(x,t-\tau)d\Gamma dt. \tag{3.14}$$

Denote $\overline{T} := \max\{\tau, T_0\}$. Then, from (3.12) and (3.14), for any $T > \overline{T}$ we have

$$\begin{aligned} E(0) &= \mathcal{E}(0) + E_N(0) \\ &\le c\int_0^T \int_{\Gamma_N} \{u_t^2(x,t) + u_t^2(x,t-\tau)\} d\Gamma dt + c\|u\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))} \end{aligned} \tag{3.15}$$

for a suitable positive constant c depending on T.

In order to obtain (3.10) we need to absorb the lower order term $\|u\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))}$. To do this, we argue by contradiction. Suppose that (3.10) is not true. Then, there exists a sequence $\{u_n\}_n$ of solutions of problem (1.1)–(1.5) such that, denoting by $E^n(0)$ the energy E related to u_n at the time 0,

$$E^n(0) > n\int_0^T \int_{\Gamma_N} \{u_{nt}^2(x,t) + u_{nt}^2(x,t-\tau)\} d\Gamma dt. \tag{3.16}$$

From (3.15), we have

$$E^n(0) \le c\left\{\int_0^T \int_{\Gamma_N} \{u_{nt}^2(x,t) + u_{nt}^2(x,t-\tau)\} d\Gamma dt + \|u_n\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))}\right\}. \tag{3.17}$$

Then, from (3.16) and (3.17),

$$\begin{aligned} n\int_0^T \int_{\Gamma_N} &\{u_{nt}^2(x,t) + u_{nt}^2(x,t-\tau)\} d\Gamma dt \\ &< c\left\{\int_0^T \int_{\Gamma_N} \{u_{nt}^2(x,t) + u_{nt}^2(x,t-\tau)\} d\Gamma dt + \|u_n\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))}\right\}, \end{aligned}$$

that is,

$$(n-c)\int_0^T \int_{\Gamma_N} \{u_{nt}^2(x,t) + u_{nt}^2(x,t-\tau)\} d\Gamma dt < c\|u_n\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))}. \tag{3.18}$$

Renormalizing, we obtain a sequence $\{w_n\}_n$ of solutions of problem (1.1)–(1.5) verifying

$$\|w_n\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))} = 1, \tag{3.19}$$

and

$$\int_0^T \int_{\Gamma_N} \{w_{nt}^2(x,t) + w_{nt}^2(x,t-\tau)\} d\Gamma dt < \frac{c}{n-c}. \tag{3.20}$$

From (3.19), (3.20), and (3.17), it follows that the sequence $\{w_n\}_n$ is bounded in $H^1(\Omega\times(0,T))$. Since $H^1(\Omega\times(0,T))$ is compactly embedded in $H^{1/2+\varepsilon}(\Omega\times(0,T))$, there exists a subsequence which, for simplicity of notation, we still denote by $\{w_n\}_n$ such that

$$w_n \longrightarrow w \qquad \text{strongly in} \quad H^{1/2+\varepsilon}(\Omega\times(0,T)).$$

Then, from (3.19),

$$\|w\|_{H^{1/2+\varepsilon}(\Omega\times(0,T))} = 1. \tag{3.21}$$

Moreover, by (3.20),

$$\int_0^T \int_{\Gamma_N} \{w_t^2(x,t) + w_t^2(x,t-\tau)\} d\Gamma dt = 0.$$

Therefore, we have that

$$w_t = 0 \quad \text{on } \Gamma_N \times (0,T)$$

and

$$\frac{\partial w}{\partial \nu} = 0 \quad \text{on } \Gamma_N \times (0,T).$$

Putting $v := w_t$, v solves in a distributional sense

$$v'' - \Delta v = 0 \quad \text{in } \Omega \times (0,T),$$

with

$$v = 0 \quad \text{on } \Gamma \times (0,T), \quad \frac{\partial v}{\partial \nu} = 0 \quad \text{on } \Gamma_N \times (0,T).$$

Therefore, from Holmgren's uniqueness theorem (see [15, Chap. I, Thm. 8.2, p. 92]) $v \equiv 0$. This implies that

$$w(x,t) = w(x).$$

Thus, w verifies

$$\begin{cases} -\Delta w = 0 & \text{in } \Omega, \\ w = 0 & \text{on } \Gamma_D, \\ \dfrac{\partial w}{\partial \nu} = 0 & \text{on } \Gamma_N, \end{cases}$$

and so $w \equiv 0$. This is in contradiction with (3.21). Then, the observability inequality (3.10) is proved. □

From (3.10) easily follows the stability result.

Proof of Theorem 1.1. From (3.1), we have

$$E(T) - E(0) \le -C \int_0^T \int_{\Gamma_N} \{u_t^2(x,t) + u_t^2(x,t-\tau)\} d\Gamma dt. \tag{3.22}$$

By (3.22) and the observability estimate (3.10), we obtain

$$E(T) \le E(0) \le C_0 \int_0^T \int_{\Gamma_N} \{u_t^2(x,t) + u_t^2(x,t-\tau)\} d\Gamma dt \le C_0 C^{-1} (E(0) - E(T));$$

then

$$E(T) \le \tilde{C} E(0),$$

with $\tilde{C} < 1$. This easily implies the stability estimate (1.11), since our system (1.1)–(1.5) is invariant by translation and the energy E is decreasing. □

Remark 3.3. Analogous arguments apply if we have more than one delay term in the boundary feedback, that is, if condition (1.3) is substituted by

$$\frac{\partial u}{\partial \nu}(x,t) = -\mu_0 u_t(x,t) - \sum_{i=1}^{k} \mu_i u_t(x,t-\tau_i) \quad \text{on} \quad \Gamma_N \times (0,+\infty),$$

with $\mu_0, \mu_i, \tau_i, i = 1, \dots, k$, positive parameters. In this case, the right energy for our problem is

$$E(t) := \frac{1}{2}\int_\Omega \{u_t^2(x,t) + |\nabla u(x,t)|^2\}dx + \sum_{i=1}^{k} \frac{\xi_i}{2}\int_{\Gamma_N}\int_0^1 u_t^2(x,t-\rho\tau_i)d\rho d\Gamma,$$

with suitable positive constants $\xi_i, i = 1, \dots, k$. Indeed, if

$$\mu_0 > \sum_{i=1}^{k} \mu_i,$$

choosing ξ_i such that

$$\mu_i < \xi_i \tau_i^{-1}, \ i = 1, \dots, k, \quad \text{and} \quad \sum_{i=1}^{k} \xi_i \tau_i^{-1} < 2\mu_0 - \sum_{i=1}^{k} \mu_i,$$

we can prove that the energy is exponentially decaying to zero.

Remark 3.4. If $\overline{\Gamma}_D \cap \overline{\Gamma}_N \neq \emptyset$, the strong solution of (1.1)–(1.5) has a singular behavior along $\overline{\Gamma}_D \cap \overline{\Gamma}_N$ since $u(t) \in E(\Delta, L^2(\Omega))$ with $\frac{\partial u}{\partial \nu} \in L^2(\Gamma_N)$ and $u = 0$ on Γ_D (see, for instance, [7] for two-dimensional domains). Therefore the results from [14] cannot be invoked. For the standard feedback law (i.e., the case $\mu_2 = 0$), the multiplier method has been used as an alternative in [1] to obtain stability results under strong geometrical conditions. Unfortunately that approach cannot be directly applied here because for $\mu_2 > 0$, we have only $\frac{\partial u}{\partial \nu} \in L^2(\Gamma_N)$, which forbids the use of the multiplier identity.

4. Internal stability result. In this section, under assumption (1.8) we want to prove exponential stability for problem (1.12)–(1.16).

We first show that the energy $\mathcal{F}$, defined by (1.19) and (1.10), is decreasing.

PROPOSITION 4.1. *For any regular solution of problem* (1.12)–(1.16)*, the energy is decreasing and there exists a positive constant C such that*

$$\mathcal{F}'(t) \leq -C\int_\Omega a(x)\{u_t^2(x,t) + u_t^2(x,t-\tau)\}dx. \tag{4.1}$$

Proof. Differentiating (1.19), we obtain

$$\mathcal{F}'(t) = \int_\Omega \{u_t u_{tt} + \nabla u \nabla u_t\}dx + \xi\int_\Omega a(x)\int_0^1 u_t(x,t-\tau\rho)u_{tt}(x,t-\tau\rho)d\rho dx,$$

and then, applying Green's formula,

$$\mathcal{F}'(t) = \int_\Omega u_t(u_{tt} - \Delta u)dx + \xi\int_\Omega a(x)\int_0^1 u_t(x,t-\tau\rho)u_{tt}(x,t-\tau\rho)d\rho dx, \tag{4.2}$$

where we have used the boundary conditions (1.13) and (1.14).

As in the proof of Proposition 3.1, we can compute

$$\begin{aligned}\int_\Omega a(x)\int_0^1 u_t(x,t-\tau\rho)u_{tt}(x,t-\tau\rho)d\rho dx&\\ &= -\tau^{-3}\int_\Omega a(x)\int_0^1 u_\rho(x,t-\tau)u_{\rho\rho}(x,t-\tau\rho)d\rho dx\\ &= \frac{\tau^{-3}}{2}\int_\Omega a(x)\{u_\rho^2(x,t)-u_\rho^2(x,t-\tau)\}dx\\ &= \frac{\tau^{-1}}{2}\int_\Omega a(x)\{u_t^2(x,t)-u_t^2(x,t-\tau)\}d\Gamma.\end{aligned} \tag{4.3}$$

Now, using (4.3) and (1.12) in identity (4.2), we have

$$\begin{aligned}\mathcal{F}'(t) = &-\mu_1\int_\Omega a(x)u_t^2(x,t)dx-\mu_2\int_\Omega a(x)u_t(x,t)u_t(x,t-\tau)dx\\ &+\frac{\xi\tau^{-1}}{2}\int_\Omega a(x)u_t^2(x,t)dx-\frac{\xi\tau^{-1}}{2}\int_\Omega a(x)u_t^2(x,t-\tau)dx.\end{aligned} \tag{4.4}$$

From (4.4), applying the Cauchy–Schwarz inequality and recalling (1.10), we obtain estimate (4.1). □

We can write

$$\mathcal{F}(t) = \mathcal{E}(t) + E^0(t),$$

where $\mathcal{E}(t)$ is the standard energy for the wave equation defined in (3.7) and where

$$E^0(t) := \frac{\xi}{2}\int_\Omega a(x)\int_0^1 u_t^2(x,t-\tau\rho)d\rho dx. \tag{4.5}$$

With a change of variable, we can rewrite

$$E^0(t) = \frac{\xi}{2\tau}\int_\Omega a(x)\int_{t-\tau}^t u_t^2(x,s)dsdx. \tag{4.6}$$

Let w be the solution of the homogeneous problem for the wave equation with mixed Dirichlet–Neumann boundary condition,

$$w_{tt}(x,t)-\Delta w(x,t)=0 \quad \text{in} \quad \Omega\times(0,+\infty), \tag{4.7}$$

$$w(x,t)=0 \quad \text{on} \quad \Gamma_D\times(0,+\infty), \tag{4.8}$$

$$\frac{\partial w}{\partial\nu}(x,t)=0 \quad \text{on} \quad \Gamma_N\times(0,+\infty), \tag{4.9}$$

$$w(x,0)=w_0(x) \quad \text{and} \quad w_t(x,0)=w_1(x) \quad \text{in} \quad \Omega. \tag{4.10}$$

Denote by $\mathcal{E}_w(t)$ the standard energy for the wave equation corresponding to w, that is,

$$\mathcal{E}_w(t) = \frac{1}{2}\int_\Omega\{w_t^2(x,t)+|\nabla w(x,t)|^2\}dx. \tag{4.11}$$

Note that $\mathcal{E}_w(t)$ is constant.

We can give an observability inequality for problem (4.7)–(4.10).

PROPOSITION 4.2. *There exists a time T_0 such that for all times $T > T_0$, there exists a positive constant C_1 (depending on T) for which*

$$\mathcal{E}_w(0) \le C_1 \int_0^T \int_\omega w_t^2(x,t)dxdt \tag{4.12}$$

for any regular solution w of problem (4.7)–(4.10).

Proof. Inequality (4.12) easily follows from some estimates of [14] and standard arguments with multipliers. We give the proof for the reader's convenience.

Let $\omega_0,\ \omega_1$ be open neighborhoods of Γ_N such that

$$\omega \supset\supset \omega_0 \supset\supset \omega_1 \supset \Gamma_N. \tag{4.13}$$

Let φ be a smooth function such that

$$0 \le \varphi(x) \le 1, \quad \varphi \equiv 0 \quad \text{on } \Omega \setminus \omega_0, \quad \varphi \equiv 1 \quad \text{on } \omega_1. \tag{4.14}$$

Then, the function φw verifies

$$(\varphi w)_{tt} - \Delta(\varphi w) = F(w),$$

where $F(w) = w\Delta\varphi + 2\nabla\varphi\nabla w$, with the same boundary conditions as w. Therefore, we can apply to φw the result of Proposition 4.2.1 in [14].

Let us recall some notation from [14]. Without loss of generality, we can suppose that the function v satisfying assumptions (1.6) and (1.7) is nonnegative on Ω.

Denote

$$T_0 = 2\left(\frac{\max_{x\in\overline{\Omega}}\ v(x)}{\alpha}\right)^{1/2}, \tag{4.15}$$

with α as in (1.6). Define the function $\phi : \Omega \times \mathbb{R} \to \mathbb{R}$ by

$$\phi(x,t) := v(x) - c_*\left(t - \frac{T}{2}\right)^2, \tag{4.16}$$

where $T > T_0$ is fixed and the constant c_* is chosen as follows. From (4.15), there exists a constant $\delta > 0$ such that

$$\alpha T^2 > 4 \max_{x\in\overline{\Omega}}\ v(x) + 4\delta.$$

For fixed δ, there is c_* such that

$$c_* T^2 > 4 \max_{x\in\overline{\Omega}}\ v(x) + 4\delta, \quad c_* \in (0,\alpha). \tag{4.17}$$

Note that

$$\phi(x,0) < -\delta \quad \text{and} \quad \phi(x,T) < -\delta \quad \text{uniformly in} \quad \Omega. \tag{4.18}$$

Set

$$\begin{aligned} BTw|_{\Gamma\times(0,T)} &= \frac{1}{2}\int_0^T \int_{\Gamma_D} e^{\gamma\phi} H\cdot\nu \left(\frac{\partial w}{\partial\nu}\right)^2 d\Gamma dt \\ &\quad + \frac{1}{2}\int_0^T \int_{\Gamma_N} e^{\gamma\phi} H\cdot\nu (w_t^2 - |\nabla_{\mathcal{T}} w|^2) d\Gamma dt, \end{aligned} \tag{4.19}$$

where $\nabla_{\mathcal{T}} w$ denotes the tangential gradient of w.

Then, from Proposition 4.2.1 of [14], using (4.18), and recalling that $\mathcal{E}_w(t)$ is constant, we have

$$\begin{aligned} BTw|_{\Gamma\times(0,T)} &\le c\left\{\int_0^T\int_\Omega e^{\gamma\phi}|\nabla(\varphi w)|^2dxdt+\int_0^T\int_\Omega \varphi^2 w_t^2dxdt\right.\\ &\qquad\left.+\int_0^T\int_\Omega \varphi^2 w^2dxdt+e^{-\gamma\delta}\mathcal{E}_w(0)\right\}\\ &\le c\left\{\int_0^T\int_{\omega_0} e^{\gamma\phi}|\nabla w|^2dxdt+\int_0^T\int_\omega w_t^2dxdt\right.\\ &\qquad\left.+\int_0^T\int_\Omega w^2dxdt+e^{-\gamma\delta}\mathcal{E}_w(0)\right\} \end{aligned} \tag{4.20}$$

for a suitable positive constant c, where the parameter γ can be chosen sufficiently large in order to have the desired inequality.

Now, consider another smooth cut-off function ψ such that

$$0\le\psi(x)\le 1,\quad \psi\equiv 0\quad \text{on } \Omega\setminus\omega,\quad \psi\equiv 1\quad \text{on } \omega_0. \tag{4.21}$$

Integrating by parts,

$$\begin{aligned} \int_0^T\int_\Omega (w_{tt}-\Delta w)\psi we^{\gamma\phi}dxdt &= \left[\int_\Omega \psi w w_t e^{\gamma\phi}dx\right]_0^T\\ &\quad+\int_0^T\int_\Omega \nabla w\nabla(\psi we^{\gamma\phi})dxdt-\int_0^T\int_\Omega w_t(\psi we^{\gamma\phi})_t dxdt\\ &= \left[\int_\Omega \psi w w_t e^{\gamma\phi}dx\right]_0^T-\int_0^T\int_\Omega \psi w_t(w_t e^{\gamma\phi}+we^{\gamma\phi}\gamma\phi_t)dxdt\\ &\quad+\int_0^T\int_\Omega \psi e^{\gamma\phi}|\nabla w|^2dxdt+\int_0^T\int_\Omega w\nabla w\nabla(\psi e^{\gamma\phi})dxdt. \end{aligned} \tag{4.22}$$

Then, from (4.22), recalling that w satisfies (4.7), we have

$$\begin{aligned} \int_0^T\int_\Omega \psi e^{\gamma\phi}|\nabla w|^2dxdt &= \int_0^T\int_\Omega(\psi w_t^2e^{\gamma\phi}+ww_t\psi e^{\gamma\phi}\gamma\phi_t)dxdt\\ &\quad-\left[\int_\Omega \psi w w_t e^{\gamma\phi}dx\right]_0^T-2\int_0^T\int_\Omega \sqrt{\psi}w\nabla w\nabla(\sqrt{\psi})e^{\gamma\phi}dxdt\\ &\quad-\int_0^T\int_\Omega \psi w\nabla w\nabla e^{\gamma\phi}dxdt. \end{aligned} \tag{4.23}$$

Since $\mathcal{E}_w(t)$ is constant, by using the Cauchy–Schwarz inequality and Poincaré's theorem, we can estimate

$$\left[\int_\Omega \psi w w_t e^{\gamma\phi}dx\right]_0^T\le ce^{-\delta\gamma}\mathcal{E}_w(0),$$

and so, from (4.23), we obtain

$$\begin{aligned} \int_0^T\int_\Omega \psi e^{\gamma\phi}|\nabla w|^2dxdt &\le ce^{-\delta\gamma}\mathcal{E}_w(0)+\frac{1}{2}\int_0^T\int_\Omega \psi e^{\gamma\phi}|\nabla w|^2dxdt\\ &\quad+c\left\{\int_0^T\int_\omega w_t^2dxdt+\int_0^T\int_\Omega w^2dxdt\right\} \end{aligned} \tag{4.24}$$

for a suitable positive constant c. By (4.24) we deduce

$$\int_0^T \int_{\omega_0} e^{\gamma\phi}|\nabla w|^2 dxdt \le c\left\{\int_0^T \int_\omega w_t^2 dxdt + \int_0^T \int_\Omega w^2 dxdt + e^{-\delta\gamma}\mathcal{E}_w(0)\right\},$$

which, used in (4.20), gives

$$BTw|_{\Gamma\times(0,T)} \le c\left\{\int_0^T \int_\omega w_t^2 dxdt + \int_0^T \int_\Omega w^2 dxdt + e^{-\delta\gamma}\mathcal{E}_w(0).\right\} \tag{4.25}$$

Then, from (4.25) and Theorem 3.4 of [14] (Carleman estimate (3.14)), taking γ sufficiently large, we obtain

$$\mathcal{E}_w(0) \le c\int_0^T \int_\omega w_t^2(x,t)dxdt + c\|w\|^2_{L^2(\Omega\times(0,T))}.$$

Now, estimate (4.12) follows from compactness-uniqueness arguments. □

PROPOSITION 4.3. *There exists a time $\overline{T}$ such that for all times $T > \overline{T}$, there exists a positive constant C_0 (depending on T) for which*

$$\mathcal{F}(0) \le C_0 \int_0^T \int_\Omega a(x)\{u_t^2(x,t) + u_t^2(x,t-\tau)\}dxdt \tag{4.26}$$

for any regular solution u of problem (1.12)–(1.16).

Proof. Following Zuazua [18], we can decompose the solution u of problem (1.12)–(1.16) as

$$u = w + \tilde{w},$$

where w solves (4.7)–(4.9) with initial condition

$$w(x,0) = u_0(x), \quad w_t(x,0) = u_1(x) \quad \text{in } \Omega,$$

and $\tilde{w}$ verifies

$$\tilde{w}_{tt} - \Delta\tilde{w} = -a(x)[\mu_1 u_t(x,t) + \mu_2 u_t(x,t-\tau)] \quad \text{in} \quad \Omega\times(0,+\infty), \tag{4.27}$$

$$\tilde{w}(x,t) = 0 \quad \text{on} \quad \Gamma_D\times(0,+\infty), \tag{4.28}$$

$$\frac{\partial\tilde{w}}{\partial\nu}(x,t) = 0 \quad \text{on} \quad \Gamma_N\times(0,+\infty), \tag{4.29}$$

$$\tilde{w}(x,0) = 0 \quad \text{and} \quad \tilde{w}_t(x,0) = 0 \quad \text{in} \quad \Omega. \tag{4.30}$$

Then, from (4.5) and (4.11),

$$\mathcal{F}(0) = \mathcal{E}(0) + E^0(0) = \mathcal{E}_w(0) + \frac{\xi}{2}\int_\Omega a(x)\int_0^1 u_t^2(x,-\rho\tau)d\rho dx. \tag{4.31}$$

If we take $T > \overline{T} := \max\{T_0, \tau\}$, from (4.31) with a change of variable we obtain

$$\mathcal{F}(0) \le \mathcal{E}_w(0) + c\int_\Omega a(x)\int_0^T u_t^2(x,t-\tau)dtdx,$$

and then, from (4.12),

$$\begin{aligned}\mathcal{F}(0) &\le c\int_\Omega a(x)\int_0^T \{w_t^2(x,t)+u_t^2(x,t-\tau)\}dtdx\\ &\le c\int_\Omega a(x)\int_0^T \{\tilde{w}_t^2(x,t)+u_t^2(x,t)+u_t^2(x,t-\tau)\}dtdx\end{aligned} \tag{4.32}$$

for a suitable positive constant c. Therefore, from standard energy estimates for $\tilde{w}$, we obtain

$$\mathcal{F}(0) \le C_0\int_0^T\int_\Omega a(x)\{u_t^2(x,t)+u_t^2(x,t-\tau)\}dxdt. \qquad \square$$

Now, using estimate (4.26), as in the case of boundary feedback we obtain the exponential stability result of Theorem 1.3.

Remark 4.4. Analogous arguments apply if we have more than one delay term in the internal feedback, that is, if (1.12) is replaced with

$$u_{tt}(x,t)-\Delta u(x,t)+a(x)\Big[\mu_0 u_t(x,t)+\sum_{i=1}^k \mu_i u_t(x,t-\tau_i)\Big]=0 \quad \text{in } \Omega\times(0,+\infty),$$

with $\mu_0,\mu_i,\tau_i, i=1,\dots,k$, positive parameters. In this case, if

$$\mu_0 > \sum_{i=1}^k \mu_i,$$

the right energy to consider, in order to prove exponential decay, is

$$E(t) := \frac{1}{2}\int_\Omega \{u_t^2(x,t)+|\nabla u(x,t)|^2\}dx + \sum_{i=1}^k \frac{\xi_i}{2}\int_\Omega a(x)\int_0^1 u_t^2(x,t-\rho\tau_i)d\rho dx,$$

with constants $\xi_i, i=1,\dots,k$, chosen as in Remark 3.3.

Remark 4.5. In the case $\overline{\Gamma}_D\cap\overline{\Gamma}_N\neq\emptyset$, since for internal feedbacks we have $\frac{\partial u}{\partial\nu}=0$ on Γ_N, we can use the multiplier identity from [1] and then obtain stability results under the same geometrical conditions as those from [1].

5. Some instability examples. In this section we will give some instability examples for the case $\mu_2\ge\mu_1$.

5.1. Boundary feedback. In this subsection we consider problem (1.1)–(1.5) with boundary feedback, and we prove Theorem 1.2.

Let us consider the spectral problem for the system

$$\begin{cases} u_{tt}(x,t)-\Delta u(x,t)=0 \quad \text{in } \Omega\times(0,+\infty),\\ u(x,t)=0 \quad \text{on } \Gamma_D\times(0,+\infty),\\ \dfrac{\partial u}{\partial\nu}(x,t)=-\mu_1 u_t(x,t)-\mu_2 u_t(x,t-\tau) \quad \text{on } \Gamma_N\times(0,+\infty).\end{cases} \tag{5.1}$$

We seek a solution of (5.1) in the form

$$u(x,t)=e^{\lambda t}\varphi(x), \quad \lambda\in\mathbb{C}.$$

Then, φ has to be a solution of the eigenvalue problem

$$\begin{cases} -\Delta\varphi + \lambda^2\varphi = 0 \quad \text{in } \Omega, \\ \varphi = 0 \quad \text{on } \Gamma_D, \\ \dfrac{\partial\varphi}{\partial\nu} = -(\mu_1 + \mu_2 e^{-\lambda\tau})\lambda\varphi \quad \text{on } \Gamma_N, \end{cases} \tag{5.2}$$

which can be reformulated, in a variational form, as

$$\int_\Omega \nabla\varphi\nabla\overline{v}dx + \lambda^2\int_\Omega \varphi\overline{v}dx + (\mu_1 + \mu_2 e^{-\lambda\tau})\lambda\int_{\Gamma_N}\varphi\overline{v}d\Gamma = 0 \quad \forall v \in H^1_{\Gamma_D}(\Omega). \tag{5.3}$$

We want to find a solution for $\lambda := ib$, with $b \in \mathbb{R}$. For this choice of λ, problem (5.3) can be rewritten as

$$\int_\Omega \nabla\varphi\nabla\overline{v}dx - b^2\int_\Omega \varphi\overline{v}dx + (\mu_1 + \mu_2 e^{-ib\tau})ib\int_{\Gamma_N}\varphi\overline{v}d\Gamma = 0 \quad \forall v \in H^1_{\Gamma_D}(\Omega). \tag{5.4}$$

Assume that

$$\cos(b\tau) = -\frac{\mu_1}{\mu_2}. \tag{5.5}$$

Note that, since we are considering the case $\mu_2 \geq \mu_1$, there exist b, τ such that (5.5) holds. Then, we choose

$$\mu_2 \sin(b\tau) = \sqrt{\mu_2^2 - \mu_1^2}. \tag{5.6}$$

Under these assumptions, (5.4) becomes

$$\int_\Omega \nabla\varphi\nabla\overline{v}dx - b^2\int_\Omega \varphi\overline{v}dx + b\sqrt{\mu_2^2 - \mu_1^2}\int_{\Gamma_N}\varphi\overline{v}d\Gamma = 0 \quad \forall v \in H^1_{\Gamma_D}(\Omega). \tag{5.7}$$

In particular, for $v = \varphi$, (5.7) gives

$$\int_\Omega |\nabla\varphi|^2dx - b^2\int_\Omega |\varphi|^2dx + b\sqrt{\mu_2^2 - \mu_1^2}\int_{\Gamma_N}|\varphi|^2d\Gamma = 0. \tag{5.8}$$

Without loss of generality, we can assume

$$\|\varphi\|_2^2 := \int_\Omega |\varphi|^2dx = 1, \tag{5.9}$$

and then the identity (5.8) can be rewritten as

$$b^2 - b\sqrt{\mu_2^2 - \mu_1^2}q_0(\varphi) - q_1(\varphi) = 0, \tag{5.10}$$

where

$$q_0(\varphi) := \int_{\Gamma_N}|\varphi|^2d\Gamma, \quad q_1(\varphi) := \int_\Omega |\nabla\varphi|^2dx. \tag{5.11}$$

Now we distinguish two cases.

Case (a): $\mu_1 = \mu_2$. In this case, under our assumptions, (5.10) becomes

$$b^2 = q_1(\varphi). \tag{5.12}$$

Define

$$b^2 := \min_{\substack{w\in H^1_{\Gamma_D}(\Omega)\\ \|w\|_2=1}} q_1(w). \tag{5.13}$$

If φ verifies

$$q_1(\varphi) = \min_{\substack{w\in H^1_{\Gamma_D}(\Omega)\\ \|w\|_2=1}} q_1(w),$$

then it easy to see that φ is a solution of (5.4) with b as in (5.13). Then φ verifies (5.2), and so

$$u(x,t) := e^{ibt}\varphi(x) \tag{5.14}$$

is a solution of problem (5.1). Therefore, we have found a solution of our boundary problem, whose energy is constant. Indeed, an easy computation shows that, for the function u defined in (5.14),

$$\int_\Omega (|\nabla u(x,t)|^2 + |u_t(x,t)|^2)dx = 2b^2 > 0 \quad \forall t \geq 0.$$

Note that, from our assumptions ($\lambda = ib$, $\cos(b\tau) = -1$, $\sin(b\tau) = 0$), problem (5.2) becomes the classical eigenvalue problem for the Laplace operator with a mixed Dirichlet–Neumann boundary condition. So, we can take a sequence $\{b_n\}_n$ of positive real numbers defined by

$$b_n^2 = \Lambda_n^2, \quad n \in \mathbb{N},$$

where Λ_n^2, $n \in \mathbb{N}$, are the eigenvalues for the Laplace operator. Then, putting

$$b_n\tau = (2l+1)\pi, \quad l \in \mathbb{N},$$

we obtain a sequence of delays

$$\tau_{n,l} = \frac{(2l+1)\pi}{b_n}, \quad l, n \in \mathbb{N},$$

which become arbitrarily small (or large) for suitable choices of the indices $n, l \in \mathbb{N}$. Therefore, in the case $\mu_1 = \mu_2$, we have found a set of time delays for which problem (1.1)–(1.5) is not asymptotically stable.

Case (b): $\mu_2 > \mu_1$. In this case, from (5.10) we have

$$b = \frac{1}{2}\left(\sqrt{\mu_2^2-\mu_1^2}q_0(\varphi) \pm \sqrt{(\mu_2^2-\mu_1^2)q_0^2(\varphi) + 4q_1(\varphi)}\right).$$

Define

$$b := \frac{1}{2}\min_{\substack{w\in H^1_{\Gamma_D}(\Omega)\\ \|w\|_2=1}} \left(\sqrt{\mu_2^2-\mu_1^2}q_0(w) + \sqrt{(\mu_2^2-\mu_1^2)q_0^2(w) + 4q_1(w)}\right). \tag{5.15}$$

We now prove that if the minimum in the right–hand side of (5.15) is attained at φ, that is,

(5.16)
$$\sqrt{\mu_2^2-\mu_1^2}q_0(\varphi)+\sqrt{(\mu_2^2-\mu_1^2)q_0^2(\varphi)+4q_1(\varphi)} := \min_{\substack{w\in H^1_{\Gamma_D}(\Omega)\\ \|w\|_2=1}} \left(\sqrt{\mu_2^2-\mu_1^2}q_0(w)+\sqrt{(\mu_2^2-\mu_1^2)q_0^2(w)+4q_1(w)}\right),$$

then φ is a solution of (5.7) with b as in (5.15). To show this, take for $\varepsilon\in\mathbb{R}$,

$$w=\varphi+\varepsilon v, \quad \text{with} \quad v\in H^1_{\Gamma_D}(\Omega) \ \text{ such that } \ \int_\Omega \varphi\overline{v}dx=0. \tag{5.17}$$

Then,

$$\|w\|_2^2=\|\varphi\|_2^2+\varepsilon^2\|v\|_2^2=1+\varepsilon^2\|v\|_2^2.$$

If we denote

(5.18)
$$g(\varepsilon):=\frac{1}{1+\varepsilon^2\|v\|_2^2}\left(\sqrt{\mu_2^2-\mu_1^2}q_0(\varphi+\varepsilon v)+\sqrt{(\mu_2^2-\mu_1^2)q_0^2(\varphi+\varepsilon v)+4q_1(\varphi+\varepsilon v)}\right),$$

then, by definition (5.16),

$$g(\varepsilon)\geq g(0)=\left(\sqrt{\mu_2^2-\mu_1^2}q_0(\varphi)+\sqrt{(\mu_2^2-\mu_1^2)q_0^2(\varphi)+4q_1(\varphi)}\right).$$

So, we have that

$$\left.\frac{dg(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}=0,$$

which, after an easy computation, gives

$$\int_\Omega \nabla\varphi\nabla\overline{v}dx+b\sqrt{\mu_2^2-\mu_1^2}\int_{\Gamma_N}\varphi\overline{v}d\Gamma=0. \tag{5.19}$$

Since any function $\tilde{v}\in H^1_{\Gamma_D}(\Omega)$ can be decomposed as

$$\tilde{v}=\gamma\varphi+v, \qquad \gamma\in\mathbb{R}, \quad v\in H^1_{\Gamma_D}(\Omega) \ \text{ with } \ \int_\Omega\varphi\overline{v}dx=0,$$

from (5.19) and (5.8) we obtain that φ satisfies (5.7) with b defined in (5.15). So, for such positive b,

$$b\tau=\arccos\left(-\frac{\mu_1}{\mu_2}\right)+2l\pi, \quad l\in\mathbb{N},$$

defines a sequence of time delays for which problem (1.1)–(1.5) is not asymptotically stable.

The above examples prove Theorem 1.2.

5.2. Internal feedback. In this subsection we will give instability examples for problem (1.12)–(1.16) with internal feedback, proving Theorem 1.4.

Let us consider the spectral problem for the system

(5.20)
$$\begin{cases} u_{tt}(x,t) - \Delta u(x,t) + a(x)[\mu_1 u_t(x,t) + \mu_2 u_t(x,t-\tau)] = 0 \quad \text{in } \Omega \times (0,+\infty), \\ u(x,t) = 0 \quad \text{on } \Gamma_D \times (0,+\infty), \\ \dfrac{\partial u}{\partial \nu}(x,t) = 0 \quad \text{on } \Gamma_N \times (0,+\infty). \end{cases}$$

We restrict our analysis to the case $a(x) \equiv 1$ in Ω.

We seek a solution of (5.20) in the form

$$u(x,t) = e^{\lambda t}\varphi(x), \quad \lambda \in \mathbb{C}.$$

Then, φ has to solve the eigenvalue problem

(5.21)
$$\begin{cases} \Delta\varphi = [\lambda^2 + (\mu_1 + \mu_2 e^{-\lambda\tau})\lambda]\varphi \quad \text{in } \Omega, \\ \varphi = 0 \quad \text{on } \Gamma_D, \\ \dfrac{\partial \varphi}{\partial \nu} = 0 \quad \text{on } \Gamma_N. \end{cases}$$

Let us consider the standard problem for the Laplace operator with a mixed Dirichlet–Neumann boundary condition

(5.22)
$$\begin{cases} \Delta\varphi = -\mu^2\varphi \quad \text{in } \Omega, \\ \varphi = 0 \quad \text{on } \Gamma_D, \\ \dfrac{\partial \varphi}{\partial \nu} = 0 \quad \text{on } \Gamma_N. \end{cases}$$

We want to show that for any Λ^2 eigenvalue of problem (5.22), there exists a $\lambda \in \mathbb{C}$ solution of the equation

$$\lambda^2 + (\mu_1 + \mu_2 e^{-\lambda\tau})\lambda = -\Lambda^2. \tag{5.23}$$

We seek a solution $\lambda = \alpha + i\beta,\ \alpha, \beta \in \mathbb{R}$, with

$$\beta\tau = (2l+1)\pi, \quad l \in \mathbb{N}. \tag{5.24}$$

Under this assumption, (5.23) becomes

(5.25)
$$\begin{cases} \alpha^2 + \beta^2 = \Lambda^2, \\ \mu_2 e^{-\alpha\tau} = 2\alpha + \mu_1. \end{cases}$$

Now we distinguish two cases.

Case (a): $\mu_1 = \mu_2$. In this case, from (5.25) we have

$$\alpha = 0, \quad \beta^2 = \Lambda^2.$$

Therefore, for any Λ_n^2 eigenvalue of problem (5.22), if $\beta_n \in \mathbb{R}$ verifies

$$\beta_n^2 = \Lambda_n^2,$$

then for $\lambda = i\beta_n$ problem (5.21) admits a nonzero solution.

Take β_n positive. From our assumption (5.24),

$$\tau_{n,l} = \frac{(2l+1)\pi}{\beta_n}, \quad n, l \in \mathbb{N},$$

is a set of time delays that become arbitrarily small (or large) for suitable choices of the indices $n, l \in \mathbb{N}$. For such delays, problem (1.12)–(1.16) admits solutions in the form

$$u(x,t) = e^{i\beta t}\varphi(x),$$

whose energy is constant and strictly positive. So, system (1.12)–(1.16) is not asymptotically stable.

Case (b): $\mu_2 > \mu_1$. For a fixed $\alpha > 0$, from the second equation of (5.25), we obtain

$$\tau(\alpha) = \frac{1}{\alpha} \ln\left(\frac{\mu_2}{\mu_1 + 2\alpha}\right), \tag{5.26}$$

and so, in order to have $\tau(\alpha) > 0$, we consider

$$0 < \alpha < \frac{1}{2}(\mu_2 - \mu_1).$$

From (5.24), the first equation of (5.25) becomes

$$\alpha^2 + \frac{(2l+1)^2\pi^2}{\tau^2(\alpha)} = \Lambda^2, \tag{5.27}$$

where $\tau(\alpha)$ is given by (5.26).

Denoting

$$g(\alpha) := \alpha^2 + \frac{(2l+1)^2\pi^2}{\tau^2(\alpha)} \qquad \forall \alpha \in \left(0, \frac{\mu_2 - \mu_1}{2}\right),$$

we have

$$\tau(\alpha) \to +\infty \quad \text{and} \quad g(\alpha) \to 0^+ \text{ as } \alpha \to 0^+,$$

while

$$\tau(\alpha) \to 0^+ \quad \text{and} \quad g(\alpha) \to +\infty \text{ as } \alpha \to \frac{1}{2}(\mu_2 - \mu_1)^-.$$

Since g is a continuous function of α, for any fixed Λ^2 eigenvalue of problem (5.22) there exists α $(0 < \alpha < (\mu_2 - \mu_1)/2)$ such that (5.27) is verified. Therefore, for such α there exists a delay $\tau(\alpha)$ (defined by (5.26)) such that a function of the form

$$e^{\alpha + i\beta}\varphi(x)$$

solves problem (1.12)–(1.16). Since $\alpha \geq 0$, the energy of such a solution is not decaying to zero. So, this solution is not asymptotically stable.

Note that, for any Λ_n^2 eigenvalue of problem (5.22) and for any $l \in \mathbb{N}$, there exist $\alpha_{n,l}$ and a delay $\tau_{n,l} = \tau(\alpha_{n,l})$ such that (5.25) is verified with

$$\beta_{n,l} = \frac{(2l+1)\pi}{\tau_{n,l}}.$$

From the first equation of (5.25),

$$\frac{(2l+1)^2\pi^2}{\tau_{n,l}^2} \leq \Lambda_n^2.$$

Then, for a fixed $l \in \mathbb{N}$, if $n \to +\infty$, then $\tau_{n,l} \to 0^+$. On the contrary, for a fixed $n \in \mathbb{N}$, if $l \to +\infty$, then $\tau_{n,l} \to +\infty$. Therefore, we have instability phenomena for a sequence of arbitrarily small or large time delays.

The examples of Case (a) and (b) prove Theorem 1.4.

Acknowledgments. We thank E. Zuazua for bringing our attention to reference [17] and suggesting we consider the stabilization of the wave equation with delay in general domains of $\mathbb{R}^n$.

REFERENCES

[1] R. Bey, A. Heminna, and M. Moussaoui, *Singularities of the solution of a mixed problem for a general second order elliptic equation and boundary stabilization of the wave equation*, J. Math. Pures Appl., 78 (1999), pp. 1043–1067.

[2] G. Chen, *Control and stabilization for the wave equation in a bounded domain, Part* I, SIAM J. Control Optim., 17 (1979), pp. 66–81.

[3] G. Chen, *Control and stabilization for the wave equation in a bounded domain, Part* II, SIAM J. Control Optim., 19 (1981), pp. 114–122.

[4] R. Datko, *Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks*, SIAM J. Control Optim., 26 (1988), pp. 697–713.

[5] R. Datko, *Two examples of ill-posedness with respect to time delays revisited*, IEEE Trans. Automat. Control, 42 (1997), pp. 511–515.

[6] R. Datko, J. Lagnese, and M. P. Polis, *An example on the effect of time delays in boundary feedback stabilization of wave equations*, SIAM J. Control Optim., 24 (1986), pp. 152–156.

[7] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 21, Pitman, Boston–London–Melbourne, 1985.

[8] V. Komornik, *Rapid boundary stabilization of the wave equation*, SIAM J. Control Optim., 29 (1991), pp. 197–208.

[9] V. Komornik, *Exact Controllability and Stabilization. The Multiplier Method*, RAM: Res. Appl. Math. 36, Masson, Paris; John Wiley, Chichester, 1994.

[10] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.

[11] J. Lagnese, *Decay of solutions of the wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.

[12] J. Lagnese, *Note on boundary stabilization of wave equations*, SIAM J. Control and Optim., 26 (1988), pp. 1250–1256.

[13] I. Lasiecka and R. Triggiani, *Uniform exponential energy decay of wave equations in a bounded region with $L_2(0,\infty;L_2(\Gamma))$-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340–390.

[14] I. Lasiecka, R. Triggiani, and P. F. Yao, *Inverse/observability estimates for second-order hyperbolic equations with variable coefficients*, J. Math. Anal. Appl., 235 (1999), pp. 13–57.

[15] J. L. Lions, *Contrôlabilité Exacte, Stabilisation et Perturbations des Systèmes Distribués*, Vol. 1, Masson, Paris, 1988.

[16] K. Liu, *Locally distributed control and damping for the conservative systems*, SIAM J. Control and Optim., 35 (1997), pp. 1574–1590.

[17] G. Q. Xu, S. P. Yung, and L. K. Li, *Stabilization of wave systems with input delay in the boundary control*, ESAIM Control Optim. Calc. Var., in press.

[18] E. Zuazua, *Exponential decay for the semi-linear wave equation with locally distributed damping*, Comm. Partial Differential Equations, 15 (1990), pp. 205–235.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1586–1611

ERROR ESTIMATES FOR THE NUMERICAL APPROXIMATION OF DIRICHLET BOUNDARY CONTROL FOR SEMILINEAR ELLIPTIC EQUATIONS*

EDUARDO CASAS† AND JEAN-PIERRE RAYMOND‡

Abstract. We study the numerical approximation of boundary optimal control problems governed by semilinear elliptic partial differential equations with pointwise constraints on the control. The control is the trace of the state on the boundary of the domain, which is assumed to be a convex, polygonal, open set in $\mathbb{R}^2$. Piecewise linear finite elements are used to approximate the control as well as the state. We prove that the error estimates are of order $O(h^{1-1/p})$ for some $p > 2$, which is consistent with the $W^{1-1/p,p}(\Gamma)$-regularity of the optimal control.

Key words. Dirichlet control, semilinear elliptic equation, numerical approximation, error estimates

AMS subject classifications. 65N30, 65N15, 49M05, 49M25

DOI. 10.1137/050626600

1. Introduction.

In this paper we study an optimal control problem governed by a semilinear elliptic equation. The control is the Dirichlet datum on the boundary of the domain. Bound constraints are imposed on the control. The cost functional involves the control in a quadratic way and the state in a general way. The goal is to derive error estimates for the discretization of the control problem.

There are not many papers devoted to the derivation of error estimates for the discretization of control problems governed by partial differential equations; see the pioneering works by Falk [19] and Geveci [21]. However, recently some papers have appeared, providing new methods and ideas. Arada, Casas, and Tröltzsch [1] derived error estimates for the controls in the L^∞ and L^2 norms for distributed control problems. Similar results for an analogous problem, but also including integral state constraints, were obtained by Casas [8]. The case of a Neumann boundary control problem has been studied by Casas, Mateos, and Tröltzsch [11]. The novelty of our paper with respect to the previous ones is twofold. First, here we deal with a Dirichlet problem, the control being the value of the state on the boundary. Second, we consider piecewise linear continuous functions to approximate the optimal control, which is necessary because of the Dirichlet nature of the control, but it introduces some new difficulties. In the previous papers the controls were always approximated by piecewise constant functions. In the present situation we have developed new methods, which can be used in the framework of distributed or Neumann controls to consider piecewise linear approximations. This could lead to better error estimates than those deduced for piecewise controls.

As far as we know, there is another paper dealing with the numerical approximation of a Dirichlet control problem of Navier–Stokes equations, by Gunzburger,

*Received by the editors March 11, 2005; accepted for publication (in revised form) May 1, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/sicon/45-5/62660.html

†Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39071 Santander, Spain (eduardo.casas@unican.es). This author was supported by the Ministerio de Educación y Ciencia (Spain).

‡Laboratoire MIP, UMR CNRS 5640, Université Paul Sabatier, 31062 Toulouse Cedex 4, France (raymond@mip.ups-tlse.fr).

Hou, and Svobodny [23]. Their procedure of proof does not work when the controls are subject to bound constraints, as considered in our problem. To deal with this difficulty we assume that sufficient second order optimality conditions are satisfied. We also see that the gap between the necessary and sufficient optimality conditions is very narrow. It is of the same type as in the finite dimensional case.

Let us mention some recent papers which provide some new ideas for deriving optimal error estimates. Hinze [26] suggested discretizing the state equation but not the control space. In some cases, including the case of semilinear equations, it is possible to solve the incompletely discretized problem on a computer. However, we believe this process offers no advantages for our problem because the discretization of the states forces the discretization of the controls. Another idea, due to Meyer and Rösch [33], works for linear-quadratic control problems in the distributed case, but we do not know if it is possible to adapt it to the general case.

In the case of parabolic problems, the theory is far from being complete, but some research has been carried out; see Knowles [27], Lasiecka [28], [29], McKnight and Bosarge [32], Tiba and Tröltzsch [36], and Tröltzsch [38], [39], [40], [41].

In the context of control problems of ordinary differential equations, great work has been done by Hager [24], [25] and Dontchev and Hager [16], [17]; see also the work by Malanowski, Büskens, and Maurer [31]. The reader is also referred to the detailed bibliography in [17].

The plan of the paper is as follows. In section 2 we set the optimal control problem and we establish the results we need for the state equation. In section 3 we write the first and second order optimality conditions. The first order conditions allow us to deduce some regularity results of the optimal control, which are necessary to derive the error estimates of the discretization. The second order conditions are also essential to prove the error estimates. The discrete optimal control problem is formulated in section 4 and the first order optimality conditions are given. To write these conditions we have defined a discrete normal derivative for piecewise linear functions, which are solutions of some discrete equation. Sections 6 and 7 are devoted to the analysis of the convergence of the solutions of the discrete optimal control problems and to the proof of error estimates. The main result is Theorem 7.1, where we establish $\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} = O(h^{1-1/p})$.

The numerical tests we have performed confirm our theoretical estimates. For a detailed report we refer to [12]. A simple example is reported in section 8.

2. The control problem. Throughout this paper, Ω denotes an open convex bounded polygonal set of $\mathbb{R}^2$, and Γ denotes its boundary. In this domain we formulate the following control problem:

$$
(\mathrm{P}) \begin{cases} \inf \quad J(u) = \displaystyle\int_\Omega L(x, y_u(x))\,dx \; + \frac{N}{2} \int_\Gamma u^2(x)\,dx \\ \qquad \text{subject to } (y_u, u) \in L^\infty(\Omega) \times L^\infty(\Gamma), \\ u \in U^{ad} = \{u \in L^\infty(\Gamma) \mid \alpha \le u(x) \le \beta \;\; \text{a.e. } x \in \Gamma\}, \\ (y_u, u) \;\; \text{satisfying the state equation (2.1)}, \end{cases}
$$

$$
-\Delta y_u(x) = f(x, y_u(x)) \;\; \text{in } \Omega, \qquad y_u(x) = u(x) \;\; \text{on } \Gamma, \tag{2.1}
$$

where $-\infty < \alpha < \beta < +\infty$ and $N > 0$. Here u is the control, while y_u is the associated state. The following hypotheses are assumed about the functions involved in the control problem (P).

(A1) The function $L : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is measurable with respect to the first component and is of class C^2 with respect to the second one, $L(\cdot, 0) \in L^1(\Omega)$, and for all $M > 0$ there exist a function $\psi_{L,M} \in L^{\bar{p}}(\Omega)$ ($\bar{p} > 2$) and a constant $C_{L,M} > 0$ such that

$$\left|\frac{\partial L}{\partial y}(x,y)\right| \le \psi_{L,M}(x), \quad \left|\frac{\partial^2 L}{\partial y^2}(x,y)\right| \le C_{L,M},$$
$$\left|\frac{\partial^2 L}{\partial y^2}(x,y_2) - \frac{\partial^2 L}{\partial y^2}(x,y_1)\right| \le C_{L,M}|y_2 - y_1|,$$

for a.e. $x \in \Omega$ and $|y|, |y_i| \le M$, $i = 1, 2$.

(A2) The function $f : \Omega \times \mathbb{R} \longrightarrow \mathbb{R}$ is measurable with respect to the first variable and is of class C^2 with respect to the second one,

$$f(\cdot, 0) \in L^{\bar{p}}(\Omega) \ \ (\bar{p} > 2), \quad \frac{\partial f}{\partial y}(x,y) \le 0 \ \text{ a.e. } x \in \Omega \text{ and } y \in \mathbb{R}.$$

For all $M > 0$ there exists a constant $C_{f,M} > 0$ such that

$$\left|\frac{\partial f}{\partial y}(x,y)\right| + \left|\frac{\partial^2 f}{\partial y^2}(x,y)\right| \le C_{f,M} \text{ a.e. } x \in \Omega \text{ and } |y| \le M,$$
$$\left|\frac{\partial^2 f}{\partial y^2}(x,y_2) - \frac{\partial^2 f}{\partial y^2}(x,y_1)\right| < C_{f,M}|y_2 - y_1| \ \text{ a.e. } x \in \Omega \text{ and } |y_1|, |y_2| \le M.$$

Let us finish this section by proving that problem (P) is well defined. We will say that an element $y_u \in L^\infty(\Omega)$ is a solution of (2.1) if

(2.2)
$$\int_\Omega -\Delta w\, y\, dx = \int_\Omega f(x, y(x))w(x)dx - \int_\Gamma u(x)\partial_\nu w(x)dx, \quad \forall w \in H^2(\Omega) \cap H_0^1(\Omega),$$

where ∂_ν denotes the normal derivative on the boundary Γ. This is the classical definition in the transposition sense. To study (2.1), we state an estimate for the linear equation

$$-\Delta z(x) = b(x)z(x) \ \text{ in } \Omega, \qquad z(x) = u(x) \ \text{ on } \Gamma, \tag{2.3}$$

where b is a nonpositive function belonging to $L^\infty(\Omega)$.

LEMMA 2.1. *For every $u \in L^\infty(\Gamma)$, the linear equation* (2.3) *has a unique solution $z \in L^\infty(\Omega)$ (defined in the transposition sense), and it satisfies*

(2.4)
$$\|z\|_{L^2(\Omega)} \le C\|u\|_{H^{-1/2}(\Gamma)}, \ \|z\|_{H^{1/2}(\Omega)} \le C\|u\|_{L^2(\Gamma)} \text{ and } \|z\|_{L^\infty(\Omega)} \le \|u\|_{L^\infty(\Gamma)}.$$

The proof is standard: the first inequality is obtained by using the transposition method (see Lions and Magenes [30]), the second inequality is deduced by interpolation, and the last one is obtained by applying the maximum principle.

THEOREM 2.2. *For every $u \in L^\infty(\Gamma)$, the state equation* (2.1) *has a unique solution $y_u \in L^\infty(\Omega) \cap H^{1/2}(\Omega)$. Moreover the following Lipschitz properties hold:*

$$\begin{aligned} &\|y_u - y_v\|_{L^\infty(\Omega)} \le \|u - v\|_{L^\infty(\Gamma)}, \\ &\|y_u - y_v\|_{H^{1/2}(\Omega)} \le C\|u - v\|_{L^2(\Gamma)} \quad \forall u, v \in L^\infty(\Gamma). \end{aligned} \tag{2.5}$$

Finally if $u_n \rightharpoonup u$ *weakly** *in* $L^\infty(\Gamma)$, *then* $y_{u_n} \to y_u$ *strongly in* $L^r(\Omega)$ *for all* $r < +\infty$.

Proof. Let us introduce the following problems:

$$-\Delta z = 0 \ \text{ in } \Omega, \qquad z = u \ \text{ on } \Gamma, \tag{2.6}$$

and

$$-\Delta \zeta = g(x,\zeta) \ \text{ in } \Omega, \qquad \zeta = 0 \ \text{ on } \Gamma, \tag{2.7}$$

where $g : \Omega \times \mathbb{R} \mapsto \mathbb{R}$ is given by $g(x,t) = f(x, z(x) + t)$, with z being the solution of (2.6). Lemma 2.1 implies that (2.6) has a unique solution in $L^\infty(\Omega) \cap H^{1/2}(\Omega)$. It is obvious that assumption (A2) is fulfilled by g and that (2.7) is a classically well-set problem having a unique solution in $H_0^1(\Omega) \cap L^\infty(\Omega)$. Moreover, since Ω is convex, we know that $\zeta \in H^2(\Omega)$; see Grisvard [22]. Finally the solution y_u of (2.1) can be written as $y_u = z + \zeta$. Estimates (2.5) follow from Lemma 2.1; see Arada and Raymond [2] for a detailed proof in the parabolic case. The continuous dependence in $L^r(\Omega)$ follows in a standard way by using (2.5) and the compactness of the inclusion $H^{1/2}(\Omega) \subset L^2(\Omega)$ along with the fact that $\{y_{u_n}\}$ is bounded in $L^\infty(\Omega)$, as deduced from the first inequality of (2.5). ∎

Now the following theorem can be proved by standard arguments.

THEOREM 2.3. *Problem* (P) *has at least one solution.*

3. Optimality conditions. Before writing the optimality conditions for (P) let us state the differentiability properties of J.

THEOREM 3.1. *The mapping* $G : L^\infty(\Gamma) \longrightarrow L^\infty(\Omega) \cap H^{1/2}(\Omega)$ *defined by* $G(u) = y_u$ *is of class* C^2. *Moreover, for all* $u, v \in L^\infty(\Gamma)$, $z_v = G'(u)v$ *is the solution of*

$$-\Delta z_v = \frac{\partial f}{\partial y}(x, y_u) z_v \ \text{ in } \Omega, \qquad z_v = v \ \text{ on } \Gamma, \tag{3.1}$$

and for every $v_1, v_2 \in L^\infty(\Omega)$, $z_{v_1 v_2} = G''(u)v_1 v_2$ *is the solution of*

$$\begin{cases} -\Delta z_{v_1 v_2} &= \dfrac{\partial f}{\partial y}(x, y_u) z_{v_1 v_2} + \dfrac{\partial^2 f}{\partial y^2}(x, y_u) z_{v_1} z_{v_2} & \text{in} \quad \Omega, \\ z_{v_1 v_2} &= 0 & \text{on} \quad \Gamma, \end{cases} \tag{3.2}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$.

Proof. Let us define the space

$$V = \{ y \in H^{1/2}(\Omega) \cap L^\infty(\Omega) : \Delta y \in L^2(\Omega) \}$$

endowed with the natural graph norm. Now we consider the function $F : L^\infty(\Gamma) \times V \longrightarrow L^\infty(\Gamma) \times L^2(\Omega)$ defined by $F(u,y) = (y|_\Gamma - u, \Delta y + f(x,y))$. It is obvious that F is of class C^2 and that for every pair (u,y) satisfying (2.1) we have $F(u,y) = (0,0)$. Furthermore

$$\frac{\partial F}{\partial y}(u,y) \cdot z = \left(z|_\Gamma, \Delta z + \frac{\partial f}{\partial y}(x,y) z \right).$$

By using Lemma 2.1 we deduce that $(\partial F/\partial y)(u,y) : V \longrightarrow L^\infty(\Gamma) \times L^2(\Omega)$ is an isomorphism. Then the implicit function theorem allows us to conclude that G is of class C^2, and now the rest of the theorem follows easily. ∎

Theorem 3.1, along with the chain rule, leads to the following result.

THEOREM 3.2. *The functional $J : L^\infty(\Gamma) \to \mathbb{R}$ is of class C^2. Moreover, for every $u, v, v_1, v_2 \in L^\infty(\Gamma)$,*

$$J'(u)v = \int_\Gamma (Nu - \partial_\nu \phi_u)\, v\, dx \tag{3.3}$$

and

$$J''(u)v_1v_2 = \int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x, y_u)z_{v_1}z_{v_2} + \phi_u \frac{\partial^2 f}{\partial y^2}(x, y_u)z_{v_1}z_{v_2}\right] dx + \int_\Gamma N v_1 v_2\, dx, \tag{3.4}$$

where $z_{v_i} = G'(u)v_i$, $i = 1, 2$, $y_u = G(u)$, and the adjoint state $\phi_u \in H^2(\Omega)$ is the unique solution of the problem

$$-\Delta\phi = \frac{\partial f}{\partial y}(x, y_u)\phi + \frac{\partial L}{\partial y}(x, y_u) \quad \text{in } \Omega, \qquad \phi = 0 \quad \text{on } \Gamma. \tag{3.5}$$

The first order optimality conditions for problem (P) follow readily from Theorem 3.2.

THEOREM 3.3. *Assume that $\bar u$ is a local solution of problem* (P) *and let $\bar y$ be the corresponding state. Then there exists $\bar\phi \in H^2(\Omega)$ such that*

$$-\Delta\bar\phi = \frac{\partial f}{\partial y}(x, \bar y)\bar\phi + \frac{\partial L}{\partial y}(x, \bar y) \text{ in } \Omega, \qquad \bar\phi = 0 \text{ on } \Gamma, \tag{3.6}$$

and

$$\int_\Gamma \left(N\bar u - \partial_\nu \bar\phi\right)(u - \bar u)\, dx \ge 0, \quad \forall u \in U^{ad}, \tag{3.7}$$

which is equivalent to

$$\bar u(x) = \mathrm{Proj}_{[\alpha,\beta]}\Big(\frac{1}{N}\partial_\nu\bar\phi(x)\Big) = \max\Big\{\alpha, \min\Big\{\beta, \frac{1}{N}\partial_\nu\bar\phi(x)\Big\}\Big\}. \tag{3.8}$$

THEOREM 3.4. *Assume that $\bar u$ is a local solution of problem* (P) *and let $\bar y$ and $\bar\phi$ be the corresponding state and adjoint state. Then there exists $p \in (2, \bar p]$ (with $\bar p > 2$ as introduced in assumptions* (A1) *and* (A2)*) depending on the measure of the angles of the polygon Ω such that $\bar y \in W^{1,p}(\Omega)$, $\bar\phi \in W^{2,p}(\Omega)$, and $\bar u \in W^{1-1/p,p}(\Gamma) \subset C(\Gamma)$.*

Proof. From assumption (A1) and using elliptic regularity results, it follows that $\bar\phi$ belongs to $W^{2,p}(\Omega)$ for some $p \in (2, \bar p]$ depending on the measure of the angles of Γ; see Grisvard [22, Chapter 4]. To prove that $\bar u$ belongs to $W^{1-1/p,p}(\Gamma)$ we recall the norm in this space,

$$\|\bar u\|_{W^{1-1/p,p}(\Gamma)} = \left\{\int_\Gamma |\bar u(x)|^p dx + \int_\Gamma\int_\Gamma \frac{|\bar u(x) - \bar u(\xi)|^p}{|x - \xi|^p}\, dx\, d\xi\right\}^{1/p},$$

where we have used the fact that $\Omega \subset \mathbb{R}^2$. Due to [22, Theorem 1.5.2.3] and the fact that $\bar\phi = 0$ on Γ, it can be shown that $\partial_\nu\bar\phi$ belongs to $W^{1-1/p,p}(\Gamma)$. With the relation (3.8) and

$$\left|\mathrm{Proj}_{[\alpha,\beta]}\left(\frac{1}{N}\partial_\nu\bar\phi(x)\right) - \mathrm{Proj}_{[\alpha,\beta]}\left(\frac{1}{N}\partial_\nu\bar\phi(\xi)\right)\right| \le \frac{1}{N}|\partial_\nu\bar\phi(x) - \partial_\nu\bar\phi(\xi)|,$$

one can prove that the integrals in the above norm are finite.

Finally, decomposing (2.1) into two problems as in the proof of Theorem 2.3, we get that $\bar{y} = \bar{z} + \bar{\zeta}$, with $\bar{\zeta} \in H^2(\Omega)$ and $\bar{z} \in W^{1,p}(\Omega)$, which completes the proof. □

In order to establish the second order optimality conditions, we define the cone of critical directions

$$C_{\bar{u}} = \{v \in L^2(\Gamma) \text{ satisfying (3.9) and } v(x) = 0 \text{ if } |\bar{d}(x)| > 0\},$$

$$v(x) = \begin{cases} \geq 0 & \text{where } \bar{u}(x) = \alpha \\ \leq 0 & \text{where } \bar{u}(x) = \beta \end{cases} \quad \text{for a.e. } x \in \Gamma, \tag{3.9}$$

where $\bar{d}$ denotes the derivative $J'(\bar{u})$,

$$\bar{d}(x) = N\bar{u}(x) - \partial_\nu \bar{\phi}(x). \tag{3.10}$$

Now we formulate the second order necessary and sufficient optimality conditions.

THEOREM 3.5. *If $\bar{u}$ is a local solution of* (P), *then $J''(\bar{u})v^2 \geq 0$ holds for all $v \in C_{\bar{u}}$. Conversely, if $\bar{u} \in U^{ad}$ satisfies the first order optimality conditions provided by Theorem* 3.3 *and the coercivity condition*

$$J''(\bar{u})v^2 > 0, \quad \forall v \in C_{\bar{u}} \setminus \{0\}, \tag{3.11}$$

then there exist $\mu > 0$ and $\varepsilon > 0$ such that $J(u) \geq J(\bar{u}) + \mu\|u - \bar{u}\|^2_{L^2(\Gamma)}$ is satisfied for every $u \in U^{ad}$ obeying $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$.

The necessary condition provided in the theorem is quite easy to get. The sufficient conditions are proved by Casas and Mateos [9, Theorem 4.3] for distributed control problems with integral state constraints. The proof can be translated in a straightforward way into the case of boundary controls; see also Bonnans and Zidani [4].

Remark 3.6. It can be proved (see Casas and Mateos [9, Theorem 4.4]) that the following two conditions are equivalent:

(1) $J''(\bar{u})v^2 > 0$ for every $v \in C_{\bar{u}} \setminus \{0\}$.

(2) There exist $\delta > 0$ and $\tau > 0$ such that $J''(\bar{u})v^2 \geq \delta\|v\|^2_{L^2(\Gamma)}$ for every $v \in C^\tau_{\bar{u}}$, where

$$C^\tau_{\bar{u}} = \{v \in L^2(\Gamma) \text{ satisfying (3.9) and } v(x) = 0 \text{ if } |\bar{d}(x)| > \tau\}.$$

It is clear that $C^\tau_{\bar{u}}$ contains strictly $C_{\bar{u}}$, so condition (2) seems to be stronger than (1), but in fact they are equivalent. For the proof of this equivalence, we use the fact that u appears linearly in the state equation and quadratically in the cost functional.

4. Numerical approximation of (P). Let us consider a family of triangulations $\{\mathcal{T}_h\}_{h>0}$ of $\bar{\Omega}$: $\bar{\Omega} = \cup_{T\in\mathcal{T}_h} T$. With each element $T \in \mathcal{T}_h$, we associate two parameters $\rho(T)$ and $\sigma(T)$, where $\rho(T)$ denotes the diameter of the set T, and $\sigma(T)$ is the diameter of the largest ball contained in T. Let us define the size of the mesh by $h = \max_{T\in\mathcal{T}_h} \rho(T)$. For fixed $h > 0$, we denote by $\{T_j\}_{j=1}^{N(h)}$ the family of triangles belonging to $\mathcal{T}_h$ and having a side included in the boundary Γ. If the vertices of $T_j \cap \Gamma$ are x^j_Γ and x^{j+1}_Γ, then $[x^j_\Gamma, x^{j+1}_\Gamma] := T_j \cap \Gamma$, $1 \leq j \leq N(h)$, with $x^{N(h)+1}_\Gamma = x^1_\Gamma$. We will also follow the notation $x^0_\Gamma = x^{N(h)}_\Gamma$. We assume that every vertex of the polygon Ω is one of these boundary points x^j_Γ of the triangulation and that the numbering of

the nodes $\{x_\Gamma^j\}_{j=1}^{N(h)}$ is made counterclockwise. The length of the interval $[x_\Gamma^j, x_\Gamma^{j+1}]$ is denoted by $h_j = |x_\Gamma^{j+1} - x_\Gamma^j|$. The following hypotheses on the triangulation are also assumed.

(H1) There exists a constant $\rho > 0$ such that $h/\rho(T) \leq \rho$ for all $T \in \mathcal{T}_h$ and $h > 0$.

(H2) All the angles of all triangles are less than or equal to $\pi/2$.

The first assumption is not a restriction in practice and it is the usual one. The second assumption is going to allow us to use the discrete maximum principle and it is actually not too restrictive.

Given two points ξ_1 and ξ_2 of Γ, we denote by $[\xi_1, \xi_2]$ the part of Γ obtained by running the boundary from ξ_1 to ξ_2 counterclockwise. With this convention we have $(\xi_2, \xi_1) = \Gamma \setminus [\xi_1, \xi_2]$. According to this notation,

$$\int_{\xi_1}^{\xi_2} u(x)\,dx \quad \text{and} \quad \int_{\xi_2}^{\xi_1} u(x)\,dx$$

denote the integrals of a function $u \in L^1(\Gamma)$ on the parts of Γ defined by $[\xi_1, \xi_2]$ and $[\xi_2, \xi_1]$, respectively. In particular we have

$$\int_{\xi_1}^{\xi_2} u(x)\,dx = \int_\Gamma u(x)\,dx - \int_{\xi_2}^{\xi_1} u(x)\,dx.$$

Associated with this triangulation, we consider the sets

$$U_h = \Big\{u_h \in C(\Gamma) : u_h|_{[x_\Gamma^j, x_\Gamma^{j+1}]} \in \mathcal{P}_1 \text{ for } 1 \leq j \leq N(h)\Big\},$$
$$Y_h = \Big\{y_h \in C(\bar\Omega) : y_h|_T \in \mathcal{P}_1 \ \forall\, T \in \mathcal{T}_h\Big\},$$
$$Y_{h0} = \Big\{y_h \in Y_h : y_h|_\Gamma = 0\Big\},$$

where $\mathcal{P}_1$ is the space of polynomials of degree less than or equal to 1. The space U_h is formed by the restrictions to Γ of the functions of Y_h.

Let us consider the projection operator $\Pi_h : L^2(\Gamma) \longmapsto U_h$,

$$(\Pi_h v, u_h)_{L^2(\Gamma)} = (v, u_h)_{L^2(\Gamma)} \quad \forall u_h \in U_h.$$

The following approximation property of Π_h is well known (see for instance [20, Lemma 3.1]):

$$\|y - \Pi_h y\|_{L^2(\Gamma)} + h^{1/2}\|y - \Pi_h y\|_{H^{1/2}(\Gamma)} \leq Ch^{s-1/2}\|y\|_{H^s(\Omega)}, \quad \forall y \in H^s(\Omega),$$

and for every $1 \leq s \leq 2$. Observing that, for $1/2 < s \leq 3/2$,

$$u \longmapsto \inf_{y|_\Gamma = u} \|y\|_{H^s(\Omega)}$$

is a norm equivalent to the usual one of $H^{s-1/2}(\Gamma)$, we deduce from the above inequality that

$$\|u - \Pi_h u\|_{L^2(\Gamma)} + h^{1/2}\|u - \Pi_h u\|_{H^{1/2}(\Gamma)} \leq Ch^s\|u\|_{H^s(\Gamma)}, \quad \forall u \in H^s(\Gamma), \tag{4.1}$$

and for every $1/2 < s \leq 3/2$.

Let $a : Y_h \times Y_h \longmapsto \mathbb{R}$ be the bilinear form given by

$$a(y_h, z_h) = \int_\Omega \nabla y_h(x) \nabla z_h(x)\, dx.$$

For all $u \in L^\infty(\Gamma)$, we consider the following problem:

$$\text{(4.2)} \qquad \begin{cases} \text{Find } y_h(u) \in Y_h, \text{ such that } y_h = \Pi_h u \text{ on } \Gamma, \text{ and} \\ a(y_h(u), w_h) = \displaystyle\int_\Omega f(x, y_h(u)) w_h\, dx \quad \forall w_h \in Y_{h0}. \end{cases}$$

PROPOSITION 4.1. *For every $u \in L^\infty(\Gamma)$, (4.2) admits a unique solution $y_h(u)$.*

Proof. Let z_h be the unique element in Y_h satisfying $z_h = \Pi_h u$ on Γ, and let $z_h(x_i) = 0$ for all vertices x_i of the triangulation $\mathcal{T}_h$ not belonging to Γ. The equation

$$\zeta_h \in Y_{h0}, \quad a(\zeta_h, w_h) = -a(z_h, w_h) + \int_\Omega f(x, z_h + \zeta_h) w_h\, dx, \quad \forall w_h \in Y_{h0},$$

admits a unique solution (it is a consequence of the Minty–Browder theorem; see Brézis [7]). The function $z_h + \zeta_h$ is clearly a solution of (4.2). The uniqueness of the solution to (4.2) also follows from the Minty–Browder theorem. □

Due to Proposition 4.1, we can define a functional J_h in $L^\infty(\Gamma)$ by

$$J_h(u) = \int_\Omega L(x, y_h(u)(x))\, dx + \frac{N}{2} \int_\Gamma u^2(x)\, dx.$$

The finite dimensional control problem approximating (P) is

$$(\mathrm{P}_h) \begin{cases} \min J_h(u_h) = \displaystyle\int_\Omega L(x, y_h(u_h)(x))\, dx + \frac{N}{2} \int_\Gamma u_h^2(x)\, dx \\ \text{subject to } u_h \in U_h^{ad}, \end{cases}$$

where

$$U_h^{ad} = U_h \cap U^{ad} = \{u_h \in U_h \mid \alpha \le u_h(x) \le \beta \ \forall\ x \in \Gamma\}.$$

The existence of a solution of (P_h) follows from the continuity of J_h in U_h and the fact that U_h^{ad} is a nonempty compact subset of U_h. Our next goal is to write the conditions for optimality satisfied by any local solution $\bar{u}_h$. First, we have to obtain an expression for the derivative of $J_h : L^\infty(\Gamma) \to \mathbb{R}$ analogous to the one of J given by formula (3.3). Given $u \in L^\infty(\Gamma)$ we consider the adjoint state $\phi_h(u) \in Y_{h0}$ solution of the equation

$$\text{(4.3)} \qquad a(w_h, \phi_h(u)) = \int_\Omega \left[\frac{\partial f}{\partial y}(x, y_h(u)) \phi_h(u) + \frac{\partial L}{\partial y}(x, y_h(u)) \right] w_h\, dx \quad \forall w_h \in Y_{h0}.$$

To obtain the analogous expression to (3.3) we have to define a discrete normal derivative $\partial_\nu^h \phi_h(u)$.

PROPOSITION 4.2. *Let u belong to $L^\infty(\Gamma)$ and let $\phi_h(u)$ be the solution of* (4.3). *There exists a unique element $\partial^h_\nu \phi_h(u) \in U_h$ verifying*

$$
\begin{aligned}
&(\partial^h_\nu \phi_h(u), w_h)_{L^2(\Gamma)} = a(w_h, \phi_h(u)) \\
&- \int_\Omega \left[\frac{\partial f}{\partial y}(x, y_h(u))\phi_h(u) + \frac{\partial L}{\partial y}(x, y_h(u)) \right] w_h \, dx \quad \forall w_h \in Y_h.
\end{aligned} \tag{4.4}
$$

Proof. The trace mapping is a surjective mapping from Y_h on U_h; therefore the linear form

$$
L(w_h) = a(w_h, \phi_h(u)) - \int_\Omega \left[\frac{\partial f}{\partial y}(x, y_h(u))\phi_h(u) + \frac{\partial L}{\partial y}(x, y_h(u)) \right] w_h \, dx
$$

is well defined on U_h, and it is continuous on U_h. Let us remark that if in (4.4) the trace of w_h on Γ is zero, then (4.3) leads to

$$
L(w_h) = 0.
$$

Hence L can be identified with a unique element of U_h, which proves the above proposition. □

Now the function G introduced in Theorem 3.1 is approximated by the function $G_h : L^\infty(\Gamma) \longmapsto Y_h$ defined by $G_h(u) = y_h(u)$. We can easily verify that G_h is of class C^2 and that for $u, v \in L^\infty(\Gamma)$, the derivative $z_h = G_h'(u)v \in Y_h$ is the unique solution of

$$
\begin{cases}
a(z_h, w_h) = \displaystyle\int_\Omega \frac{\partial f}{\partial y}(x, y_h(u)) z_h w_h \, dx, \quad \forall w_h \in Y_{h0}, \\
z_h = \Pi_h v \quad \text{on } \Gamma.
\end{cases} \tag{4.5}
$$

From here we deduce

$$
J_h'(u)v = \int_\Omega \frac{\partial L}{\partial y}(x, y_h(u)) z_h \, dx + N \int_\Gamma uv \, dx.
$$

Now (4.4) and the definition of Π_h lead to

$$
J_h'(u)v = N \int_\Gamma uv \, dx - \int_\Gamma \partial^h_\nu \phi_h(u) \Pi_h v \, dx = \int_\Gamma (Nu - \partial^h_\nu \phi_h(u)) v \, dx \tag{4.6}
$$

for all $u, v \in L^\infty(\Gamma)$.

Finally, we can write the first order optimality conditions.

THEOREM 4.3. *Let us assume that $\bar u_h$ is a local solution of* (P_h) *and $\bar y_h$ the corresponding state; then there exists $\bar\phi_h \in Y_{h0}$ such that*

$$
a(w_h, \bar\phi_h) = \int_\Omega \left[\frac{\partial f}{\partial y}(x, \bar y_h)\bar\phi_h + \frac{\partial L}{\partial y}(x, \bar y_h) \right] w_h \, dx \quad \forall w_h \in Y_{h0} \tag{4.7}
$$

and

$$
\int_\Gamma (N\bar u_h - \partial^h_\nu \bar\phi_h)(u_h - \bar u_h) \, dx \geq 0 \quad \forall u_h \in U_h^{ad}. \tag{4.8}
$$

This theorem follows readily from (4.6).

Remark 4.4. The reader could think that a projection property for $\bar{u}_h$ similar to that obtained for $\bar{u}$ in (3.8) can be deduced from (4.8). Unfortunately this property does not hold because $u_h(x)$ cannot be taken arbitrarily in $[\alpha, \beta]$. Functions $u_h \in U_h$ are determined by their values at the nodes $\{x_\Gamma^j\}_{j=1}^{N(h)}$. If we consider the basis of U_h $\{e_j\}_{j=1}^{N(h)}$ defined by $e_j(x_\Gamma^i) = \delta_{ij}$, then we have

$$u_h = \sum_{j=1}^{N(h)} u_{h,j} e_j \quad \text{with} \quad u_{h,j} = u_h(x_\Gamma^j), \quad 1 \le j \le N(h).$$

Now (4.8) can be written

$$\sum_{j=1}^{N(h)} \int_\Gamma (N\bar{u}_h - \partial_\nu^h \bar{\phi}_h) e_j \, dx (u_{h,j} - \bar{u}_{h,j}) \ge 0, \quad \forall \{u_{h,j}\}_{j=1}^{N(h)} \subset [\alpha, \beta], \tag{4.9}$$

where $\bar{u}_{h,j} = \bar{u}_h(x_\Gamma^j)$. Then (4.9) leads to

$$\bar{u}_{h,j} = \begin{cases} \alpha & \text{if} \quad \int_\Gamma (N\bar{u}_h - \partial_\nu^h \bar{\phi}_h) e_j \, dx > 0, \\ \beta & \text{if} \quad \int_\Gamma (N\bar{u}_h - \partial_\nu^h \bar{\phi}_h) e_j \, dx < 0. \end{cases} \tag{4.10}$$

In order to characterize $\bar{u}_h$ as the projection of $\partial_\nu^h \bar{\phi}_h / N$, let us introduce the operator $\mathrm{Proj}_h : L^2(\Gamma) \longmapsto U_h^{ad}$ as follows. Given $u \in L^2(\Gamma)$, $\mathrm{Proj}_h u$ denotes the unique solution of the problem

$$\inf_{v_h \in U_h^{ad}} \|u - v_h\|_{L^2(\Gamma)},$$

which is characterized by the relation

$$\int_\Gamma (u(x) - \mathrm{Proj}_h u(x))(v_h(x) - \mathrm{Proj}_h u(x)) \, dx \le 0 \quad \forall v_h \in U_h^{ad}. \tag{4.11}$$

Then (4.8) is equivalent to

$$\bar{u}_h = \mathrm{Proj}_h \left(\frac{1}{N} \partial_\nu^h \bar{\phi}_h \right). \tag{4.12}$$

Let us recall the result in [13, Lemma 3.3], where a characterization of $\mathrm{Proj}_h(u_h)$ is stated. Given $u_h \in U_h$ and $\bar{u}_h = \mathrm{Proj}_h(u_h)$, $\bar{u}_h$ is then characterized by the inequalities

$$\begin{aligned} &h_{j-1}[(u_{h,j-1} - \bar{u}_{h,j-1}) + 2(u_{h,j} - \bar{u}_{h,j})](t - \bar{u}_{h,j}) \\ &\quad + h_j[2(u_{h,j} - \bar{u}_{h,j}) + (u_{h,j+1} - \bar{u}_{h,j+1})](t - \bar{u}_{h,j}) \le 0 \end{aligned}$$

for all $t \in [\alpha, \beta]$ and $1 \le j \le N(h)$.

5. Numerical analysis of the state and adjoint equations. Throughout the following, the operator $I_h \in \mathcal{L}(W^{1,p}(\Omega), Y_h)$ denotes the classical interpolation operator [6]. We also need the interpolation operator $I_h^\Gamma \in \mathcal{L}(W^{1-1/p,p}(\Gamma), U_h)$. Since we have

$$I_h^\Gamma(y|_\Gamma) = (I_h y)|_\Gamma, \quad \forall y \in W^{1,p}(\Omega),$$

we shall use the same notation for both interpolation operators. The reader can observe that this abuse of notation does not lead to any confusion.

The goal of this section is to obtain the error estimates of the approximations $y_h(u)$ given by (4.2) to the solution y_u of (2.1). In order to carry out this analysis we decompose (2.1) into two problems, as in the proof of Theorem 2.3. We take $z \in H^{1/2}(\Omega) \cap L^\infty(\Omega)$ and $\zeta \in H_0^1(\Omega) \cap H^2(\Omega)$ as the solutions of (2.6) and (2.7), respectively. Then we have $y_u = z + \zeta$.

Let us consider now the discretizations of (2.6) and (2.7):

$$\begin{cases} \text{Find } z_h \in \ Y_h \text{ such that } \ z_h = \Pi_h u \text{ on } \Gamma \text{ and} \\ a(z_h, w_h) = 0, \quad \forall w_h \in Y_{h0}, \end{cases} \tag{5.1}$$

and

$$\begin{cases} \text{Find } \zeta_h \in Y_{h0} \text{ such that} \\ a(\zeta_h, w_h) = \displaystyle\int_\Omega g_h(x, \zeta_h(x)) w_h(x)\, dx, \quad \forall w_h \in Y_{h0}, \end{cases} \tag{5.2}$$

where $g_h(x,t) = f(x, z_h(x) + t)$. Now the solution $y_h(u)$ of (4.2) is decomposed as follows: $y_h(u) = z_h + \zeta_h$. The following lemma provides the estimates for $z - z_h$.

LEMMA 5.1. *Let $u \in U^{ad}$, and let z and z_h be the solutions of* (2.6) *and* (5.1), *respectively; then*

$$\|z_h\|_{L^\infty(\Omega)} \le \|\Pi_h u\|_{L^\infty(\Gamma)} \le C(\alpha,\beta) \text{ and } \|z_h\|_{W^{1,r}(\Omega)} \le C\|\Pi_h u\|_{W^{1-1/r,r}(\Gamma)}, \tag{5.3}$$

$$\|z_h\|_{L^2(\Omega)} \le C\|\Pi_h u\|_{H^{-1/2}(\Gamma)}, \tag{5.4}$$

where $1 < r \le p$ is arbitrary, with p being as given in Theorem 3.4. *If, in addition, $u \in H^s(\Gamma) \cap U^{ad}$, with $0 \le s \le 1$, then we also have*

$$\|z - z_h\|_{L^2(\Omega)} \le Ch^{s+1/2}\|u\|_{H^s(\Gamma)} \quad \forall h > 0 \text{ and } 0 \le s \le 1. \tag{5.5}$$

Proof. The first inequality of (5.3) is proved in Ciarlet and Raviart [14]; we have only to notice that

$$\|\Pi_h u\|_{L^\infty(\Gamma)} \le C\|u\|_{L^\infty(\Gamma)} \le C(\alpha,\beta), \tag{5.6}$$

where C is independent of h and $u \in U^{ad}$; see Douglas, Dupont, and Wahlbin [18].

Inequality (5.5) can be found in French and King [20, Lemma 3.3] by just taking into account that

$$\|z\|_{H^{s+1/2}(\Omega)} \le C\|u\|_{H^s(\Gamma)}.$$

The second inequality of (5.3) is established in Bramble, Pasciak, and Schatz [5, Lemma 3.2] for $r = 2$. Let us prove it for all r in the range $(1, p]$. Let us consider the $z^h \in H^1(\Omega)$ solution of the problem

$$-\Delta z^h = 0 \quad \text{in } \Omega, \qquad z^h = \Pi_h u \quad \text{on } \Gamma.$$

This is a standard Dirichlet problem with the property (see Dauge [15])

$$\|z^h\|_{W^{1,r}(\Omega)} \le C\|\Pi_h u\|_{W^{1-1/r,r}(\Gamma)}.$$

Let us denote by $\hat{I}_h : W^{1,r}(\Omega) \longmapsto Y_h$ the generalized interpolation operator, due to Scott and Zhang [35], that preserves piecewise-affine boundary conditions. More

precisely, it has the properties $\hat{I}_h(y_h) = y_h$ for all $y_h \in Y_h$ and $\hat{I}_h(W_0^{1,r}(\Omega)) \subset Y_{h0}$. These properties imply that $\hat{I}_h(z^h) = \Pi_h u$ on Γ. Thus we have

$$-\Delta(z^h - \hat{I}_h(z^h)) = \Delta \hat{I}_h(z^h) \quad \text{in } \Omega, \qquad z^h - \hat{I}_h(z^h) = 0 \quad \text{on } \Gamma,$$

and $z_h - \hat{I}_h(z^h) \in Y_{h0}$ satisfies

$$a(z_h - \hat{I}_h(z^h), w_h) = -a(\hat{I}_h(z^h), w_h) \ \ \forall w_h \in Y_{h0}.$$

Then by using the L^p estimates (see, for instance, Brenner and Scott [6, Theorem 7.5.3]), we get

$$\begin{aligned}&\|z_h - \hat{I}_h(z^h)\|_{W^{1,r}(\Omega)} \le C\|z^h - \hat{I}_h(z^h)\|_{W^{1,r}(\Omega)}\\ &\le C(\|z^h\|_{W^{1,r}(\Omega)} + \|\hat{I}_h(z^h)\|_{W^{1,r}(\Omega)}) \le C\|z^h\|_{W^{1,r}(\Omega)} \le C\|\Pi_h u\|_{W^{1-1/r,r}(\Gamma)}.\end{aligned}$$

Then we conclude the proof as follows:

$$\|z_h\|_{W^{1,r}(\Omega)} \le \|\hat{I}_h(z^h)\|_{W^{1,r}(\Omega)} + \|z_h - \hat{I}_h(z^h)\|_{W^{1,r}(\Omega)} \le C\|\Pi_h u\|_{W^{1-1/r,r}(\Gamma)}.$$

Finally, let us prove (5.4). Using (5.5) with $s = 0$, (2.4), and an inverse inequality, we get

$$\begin{aligned}&\|z_h\|_{L^2(\Omega)} \le \|z^h - z_h\|_{L^2(\Omega)} + \|z^h\|_{L^2(\Omega)}\\ &\le C(h^{1/2}\|\Pi_h u\|_{L^2(\Gamma)} + \|\Pi_h u\|_{H^{-1/2}(\Gamma)}) \le C\|\Pi_h u\|_{H^{-1/2}(\Gamma)}. \qquad \square\end{aligned}$$

Remark 5.2. The inverse estimate used in the proof,

$$\|u\|_{L^2(\Gamma)} \le Ch^{-1/2}\|u\|_{H^{-1/2}(\Gamma)}, \quad \forall u \in U_h,$$

can be derived from the well-known inverse estimate [3],

$$\|u\|_{H^{1/2}(\Gamma)} \le Ch^{-1/2}\|u\|_{L^2(\Gamma)}, \quad \forall u \in U_h,$$

and from the equality

$$\|u\|^2_{L^2(\Gamma)} = \|u\|_{H^{1/2}(\Gamma)}\|u\|_{H^{-1/2}(\Gamma)}.$$

Now we obtain the estimates for $\zeta - \zeta_h$.

LEMMA 5.3. *There exist constants $C_i = C_i(\alpha, \beta) > 0$ $(i = 1, 2)$ such that, for all $u \in U^{ad} \in H^s(\Gamma)$, the following estimates hold:*

$$\|\zeta_h\|_{L^\infty(\Omega)} \le C_1, \quad \forall h > 0 \text{ and } s = 0, \tag{5.7}$$

$$\|\zeta - \zeta_h\|_{L^2(\Omega)} \le C_2 h^{s+1/2}(1 + \|u\|_{H^s(\Gamma)}), \quad \forall h > 0 \text{ and } 0 \le s \le 1, \tag{5.8}$$

where ζ and ζ_h are the solutions of (2.7) *and* (5.2)*, respectively.*

Proof. We are going to introduce an intermediate function $\zeta^h \in H^2(\Omega)$ satisfying

$$-\Delta\zeta^h = g_h(x, \zeta^h(x)) \quad \text{in } \Omega, \qquad \zeta^h = 0 \quad \text{on } \Gamma. \tag{5.9}$$

By using classical methods (see for instance Stampacchia [34]), we get the boundedness of ζ and ζ^h in $L^\infty(\Omega)$ for some constants depending on $\|u\|_{L^\infty(\Gamma)}$ and $\|\Pi_h u\|_{L^\infty(\Gamma)}$,

which are uniformly estimated by a constant depending only on α and β; see (5.6). On the other hand, from (2.7), (5.9), and assumption (A2), we deduce

$$
\begin{aligned}
&C_1\|\zeta-\zeta^h\|^2_{H^1(\Omega)} \le a(\zeta-\zeta^h,\zeta-\zeta^h)\\
&= \int_\Omega [g(x,\zeta(x)) - g_h(x,\zeta^h(x))](\zeta(x)-\zeta^h(x))\,dx\\
&= \int_\Omega [g(x,\zeta(x)) - g(x,\zeta^h(x))](\zeta(x)-\zeta^h(x))\,dx\\
&\quad + \int_\Omega [g(x,\zeta^h(x)) - g_h(x,\zeta^h(x))](\zeta(x)-\zeta^h(x))\,dx\\
&\le \int_\Omega [g(x,\zeta^h(x)) - g_h(x,\zeta^h(x))](\zeta(x)-\zeta^h(x))\,dx \le C_2\|z-z_h\|_{L^2(\Omega)}\|\zeta-\zeta^h\|_{L^2(\Omega)}\\
&\le C_3\|z-z_h\|^2_{L^2(\Omega)} + \frac{C_1}{2}\|\zeta-\zeta^h\|^2_{L^2(\Omega)}.
\end{aligned}
$$

This inequality, along with (5.5), implies

$$
\|\zeta-\zeta^h\|_{H^1(\Omega)} \le Ch^{s+1/2}\|u\|_{H^s(\Gamma)}. \tag{5.10}
$$

Thanks to the convexity of Ω, ζ^h belongs to $H^2(\Omega)$ (see Grisvard [22]) and

$$
\|\zeta^h\|_{H^2(\Omega)} \le C\|g_h(x,\zeta^h)\|_{L^2(\Omega)} = C(\|u\|_{L^\infty(\Gamma)}, \|\Pi_h u\|_{L^\infty(\Gamma)}).
$$

Now using the results of Casas and Mateos [10, Lemma 4 and Theorem 1] we deduce that

$$
\|\zeta^h-\zeta_h\|_{L^2(\Omega)} \le Ch^2, \tag{5.11}
$$

$$
\|\zeta^h-\zeta_h\|_{L^\infty(\Omega)} \le Ch. \tag{5.12}
$$

Finally (5.8) follows from (5.10) and (5.11), and (5.7) is a consequence of the boundedness of $\{\zeta^h\}_{h>0}$ and (5.12). ∎

THEOREM 5.4. *There exist constants $C_i = C_i(\alpha,\beta) > 0$ $(i=1,2)$ such that for every $u \in U^{ad} \cap H^s(\Gamma)$, with $0 \le s \le 1$, the following inequalities hold:*

$$
\|y_h(u)\|_{L^\infty(\Omega)} \le C_1, \quad \forall h>0 \text{ and } s=0, \tag{5.13}
$$

$$
\|y_u - y_h(u)\|_{L^2(\Omega)} \le C_2 h^{s+1/2}(1+\|u\|_{H^s(\Gamma)}) \quad \forall h>0 \text{ and } 0\le s\le 1. \tag{5.14}
$$

Furthermore if $u_h \rightharpoonup u$ weakly in $L^2(\Gamma)$, $\{u_h\}_{h>0} \subset U^{ad}$, then $y_h(u_h) \to y_u$ strongly in $L^r(\Omega)$ for every $r < +\infty$.

Proof. Remembering that $y_u = z+\zeta$ and $y_h(u) = z_h+\zeta_h$, we see that (5.3), (5.5), (5.7), and (5.8) lead readily to inequalities (5.13) and (5.14). To prove the last part of the theorem, it is enough to use Theorem 2.2 and (5.14) with $s=0$ as follows:

$$
\|y_u - y_h(u_h)\|_{L^2(\Omega)} \le \|y_u - y_{u_h}\|_{L^2(\Omega)} + \|y_{u_h} - y_h(u_h)\|_{L^2(\Omega)} \longrightarrow 0 \quad \text{as } h \longrightarrow 0.
$$

The convergence in $L^r(\Omega)$ follows from (5.13). ∎

COROLLARY 5.5. *There exists a constant $C = C(\alpha,\beta) > 0$ such that, for all $u \in U^{ad}$ and $v \in U^{ad} \cap H^s(\Gamma)$, with $0 \le s \le 1$, we have*

$$
\|y_u - y_h(v)\|_{L^2(\Omega)} \le C\big\{\|u-v\|_{L^2(\Gamma)} + h^{s+1/2}(1+\|v\|_{H^s(\Gamma)})\big\}. \tag{5.15}
$$

This corollary is an immediate consequence of the second estimate in (2.5) and of (5.14).

Let us finish this section by establishing some estimates for the adjoint states.

THEOREM 5.6. *Given $u, v \in U^{ad}$, let ϕ_u and $\phi_h(v)$ be the solutions of* (3.5) *and* (4.3) *with u replaced by v in the last equation. Then there exist some constants $C_i = C_i(\alpha, \beta) > 0$ $(1 \le i \le 3)$ such that*

$$\|\phi_h(v)\|_{L^\infty(\Omega)} \le C_1 \quad \forall h > 0, \tag{5.16}$$

$$\|\phi_u - \phi_h(v)\|_{L^2(\Omega)} \le C_2(\|u - v\|_{L^2(\Gamma)} + h^2), \tag{5.17}$$

$$\|\phi_u - \phi_h(v)\|_{L^\infty(\Omega)} + \|\phi_u - \phi_h(v)\|_{H^1(\Omega)} \le C_3(\|u - v\|_{L^2(\Gamma)} + h). \tag{5.18}$$

Proof. All the inequalities follow from the results of Casas and Mateos [10] just by taking into account that

$$\|\phi_u - \phi_h(v)\|_X \le \|\phi_u - \phi_v\|_X + \|\phi_v - \phi_h(v)\|_X \le C(\|y_u - y_v\|_{L^2(\Omega)} + \|\phi_v - \phi_h(v)\|_X),$$

with X equal to $L^\infty(\Omega)$, $L^2(\Omega)$, and $H^1(\Omega)$, respectively. ∎

Now we provide an error estimate for the discrete normal derivative of the adjoint state defined by Proposition 4.2.

THEOREM 5.7. *There exists a constant $C = C(\alpha, \beta) > 0$ such that the following estimate holds:*

$$\|\partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\|_{L^2(\Gamma)} \le \begin{cases} Ch^{1/2} & \forall u \in U^{ad}, \\ C(\|u\|_{H^{1/2}(\Gamma)} + 1)h^{1-1/p} & \forall u \in U^{ad} \cap H^{1/2}(\Gamma). \end{cases} \tag{5.19}$$

Proof. First, let us remember that $\phi_u \in H^2(\Omega)$ and therefore $\partial_\nu \phi_u \in H^{1/2}(\Gamma)$. Observe that the definition of the projection operator Π_h leads to

$$\int_\Gamma \left|\partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\right|^2 = \int_\Gamma \left|\partial_\nu \phi_u - \Pi_h \partial_\nu \phi_u\right|^2 + \int_\Gamma \left|\Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\right|^2 = I_1 + I_2.$$

Since $\partial_\nu^h \phi_h(u)$ belongs to U_h, we can write

$$I_2 = \int_\Gamma (\partial_\nu \phi_u - \partial_\nu^h \phi_h(u))(\Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u)).$$

Let us introduce $z_h \in Y_h$ as the solution to the variational equation

$$\begin{cases} a(z_h, w_h) = 0 & \forall w_h \in Y_{h0}, \\ z_h = \Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u) & \text{on } \Gamma. \end{cases}$$

From (5.3) it follows that

$$\|z_h\|_{H^1(\Omega)} \le C\|\Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\|_{H^{1/2}(\Gamma)}. \tag{5.20}$$

Now using the definition of $\partial_\nu^h \phi_h(u)$ stated in Proposition 4.2 and a Green formula for ϕ_u, we can write

$$\begin{aligned} I_2 = a(z_h, \phi_u - \phi_h(u)) &+ \int_\Omega \left(\frac{\partial f}{\partial y}(x, y_h(u))\phi_h(u) - \frac{\partial f}{\partial y}(x, y_u)\phi_u\right) z_h \\ &+ \int_\Omega \left(\frac{\partial L}{\partial y}(x, y_h(u)) - \frac{\partial L}{\partial y}(x, y_u)\right) z_h. \end{aligned} \tag{5.21}$$

Due to the equation satisfied by z_h,

$$a(z_h, I_h\phi_u) = a(z_h, \phi_h(u)) = 0,$$

we also have

(5.22)
$$I_2 = a(z_h, \phi_u - I_h\phi_u) + \int_\Omega \left(\frac{\partial f}{\partial y}(x, y_h(u)) - \frac{\partial f}{\partial y}(x, y_u)\right)\phi_u z_h + \int_\Omega \frac{\partial f}{\partial y}(x, y_h(u))(\phi_h(u) - \phi_u)z_h + \int_\Omega \left(\frac{\partial L}{\partial y}(x, y_h(u)) - \frac{\partial L}{\partial y}(x, y_u)\right)z_h.$$

From well-known interpolation estimates, the second inequality of (5.3), and an inverse inequality, it follows that

$$\begin{aligned} &a(z_h, \phi_u - I_h\phi_u) \le \|z_h\|_{W^{1,p'}(\Omega)}\|\phi_u - I_h\phi_u\|_{W^{1,p}(\Omega)} \\ &\le Ch\|\phi_u\|_{W^{2,p}(\Omega)}\|z_h|_\Gamma\|_{W^{1-1/p',p'}(\Gamma)} \le Ch\|z_h|_\Gamma\|_{H^{1-1/p'}(\Gamma)} \\ &\le Ch^{1/p'}\|z_h|_\Gamma\|_{L^2(\Gamma)} = Ch^{1/p'}\sqrt{I_2}, \end{aligned} \tag{5.23}$$

where $p' = p/(p-1)$.

From assumptions (A1) and (A2) and inequalities (5.13), (5.14) with $s = 0$, (5.16), and (5.17), we get

$$\left|\int_\Omega \left(\frac{\partial f}{\partial y}(x, y_h(u)) - \frac{\partial f}{\partial y}(x, y_u)\right)\phi_u z_h\right| \le Ch^{1/2}\|z_h\|_{L^2(\Omega)}, \tag{5.24}$$

$$\begin{aligned} &\left|\int_\Omega \frac{\partial f}{\partial y}(x, y_h(u))(\phi_h(u) - \phi_u)z_h\right| \le C\|\phi_h(u) - \phi_u\|_{L^2(\Omega)}\|z_h\|_{L^2(\Omega)} \\ &\le Ch^2\|z_h\|_{L^2(\Omega)}, \end{aligned} \tag{5.25}$$

and

$$\left|\int_\Omega \left(\frac{\partial L}{\partial y}(x, y_h(u)) - \frac{\partial L}{\partial y}(x, y_u)\right)z_h\right| \le Ch^{1/2}\|z_h\|_{L^2(\Omega)}. \tag{5.26}$$

Collecting together the estimates (5.23)–(5.26) and using (5.20) and the fact that $p' < 2$, we obtain

$$\begin{aligned} I_2 &\le Ch^{1/p'}\sqrt{I_2} + Ch^{1/2}\|z_h\|_{L^2(\Omega)} \\ &\le C(h^{1/p'}\sqrt{I_2} + h^{1/2}\|\Pi_h\partial_\nu\phi_u - \partial_\nu^h\phi_h(u)\|_{L^2(\Gamma)}) \le Ch^{1/2}\sqrt{I_2}, \end{aligned} \tag{5.27}$$

which implies that

$$I_2 \le Ch. \tag{5.28}$$

Using again that $\phi_u \in W^{2,p}(\Omega)$, we get that $\partial_\nu\phi_u \in W^{1-1/p,p}(\Gamma) \subset H^{1-1/p}(\Gamma)$. Hence from (4.1) with $s = 1 - 1/p$, we can derive

$$I_1 \le Ch\|\partial_\nu\phi_u\|^2_{H^{1/2}(\Gamma)} \le Ch\|\phi_u\|^2_{H^2(\Omega)} \le Ch^{2(1-1/p)}. \tag{5.29}$$

So the first estimate in (5.19) is proved.

To complete the proof let us assume that $u \in H^{1/2}(\Gamma)$; then we can use (5.14) with $s = 1/2$ to estimate $y_u - y_h(u)$ in $L^2(\Omega)$ by Ch. This allows us to change $h^{1/2}$ in (5.24) and (5.26) by h. Therefore (5.27) can be replaced with $I_2 \le Ch^{1/p'} = Ch^{1-1/p}$; thus $I_2 \le Ch^{2(1-1/p)}$. So the second estimate in (5.19) is proved. ∎

COROLLARY 5.8. *There exists a constant C independent of h such that*

$$\begin{cases} \|\partial_\nu^h \phi_h(u)\|_{H^{1/2}(\Gamma)} \le C \quad \forall u \in U^{ad}, \\ \|\partial_\nu^h \phi_h(u)\|_{W^{1-1/p,p}(\Gamma)} \le C(\|u\|_{H^{1/2}(\Gamma)} + 1) \quad \forall u \in U^{ad} \cap H^{1/2}(\Gamma), \\ \|\partial_\nu \phi_u - \partial_\nu^h \phi_h(v)\|_{L^2(\Gamma)} \le C\left\{\|u - v\|_{L^2(\Gamma)} + h^\kappa\right\} \quad \forall u, v \in U^{ad}, \end{cases} \tag{5.30}$$

where $\kappa = 1 - 1/p$ if $v \in H^{1/2}(\Gamma)$ and $\kappa = 1/2$ otherwise.

Proof. Let us make the proof in the case when $u \in U^{ad} \cap H^{1/2}(\Gamma)$. The case when $u \in U^{ad}$ can be treated similarly. We know that

$$\|\partial_\nu \phi_u\|_{W^{1-1/p,p}(\Gamma)} \le C\|\phi_u\|_{W^{2,p}(\Omega)} \le C \quad \forall u \in U^{ad}.$$

On the other hand, the projection operator Π_h is stable in the Sobolev spaces $W^{s,q}(\Gamma)$, for $1 \le q \le \infty$ and $0 \le s \le 1$ (see Casas and Raymond [13]); therefore

$$\|\Pi_h \partial_\nu \phi_u\|_{W^{1-1/p,p}(\Gamma)} \le C\|\partial_\nu \phi_u\|_{W^{1-1/p,p}(\Gamma)}.$$

Finally, with an inverse inequality and the estimate $I_2 \le Ch^{2-2/p}$ obtained in the previous proof, we deduce

$$\begin{aligned} &\|\partial_\nu^h \phi_h(u)\|_{W^{1-1/p,p}(\Gamma)} \le \|\Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\|_{W^{1-1/p,p}(\Gamma)} + \|\Pi_h \partial_\nu \phi_u\|_{W^{1-1/p,p}(\Gamma)} \\ &\le C\|\Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\|_{H^{1-1/p}(\Gamma)} + \|\Pi_h \partial_\nu \phi_u\|_{W^{1-1/p,p}(\Gamma)} \\ &\le Ch^{-1+1/p}\|\Pi_h \partial_\nu \phi_u - \partial_\nu^h \phi_h(u)\|_{L^2(\Gamma)} + \|\partial_\nu \phi_u\|_{W^{1-1/p,p}(\Gamma)} \le C. \end{aligned}$$

The third inequality of (5.30) is an immediate consequence of Theorem 5.7. ∎

6. Convergence analysis for (P_h). In this section we will prove the strong convergence in $L^2(\Gamma)$ of the solutions $\bar u_h$ of discrete problems (P_h) to the solutions of (P). Moreover, we will first prove that $\{\bar u_h\}_h$ remains bounded in $H^{1/2}(\Gamma)$, and then that it is also bounded in $W^{1-1/p,p}(\Gamma)$. Finally, we will prove the strong convergence of the solutions $\bar u_h$ of discrete problems (P_h) to the solutions of (P) in $C(\Gamma)$.

THEOREM 6.1. *For every $h > 0$ let $\bar u_h$ be a global solution of problem (P_h). Then there exist weakly*-converging subsequences of $\{\bar u_h\}_{h>0}$ in $L^\infty(\Gamma)$ (still indexed by h). If the subsequence $\{\bar u_h\}_{h>0}$ is converging weakly* in $L^\infty(\Gamma)$ to some $\bar u$, then $\bar u$ is a solution of* (P),

$$\lim_{h\to 0} J_h(\bar u_h) = J(\bar u) = \inf(P) \quad \textit{and} \quad \lim_{h\to 0} \|\bar u - \bar u_h\|_{L^2(\Gamma)} = 0. \tag{6.1}$$

Proof. Since $U_h^{ad} \subset U^{ad}$ holds for every $h > 0$ and U^{ad} is bounded in $L^\infty(\Gamma)$, $\{\bar u_h\}_{h>0}$ is also bounded in $L^\infty(\Gamma)$. Therefore, there exist weakly*-converging subsequences as claimed in the statement of the theorem. Let $\{\bar u_h\}$ be one of these subsequences and let $\bar u$ be the weak* limit. It is obvious that $\bar u \in U^{ad}$. Let us prove that $\bar u$ is a solution of (P). Let us take a solution of (P), $\tilde u \in U^{ad}$; therefore $\tilde u \in W^{1-1/p,p}(\Gamma)$ for some $p > 2$; see Theorem 3.4. Let us take $u_h = I_h \tilde u$. Then

$u_h \in U_h^{ad}$ and $\{u_h\}_h$ tends to $\tilde{u}$ in $L^\infty(\Gamma)$; see Brenner and Scott [6]. By taking $u = \tilde{u}$, $v = u_h$, and $s = 0$ in (5.15) we deduce that $y_h(u_h) \to y_{\tilde{u}}$ in $L^2(\Omega)$. Moreover, (5.13) implies that $\{y_h(u_h)\}_{h>0}$ is bounded in $L^\infty(\Omega)$. On the other hand, Theorem 5.4 implies that $\bar{y}_h = y_h(\bar{u}_h) \to \bar{y} = y_{\bar{u}}$ strongly in $L^2(\Omega)$, and $\{\bar{y}_h\}_{h>0}$ is also bounded in $L^\infty(\Omega)$. Then we have

$$J(\bar{u}) \le \liminf_{h\to 0} J_h(\bar{u}_h) \le \limsup_{h\to 0} J_h(\bar{u}_h) \le \limsup_{h\to 0} J_h(I_h\tilde{u}) = J(\tilde{u}) = \inf(P).$$

This proves that $\bar{u}$ is a solution of (P) as well as the convergence of the optimal costs, which leads to $\|\bar{u}_h\|_{L^2(\Gamma)} \longrightarrow \|\bar{u}\|_{L^2(\Gamma)}$; hence we deduce the strong convergence of the controls in $L^2(\Gamma)$. □

THEOREM 6.2. *Let $p > 2$ be as in Theorem* 3.4*, and for every h let $\bar{u}_h$ denote a local solution of* (P_h)*. Then there exists a constant $C > 0$ independent of h such that*

$$\|\bar{u}_h\|_{W^{1-1/p,p}(\Gamma)} \le C \quad \forall h > 0. \tag{6.2}$$

Moreover, the convergence of $\{\bar{u}_h\}_{h>0}$ to $\bar{u}$ stated in Theorem 6.1 *holds in $C(\Gamma)$.*

Proof. By using the stability in $H^{1/2}(\Gamma)$ of the $L^2(\Gamma)$-projections on the sets U_h^{ad} (see Casas and Raymond [13]) along with (4.12) and the first inequality of (5.30), we get that $\{\bar{u}_h\}_{h>0}$ is uniformly bounded in $H^{1/2}(\Gamma)$. Using now the second inequality of (5.30) and the stability of Π_h in $W^{1-1/p,p}(\Gamma)$, we deduce (6.2). Finally, the convergence is a consequence of the compactness of the imbedding $W^{1-1/p,p}(\Gamma) \subset C(\Gamma)$ for $p > 2$. □

7. Error estimates. The goal in this section is to prove the following theorem.

THEOREM 7.1. *Let us assume that $\bar{u}$ is a local solution of* (P) *satisfying the sufficient second order optimality conditions provided in Theorem* 3.5*, and let $\bar{u}_h$ be a local solution of* (P_h) *such that $\bar{u}_h \to \bar{u}$ in $L^2(\Gamma)$; see Theorem* 6.1*. Then the following inequality holds:*

$$\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} \le Ch^{1-1/p}, \tag{7.1}$$

where $p > 2$ is given by Theorem 3.4*.*

We will prove the theorem arguing by contradiction. The statement of the theorem can be stated as follows. There exists a positive constant C such that for all $0 < h < 1/C$, we have

$$\frac{\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)}}{h^{1-1/p}} \le C.$$

Thus if (7.1) is false, for all $k > 0$, there exists $0 < h_k < 1/k$ such that

$$\frac{\|\bar{u} - \bar{u}_{h_k}\|_{L^2(\Gamma)}}{h_k^{1-1/p}} > k.$$

Therefore there exists a sequence of h such that

$$\lim_{h\to 0} \frac{1}{h^{1-1/p}} \|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} = +\infty. \tag{7.2}$$

We will obtain a contradiction for this sequence. For the proof of this theorem, we need some lemmas.

LEMMA 7.2. *Let us assume that* (7.1) *is false. Let* $\delta > 0$ *as given by Remark* 3.6(2). *Then there exists* $h_0 > 0$ *such that*

$$\frac{1}{2}\min\{\delta, N\}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Gamma)} \le (J'(\bar{u}_h) - J'(\bar{u}))(\bar{u}_h - \bar{u}) \quad \forall h < h_0. \tag{7.3}$$

Proof. Let $\{\bar{u}_h\}_h$ be a sequence satisfying (7.2). By applying the mean value theorem, we get for some $\hat{u}_h = \bar{u} + \theta_h(\bar{u}_h - \bar{u})$,

$$(J'(\bar{u}_h) - J'(\bar{u}))(\bar{u}_h - \bar{u}) = J''(\hat{u}_h)(\bar{u}_h - \bar{u})^2. \tag{7.4}$$

Let us take

$$v_h = \frac{1}{\|\bar{u}_h - \bar{u}\|_{L^2(\Gamma)}}(\bar{u}_h - \bar{u}).$$

Taking a subsequence if necessary, we can assume that $v_h \rightharpoonup v$ in $L^2(\Gamma)$. Let us prove that v belongs to the critical cone $C_{\bar{u}}$ defined in section 3. First, we remark that every v_h satisfies the sign condition (3.9); hence v also does. Let us prove that $v(x) = 0$ if $\bar{d}(x) \neq 0$, with $\bar{d}$ being defined by (3.10). We will use the interpolation operator $I_h \in \mathcal{L}(W^{1-1/p,p}(\Gamma), U_h)$, with $p > 2$ given in Theorem 3.4. Since $\bar{u} \in U^{ad}$, it is obvious that $I_h\bar{u} \in U_h^{ad}$. For any $y \in W^{1,p}(\Omega)$ such that $y|_\Gamma = \bar{u}$, it is clear that $I_h\bar{u}$ is the trace of $I_h y$ (see the beginning of section 5). Now, by using a result of Grisvard [22, Chapter 1], we get

$$\|\bar{u} - I_h\bar{u}\|^p_{L^p(\Gamma)} \le C\left(\varepsilon^{1-1/p}\|y - I_h y\|^p_{W^{1,p}(\Omega)} + \varepsilon^{-1/p}\|y - I_h y\|^p_{L^p(\Omega)}\right)$$

for every $\varepsilon > 0$ and for some constant $C > 0$ independent of ε and y. Setting $\varepsilon = h^p$ and using that (see, for instance, Brenner and Scott [6])

$$\|y - I_h y\|_{L^p(\Omega)} \le C_1 h\|y\|_{W^{1,p}(\Omega)}, \quad \|I_h y\|_{W^{1,p}(\Omega)} \le C_2\|y\|_{W^{1,p}(\Omega)},$$

and

$$\inf_{y|_\Gamma = \bar{u}} \|y\|_{W^{1,p}(\Omega)} \le C_3\|\bar{u}\|_{W^{1-1/p}(\Gamma)},$$

we conclude that

$$\|\bar{u} - I_h\bar{u}\|_{L^2(\Gamma)} \le |\Gamma|^{\frac{p-2}{2p}}\|\bar{u} - I_h\bar{u}\|_{L^p(\Gamma)} \le Ch^{1-1/p}\|\bar{u}\|_{W^{1-1/p,p}(\Gamma)}. \tag{7.5}$$

Let us define

$$\bar{d}_h(x) = N\bar{u}_h(x) - \partial^h_\nu\bar{\phi}_h(x). \tag{7.6}$$

The third inequality of (5.30) implies that $\bar{d}_h \to \bar{d}$ in $L^2(\Gamma)$. Now we have

$$\begin{aligned}\int_\Gamma \bar{d}(x)v(x)\,dx &= \lim_{h\to 0}\int_\Gamma \bar{d}_h(x)v_h(x)\,dx \\ &= \lim_{h\to 0}\frac{1}{\|\bar{u}_h - \bar{u}\|_{L^2(\Gamma)}}\left\{\int_\Gamma \bar{d}_h(I_h\bar{u} - \bar{u})\,dx + \int_\Gamma \bar{d}_h(\bar{u}_h - I_h\bar{u})\,dx\right\}.\end{aligned}$$

From (4.8), (7.2), and (7.5) we deduce

$$\int_\Gamma \bar d(x)v(x)\,dx \le \lim_{h\to 0} \frac{1}{\|\bar u_h-\bar u\|_{L^2(\Gamma)}} \int_\Gamma \bar d_h(x)(I_h\bar u(x)-\bar u(x))\,dx$$
$$\le \lim_{h\to 0} \frac{Ch^{1-1/p}}{\|\bar u_h-\bar u\|_{L^2(\Gamma)}} = 0.$$

Since v satisfies the sign condition (3.9), then $\bar d(x)v(x)\ge 0$; hence the above inequality proves that v is zero whenever $\bar d$ is not, which allows us to conclude that $v\in C_{\bar u}$. Now from the definition of v_h, (3.4), and (3.11) we get

$$\lim_{h\to 0} J''(\hat u_h)v_h^2 = \lim_{h\to 0}\left\{\int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x,y_{\hat u_h})+\phi_{\hat u_h}\frac{\partial^2 f}{\partial y^2}(x,y_{\hat u_h})\right] z_{v_h}^2\,dx + N\right\}$$
$$= \int_\Omega \left[\frac{\partial^2 L}{\partial y^2}(x,\bar y)+\bar\phi\frac{\partial^2 f}{\partial y^2}(x,\bar y)\right] z_v^2\,dx + N$$
$$= J''(\bar u)v^2 + N(1-\|v\|^2_{L^2(\Gamma)}) \ge N + (\delta - N)\|v\|^2_{L^2(\Gamma)}.$$

Taking into account that $\|v\|_{L^2(\Gamma)}\le 1$, these inequalities lead to

$$\lim_{h\to 0} J''(\hat u_h)v_h^2 \ge \min\{\delta, N\} > 0,$$

which proves the existence of $h_0>0$ such that

$$J''(\hat u_h)v_h^2 \ge \frac{1}{2}\min\{\delta,N\} \quad \forall h<h_0.$$

From this inequality, the definition of v_h, and (7.4) we deduce (7.3). ∎

LEMMA 7.3. *There exists a constant $C>0$ independent of h such that for every $v\in L^\infty(\Gamma)$,*

$$|(J_h'(\bar u_h)-J'(\bar u_h))v| \le Ch^{1-1/p}\|v\|_{L^2(\Gamma)}. \tag{7.7}$$

Proof. From (3.3), (4.6), (7.6), (6.2), and Theorem 5.7 we get

$$(J_h'(\bar u_h)-J'(\bar u_h))v = \int_\Gamma (\partial_\nu\phi_{\bar u_h}-\partial_\nu^h\bar\phi_h)v\,dx \le \|\partial_\nu\phi_{\bar u_h}-\partial_\nu^h\bar\phi_h\|_{L^2(\Gamma)}\|v\|_{L^2(\Gamma)}$$
$$\le C(\|\bar u_h\|_{H^{1/2}(\Gamma)}+1)h^{(1-1/p)}\|v\|_{L^2(\Gamma)} \le Ch^{(1-1/p)}\|v\|_{L^2(\Gamma)}. \quad ∎$$

LEMMA 7.4. *There exists a constant $C>0$ independent of h such that for every $v\in L^\infty(\Gamma)$,*

$$|(J_h'(\bar u_h)-J'(\bar u))v| \le \Big(N\|\bar u-\bar u_h\|_{L^2(\Gamma)}+Ch^{1-1/p}\Big)\|v\|_{L^2(\Gamma)}. \tag{7.8}$$

Proof. Arguing in a way similar to the previous proof, and using (5.30) and (6.2), we have

$$(J_h'(\bar u_h)-J'(\bar u))v = \int_\Gamma \Big(N\bar u_h-\partial_\nu^h\bar\phi_h\Big)\Pi_h v\,dx - \int_\Gamma \Big(N\bar u-\partial_\nu\bar\phi\Big)v\,dx$$
$$= N\int_\Gamma(\bar u_h-\bar u)v\,dx + \int_\Gamma\Big(\partial_\nu\bar\phi-\partial_\nu^h\bar\phi_h\Big)v\,dx$$
$$\le \Big(N\|\bar u_h-\bar u\|_{L^2(\Gamma)}+Ch^{(1-1/p)}\Big)\|v\|_{L^2(\Gamma)}. \quad ∎$$

One key point in the proof of the error estimates is to get a discrete control $u_h \in U_h^{ad}$ that approximates $\bar{u}$ conveniently and satisfies $J'(\bar{u})\bar{u} = J'(\bar{u})u_h$. Let us find such a control. For every $1 \le j \le N(h)$, let us set

$$I_j = \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \bar{d}(x)e_j(x)\,dx.$$

Now we define $u_h \in U_h$ with $u_h(x_\Gamma^j) = u_{h,j}$ for every node $x_\Gamma^j \in \Gamma$ by the expression

$$u_{h,j} = \begin{cases} \dfrac{1}{I_j}\displaystyle\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \bar{d}(x)\bar{u}(x)e_j(x)\,dx & \text{if } I_j \neq 0, \\ \dfrac{1}{h_{j-1}+h_j}\displaystyle\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \bar{u}(x)\,dx & \text{if } I_j = 0. \end{cases} \tag{7.9}$$

Remember that the measure of $[x_\Gamma^{j-1}, x_\Gamma^{j+1}]$ is $h_{j-1}+h_j = |x_\Gamma^j - x_\Gamma^{j-1}| + |x_\Gamma^{j+1} - x_\Gamma^j|$, which coincides with $|x_\Gamma^{j+1} - x_\Gamma^{j-1}|$ if x_Γ^j is not a vertex of Ω.

In the following lemma, we state that the function u_h defined by (7.9) satisfies our requirements.

LEMMA 7.5. *There exists $h_0 > 0$ such that, for every $0 < h < h_0$, the element $u_h \in U_h$ defined by* (7.9) *obeys the following properties:*

1. $u_h \in U_h^{ad}$.
2. $J'(\bar{u})\bar{u} = J'(\bar{u})u_h$.
3. *The approximation property*

$$\|\bar{u} - u_h\|_{L^2(\Gamma)} \le Ch^{1-1/p} \tag{7.10}$$

is fulfilled for some constant $C > 0$ independent of h.

Proof. Since $\bar{u}$ is continuous on Γ, there exists $h_0 > 0$ such that

$$|\bar{u}(\xi_2) - \bar{u}(\xi_1)| \le \frac{\beta - \alpha}{2}, \quad \forall h < h_0, \quad \forall \xi_1, \xi_2 \in [x_\Gamma^{j-1}, x_\Gamma^{j+1}],\ 1 \le j \le N(h),$$

which implies that $\bar{u}$ cannot admit both the values α and β on one segment $[x_\Gamma^{j-1}, x_\Gamma^{j+1}]$ for any $h < h_0$. Hence the sign of $\bar{d}$ on $[x_\Gamma^{j-1}, x_\Gamma^{j+1}]$ must be constant due to (3.7). Therefore, $I_j = 0$ if and only if $\bar{d}(x) = 0$ for all $x \in [x_\Gamma^{j-1}, x_\Gamma^{j+1}]$. Moreover if $I_j \neq 0$, then $\bar{d}(x)/I_j \ge 0$ for every $x \in [x_\Gamma^{j-1}, x_\Gamma^{j+1}]$. As a first consequence of this we get that $\alpha \le u_{h,j} \le \beta$, which means that $u_h \in U_h^{ad}$. On the other hand,

$$J'(\bar{u})u_h = \sum_{j=1}^{N(h)} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \bar{d}(x)e_j(x)\,dx\,u_{h,j} = \sum_{j=1}^{N(h)} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \bar{d}(x)\bar{u}(x)e_j(x)\,dx = J'(\bar{u})\bar{u}.$$

Finally, let us prove (7.10). Let us remember that $\bar{u} \in W^{1-1/p,p}(\Gamma) \subset H^{1-1/p}(\Gamma)$ and $p > 2$. We note that the norm in $H^s(\Gamma)$, $0 < s < 1$, is given by

$$\|u\|_{H^s(\Gamma)} = \left(\|u\|_{L^2(\Gamma)}^2 + \int_\Gamma \int_\Gamma \frac{|u(x) - u(\xi)|^2}{|x - \xi|^{1+2s}}\,dx\,d\xi \right)^{1/2}. \tag{7.11}$$

Using that $\sum_{j=1}^{N(h)} e_j(x) = 1$ and $0 \le e_j(x) \le 1$ we get

$$
\begin{aligned}
\|\bar u - u_h\|^2_{L^2(\Gamma)} &= \int_\Gamma \Big| \sum_{j=1}^{N(h)} (\bar u(x) - u_{h,j}) e_j(x) \Big|^2 \, dx \\
&\le \sum_{j=1}^{N(h)} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - u_{h,j}|^2 e_j(x)\, dx \le \sum_{j=1}^{N(h)} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - u_{h,j}|^2 \, dx.
\end{aligned} \tag{7.12}
$$

Let us estimate every term of the sum.

Let us start by assuming that $I_j = 0$ so that $u_{h,j}$ is defined by the second relation in (7.9). Then we have

$$
\begin{aligned}
&\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - u_{h,j}|^2 \, dx = \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \Big| \frac{1}{h_{j-1} + h_j} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} (\bar u(x) - \bar u(\xi))\, d\xi \Big|^2 \, dx \\
&\le \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \frac{1}{h_{j-1} + h_j} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - \bar u(\xi)|^2 \, d\xi \, dx \\
&\le (h_{j-1} + h_j)^{2(1-1/p)} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \frac{|\bar u(x) - \bar u(\xi)|^2}{|x - \xi|^{1+2(1-1/p)}} \, dx \, d\xi \\
&\le (2h)^{2(1-1/p)} \|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1}, x_\Gamma^{j+1})}.
\end{aligned} \tag{7.13}
$$

Now let us consider the case $I_j \ne 0$:

$$
\begin{aligned}
&\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - u_{h,j}|^2 \, dx = \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \Big| \frac{1}{I_j} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \bar d(\xi) e_j(\xi) (\bar u(x) - \bar u(\xi))\, d\xi \Big|^2 \, dx \\
&\le \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \Big| \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \sqrt{\frac{\bar d(\xi) e_j(\xi)}{I_j}} |\bar u(x) - \bar u(\xi)| \sqrt{\frac{\bar d(\xi) e_j(\xi)}{I_j}}\, d\xi \Big|^2 \, dx \\
&\le \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - \bar u(\xi)|^2 \frac{\bar d(\xi) e_j(\xi)}{I_j} \, d\xi \, dx \\
&\le \Big(\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \frac{\bar d(\xi) e_j(\xi)}{I_j} \, d\xi \Big) \sup_{\xi \in [x_\Gamma^{j-1}, x_\Gamma^{j+1}]} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - \bar u(\xi)|^2 \, dx \\
&= \sup_{\xi \in [x_\Gamma^{j-1}, x_\Gamma^{j+1}]} \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - \bar u(\xi)|^2 \, dx.
\end{aligned} \tag{7.14}
$$

To obtain the estimate for the last term we are going to use Lemma 7.6 stated below, with

$$
f(\xi) = \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x) - \bar u(\xi)|^2 \, dx.
$$

Since $H^{1-1/p}(\Gamma) \subset C^{0,\theta}(\Gamma)$ for $\theta = 1/2 - 1/p$ (see, e.g., [37, Theorem 2.8.1]), it is

easy to check that

$$|f(\xi_2)-f(\xi_1)| \le \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \Big|[\bar u(x)-\bar u(\xi_1)]+[\bar u(x)-\bar u(\xi_2)]\Big|\,\Big|\bar u(\xi_2)-\bar u(\xi_1)\Big|\,dx$$
$$\le 2(h_{j-1}+h_j)^{1+2\theta} C_{\theta,p}\|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})}.$$

On the other hand, we have

$$\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} f(\xi)\,d\xi = \int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}}\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \frac{|\bar u(x)-\bar u(\xi)|^2}{|x-\xi|^{1+2(1-1/p)}}|x-\xi|^{1+2(1-1/p)}\,dx\,d\xi$$
$$\le (h_{j-1}+h_j)^{1+2(1-1/p)}\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}}\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} \frac{|\bar u(x)-\bar u(\xi)|^2}{|x-\xi|^{1+2(1-1/p)}}\,dx\,d\xi$$
$$\le (h_{j-1}+h_j)^{2+(1-2/p)}\|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})}.$$

Then we can apply Lemma 7.6 to the function f, with

$$M=(h_{j-1}+h_j)^{2\theta}\max\{4C_{\theta,p},1\}\|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})} \le Ch^{2\theta}\|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})},$$

to deduce that

$$f(\xi) \le C\|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})}h^{1+2\theta}. \tag{7.15}$$

This inequality, along with (7.14), leads to

$$\int_{x_\Gamma^{j-1}}^{x_\Gamma^{j+1}} |\bar u(x)-u_{h,j}|^2\,dx \le C\|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})}h^{1+2\theta} \tag{7.16}$$

in the case when $I_j \neq 0$.

Since

$$\sum_{j=1}^{N(h)} \|\bar u\|^2_{H^{1-1/p}(x_\Gamma^{j-1},x_\Gamma^{j+1})} \le 2\|\bar u\|^2_{H^{1-1/p}(\Gamma)},$$

inequality (7.10) follows from (7.12), (7.13), (7.16), and the fact that $1+2\theta = 2(1-1/p)$. ∎

LEMMA 7.6. *Given $-\infty < a < b < +\infty$ and $f : [a,b] \longmapsto \mathbb{R}^+$, a function satisfying*

$$|f(x_2)-f(x_1)| \le \frac{M}{2}(b-a) \quad \textit{and} \quad \int_a^b f(x)\,dx \le M(b-a)^2,$$

we have that $f(x) \le 2M(b-a)$ for all $x \in [a,b]$.

Proof. We argue by contradiction and we assume that there exists a point $\xi \in [a,b]$ such that $f(\xi) > 2M(b-a)$; then

$$\int_a^b f(x)\,dx = \int_a^b \{[f(x)-f(\xi)]+f(\xi)\}\,dx > -\frac{M}{2}(b-a)^2 + 2M(b-a)^2 = \frac{3M}{2}(b-a)^2,$$

which contradicts the second assumption on f. ∎

Proof of Theorem 7.1. Setting $u = \bar{u}_h$ in (3.7), we get

$$J'(\bar{u})(\bar{u}_h - \bar{u}) = \int_\Gamma \left(N\bar{u} - \partial_\nu \bar{\phi}\right)(\bar{u}_h - \bar{u})\, dx \geq 0. \tag{7.17}$$

From (4.8) with u_h defined by (7.9), it follows that

$$J_h'(\bar{u}_h)(u_h - \bar{u}_h) = \int_\Gamma \left(N\bar{u}_h - \partial_\nu^h \bar{\phi}_h\right)(u_h - \bar{u}_h)\, dx \geq 0$$

and then that

$$J_h'(\bar{u}_h)(\bar{u} - \bar{u}_h) + J_h'(\bar{u}_h)(u_h - \bar{u}) \geq 0. \tag{7.18}$$

By adding (7.17) and (7.18) and using Lemma 7.5(2), we derive

$$(J'(\bar{u}) - J_h'(\bar{u}_h))(\bar{u} - \bar{u}_h) \leq J_h'(\bar{u}_h)(u_h - \bar{u}) = (J_h'(\bar{u}_h) - J'(\bar{u}))(u_h - \bar{u}).$$

For $h < h_0$, this inequality and (7.3) lead to

$$\begin{aligned} &\frac{1}{2}\min\{N, \delta\}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Gamma)} \leq (J'(\bar{u}) - J'(\bar{u}_h))(\bar{u} - \bar{u}_h) \\ &\leq (J_h'(\bar{u}_h) - J'(\bar{u}_h))(\bar{u} - \bar{u}_h) + (J_h'(\bar{u}_h) - J'(\bar{u}))(u_h - \bar{u}). \end{aligned} \tag{7.19}$$

Now from (7.7) and Young's inequality, we obtain

$$\begin{aligned} &|(J_h'(\bar{u}_h) - J'(\bar{u}_h))(\bar{u} - \bar{u}_h)| \leq Ch^{1-1/p}\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} \\ &\leq Ch^{2(1-1/p)} + \frac{1}{8}\min\{N, \delta\}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Gamma)}. \end{aligned} \tag{7.20}$$

On the other hand, using again Young's inequality, (7.8), and (7.10), we deduce

$$\begin{aligned} |(J_h'(\bar{u}_h) - J'(\bar{u}))(u_h - \bar{u})| &\leq \left(N\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} + Ch^{1-1/p}\right)\|\bar{u} - u_h\|_{L^2(\Gamma)} \\ &\leq \left(N\|\bar{u} - \bar{u}_h\|_{L^2(\Gamma)} + Ch^{1-1/p}\right)h^{1-1/p} \\ &\leq \frac{1}{8}\min\{N, \delta\}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Gamma)} + Ch^{2(1-1/p)}. \end{aligned} \tag{7.21}$$

From (7.19)–(7.21) we get

$$\frac{1}{4}\min\{N, \delta\}\|\bar{u} - \bar{u}_h\|^2_{L^2(\Gamma)} \leq Ch^{2(1-1/p)},$$

which contradicts (7.2). ∎

8. Numerical tests. In this section we present some numerical tests which illustrate our theoretical results. Let Ω be the unit square $(0,1)^2$. Consider

$$y_d(x_1, x_2) = \frac{1}{(x_1^2 + x_2^2)^{1/3}}.$$

We are going to solve the following problem:

$$\text{(P)} \begin{cases} \text{Min } J(u) = \dfrac{1}{2}\displaystyle\int_\Omega (y_u(x) - y_d(x))^2 dx + \frac{1}{2}\int_\Gamma u(x)^2 dx, \\ u \in U_{ad} = \{u \in L^2(\Gamma) :\ -1 \leq u(x) \leq 2 \text{ a.e. } x \in \Gamma\}, \\ -\Delta y_u = 0 \text{ in } \Omega,\ y_u = u \text{ on } \Gamma. \end{cases}$$

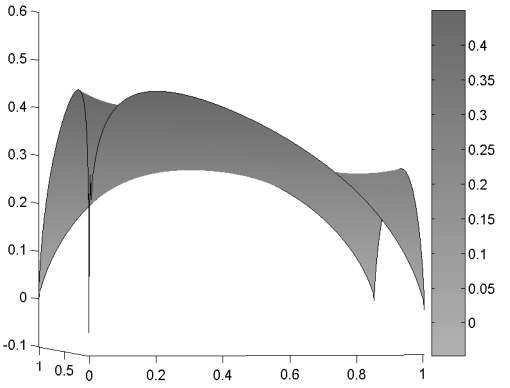


FIG. 8.1.

We remark that $y_d \in L^p(\Omega)$ for all $p < 3$, but $y_d \notin L^3(\Omega)$; therefore the optimal adjoint state $\bar{\varphi}$ is actually in $W^{2,p}(\Omega)$ for $p < 3$. Consequently we can deduce that the optimal control belongs to $W^{1-1/p,p}(\Gamma)$, but $W^{1-1/p,p}(\Gamma)$ is not included in $H^1(\Gamma)$. There is no reason for the normal derivative $\partial_\nu \bar{\varphi}$ to be more regular than $W^{1-1/p,p}(\Gamma)$. For our problem, the plot in Figure 8.1 shows that the optimal control has a singularity in the corner at the origin, and it seems that $\bar{u} \notin H^1(\Gamma)$. So we cannot hope to have a convergence order of $O(h)$. Instead of that, we have a convergence of order $O(h^{1-1/p})$ for some $p > 2$, as predicted by the theory.

Since we do not have an exact solution for (P), we have solved it numerically for $h = 2^{-9}\sqrt{2}$, and we have used this solution for comparison with other solutions for bigger values of h. We have solved it using an active set strategy, as is explained in [11]. Figure 8.1 shows a plot of the optimal solution. The control constraints are not active at the optimal control. In Table 8.1 we show the norm in $L^2(\Gamma)$ of the error of the control and the order of convergence step by step. The order of convergence is measured as

$$o_i = \frac{\log(\|\bar{u}_{h_i} - \bar{u}\|_{L^2(\Gamma)}) - \log(\|\bar{u}_{h_{i-1}} - \bar{u}\|_{L^2(\Gamma)})}{\log(h_i) - \log(h_{i-1})}.$$

Let us remark that $1 - 1/p < 2/3$ for $p < 3$. The values o_i are approximately $2/3$. We believe that the order of convergence could be closer to $2/3$ if we could compare the computed controls with the true optimal control instead of with its numerical approximation. We refer to [12] for more details and numerical tests.

TABLE 8.1

$h_i/\sqrt{2}$	$\|\bar{u}_{h_i} - \bar{u}\|_{L^2(\Gamma)}$	o_i
2^{-3}	0.1055	—
2^{-4}	0.0652	0.6944
2^{-5}	0.0393	0.7302
2^{-6}	0.0237	0.7314
2^{-7}	0.0146	0.7008
2^{-8}	0.0093	0.6493

Acknowledgments. The authors thank the referees for their valuable suggestions that have improved the first version of the paper. They also thank Professor Mariano Mateos for his collaboration in the numerical tests.

REFERENCES

[1] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Comput. Optim. Appl., 23 (2002), pp. 201–229.
[2] N. Arada and J.-P. Raymond, *Dirichlet boundary control of semilinear parabolic equations. Part* 1: *Problems with no state constraints*, Appl. Math. Optim., 45 (2002), pp. 125–143.
[3] M. Berggren, *Approximations of very weak solutions to boundary-value problems*, SIAM J. Numer. Anal., 42 (2004), pp. 860–877.
[4] J. F. Bonnans and H. Zidani, *Optimal control problems with partially polyhedric constraints*, SIAM J. Control Optim., 37 (1999), pp. 1726–1741.
[5] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring.* I, Math. Comp., 47 (1986), pp. 103–134.
[6] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, Berlin, Heidelberg, 1994.
[7] H. Brézis, *Analyse Fonctionnelle, Théorie et Applications*, Masson, Paris, 1983.
[8] E. Casas, *Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 345–374.
[9] E. Casas and M. Mateos, *Second order optimality conditions for semilinear elliptic control problems with finitely many state constraints*, SIAM J. Control Optim., 40 (2002), pp. 1431–1454.
[10] E. Casas and M. Mateos, *Uniform convergence of the FEM. Applications to state constrained control problems*, Comput. Appl. Math., 21 (2002), pp. 67–100.
[11] E. Casas, M. Mateos, and F. Tröltzsch, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–219.
[12] E. Casas, M. Mateos, and J.-P. Raymond, *Robin approximation of Dirichlet optimal control problems for semilinear elliptic equations*, in preparation.
[13] E. Casas and J.-P. Raymond, *The stability in $W^{s,p}(\Gamma)$ spaces of L^2-projections on some convex sets*, Numer. Funct. Anal. Optim., 27 (2006), pp. 117–137.
[14] P. Ciarlet and P. Raviart, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31.
[15] M. Dauge, *Neumann and mixed problems on curvilinear polyhedra*, Integral Equations Operator Theory, 15 (1992), 227–261.
[16] A. Dontchev and W. Hager, *The Euler approximation in state constrained optimal control*, Math. Comp., 70 (2000), pp. 173–203.
[17] A. L. Dontchev, W. Hager, and V. M. Veliov, *Second-order Runge–Kutta approximations in control constrained optimal control*, SIAM J. Numer. Anal., 38 (2000), pp. 202–226.
[18] J. J. Douglas, T. Dupont, and L. Wahlbin, *The stability in L^q of the L^2 projection into finite element function spaces*, Numer. Math., 23 (1975), pp. 193–197.
[19] R. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.
[20] D. French and J. King, *Approximation of an elliptic control problem by the finite element method*, Numer. Funct. Anal. Optim., 12 (1991), pp. 299–314.
[21] T. Geveci, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numér., 13 (1979), pp. 313–328.
[22] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, London, Melbourne, 1985.
[23] M. D. Gunzburger, L. S. Hou, and Th. P. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 711–748.
[24] W. W. Hager, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.
[25] W. Hager, *Numerical analysis in optimal control*, in Optimal Control of Complex Structures, Internat. Ser. Numer. Math. 139, Birkhäuser, Basel, 2002, pp. 83–93.
[26] M. Hinze, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.

[27] G. Knowles, *Finite element approximation of parabolic time optimal control problems*, SIAM J. Control Optim., 20 (1982), pp. 414–427.
[28] I. Lasiecka, *Boundary control of parabolic systems: Finite-element approximations*, Appl. Math. Optim., 6 (1980), pp. 287–333.
[29] I. Lasiecka, *Ritz–Galerkin approximation of the time optimal boundary control problem for parabolic systems with Dirichlet boundary conditions*, SIAM J. Control Optim., 22 (1984), pp. 477–500.
[30] J. L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes*, Dunod, Paris, 1968.
[31] K. Malanowski, C. Büskens, and H. Maurer, *Convergence of approximations to nonlinear control problems*, in Mathematical Programming with Data Perturbation, A. V. Fiacco, ed., Marcel Dekker, Inc., New York, 1997, pp. 253–284.
[32] R. S. McKnight and W. E. Bosarge, Jr., *The Ritz–Galerkin procedure for parabolic control problems*, SIAM J. Control Optim., 11 (1973), pp. 510–524.
[33] C. Meyer and A. Rösch, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.
[34] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.
[35] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[36] D. Tiba and F. Tröltzsch, *Error estimates for the discretization of state constrained convex control problems*, Numer. Funct. Anal. Optim., 17 (1996), pp. 1005–1028.
[37] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.
[38] F. Tröltzsch, *Semidiscrete finite element approximation of parabolic boundary control problems—convergence of switching points*, in Optimal Control of Partial Differential Equations II, Internat. Ser. Numer. Math. 78, Birkhäuser, Basel, 1987, pp. 219–232.
[39] F. Tröltzsch, *Approximation of nonlinear parabolic boundary problems by the Fourier method—convergence of optimal controls*, Optimization, 2 (1991), pp. 83–98.
[40] F. Tröltzsch, *On convergence of semidiscrete Ritz-Galerkin schemes applied to the boundary control of parabolic equations with nonlinear boundary condition*, Z. Angew. Math. Mech., 72 (1992), pp. 291–301.
[41] F. Tröltzsch, *Semidiscrete Ritz-Galerkin approximation of nonlinear parabolic boundary control problems—strong convergence of optimal controls*, Appl. Math. Optim., 29 (1994), pp. 309–329.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1612–1632

OPTIMAL ENERGY DECAY RATE FOR PARTIALLY DAMPED SYSTEMS BY SPECTRAL COMPENSATION*

PAOLA LORETI† AND BOPENG RAO‡

Abstract. We study the stability of weakly coupled and partially damped systems by means of the Riesz basis approach in higher dimensional spaces. We propose a weaker structural damping that compensates for the behavior of the eigenvalues of the system, therefore giving the optimal polynomial energy decay rate for smooth initial data.

Key words. weaker damping, spectral compensation, optimal energy decay rate, Riesz basis in higher dimensional space

AMS subject classification. 93D20

DOI. 10.1137/S0363012903437319

1. Introduction.

The aim of this paper, the results of which were announced in [25], is to investigate the energy decay rate of the weakly coupled and partially damped system

$$\begin{cases} y_{tt} + Ay + By_t + au = 0, \\ u_{tt} + Au + ay = 0, \end{cases} \tag{1.1}$$

where a is a real number, A is a self-adjoint coercive operator, and B is a linear bounded positive operator in a separated Hilbert space H. Assume furthermore that the resolvent of A is compact in H. Then there exists an increasing sequence $\mu_n^2 \to +\infty$ and an orthonormal sequence $e_n \in H$ such that

$$Ae_n = \mu_n^2 e_n \qquad \forall n \geq 1. \tag{1.2}$$

Write (1.1) as

$$\frac{d}{dt}\begin{pmatrix} y \\ z \\ u \\ v \end{pmatrix} = \begin{pmatrix} z \\ -Ay - Bz - au \\ v \\ -Au - ay \end{pmatrix} =: \mathcal{A}\begin{pmatrix} y \\ z \\ u \\ v \end{pmatrix}. \tag{1.3}$$

For the case when a is small enough, it was shown in [3] that (1.3) generates a C_0-semigroup of contractions on the Hilbert space

$$\mathcal{H} = D(A^{\frac{1}{2}}) \times H \times D(A^{\frac{1}{2}}) \times H.$$

Moreover, let

$$w_n = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ 0 \\ \frac{e_n}{i\mu_n} \\ e_n \end{pmatrix}.$$

*Received by the editors November 6, 2003; accepted for publication (in revised form) May 22, 2006; published electronically November 14, 2006.
http://www.siam.org/journals/sicon/45-5/43731.html

†Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università degli Studi di Roma "La Sapienza", Via A. Scarpa n. 16, 00161 Roma, Italy (loreti@dmmm.unironme1.it).

‡Institut de Recherche Mathématique Avancée, Université Louis Pasteur de Strasbourg, 7 Rue René-Descartes, 67084 Strasbourg, France (rao@math.u-strasbg.fr).

Then using (1.2) a straightforward computation gives

$$\|w_n\|_{\mathcal{H}} = 1, \qquad \|(i\mu_n - \mathcal{A})w_n\|_{\mathcal{H}}^2 = \frac{a^2}{2\mu_n^2} \to 0.$$

This shows that the resolvent of $\mathcal{A}$ is not uniformly bounded on the imaginary axis. Following [15] and [27] (see also [24] for applications) the system (1.3) is not uniformly stable in $\mathcal{H}$.

The fractional order damping of the type $B = A^\gamma$, arising from the material property, has been introduced in [7]. The cases $\gamma = 0, 1/2, 1$ include so-called viscous damping, "square-root" damping or structural damping, and Kelvin–Voigt damping. Especially for $\gamma = 1/2$ it was shown in [7] that the semigroup corresponding to the damped elastic model

$$y_{tt} + Ay + A^\gamma y_t = 0 \tag{1.4}$$

is analytical, while the subsequent works in [8] and [9] showed that the semigroup is still analytical for $1/2 \le \gamma \le 1$ but only of Gevrey class for $0 < \gamma < 1/2$. Moreover, the optimal regularity of solutions of system (1.8) with boundary control was also obtained in [14].

Now assume that $B = I$; then the spectrum of $\mathcal{A}$ has asymptotic expansions

$$\lambda_{1,n} \sim \pm i\mu_n - \frac{1}{2}, \quad \lambda_{2,n} \sim \pm i\mu_n - \frac{a^2}{2\mu_n^2}. \tag{1.5}$$

Then the energy corresponding to the first branch of eigenvalues decays exponentially and the energy corresponding to the second branch of eigenvalues decays only at the rate of $1/t$. Therefore the total energy decays at the rate of $1/t$.

Inspired by this remark, we look for a weaker damping operator for $\gamma < 0$. In that case, we have the following asymptotic expansions of the eigenvalues:

$$\begin{cases} \lambda_{1,n}^{\pm} \sim \pm i\mu_n - \frac{\mu_n^{2\gamma}}{2}, \quad \lambda_{2,n}^{\pm} \sim \pm i\mu_n - \frac{a^2}{2\mu_n^{2\gamma+2}}, & 2\gamma + 1 > 0, \\ \lambda_{1,n}^{\pm} \sim \pm i\mu_n - \frac{\mu_n^{2\gamma}}{2}, \quad \lambda_{2,n}^{\pm} \sim \pm i\mu_n - \frac{\mu_n^{2\gamma}}{2}, & 2\gamma + 1 < 0, \end{cases} \tag{1.6}$$

and the eigenvectors

$$e_{1,n}^{\pm} \sim \frac{1}{\sqrt{2}} \begin{pmatrix} \frac{e_n}{\pm i\mu_n} \\ e_n \\ 0 \\ 0 \end{pmatrix}, \quad e_{2,n}^{\pm} \sim \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 0 \\ \frac{e_n}{\pm i\mu_n} \\ e_n \end{pmatrix}, \quad 2\gamma + 1 > 0, \tag{1.7}$$

$$e_{1,n}^{\pm} \sim \frac{1}{2} \begin{pmatrix} \frac{e_n}{\pm i\mu_n} \\ e_n \\ \frac{e_n}{\pm i\mu_n} \\ e_n \end{pmatrix}, \qquad e_{2,n}^{\pm} \sim \frac{1}{2} \begin{pmatrix} \frac{e_n}{\mp i\mu_n} \\ -e_n \\ \frac{e_n}{\pm i\mu_n} \\ e_n \end{pmatrix}, \qquad 2\gamma + 1 < 0. \tag{1.8}$$

If $2\gamma + 1 > 0$, then the real part of $\lambda_{1,n}^{\pm}$ is of order $\mu_n^{2\gamma}$, and $\lambda_{2,n}^{\pm}$ is of order $\frac{1}{\mu_n^{2\gamma+2}}$. Therefore the energy corresponding to the first branch decays at the rate of $t^{\frac{1}{\gamma}}$, and the energy corresponding to the second branch decays at the rate of $t^{-\frac{1}{\gamma+1}}$. Since $-\frac{1}{\gamma+1} > \frac{1}{\gamma}$, the total energy decays only at the rate of $t^{-\frac{1}{\gamma+1}}$. In this case, the eigenvectors $e_{1,n}^{\pm}, e_{2,n}^{\pm}$ are asymptotically decoupled, and the two equations of (1.1) are very weakly coupled or almost independent.

If $2\gamma + 1 < 0$, then the real parts of $\lambda_{1,n}^{\pm}$ and $\lambda_{2,n}^{\pm}$ are of the same order $\mu_n^{2\gamma}$. The total energy decays at the rate of $t^{\frac{1}{\gamma}}$ and achieves the maximum decay $1/t^2$ for $2\gamma + 1 = 0$. In that case, the eigenvectors $e_{1,n}^{\pm}$ are effectively involved with the eigenvectors $e_{2,n}^{\pm}$, and the two equations are really coupled.

The unbalanced distribution of energy is a natural phenomenon for partially damped systems. The essential question is, how is the dissipation transmitted from one equation to another? Using the spectral compensation, we show that this depends on the damping and the coupling. Lebeau and Zuazua [17] (see also [6] for the generalization) estimated the total energy of the three-dimensional thermoelastic system by means of the energy concentrated on its longitudinal component. By a construction of geometric optics, they shown that there exist solutions whose energy is mostly concentrated on the transversal component. Consequently, the energy concentrated on the transversal component decays very slowly. From this point of view, the nonuniform decay is due to the unbalanced distribution of energy. The same phenomenon was observed for a magneto-elastic system [11]. Also the transmission of the dissipation plays an important role in the theory of control of networks as the controllers are enacted on only a part of the network. We refer to [10] for the network of elastic strings and to [18] for the network of Saint-Venant system of unsteady flows.

The paper is organized as follows. In section 2, we establish a general result on a polynomial decay rate of energy by a spectral approach. Section 3 is devoted to the study of the optimal decay rate of system (1.1) with a weaker damping $B = A^{\gamma}$ for $\gamma < 0$. Under a suitable framework, we can establish the asymptotic expansions for the eigenvalues and the eigenvectors as in (1.6)–(1.7). But these expansions could not give the desired quadratical convergence for the usual Riesz basis approach, except in one dimension. We will construct a sequence of pairwise orthogonal subspaces. In this way, we reduced the quadratical convergence to the Parseval equality. In order to obtain a Riesz basis we need only some weaker asymptotic expansions on the subspaces. In section 4, we give some examples of application.

There are many results concerning the polynomial decay rate. The majority were obtained by spectral approaches [16], [19], [33], [34]. Others results were obtained by frequency domain methods [5], [22], [23] and multiplier methods [3], [29], [30]. In particular, the method of weak observability inequality used in [4] and [31] works also, but it yields a lower decay rate than that obtained with our spectral method. Finally, we mention [32] for a general formulation for indirectly damped systems and [21] for the exact controllability and the observability of coupled distributed systems.

2. Optimal polynomial energy decay rate by spectral approach. In this section, we give a spectral approach for the polynomial energy decay rate of C_0-semigroups.

THEOREM 2.1. *Let $S(t)$ be a C_0-semigroup of contractions generated by the operator $\mathcal{A}$ on a Hilbert space $\mathcal{H}$. Let $\lambda_{k,n}(1 \le k \le K)$ denote the kth branch of eigenvalues of $\mathcal{A}$ and $\{e_{k,n}\}_{1\le k\le K, n\ge 1}$ the system of eigenvectors which forms a Riesz basis in $\mathcal{H}$. Assume that for each $1 \le k \le K$ there exist a positive sequence $\mu_{k,n} \to +\infty$ as $n \to +\infty$ and two positive constants $\alpha_k \ge 0, \beta_k > 0$ such that*

$$Re\lambda_{k,n} \le -\frac{\beta_k}{\mu_{k,n}^{\alpha_k}} \quad \text{and} \quad \left|Im\lambda_{k,n}\right| \ge \mu_{k,n} \qquad \forall n \ge 1. \tag{2.1}$$

Then for any $u_0 \in D(\mathcal{A}^{\theta})$ with $\theta > 0$, there exists a constant $M > 0$ independent of u_0 such that

$$\|S(t)u_0\|_{\mathcal{H}}^2 \le \|\mathcal{A}^{\theta}u_0\|_{\mathcal{H}}^2 \frac{M}{t^{2\theta\delta}} \qquad \forall t > 0, \tag{2.2}$$

where the decay rate δ is given by

$$\delta := \min_{1\le k\le K} \frac{1}{\alpha_k} = \frac{1}{\alpha_l}. \tag{2.3}$$

Moreover, if there exists a constant $c_1 > 0, c_2 > 0$ such that

$$Re\lambda_{l,n} \ge -\frac{c_1}{\mu_{l,n}^{\alpha_l}} \quad \text{and} \quad \left|Im\lambda_{l,n}\right| \le c_2\mu_{l,n} \qquad \forall n \ge 1, \tag{2.4}$$

then the decay rate δ given in (2.3) *is optimal.*

Proof. Since $\{e_{k,n}\}_{1\le k\le K, n\ge 1}$ is a Riesz basis in $\mathcal{H}$, any $u_0 \in D(\mathcal{A}^\theta)$ can be written as

$$u_0 = \sum_{k=1}^{K}\sum_{n=1}^{+\infty} a_{k,n}e_{k,n}. \tag{2.5}$$

Moreover, there exist two constants $C_1 > 0, C_2 > 0$ such that

$$C_1\sum_{k=1}^{K}\sum_{n=1}^{+\infty}|a_{k,n}|^2 \le \|u_0\|_{\mathcal{H}}^2 \le C_2\sum_{k=1}^{K}\sum_{n=1}^{+\infty}|a_{k,n}|^2. \tag{2.6}$$

Using expansion (2.5) and the continuity of the semigroup $S(t)$, we get

$$S(t)u_0 = \sum_{k=1}^{K}\sum_{n=1}^{+\infty} a_{k,n}e^{\lambda_{k,n}t}e_{k,n}. \tag{2.7}$$

Then using the first conditions of (2.1) and the second inequality of (2.6) in (2.7), we obtain

$$\|S(t)u_0\|_{\mathcal{H}}^2 \le \sup_{1\le k\le K}\sup_{n\ge 1}\left(\frac{C_2}{\mu_{k,n}^{2\theta}\exp\left(\frac{2\beta_k t}{\mu_{k,n}^{\alpha_k}}\right)}\right)\sum_{k=1}^{K}\sum_{n=1}^{+\infty}|a_{k,n}|^2\mu_{k,n}^{2\theta}. \tag{2.8}$$

On the other hand, the conditions of (2.1) imply that there exists a positive constant $C_3 > 0$ such that

$$\sum_{k=1}^{K}\sum_{n=1}^{+\infty}\mu_{k,n}^{2\theta}|a_{k,n}|^2 \le C_3\sum_{k=1}^{K}\sum_{n=1}^{+\infty}|\lambda_{k,n}|^{2\theta}|a_{k,n}|^2 \le \frac{C_3}{C_1}\|\mathcal{A}^\theta u_0\|_{\mathcal{H}}^2. \tag{2.9}$$

Then inserting (2.9) into (2.8), we get

$$\|S(t)u_0\|_{\mathcal{H}}^2 \le \sup_{1\le k\le K}\sup_{n\ge 1}\left(\frac{C_2C_3}{C_1\mu_{k,n}^{2\theta}\exp\left(\frac{2\beta_k t}{\mu_{k,n}^{\alpha_k}}\right)}\right)\|\mathcal{A}^\theta u_0\|_{\mathcal{H}}^2. \tag{2.10}$$

If $\alpha_k = 0$ for some $1 \le k \le K$, since $\inf_{n\ge 1}\mu_{k,n} > 0$, then it is easy to find a constant $M_k > 0$ such that

$$\frac{1}{\mu_{k,n}^{2\theta}\exp\left(2\beta_k t\right)} \le \frac{M_k}{t^{2\theta\delta}} \qquad \forall n \ge 1. \tag{2.11}$$

If $\alpha_k > 0$ for some $1 \le k \le K$, then putting

$$\frac{t}{\mu_{k,n}^{\alpha_k}} = \frac{1}{x}, \qquad f_k(x) = x^{\frac{2\theta}{\alpha_k}} e^{\frac{2\beta_k}{x}},$$

we rewrite

$$\mu_{k,n}^{2\theta} \exp\Big(\frac{2\beta_k t}{\mu_{k,n}^{\alpha_k}}\Big) = f_k(x) t^{\frac{2\theta}{\alpha_k}}. \tag{2.12}$$

Then a straightforward computation gives

$$f_k'(x) = 2e^{\frac{\beta_k}{x}} x^{\frac{2\theta}{\alpha_k}-2}\Big(\frac{\theta}{\alpha_k}x - \beta_k\Big) \quad \begin{cases} > 0, & x > \frac{\alpha_k\beta_k}{\theta}, \\ < 0, & x < \frac{\alpha_k\beta_k}{\theta}. \end{cases}$$

It follows that

$$\inf_{x>0} f_k(x) = f_k\Big(\frac{\alpha_k\beta_k}{\theta}\Big) = \Big(\frac{\alpha_k\beta_k}{\theta}\Big)^{\frac{2\theta}{\alpha_k}} e^{\frac{2\theta}{\alpha_k}} := M_k^{-1} > 0. \tag{2.13}$$

Inserting (2.13) into (2.12) gives that

$$\frac{1}{\mu_{k,n}^{2\theta} \exp\Big(\frac{2\beta_k t}{\mu_{k,n}^{\alpha_k}}\Big)} \le \frac{M_k}{t^{\frac{2\theta}{\alpha_k}}}, \qquad n \ge 1. \tag{2.14}$$

Finally, combining the cases (2.11) and (2.14), we get the polynomial energy decay rate (2.2), with the constant M given by

$$M = \frac{C_2 C_3}{C_1} \max_{1\le k\le K} M_k. \tag{2.15}$$

Now we consider the optimality of δ. To simplify the notation, we write

$$\alpha_l = \alpha, \quad \mu_{l,n} = \mu_n, \quad \lambda_{l,n} = \lambda_n, \quad e_{l,n} = e_n.$$

Since $\mu_n \to +\infty$, then for any $\epsilon > 0$, we can find a subsequence, still denoted by μ_n, such that

$$\sum_{n=1}^{\infty} \frac{1}{\mu_n^{\alpha\epsilon}} < +\infty. \tag{2.16}$$

Then putting

$$u_0 = \sum_{n=1}^{\infty} \frac{C_4}{\mu_n^{(\theta+\alpha\epsilon/2)}} e_n, \tag{2.17}$$

we see that $u_0 \in D(\mathcal{A}^\theta)$ due to the convergence (2.16). On the other hand, thanks to the second condition of (2.4), we can choose a constant $C_4 > 0$ such that $\|\mathcal{A}^\theta u_0\|_{\mathcal{H}} = 1$. Now using the first condition of (2.4), we get

$$\|S(t)u_0\|_{\mathcal{H}}^2 = \Big\| \sum_{n=1}^{\infty} \frac{C_4 e^{\lambda_n t} e_n}{\mu_n^{(\theta+\alpha\epsilon/2)}} \Big\|_{\mathcal{H}}^2 \ge \frac{C_1 C_4^2}{\mu_m^{2\theta+\alpha\epsilon} \exp\Big(\frac{2c_1 t}{\mu_m^\alpha}\Big)} \qquad \forall m \ge 1. \tag{2.18}$$

Finally, setting $t_m = \mu_m^\alpha$ in (2.18), we obtain that

$$\|S(t_m)u_0\|_{\mathcal{H}}^2 \geq \frac{C_1 C_4^2}{e^{2c_1}\mu_m^{2\theta+\alpha\epsilon}} = \frac{C_1 C_4^2}{e^{2c_1} t_m^{2\theta\delta+\epsilon}} \qquad \forall m \geq 1. \tag{2.19}$$

This means that the trajectory $S(t)u_0$ decays slower than $\frac{1}{t^{2\theta\delta+\epsilon}}$ on the time sequence $t_m \to +\infty$. Then for any $\epsilon > 0$, we cannot expect the decay rate $\frac{1}{t^{2\theta\delta+\epsilon}}$ for all initial data $u_0 \in D(\mathcal{A}^\theta)$ and for all $t > 0$. The proof is thus complete. □

Remark 2.1. If $\lambda_{k,n}$ is an eigenvalue of algebraic multiplicity $d > 1$, then the corresponding factors in (2.8) and (2.14) will be replaced by

$$\frac{|p(t)|^2}{\mu_{k,n}^{2\theta} \exp\left(\frac{2\beta_k t}{\mu_{k,n}^{\alpha_k}}\right)} \leq \frac{M_k}{t^{2\theta\delta}},$$

where $p(t)$ is a polynomial of degree $d-1$ and $M_k > 0$ is a constant. Therefore, Theorem 2.1 remains valid if the operator $\mathcal{A}$ admits a finite number of algebraically multiple eigenvalues and the system of root vectors forms a Riesz basis in $\mathcal{H}$.

Remark 2.2. Theorem 2.1 is valid for all initial data $u_0 \in D(\mathcal{A}^\theta)$ with $\theta > 0$. This is different from an earlier result of Littman and Markus in [20], where the initial data u_0 satisfies some stronger conditions.

3. Spectral compensation for weakly damped systems. Let A be a densely defined closed self-adjoint operator in a Hilbert space H such that for some positive constant c, we have

$$(Au, u) \geq c\|u\|_H^2 \qquad \forall u \in D(A). \tag{3.1}$$

Let $\gamma \leq 0$, and let a be a small real number. We consider the weakly coupled equations

$$y_{tt} + Ay + A^\gamma y_t + au = 0, \tag{3.2}$$

$$u_{tt} + Au + ay = 0. \tag{3.3}$$

In order to curry the explicit computation of the eigenvalues, we have chosen the same operator A in the two equations of system (3.2)–(3.3). The results of this section complete and improve a recent work of Alabau, Cannarsa, and Komornik [3].

Define the Hilbert space

$$\mathcal{H} = D(A^{\frac{1}{2}}) \times H \times D(A^{\frac{1}{2}}) \times H$$

equipped with the equivalent inner product

$$\begin{aligned}\langle (y,z,u,v), (f,g,p,q)\rangle_{\mathcal{H}} =& (A^{\frac{1}{2}}y, A^{\frac{1}{2}}f) + (z,g) + (A^{\frac{1}{2}}u, A^{\frac{1}{2}}p) + (v,q) \\ &+ a\big((y,p) + (u,f)\big).\end{aligned}$$

Setting

$$D(\mathcal{A}) = \{w = (y,z,u,v)^T \in \mathcal{H} \quad z, v \in D(A^{\frac{1}{2}}), \quad y, u \in D(A)\},$$

$$\mathcal{A}\begin{pmatrix} y \\ z \\ u \\ v \end{pmatrix} = \begin{pmatrix} z \\ -Ay - A^\gamma z - au \\ v \\ -Au - ay \end{pmatrix},$$

we can write system (3.2)–(3.3) as an evolution equation

$$\frac{d}{dt}w = \mathcal{A}w, \qquad w(0) = w_0. \tag{3.4}$$

Under condition (3.1), we can prove easily that $\mathcal{A}$ is a maximal dissipative operator on $\mathcal{H}$, provided that a is small enough (e.g., $|a| < c$). Therefore $\mathcal{A}$ generates a C^0-semigroup of contractions on $\mathcal{H}$ (see Theorem 1.4.3 in [26]). Moreover, if we set the energy as

$$E(t) = \frac{1}{2}\Big(\|A^{\frac{1}{2}}y\|^2 + \|y_t\|^2 + \|A^{\frac{1}{2}}u\|^2 + \|u_t\|^2 + a(y,u) + a(u,y)\Big),$$

then a straightforward computation gives that

$$\frac{d}{dt}E(t) = -\|A^{\gamma}y_t\|^2 \le 0.$$

Assume furthermore that the resolvent of A is compact in H. Then there exists an increasing sequence $\mu_n^2 \to +\infty$ and an orthonormal sequence $e_n \in H$ such that

$$Ae_n = \mu_n^2 e_n \qquad \forall n \ge 1. \tag{3.5}$$

Moreover, the system $\{e_n\}_{n\ge1}$ forms a Hilbert basis on H.

Now let λ be an eigenvalue and (y,z,u,v) be the associated eigenvector of the operator $\mathcal{A}$. Then we have

$$\begin{cases} z = \lambda y, \\ -Ay - A^{\gamma}z - au = \lambda z, \\ v = \lambda u, \\ -Au - ay = \lambda v. \end{cases} \tag{3.6}$$

We will see in Proposition 3.3 that all eigenvectors of $\mathcal{A}$ are of the form

$$y = \alpha_n e_n, \qquad z = \lambda\alpha_n e_n, \qquad u = \beta_n e_n, \qquad v = \lambda\beta_n e_n. \tag{3.7}$$

Inserting (3.7) into (3.6), we get

$$\begin{cases} (\lambda^2 + \mu_n^2 + \lambda\mu_n^{2\gamma})\alpha_n + a\beta_n = 0, \\ a\alpha_n + (\lambda^2 + \mu_n^2)\beta_n = 0, \end{cases} \tag{3.8}$$

which has nontrivial solution $(\alpha_n, \beta_n) \neq (0,0)$ if and only if λ is a solution of the equation

$$(\lambda^2 + \lambda\mu_n^{2\gamma} + \mu_n^2)(\lambda^2 + \mu_n^2) = a^2. \tag{3.9}$$

PROPOSITION 3.1. *For each $\mu_n > 0$, the corresponding eigenvalues $\lambda_{1,n}^{\pm}, \lambda_{2,n}^{\pm}$ of the system* (3.2)–(3.3) *satisfy the following asymptotic expansions.*

I. *If $-1/2 < \gamma \le 0$, then*

$$\lambda_{1,n}^{\pm} = \pm i\mu_n - \frac{\mu_n^{2\gamma}}{2} + O\Big(\frac{1}{\mu_n}\Big), \tag{3.10}$$

$$\lambda_{2,n}^{\pm} = \pm i\mu_n - \frac{a^2}{2\mu_n^{2\gamma+2}} + O\Big(\frac{1}{\mu_n^{6\gamma+4}}\Big). \tag{3.11}$$

II. *If* $\gamma < -1/2$, *then*

$$\lambda_{1,n}^{\pm} = \pm i\sqrt{\mu_n^2 + a} - \frac{\mu_n^{2\gamma}}{2} + O(\mu_n^{4\gamma+1}), \tag{3.12}$$

$$\lambda_{2,n}^{\pm} = \pm i\sqrt{\mu_n^2 - a} - \frac{\mu_n^{2\gamma}}{2} + O(\mu_n^{4\gamma+1}). \tag{3.13}$$

III. *If* $\gamma = -1/2$, *then*

$$\lambda_{1,n}^{\pm} = \pm i\mu_n - \frac{1 \pm \sqrt{1-4a^2}}{4\mu_n} + O\Big(\frac{1}{\mu_n^2}\Big), \tag{3.14}$$

$$\lambda_{2,n}^{\pm} = \pm i\mu_n - \frac{1 \mp \sqrt{1-4a^2}}{4\mu_n} + O\Big(\frac{1}{\mu_n^2}\Big). \tag{3.15}$$

Proof. First, let λ_n be one of the four solutions of (3.9). If $|\lambda_n^2 + \mu_n^2| \le 1$, then it follows that

$$\frac{\lambda_n}{\mu_n} = \pm i + O(\mu_n^{-2}). \tag{3.16}$$

If $|\lambda_n^2 + \mu_n^2| \ge 1$, then from (3.9) we get

$$\Big(\frac{\lambda_n}{\mu_n}\Big)^2 + \frac{\lambda_n}{\mu_n}\mu_n^{2\gamma-1} + 1 = O(\mu_n^{-2}).$$

It follows that

$$\frac{\lambda_n}{\mu_n} = \pm i + O(\mu_n^{2\gamma-1}). \tag{3.17}$$

Combining (3.16) and (3.17), we get

$$\frac{\lambda_n}{\mu_n} = \pm i + O(\mu_n^{2\gamma-1}) + O(\mu_n^{-2}). \tag{3.18}$$

Next, solving (3.9), we get

$$2(\lambda^2 + \mu_n^2) = -\lambda\mu_n^{2\gamma} \mp \sqrt{\lambda^2\mu_n^{4\gamma} + 4a^2}. \tag{3.19}$$

The symbol $a \sim b$ means that $a - b \to 0$. In the following computations we will use often the expansion $\sqrt{1+z} = 1 + z/2 + O(z^2)$.

Case I. $-1/2 < \gamma \le 0$. From (3.18), we have

$$|\lambda_n\mu_n^{2\gamma}| \sim \mu_n^{1+2\gamma} \to +\infty. \tag{3.20}$$

Using the asymptotic expansion (3.20) in (3.19), we get

$$\begin{cases} \lambda_{1,n}^2 = -\mu_n^2 - \lambda_{1,n}\mu_n^{2\gamma} + O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \\ \lambda_{2,n}^2 = -\mu_n^2 + \dfrac{a^2}{\lambda_{2,n}\mu_n^{2\gamma}} + O\Big(\dfrac{1}{\mu_n^{6\gamma+3}}\Big). \end{cases} \tag{3.21}$$

It follows from (3.21) that

$$\begin{cases} \lambda_{1,n}^{\pm} = \pm i\mu_n \pm \dfrac{i\lambda_{1,n}\mu_n^{2\gamma-1}}{2} + O\Big(\dfrac{1}{\mu_n}\Big), \\ \lambda_{2,n}^{\pm} = \pm i\mu_n \mp \dfrac{ia^2}{2\lambda_{2,n}\mu_n^{2\gamma+1}} + O\Big(\dfrac{1}{\mu_n^{6\gamma+4}}\Big). \end{cases} \tag{3.22}$$

In particular, we get

$$\begin{cases} \lambda_{1,n}^{\pm} = \pm i\mu_n + O(\mu_n^{2\gamma}), \\ \lambda_{2,n}^{\pm} = \pm i\mu_n + O\Big(\frac{1}{\mu_n^{2\gamma+2}}\Big). \end{cases} \tag{3.23}$$

Inserting (3.23) into (3.22) gives that

$$\lambda_{1,n}^{\pm} = \pm i\mu_n - \frac{\mu_n^{2\gamma}}{2} + O\Big(\frac{1}{\mu_n}\Big), \tag{3.24}$$

$$\lambda_{2,n}^{\pm} = \pm i\mu_n - \frac{a^2}{2\mu_n^{2\gamma+2}} + O\Big(\frac{1}{\mu_n^{6\gamma+4}}\Big). \tag{3.25}$$

Case II. $\gamma < -1/2$. From (3.18) we have

$$|\lambda_n \mu_n^{2\gamma}| \sim |\mu_n^{1+2\gamma}| \to 0. \tag{3.26}$$

Using the asymptotic expansion (3.26) in (3.19), we have

$$\begin{cases} \lambda_{1,n}^2 = -(\mu_n^2 + a) - \dfrac{\lambda_{1,n}\mu_n^{2\gamma}}{2} + O\Big(\mu_n^{4\gamma+2}\Big), \\ \lambda_{2,n}^2 = -(\mu_n^2 - a) - \dfrac{\lambda_{1,n}\mu_n^{2\gamma}}{2} + O\Big(\mu_n^{4\gamma+2}\Big). \end{cases} \tag{3.27}$$

It follows from (3.27) that

$$\begin{cases} \lambda_{1,n}^{\pm} = \pm i\sqrt{\mu_n^2 + a} \pm \dfrac{i\lambda_{1,n}\mu_n^{2\gamma}}{2\sqrt{\mu_n^2 + a}} + O(\mu_n^{4\gamma+1}), \\ \lambda_{2,n}^{\pm} = \pm i\sqrt{\mu_n^2 - a} \pm \dfrac{i\lambda_{1,n}\mu_n^{2\gamma}}{2\sqrt{\mu_n^2 - a}} + O(\mu_n^{4\gamma+1}). \end{cases} \tag{3.28}$$

In particular, we get

$$\begin{cases} \lambda_{1,n}^{\pm} = \pm i\sqrt{\mu_n^2 + a} + O(\mu_n^{2\gamma}), \\ \lambda_{2,n}^{\pm} = \pm i\sqrt{\mu_n^2 - a} + O(\mu_n^{2\gamma}). \end{cases} \tag{3.29}$$

Inserting (3.29) into (3.28), we have

$$\begin{cases} \lambda_{1,n}^{\pm} = \pm i\sqrt{\mu_n^2 + a} - \dfrac{\mu_n^{2\gamma}}{2} + O(\mu_n^{4\gamma+1}), \\ \lambda_{2,n}^{\pm} = \pm i\sqrt{\mu_n^2 - a} - \dfrac{\mu_n^{2\gamma}}{2} + O(\mu_n^{4\gamma+1}). \end{cases} \tag{3.30}$$

Case III. $\gamma = -1/2$. From (3.18) we have

$$\frac{\lambda_{1,n}^{\pm}}{\mu_n} \sim \pm i + O\Big(\frac{1}{\mu_n^2}\Big), \qquad \frac{\lambda_{2,n}^{\pm}}{\mu_n} \sim \pm i + O\Big(\frac{1}{\mu_n^2}\Big). \tag{3.31}$$

Using the asymptotic expansion (3.31) in (3.19), we have

$$\begin{cases} \lambda_{1,n}^2 = -\mu_n^2 \mp \dfrac{i}{2}(1 \pm \sqrt{1-4a^2}) + O\Big(\dfrac{1}{\mu_n^2}\Big), \\ \lambda_{2,n}^2 = -\mu_n^2 \mp \dfrac{i}{2}(1 \mp \sqrt{1-4a^2}) + O\Big(\dfrac{1}{\mu_n^2}\Big). \end{cases} \tag{3.32}$$

It follows from (3.32) that

$$\lambda_{1,n}^{\pm} = \pm i\mu_n - \frac{1 \pm \sqrt{1-4a^2}}{4\mu_n} + O\Big(\frac{1}{\mu_n^3}\Big), \tag{3.33}$$

$$\lambda_{2,n}^{\pm} = \pm i\mu_n - \frac{1 \mp \sqrt{1-4a^2}}{4\mu_n} + O\Big(\frac{1}{\mu_n^3}\Big). \tag{3.34}$$

The proof is thus complete. □

If $\lambda_{1,n}^{\pm}, \lambda_{2,n}^{\pm}$ are simple eigenvalues, then by setting

$$\beta_{1,n}^{\pm} = -\frac{a}{(\lambda_{1,n}^{\pm})^2 + \mu_n^2}\alpha_{1,n}^{\pm}, \tag{3.35}$$

$$\alpha_{2,n}^{\pm} = -\frac{(\lambda_{2,n}^{\pm})^2 + \mu_n^2}{a}\beta_{2,n}^{\pm} \tag{3.36}$$

in (3.8), we get

$$e_{1,n}^{\pm} = \alpha_{1,n}^{\pm} \begin{pmatrix} \frac{e_n}{\lambda_{1,n}^{\pm}} \\ e_n \\ -\frac{ae_n}{\lambda_{1,n}^{\pm}((\lambda_{1,n}^{\pm})^2+\mu_n^2)} \\ -\frac{ae_n}{(\lambda_{1,n}^{\pm})^2+\mu_n^2} \end{pmatrix}, \quad e_{2,n}^{\pm} = \beta_{2,n}^{\pm} \begin{pmatrix} -\frac{((\lambda_{2,n}^{\pm})^2+\mu_n^2)e_n}{a\lambda_{2,n}^{\pm}} \\ -\frac{((\lambda_{2,n}^{\pm})^2+\mu_n^2)e_n}{a} \\ \frac{e_n}{\lambda_{2,n}^{\pm}} \\ e_n \end{pmatrix}. \tag{3.37}$$

Now let us denote by V_n the eigenspace corresponding to the four eigenvalues $\lambda_{1,n}^{\pm}, \lambda_{2,n}^{\pm}$:

$$V_n = \mathrm{Sp}\{e_{1,n}^{+}, e_{2,n}^{+}, e_{1,n}^{-}, e_{2,n}^{-}\}.$$

From the expressions in (3.37), the subspaces $\{V_n\}_{n\geq 1}$ are clearly pairwise orthogonal in $\mathcal{H}$. It is possible that (3.9) has multiple solutions for some special parameters μ_n, a. However, for a small enough the Rouché theorem shows that (3.9) admits at most a double solution for the first branch of eigenvalues $\lambda_{1,n}^{\pm} = \lambda_{1,n}, e_{1,n}^{\pm} = e_{1,n}$, which satisfies the equation of derivation

$$(2\lambda + \mu_n^{2\gamma})(\lambda^2 + \mu_n^2) + 2\lambda(\lambda^2 + \lambda\mu_n^{2\gamma} + \mu_n^2) = 0. \tag{3.38}$$

In that case, we look for the corresponding root vector $\widetilde{e}_{1,n} = (\widetilde{y}_n, \widetilde{z}_n, \widetilde{u}_n, \widetilde{v}_n)$ such that

$$(I - \lambda_{1,n}\mathcal{A})\widetilde{e}_{1,n} = e_{1,n},$$

which involves that

$$\begin{cases} \widetilde{z}_n - \lambda_{1,n}\widetilde{y}_n = \frac{e_n}{\lambda_{1,n}}, \\ -A\widetilde{y}_n - A^\gamma \widetilde{z}_n - a\widetilde{u}_n - \lambda\widetilde{z}_n = e_n, \\ \widetilde{v}_n - \lambda_{1,n}\widetilde{u}_n = -\frac{ae_n}{\lambda_{1,n}(\lambda_{1,n}^2+\mu_n^2)}, \\ -A\widetilde{u}_n - a\widetilde{y}_n - \lambda\widetilde{v}_n = -\dfrac{ae_n}{(\lambda_{1,n}^2+\mu_n^2)}. \end{cases} \tag{3.39}$$

Setting

$$\widetilde{y}_n = \alpha_{1,n}e_n, \qquad \widetilde{u}_n = \beta_{1,n}e_n \tag{3.40}$$

in (3.39), we get

$$\begin{cases} \widetilde{z}_n = \alpha_{1,n}\lambda_{1,n}e_n + \dfrac{e_n}{\lambda_{1,n}}, \\ \widetilde{v}_n = \beta_{1,n}\lambda_{1,n}e_n - \dfrac{ae_n}{\lambda_{1,n}(\lambda_{1,n}^2+\mu_n^2)}, \end{cases} \tag{3.41}$$

where the constants $\alpha_{1,n}$ and $\beta_{1,n}$ satisfy

$$\begin{cases} -(\mu^2 + \lambda_{1,n}\mu^{2\gamma} + \lambda_{1,n}^2)\alpha_{1,n} - a\beta_{1,n} = \dfrac{2\lambda_{1,n} + \mu^{2\gamma}}{\lambda_{1,n}}, \\ -a\alpha_{1,n} - (\mu^2 + \lambda_{1,n}^2)\beta_{1,n} = -\dfrac{2a}{\lambda_{1,n}^2 + \mu_n^2}. \end{cases} \tag{3.42}$$

Since $\lambda_{1,n}$ satisfies (3.9) and (3.38), the first equation of (3.42) can be reduced to the second one. Therefore, choosing

$$\alpha_{1,n} = \frac{2}{\lambda_{1,n}^2 + \mu_n^2}, \qquad \beta_{1,n} = 0 \tag{3.43}$$

in (3.40)–(3.41), we get the corresponding root vector

$$\widetilde{e}_{1,n} = \begin{pmatrix} \frac{2}{\lambda_{1,n}^2+\mu_n^2}e_n \\ \left(\frac{2a\lambda_{1,n}}{\lambda_{1,n}^2+\mu_n^2} + \frac{1}{\lambda_{1,n}}\right)e_n \\ 0 \\ -\frac{a}{\lambda_{1,n}(\lambda_{1,n}^2+\mu_n^2)}e_n \end{pmatrix}. \tag{3.44}$$

Accordingly, we modify the subspace V_n as

$$V_n = \mathrm{Sp}\{e_{1,n}, e_{2,n}^+, \widetilde{e}_{1,n}, e_{2,n}^-\}.$$

Once again, expression (3.44) shows that the subspaces $\{V_n\}_{n\geq 1}$ are still pairwise orthogonal in $\mathcal{H}$.

PROPOSITION 3.2. *Let $E_{1,n}^\pm, E_{2,n}^\pm$ be the eigenvectors of the decoupled system (corresponding to $a = 0$)*

$$E_{1,n}^\pm = \frac{1}{\sqrt{2}}\begin{pmatrix} \frac{e_n}{\pm i\mu_n} \\ e_n \\ 0 \\ 0 \end{pmatrix}, \qquad E_{2,n}^\pm = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ 0 \\ \frac{e_n}{\pm i\mu_n} \\ e_n \end{pmatrix}. \tag{3.45}$$

Then the following relationship holds:

$$(e^+_{1,n}, e^+_{2,n}, e^-_{1,n}, e^-_{2,n}) = (E^+_{1,n}, E^+_{2,n}, E^-_{1,n}, E^-_{2,n})L_n, \tag{3.46}$$

where L_n is a 4×4 matrix such as

$$L_n = \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{pmatrix} + O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \quad 2\gamma+1>0, \tag{3.47}$$

$$L_n = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 & 0 & 0\\ 1 & 1 & 0 & 0\\ 0 & 0 & 1 & -1\\ 0 & 0 & 1 & 1 \end{pmatrix} + O\Big(\frac{1}{\mu_n^{\min\{2,-(2\gamma+1)\}}}\Big), \quad 2\gamma+1<0, \tag{3.48}$$

$$L_n = \sqrt{2}\begin{pmatrix} \sqrt{a}_+ & \sqrt{a}_- & 0 & 0\\ \frac{ia}{\sqrt{a}_+} & \frac{ia}{\sqrt{a}_-} & 0 & 0\\ 0 & 0 & \sqrt{a}_- & \sqrt{a}_+\\ 0 & 0 & -\frac{ia}{\sqrt{a}_-} & -\frac{ia}{\sqrt{a}_+} \end{pmatrix} + O\Big(\frac{1}{\mu_n^2}\Big), \quad 2\gamma+1=0, \tag{3.49}$$

where we have put

$$a_\pm = \frac{1\pm\sqrt{1-4a^2}}{2}. \tag{3.50}$$

Proof. For $2\gamma+1>0$, using the asymptotic expansions (3.10)–(3.11), we have

$$\frac{1}{(\lambda^\pm_{1,n})^2+\mu_n^2} = O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \qquad \frac{1}{\lambda^\pm_{1,n}} = \frac{1}{\pm i\mu_n} + O(\mu_n^{2\gamma-2}), \tag{3.51}$$

$$(\lambda^\pm_{2,n})^2+\mu_n^2 = O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \qquad \frac{1}{\lambda^\pm_{2,n}} = \frac{1}{\pm i\mu_n} + O\Big(\frac{1}{\mu_n^{2\gamma+4}}\Big). \tag{3.52}$$

Inserting (3.51)–(3.52) into (3.37) gives

$$e^\pm_{1,n} = \frac{1}{\sqrt{2}}\begin{pmatrix} \frac{e_n}{\pm i\mu_n}\\ e_n\\ 0\\ 0 \end{pmatrix} + O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \tag{3.53}$$

$$e^\pm_{2,n} = \frac{1}{\sqrt{2}}\begin{pmatrix} 0\\ 0\\ \frac{e_n}{\pm i\mu_n}\\ e_n \end{pmatrix} + O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big). \tag{3.54}$$

For $2\gamma+1<0$, using (3.12)–(3.13), we have

$$\frac{1}{(\lambda^\pm_{1,n})^2+\mu_n^2} = -\frac{1}{a} + O(\mu_n^{2\gamma+1}), \quad \frac{1}{\lambda^\pm_{1,n}} = \frac{1}{\pm i\mu_n} + O\Big(\frac{1}{\mu_n^3}\Big), \tag{3.55}$$

$$\frac{1}{\lambda^\pm_{2,n}} = \frac{1}{\pm i\mu_n} + O\Big(\frac{1}{\mu_n^3}\Big), \quad (\lambda^\pm_{2,n})^2+\mu_n^2 = a + O(\mu_n^{2\gamma+1}). \tag{3.56}$$

Inserting (3.55)–(3.56) into (3.37) gives

$$e^{\pm}_{1,n} = \frac{1}{2}\begin{pmatrix} \frac{e_n}{\pm i\mu_n} \\ e_n \\ \frac{e_n}{\pm i\mu_n} \\ e_n \end{pmatrix} + O(\mu_n^{2\gamma+1}) + O\Big(\frac{1}{\mu_n^2}\Big), \tag{3.57}$$

$$e^{\pm}_{2,n} = \frac{1}{2}\begin{pmatrix} \frac{e_n}{\mp i\mu_n} \\ -e_n \\ \frac{e_n}{\pm i\mu_n} \\ e_n \end{pmatrix} + O(\mu_n^{2\gamma+1}) + O\Big(\frac{1}{\mu_n^2}\Big). \tag{3.58}$$

For $2\gamma + 1 = 0$, using (3.14)–(3.15), we have

$$(\lambda^{\pm}_{1,n})^2 + \mu_n^2 = \mp i a_{\pm} + O\Big(\frac{1}{\mu_n^2}\Big), \quad \frac{1}{\lambda^{\pm}_{1,n}} = \frac{1}{\pm i\mu_n} + O\Big(\frac{1}{\mu_n^3}\Big), \tag{3.59}$$

$$(\lambda^{\pm}_{2,n})^2 + \mu_n^2 = \mp i a_{\mp} + O\Big(\frac{1}{\mu_n^2}\Big), \quad \frac{1}{\lambda^{\pm}_{2,n}} = \frac{1}{\pm i\mu_n} + O\Big(\frac{1}{\mu_n^3}\Big). \tag{3.60}$$

Inserting (3.59)–(3.60) into (3.37) gives

$$e^{\pm}_{1,n} = \sqrt{a}_{\pm}\begin{pmatrix} \frac{e_n}{\pm i\mu_n} \\ e_n \\ -\frac{ae_n}{a_{\pm}\mu_n} \\ \frac{ae_n}{\pm i a_{\pm}} \end{pmatrix} + O\Big(\frac{1}{\mu_n^2}\Big), \tag{3.61}$$

$$e^{\pm}_{2,n} = \sqrt{a}_{\mp}\begin{pmatrix} \frac{e_n}{\pm i\mu_n} \\ e_n \\ -\frac{ae_n}{a_{\mp}\mu_n} \\ \frac{ae_n}{\pm i a_{\mp}} \end{pmatrix} + O\Big(\frac{1}{\mu_n^2}\Big). \tag{3.62}$$

The proof is thus complete. □

Remark 3.1. If $2\gamma + 1 > 0$, the leading term of the eigenvectors $e^{\pm}_{1,n}$ and $e^{\pm}_{2,n}$ are decoupled. System (3.2)–(3.3) is overdamped and "essentially decoupled." If $2\gamma + 1 < 0$, the leading term of $e^{\pm}_{1,n}$ and $e^{\pm}_{2,n}$ are well coupled. System (3.2)–(3.3) is right-damped in that case. In particular, when $2\gamma + 1 = 0$ the parameter a appears explicitly in the leading term of the matrix L_n.

LEMMA 3.1. *Let $\{X_n\}_{n\geq 1}$ be a Riesz basis of subspaces in a Hilbert space X and $\{Y_n\}_{n\geq 1}$ a Riesz sequence of subspaces in X. Assume that there exist a sequence of isomorphisms $\{L_n\}_{n\geq 1}$ from X_n onto Y_n and positive constants $m > 0, M > 0$ independent of n such that*

$$m\|x_n\| \leq \|L_n x_n\| \leq M\|x_n\| \qquad \forall x_n \in X_n, \quad \forall n \geq 1. \tag{3.63}$$

Assume furthermore that there exist a Riesz basis $\{f_{n,i}\}_{1\leq l\leq I_n}(I_n \leq +\infty)$ in each X_n and positive constants $c > 0, C > 0$ independent of n such that

$$c\sum_{i=1}^{I_n} |\alpha_{n,i}|^2 \leq \|x_n\|^2 \leq C\sum_{i=1}^{I_n} |\alpha_{n,i}|^2 \qquad \forall x_n = \sum_{i=1}^{I_n} \alpha_{n,i} f_{n,i}. \tag{3.64}$$

Then the sequence

$$g_{n,i} = L_n f_{n,i} \qquad \forall n \geq 1, \quad 1 \leq i \leq I_n, \tag{3.65}$$

forms a Riesz basis in X.

Proof. Since $\{X_n\}_{n\geq 1}$ is a Riesz basis of subspaces in X, then for any $x \in X$ there exists a unique sequence $\{x_n\}_{n\geq 1}$ with $x_n \in X_n$ such that

$$x = \sum_{n=1}^{\infty} x_n, \qquad c' \sum_{n=1}^{\infty} \|x_n\|^2 \leq \|x\|^2 \leq C' \sum_{n=1}^{\infty} \|x_n\|^2. \tag{3.66}$$

This combined with (3.64) implies that the sequence $\{f_{n,i}\}_{n\geq 1, 1\leq i\leq I_n}$ forms a Riesz basis in X.

Now define the application L in X as

$$Lx = \sum_{n=1}^{\infty} L_n x_n, \qquad x = \sum_{n=1}^{\infty} x_n. \tag{3.67}$$

It is obvious that L is a linear application in X. Moreover, since $\{Y_n\}_{n\geq 1}$ is a Riesz sequence of subspaces in X, we get

$$c'' \sum_{n=1}^{\infty} \|L_n x_n\|^2 \leq \|Lx\|^2 \leq C'' \sum_{n=1}^{\infty} \|L_n x_n\|^2. \tag{3.68}$$

Combining (3.63), (3.66), and (3.68) we get

$$\frac{mc''}{C'} \|x\|^2 \leq \|Lx\|^2 \leq \frac{MC''}{c'} \|x\|^2 \quad \forall x \in X. \tag{3.69}$$

It follows that L is an isomorphism in X. Thus the sequence $\{g_{n,i}\}_{n\geq 1, 1\leq i\leq I_n}$, being the image of the Riesz basis $\{f_{n,i}\}_{n\geq 1, 1\leq i\leq I_n}$ by the isomorphism L, forms a Riesz basis in X. This achieves the proof. □

PROPOSITION 3.3. *Let $\gamma \leq 0$ and a be a real number small enough. Then the system of eigenvectors $\{e^+_{1,n}, e^+_{2,n}, e^-_{2,n}, e^-_{2,n}\}_{n\geq 1}$ of $\mathcal{A}$ forms a Riesz basis in $\mathcal{H}$. In particular, all eigenvectors of $\mathcal{A}$ are of the form* (3.7).

Proof. For all $n \geq 1$ let

$$W_n = \mathrm{Sp}\{E^+_{1,n}, E^+_{2,n}, E^-_{2,n}, E^-_{1,n}\}. \tag{3.70}$$

It is clear that $\{W_n\}_{n\geq 1}$ forms a Hilbert basis of subspaces and $\{V_n\}_{n\geq 1}$ is a Hilbert sequence of subspaces in $\mathcal{H}$. The condition (3.64) is trivial since $E^+_{1,n}, E^+_{2,n}, E^-_{2,n}, E^-_{2,n}$ is a Hilbert basis in the subspace W_n. Let L_n be defined as in (3.46). Following Proposition 3.2, L_n has a constant leading term, which is invertible. This, together with the fact that L_n is invertible for all $n \geq 1$, implies the condition (3.63). Then applying Lemma 3.1, we get that the system of eigenvectors $\{e^+_{1,n}, e^+_{2,n}, e^-_{2,n}, e^-_{2,n}\}_{n\geq 1}$ forms a Riesz basis in $\mathcal{H}$. In particular, all eigenvectors of $\mathcal{A}$ are of the form (3.7). The proof is thus completed. □

Remark 3.2. There are many papers based on the Riesz basis approach in which the essential part is to show that the sequence of the root vectors is quadratically close to a known Riesz basis. This approach requires a very long calculation and is limited to one-dimensional problems (see [13] and the successions).

The idea of Proposition 3.3 lies in constructing the pairwise orthogonal subspaces V_n without the eigenvectors $e^+_{1,n}, e^+_{2,n}, e^-_{1,n}, e^-_{2,n}$ being orthogonal. In this way, we reduced the quadratical convergence to the Parseval equality. In order to get condition (3.64), we need only some asymptotic expansions such as (3.47)–(3.49), which do not give any quadratical convergence.

To fix the idea, we consider the case where $A := -\Delta$ is the Laplacian on a bounded open set $\Omega \subset \mathbb{R}^N$, with the homogeneous Dirichlet boundary condition as in Example 4.1. Following a classical result of Agmon (Theorem 14.6 in [2]), we know that $\mu_n \sim n^{1/N}$. Let $2\gamma + 1 > 0$; then the series of general term

$$\|e^{\pm}_{k,n} - E^{\pm}_{k,n}\| = O\Big(\frac{1}{n^{(2\gamma+1)/N}}\Big) \tag{3.71}$$

converges in l^2 if and only if $N < 2(2\gamma + 1)$. This is never true if $-1/2 < \gamma \le -1/4$, even in one dimension. In the best case $\gamma = 0$, which corresponds to the usual damping y_t, the series of general terms in (3.71) converges in l^2 if and only if $N < 2$. This confirms that the quadratical convergence could be expected only in one dimension.

THEOREM 3.1. *Let $\gamma \le 0$ and a be a real number small enough. Then for all $y_0, z_0 \in D(A)$ and $y_1, u_1 \in D(A^{\frac{1}{2}})$, the energy of system* (3.2)–(3.3) *has the polynomial decay rate*

$$E(t) \le C(\|Ay_0\|^2_H + \|Au_0\|^2_H + \|A^{\frac{1}{2}}y_1\|^2_H + \|A^{\frac{1}{2}}u_1\|^2_H)\frac{1}{t^{\delta(\gamma)}} \qquad \forall t > 0, \tag{3.72}$$

where

$$\delta(\gamma) = \begin{cases} \frac{1}{\gamma+1}, & 2\gamma + 1 > 0, \\ -\frac{1}{\gamma}, & 2\gamma + 1 < 0, \\ 2, & 2\gamma + 1 = 0. \end{cases} \tag{3.73}$$

Proof. Following Proposition 3.1, the eigenvalues of $\mathcal{A}$ satisfy the asymptotic expansions (2.1) with

$$\begin{cases} \alpha_1 = -2\gamma, \quad \alpha_2 = 2(\gamma + 1), & \delta = \frac{1}{\gamma+1}, \quad 2\gamma + 1 > 0, \\ \alpha_1 = \alpha_2 = -2\gamma, & \delta = \frac{1}{-\gamma}, \quad 2\gamma + 1 < 0, \\ \alpha_1 = \alpha_2 = 1, & \delta = 2, \quad 2\gamma + 1 = 0. \end{cases} \tag{3.74}$$

On the other hand, following Proposition 3.3 the system of eigenvectors of $\mathcal{A}$ forms a Riesz basis in $\mathcal{H}$. Then applying Theorem 2.1, we get the polynomial energy decay rate (3.72)–(3.73). The proof is thus achieved. ∎

The following theorem gives the repartition of energy within the two equations of the system (3.2)–(3.3).

THEOREM 3.2. *Assume the same conditions as in Theorem* 3.1. *Let $w_0 \in D(\mathcal{A})$ be initial data and*

$$E_1(t) = \frac{1}{2}(\|A^{\frac{1}{2}}y(t)\|^2_H + \|z(t)\|^2_H), \tag{3.75}$$

$$E_2(t) = \frac{1}{2}(\|A^{\frac{1}{2}}u(t)|^2_H + \|v(t)\|^2_H) \tag{3.76}$$

be the corresponding energy of the first equation (respectively, the second equation). Then the following estimates hold:

$$E_1(t) \le C\|\mathcal{A}w_0\|^2\frac{1}{t^2}, \qquad E_2(t) \le C\|\mathcal{A}w_0\|^2\frac{1}{t^{\frac{1}{\gamma+1}}}, \quad 2\gamma + 1 > 0, \tag{3.77}$$

$$E_1(t) \le C\|\mathcal{A}w_0\|^2\frac{1}{t^{\frac{1}{-\gamma}}}, \quad E_2(t) \le C\|\mathcal{A}w_0\|^2\frac{1}{t^{\frac{1}{-\gamma}}}, \qquad 2\gamma + 1 \le 0. \tag{3.78}$$

Proof. Developing $e^{\pm}_{1,n}, e^{\pm}_{2,n}$ on the basis of $E^{\pm}_{1,n}, E^{\pm}_{2,n}$, we get

$$e^{\pm}_{1,n} = \alpha^{\pm}_{1,n}E^{\pm}_{1,n} + \alpha^{\pm}_{2,n}E^{\pm}_{2,n}, \quad e^{\pm}_{2,n} = \beta^{\pm}_{1,n}E^{\pm}_{1,n} + \beta^{\pm}_{2,n}E^{\pm}_{2,n}. \tag{3.79}$$

Writing

$$w_0 = \sum_{n=1}^{\infty} a_{1,n}^{\pm} e_{1,n}^{\pm} + \sum_{n=1}^{\infty} a_{2,n}^{\pm} e_{2,n}^{\pm}, \tag{3.80}$$

we then have

$$\begin{aligned} w(t) &= \sum_{n=1}^{\infty} a_{1,n}^{\pm} e^{\lambda_{1,n}^{\pm} t} e_{1,n}^{\pm} + \sum_{n=1}^{\infty} a_{2,n}^{\pm} e^{\lambda_{2,n}^{\pm} t} e_{2,n}^{\pm} \\ &= \sum_{n=1}^{\infty} (a_{1,n}^{\pm} \alpha_{1,n}^{\pm} e^{\lambda_{1,n}^{\pm} t} + a_{2,n}^{\pm} \beta_{1,n}^{\pm} e^{\lambda_{2,n}^{\pm} t}) E_{1,n}^{\pm} \\ &\quad + \sum_{n=1}^{\infty} (a_{1,n}^{\pm} \alpha_{2,n}^{\pm} e^{\lambda_{1,n}^{\pm} t} + a_{2,n}^{\pm} \beta_{2,n}^{\pm} e^{\lambda_{2,n}^{\pm} t}) E_{2,n}^{\pm}. \end{aligned} \tag{3.81}$$

The orthogonality of $E_{1,n}^{\pm}, E_{2,n}^{\pm}$ implies

$$E_1(t) = \frac{1}{2} \sum_{n=1}^{\infty} |a_{1,n}^{\pm} \alpha_{1,n}^{\pm} e^{\lambda_{1,n}^{\pm} t} + a_{2,n}^{\pm} \beta_{1,n}^{\pm} e^{\lambda_{2,n}^{\pm} t}|^2, \tag{3.82}$$

$$E_2(t) = \frac{1}{2} \sum_{n=1}^{\infty} |a_{1,n}^{\pm} \alpha_{2,n}^{\pm} e^{\lambda_{1,n}^{\pm} t} + a_{2,n}^{\pm} \beta_{2,n}^{\pm} e^{\lambda_{2,n}^{\pm} t}|^2. \tag{3.83}$$

If $2\gamma + 1 > 0$, then from (3.47) we get

$$\alpha_{1,n}^{\pm} = O(1), \quad \alpha_{2,n}^{\pm} = O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \quad \beta_{1,n}^{\pm} = O\Big(\frac{1}{\mu_n^{2\gamma+1}}\Big), \quad \beta_{2,n}^{\pm} = O(1). \tag{3.84}$$

Inserting (3.84) into (3.82)–(3.83), we get

$$E_1(t) \le C \sum_{n=1}^{\infty} \left(\frac{|\mu_n a_{1,n}^{\pm}|^2}{\mu_n^2 e^{\mu_n^{2\gamma} t}} + \frac{|\mu_n a_{2,n}^{\pm}|^2}{\mu_n^{4\gamma+4} e^{\frac{a^2 t}{\mu_n^{2\gamma+2}}}} \right), \tag{3.85}$$

$$E_2(t) \le C \sum_{n=1}^{\infty} \frac{|\mu_n a_{1,n}^{\pm}|^2}{\mu_n^{4\gamma+4} e^{\mu_n^{2\gamma} t}} + \frac{|\mu_n a_{2,n}^{\pm}|^2}{\mu_n^2 e^{\frac{a^2 t}{\mu_n^{2\gamma+2}}}}. \tag{3.86}$$

Using the estimate

$$\mu_n^{2\theta} e^{\frac{t}{\mu_n^{\alpha}}} \ge C t^{\frac{2\theta}{\alpha}} \qquad \forall\ \theta > 0, \alpha > 0 \tag{3.87}$$

in (3.85)–(3.86), it follows that

$$E_1(t) \le C \|\mathcal{A} w_0\|^2 \left(\frac{1}{t^{\frac{1}{-\gamma}}} + \frac{1}{t^2} \right) \le C \|\mathcal{A} w_0\|^2 \frac{1}{t^2}, \tag{3.88}$$

$$E_2(t) \le C \|\mathcal{A} w_0\|^2 \left(\frac{1}{t^{\frac{2(\gamma+1)}{-\gamma}}} + \frac{1}{t^{\frac{1}{\gamma+1}}} \right) \le C \|\mathcal{A} w_0\|^2 \frac{1}{t^{\frac{1}{\gamma+1}}}. \tag{3.89}$$

If $2\gamma + 1 \le 0$, then from (3.48)–(3.49) we get

$$\alpha_{1,n}^{\pm} = O(1), \quad \alpha_{2,n}^{\pm} = O(1), \quad \beta_{1,n}^{\pm} = O(1), \quad \beta_{2,n}^{\pm} = O(1). \tag{3.90}$$

Inserting (3.90) into (3.82)–(3.83), we get

$$E_{1,2}(t) \le C \sum_{n=1}^{\infty} (|\mu_n a_{1,n}^{\pm}|^2 + |\mu_n a_{2,n}^{\pm}|^2) \frac{1}{\mu_n^2 e^{\mu_n^{2\gamma} t}}. \tag{3.91}$$

Then using (3.87) in (3.91) gives (3.78). The proof is thus complete. □

Comments on Theorems 3.1 and 3.2. From (3.73) we see that the decay rate $\delta(\gamma)$ is an increasing function for $-\infty < \gamma \le -1/2$, decreasing for $-1/2 \le \gamma \le 0$, and achieves the maximum 2 at $\gamma = -1/2$. The best decay rate $\delta = 2$ occurs for $\gamma = -1/2$.

If $2\gamma + 1 > 0$, system (3.2)–(3.3) is *overdamped* in the sense that the eigenvectors $e_{1,n}^{\pm}$ are asymptotically decoupled from the eigenvectors $e_{2,n}^{\pm}$. In that case, the interferences between the two equations are very weak and the wave propagates *almost independently* in the two equations. The energy of the first equation, $E_1(t)$, decays more quickly than $E_2(t)$, the energy of the second equation.

If $2\gamma + 1 \le 0$, system (3.2)–(3.3) is *right-damped* in the sense that the eigenvectors $e_{1,n}^{\pm}$ are involved with the eigenvectors $e_{2,n}^{\pm}$. In that case, the damping applied to the first equation is well transmitted to the second equation and the system is *really coupled*. Consequently, the energies $E_1(t), E_2(t)$ of the two equations decay at the same rate. This balanced repartition of energies within the two equations is due to the compensation of real parts of the two branches of the eigenvalues.

A stronger damping $A^{\gamma} y_t$ does not necessarily give a better total decay rate of energy. A good damping should provide a compensation of the real parts of the two branches of eigenvalues and carry the transmission of the damping from one equation to another. This can be done by means of a suitably weaker damping.

We believe that the spectrum compensation is a natural phenomenon for partially damped distributed systems. It seems interesting to consider coupled systems with different operators A_1, A_2 as in [3]. Indeed the same results could be obtained without essential difficulty in the case $A_2 = A_1^2$ or other similar situations in [1]. But in general we can no long calculate explicitly the eigenvalues as in Proposition 3.1.

4. Examples of application. Let $\Omega \subset \mathbb{R}^N$ be a bounded open set with smooth boundary Γ. We denote by $\|\cdot\|_{0,\Omega},\ \|\cdot\|_{1,\Omega},\ \|\cdot\|_{2,\Omega} \cdots$ the norms of the Sobolev spaces $L^2(\Omega), H^1(\Omega), H^2(\Omega) \cdots$, respectively.

Example 4.1. Let us consider the system of weakly coupled wave equations

$$\begin{cases} y_{tt} - \Delta y + (-\Delta)^{-1/2} y_t + au = 0 & \text{in} \quad \Omega, \\ u_{tt} - \Delta u + ay = 0 & \text{in} \quad \Omega, \\ y = u = 0 & \text{on} \quad \Gamma, \end{cases} \tag{4.1}$$

where $a \in \mathbb{R}$ is small enough, and where $(-\Delta)^{-1/2}$ is a linear continuous operator from $L^2(\Omega)$ onto $H_0^1(\Omega)$.

Defining the operator A in $L^2(\Omega)$ by

$$A = -\Delta \quad \text{with} \quad D(A) = H^2(\Omega) \cap H_0^1(\Omega), \tag{4.2}$$

we easily check that A is a densely defined closed self-adjoint operator with compact resolvent in $L^2(\Omega)$. Then applying Theorem 3.1 we get the optimal polynomial energy decay rate

$$\begin{aligned} &\|y(t)\|_{1,\Omega}^2 + \|y_t(t)\|_{0,\Omega}^2 + \|u(t)\|_{1,\Omega}^2 + \|u_t(t)\|_{0,\Omega}^2 \\ &\quad \le \frac{c}{t^2} \Big(\|y_0\|_{2,\Omega}^2 + \|z_0\|_{1,\Omega}^2 + \|u_0\|_{2,\Omega}^2 + \|v_0\|_{1,\Omega}^2 \Big) \end{aligned} \tag{4.3}$$

for all smooth initial data

$$\begin{cases} y(x,0) = y_0(x) \in H^2(\Omega) \cap H_0^1(\Omega), & y_t(x,0) = z_0(x) \in H_0^1(\Omega), \\ u(x,0) = u_0(x) \in H^2(\Omega) \cap H_0^1(\Omega), & u_t(x,0) = v_0(x) \in H_0^1(\Omega). \end{cases} \tag{4.4}$$

If we had taken a stronger damping y_t as was done in [3], instead of $1/t^2$ we could get only $1/t$ as the energy decay rate for the same initial data.

Example 4.2. Now we consider the system of weakly coupled plate equations

$$\begin{cases} y_{tt} + \Delta^2 y + (\Delta^2)^{-1/2} y_t + au = 0 & \text{in} \quad \Omega, \\ u_{tt} + \Delta^2 u + ay = 0 & \text{in} \quad \Omega, \\ y = \frac{\partial y}{\partial n} = u = \frac{\partial u}{\partial n} = 0 & \text{on} \quad \Gamma, \end{cases} \tag{4.5}$$

where $a \in \mathbb{R}$ is small enough, and where $(\Delta^2)^{-1/2}$ is a linear continuous operator from $L^2(\Omega)$ onto $H_0^2(\Omega)$.

Defining the operator A in $L^2(\Omega)$ by

$$A = \Delta^2 \quad \text{with} \quad D(A) = H^4(\Omega) \cap H_0^2(\Omega), \tag{4.6}$$

we can easily check all the conditions of Theorem 3.1. Then for any smooth initial data

$$\begin{cases} y(x,0) = y_0(x) \in H^4(\Omega) \cap H_0^2(\Omega), & y_t(x,0) = z_0(x) \in H_0^2(\Omega), \\ u(x,0) = u_0(x) \in H^4(\Omega) \cap H_0^2(\Omega), & u_t(x,0) = v_0(x) \in H_0^2(\Omega), \end{cases} \tag{4.7}$$

we have

$$\|y(t)\|_{2,\Omega}^2 + \|y_t(t)\|_{0,\Omega}^2 + \|u(t)\|_{2,\Omega}^2 + \|u_t(t)\|_{0,\Omega}^2 \tag{4.8}$$

$$\le \frac{c}{t^2}\Big(\|y_0\|_{4,\Omega}^2 + \|z_0\|_{2,\Omega}^2 + \|u_0\|_{4,\Omega}^2 + \|v_0\|_{2,\Omega}^2\Big). \tag{4.9}$$

Similarly, if we had taken a stronger feedback control y_t as was done in [3], we could get only $1/t$ as the energy decay rate for the same initial data.

In [3] more general systems with different operators A_1, A_2 were considered under an artificial condition $D(A_2^{j/2}) \subset D(A_1)$, $j \ge 2$, which restricts the applications to the distributed systems of the same kind of boundary conditions (see Examples 6.1–6.5 in [3]). The following example shows that the above-mentioned condition is not necessary, and the polynomial energy decay rate should be true for more general coupled systems. In order to make the computation as clear as possible, we have chosen the special boundary conditions.

Example 4.3. Consider the following system of compactly coupled and partially damped Euler–Bernoulli beam equations:

$$\begin{cases} y_{tt} + y_{xxxx} + by_t + au_x = 0, & 0 < x < \pi, \\ u_{tt} + u_{xxxx} - ay_x = 0, & 0 < x < \pi, \\ y_x(0) = y_{xxx}(0) = y_x(\pi) = y_{xxx}(\pi) = 0, \\ u(0) = u_{xx}(0) = u(\pi) = u_{xx}(\pi) = 0. \end{cases} \tag{4.9}$$

Putting

$$\widetilde{L}^2(0,1) = \left\{ y \in L^2(0,\pi), \quad \int_0^\pi y(x)dx = 0 \right\},$$

$$V = \left\{ y \in H^2(0,\pi), \quad y'(0) = y'(\pi) = 0, \quad \int_0^\pi y(x)dx = 0 \right\},$$

$$W = \{ u \in H^2(0,\pi), \quad u(0) = u(\pi) = 0 \}$$

we define the energy space

$$\mathcal{H} = V \times \widetilde{L}^2(0,\pi) \times W \times L^2(0,\pi).$$

It is easy to prove that the system (4.9) generates a C_0-semigroup of contractions on the space $\mathcal{H}$ for $b > 0$ and a small enough. Moreover, setting the energy

$$E(t) = \frac{1}{2}\int_0^\pi (|y_{xx}|^2 + |y_t|^2 + |u_{xx}|^2 + |u_t|^2 - ay_x\overline{u} - au\overline{y}_x)dx, \tag{4.10}$$

we have

$$\frac{d}{dt}E(t) = -b\int_0^\pi |y_t|^2 dx \le 0. \tag{4.11}$$

Now we consider the associated eigenproblem

$$\begin{cases} \lambda^2 y + y_{xxxx} + b\lambda y + au_x = 0 & 0 < x < \pi, \\ \lambda^2 u + u_{xxxx} - ay_x = 0 & 0 < x < \pi, \\ y_x(0) = y_{xxx}(0) = y_x(\pi) = y_{xxx}(\pi) = 0, \\ u(0) = u_{xx}(0) = u(\pi) = u_{xx}(\pi) = 0. \end{cases} \tag{4.12}$$

Let

$$y = \alpha_n \cos nx, \qquad u = \beta_n \sin nx, \qquad n \ge 1, \tag{4.13}$$

be an eigenfunction. Then λ must satisfy the system

$$\begin{cases} \alpha_n(\lambda^2 + b\lambda + n^4) + \beta_n an = 0, \\ \beta_n(\lambda^2 + n^4) + \alpha_n an = 0, \end{cases} \tag{4.14}$$

which has a nontrivial solution if and only if

$$(\lambda^2 + n^4)(\lambda^2 + b\lambda + n^4) - a^2n^2 = 0. \tag{4.15}$$

Proceeding as in Proposition 3.1, we find easily that

$$\lambda_{1,n}^{\pm} = \pm in^2 - \frac{b}{2} + O\Big(\frac{1}{n^2}\Big), \tag{4.16}$$

$$\lambda_{2,n}^{\pm} = \pm in^2 - \frac{a^2}{2bn^2} + O\Big(\frac{1}{n^6}\Big). \tag{4.17}$$

Then for $\lambda = \lambda_{1,n}^{\pm}$, taking

$$\alpha_{1,n}^{\pm} = \frac{1}{\sqrt{2}}, \quad \beta_{1,n}^{\pm} = -\frac{an}{(\lambda_{1,n}^{\pm})^2 + n^4}\alpha_{1,n}^{\pm} = O\Big(\frac{1}{n}\Big)$$

in (4.13) we get the corresponding eigenvector

$$e_{1,n}^{\pm} = \frac{1}{\sqrt{\pi}}\begin{pmatrix} \frac{\cos nx}{\pm in^2} \\ \cos nx \\ 0 \\ 0 \end{pmatrix} + O\Big(\frac{1}{n}\Big) \quad \forall n \ge 1. \tag{4.18}$$

For $\lambda = \lambda_{2,n}^{\pm}$, taking

$$\beta_{2,n}^{\pm} = \frac{1}{\sqrt{\pi}}, \quad \alpha_{2,n}^{\pm} = -\frac{(\lambda_{2,n}^{\pm})^2 + n^4}{an}\beta_{2,n}^{\pm} = O\Big(\frac{1}{n}\Big)$$

in (4.13), we get the corresponding eigenvector

$$e_{2,n}^{\pm} = \frac{1}{\sqrt{\pi}} \begin{pmatrix} 0 \\ 0 \\ \frac{\sin n\pi x}{\pm i n^2} \\ \sin n\pi x \end{pmatrix} + O\Big(\frac{1}{n}\Big) \quad \forall n \geq 1. \tag{4.19}$$

Using the same procedure as in Proposition 3.3, we can prove that the system $\{e_{1,n}^{\pm}, e_{2,n}^{\pm}\}_{n\geq 1}$ forms a Riesz basis in $\mathcal{H}$. But in the one-dimensional case, the Riesz basis property is evident because the eigenvectors $e_{1,n}^{\pm}, e_{2,n}^{\pm}$ are quadratically close to a Hilbert basis, due to the expansions (4.18)–(4.19), and therefore form a Riesz basis in $\mathcal{H}$ (see Theorem 4.2.3 in [12] and Theorem 4.1 in [28] for the original idea on the eigenvalues of lower frequencies). Moreover, (4.16)–(4.17) show that the eigenvalues of system (4.12) satisfy condition (2.1) with $\mu_{1,n} = \mu_{2,n} = n^2$ and $\alpha_1 = 0, \alpha_2 = 1$. Then applying Theorem 2.1, we get the polynomial energy decay rate

$$E(t) \leq C(\|y_0\|_4^2 + \|z_0\|_2^2 + \|u_0\|_4^2 + \|v_0\|_2^2)\frac{1}{t^2} \qquad \forall t > 0. \tag{4.20}$$

for the solution of system (4.9) with smooth initial data $y_0 \in H^4(0,\pi) \cap V, u_0 \in H^4(0,\pi) \cap W$ and $z_0 \in V, v_0 \in W$.

Remark 4.1. The coupling terms ay_x, au_x do not appear in expression (4.11). Nevertheless, the energy decays at the rate of $1/t^2$ instead of $1/t$ for the system coupled through the displacements ay, au.

Acknowledgments. The authors would like to thank the referees for very valuable comments and suggestions. B. Rao thanks E. Zuazua for a fruitful discussion on the topic of this work.

REFERENCES

[1] M. Afilal and F. Ammar Khodja, *Stability of coupled second order equations*, Comput. Appl. Math., 19 (2000), pp. 91–107.

[2] S. Agmon, *Lectures on Elliptic Boundary Values Problems*, Van Nostrand, Princeton, NJ, 1965.

[3] F. Alabau, P. Cannarsa, and V. Komornik, *Indirect internal stabilization of weakly coupled evolution equations*, J. Evol. Equ., 2 (2002), pp. 127–150.

[4] K. Ammari and M. Tucsnak, *Stabilization of second order evolution equations by a class of unbounded feedbacks*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 361–386.

[5] A. Bátkai, K.-J. Engel, J. Prüss, and R. Schnaubelt, *Polynomial stability of operator semigroups*, Math. Nachr., to appear.

[6] N. Burq and G. Lebeau, *Mesures de défaut de compacité, application au système de Lamé*, Ann. Sci. École Norm. Sup. (4), 34 (2001), pp. 817–870.

[7] G. Chen and D. L. Russell, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 39 (1982), pp. 433–454.

[8] S. Chen and R. Triggiani, *Proof of extensions of two conjectures on structural damping for elastic systems*, Pacific J. Math., 136 (1989), pp. 15–55.

[9] S. Chen and R. Triggiani, *Gevrey class semigroups arising from elastic systems with gentle dissipation: The case* $0 < \alpha < 1/2$. Proc. Amer. Math. Soc., 110 (1990), pp. 401–415.

[10] R. Dáger and E. Zuazua, *Wave Propagation, Observation and Control in* $1-d$ *Flexible Multi-Structures*, Math. Appl. (Berlin) 50, Springer-Verlag, Berlin, 2006.

[11] Th. Duyckaerts, *Stabilization of the linear system of magnetoelasticity*, preprint.
[12] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, AMS, Providence, RI, 1969.
[13] B.-Z. Guo, *Riesz basis approach to the stabilization of a flexible beam with a tip mass*, SIAM J. Control Optim., 39 (2001), pp. 1736–1747.
[14] S. W. Hansen, *Optimal regularity results in boundary control of elastic systems with fractional order damping*, ESAIM Proc., 8 (2000), pp. 53–64.
[15] F. Huang, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.
[16] S. Jaffard, M. Tucsnak, and E. Zuazua, *Singular inter stabilization of the wave equation*, J. Differential Equations, 45 (1998), pp. 184–215.
[17] G. Lebeau and E. Zuazua, *Decay rates for the three-dimensional linear system of thermoelasticity*, Arch. Ration. Mech. Anal., 148 (1999), pp. 179–231.
[18] T.-T. Li and B. Rao, *Exact boundary controllability of unsteady flows in a tree-like network of open canals*, Methods Appl. Anal., 11 (2004), pp. 353–366.
[19] W. Littman and B. Liu, *On the spectral properties and stabilization of acoustic flow*, SIAM J. Appl. Math., 59 (1998), pp. 17–34.
[20] W. Littman, *Recent work on the SCOLE model, stabilization of flexible structure*, Lecture Notes in Control and Inform. Sci. 147, Springer, Berlin, 1990.
[21] J.-L. Lions, *Contrôlabilité Exacte et Stabilisation de Systèmes Distribués*, Masson, Paris, 1984.
[22] K. Liu, Z. Liu, and B. Rao, *Exponential stability of an abstract nondissipative linear system*, J. SIAM Control Optim., 40 (2001), pp. 149–165.
[23] Z. Liu and B. Rao, *Characterization of polynomial decay rate for the solution of linear evolution equation*, Z. Angew. Math. Phys., 56 (2005), pp. 630–644.
[24] Z. Liu and S. Zheng, *Semigroup associated with dissipative systems*, Notes in Math. 394, Chapman & Hall/CRC, Boca Raton, FL, 1999.
[25] P. Loreti and B. Rao, *Compensation spectrale et taux de décroissance optimal de l'energie de systèmes partiellement amortis*, C. R. Math. Acad. Sci. Paris, 337, (2003), pp. 531–536.
[26] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1984.
[27] J. Prüss, *On the spectrum of C_0-semigroups*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.
[28] B. Rao, *Optimal energy decay rate in the Rayleigh beam equation*, Contemp. Math., 209 (1997) pp. 211–229.
[29] B. Rao, *Stabilization of elastic plates with dynamical boundary control*, J. SIAM Control Optim., 36 (1998), pp. 148–163.
[30] B. Rao and A. Wehbe, *Polynomial energy decay rate and strong stability of Kirchhoff plates with non-compact resolvent*, J. Evol. Equ., 5 (2005), pp. 137–152.
[31] D. L. Russell, *Decay rates for weakly damped systems in Hilbert space obtained with control-theoretic methods*, J. Differential Equations, 19 (1975), pp. 344–370.
[32] D. L. Russell, *A general framework for the study of indirect damping mechanisms in elastic systems*, J. Math. Anal. Appl., 173 (1993), pp. 339–354.
[33] A. Wehbe, *Boundary stabilization of wave equations with dynamical control*, Appl. Math. Lett., 16 (2003), pp. 357–363.
[34] X. Zhang and E. Zuazua, *Polynomial decay and control of a $1-d$ hyperbolic-parabolic coupled system*, J. Differential Equations, 204 (2004), pp. 380–438.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1633–1656

SIMULATION-BASED UNIFORM VALUE FUNCTION ESTIMATES OF MARKOV DECISION PROCESSES*

RAHUL JAIN† AND PRAVIN P. VARAIYA†

Abstract. The value function of a Markov decision process (MDP) assigns to each policy its expected discounted reward. This expected reward can be estimated as the empirical average of the reward over many independent simulation runs. We derive bounds on the number of runs needed for the uniform convergence of the empirical average to the expected reward for a class of policies, in terms of the Vapnik-Chervonenkis or P-dimension of the policy class. Further, we show through a counterexample that whether we get uniform convergence or not for an MDP depends on the simulation method used. Uniform convergence results are also obtained for the average-reward case, for partially observed Markov decision processes, and can be easily extended to Markov games. The results can be viewed as a contribution to empirical process theory and as an extension of the probably approximately correct (PAC) learning theory for partially observable MDPs and Markov games.

Key words. Markov decision processes, Markov games, empirical process theory, PAC learning, value function estimation, uniform rate of convergence

AMS subject classifications. 90C40, 62C12, 62M05

DOI. 10.1137/040619508

1. Introduction.

We address the following question:

Given a Markov decision process (MDP) with an unknown policy from a given set of policies, how can we estimate its value function from computer simulations?

The question is motivated by the system identification problem in stochastic dynamical systems, such as pursuit-evasion games [38], and the estimation problem in econometric analysis [32].

The question is intimately related to empirical process theory (EPT) [35, 44], which studies the uniform behavior of a class $\mathcal{G}$ of measurable functions in the law of large numbers [34] (as well as the central limit theorem [15]) regime. In particular, EPT studies the conditions for which

$$\Pr\left\{\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i)-\mathbb{E}_P[g(X)]\right|>\epsilon\right\}\to 0 \tag{1.1}$$

and the rate of convergence. Convergence results in EPT typically use concentration of measure inequalities such as those of Chernoff [11] and Hoeffding [19]. The rate of convergence of an empirical average to the expected value depends on the exponent in the upper bound of such inequalities. Thus, there has been an effort to improve the exponent [23]. Talagrand [40] introduced new concentration of measure inequalities for product probability spaces that are significantly tighter than the Hoeffding–Chernoff type inequalities. The setting of general product probability spaces [27, 41] instead of just independent, identically distributed (i.i.d.) product probability spaces [11,

*Received by the editors November 23, 2004; accepted for publication (in revised form) May 24, 2006; published electronically November 14, 2006. This research supported by National Science Foundation Award ECS-0424445.
http://www.siam.org/journals/sicon/45-5/61950.html
†EECS Department, University of California, Berkeley, CA 94720 (rjain@eecs.berkeley.edu, varaiya@eecs.berkeley.edu).

19] has greatly expanded the applications [28, 39]. Moreover, many applications involve dependent processes. The Hoeffding inequality has been extended to various dependent cases [8, 26, 43], while an extension of Talagrand's inequality for Markov chains and certain mixing processes was provided by [37].

The goal of this paper is to extend the reach of this rich and rapidly developing theory in a new direction. We provide the beginnings of an empirical process theory for MDPs. This essentially involves considering empirical averages of iterates of functions; i.e., if f is a map from $\mathbb{R}$ to itself, then we consider $g = f^t$ for some fixed integer t, where f^t denotes $f \circ \cdots \circ f$, with iteration being done t times. This case is not subsumed in the existing results in the empirical process theory discussed above (see also [13]). Interestingly, we discover that the method used to obtain the sample trajectories of the MDPs from computer simulation affects the rate of convergence. Thus, such a theory fills an important void in the EPT [35, 44] and the stochastic control literatures [36]. It also underlines the importance of choosing a suitable computer simulation.

We now make the question above more precise and explain the contribution of this paper.

Consider an MDP with a set of policies Π. The value function assigns to each $\pi \in \Pi$ its expected discounted reward $V(\pi)$. We estimate V from independent samples of the discounted reward by the empirical mean, $\hat{V}(\pi)$. We obtain the number of samples $n(\epsilon, \delta)$ (or sample complexity) needed so that the probability

$$\Pr\left\{\sup_{\pi\in\Pi} |\hat{V}(\pi) - V(\pi)| > \epsilon\right\} < \delta. \tag{1.2}$$

Our approach is broadly inspired by [17, 45, 46] and influenced by [21]. Thus, we would like to reduce the problem in (1.2) to understanding the geometry of Π in terms of its covering number. (If the covering number is finite, it is the minimal number of elements of a set needed to approximate any element in the set Π with a given accuracy.) We first relate the covering numbers of the space of stationary stochastic policies and the space of Markov chains that they induce. We relate these to the space of simulation functions that simulate the Markov chains when the set of transition probabilities of the latter is convex. These results together yield the rate of convergence of the empirical estimate to the expected value for the discounted-reward MDPs. What makes the problem nontrivial is that obtaining an empirical discounted reward from simulation involves an iteration of simulation functions. The geometry of the space of iterated simulation functions is much more complex than that of the original space.

One of the key contributions of this paper is the observation that how we simulate an MDP matters for obtaining uniform estimates. We show through an example (see Example 5.4) that uniform estimates of an MDP may converge under one simulation model but fail to do so under another. This is a new (and surprising) observation in the Markov decision theory as well as in the EPT literature.

We then consider the average-reward case. This appears to be the first attempt at nonparametric uniform value estimation for the average-reward case when simulation is done with just one sample path. Ergodicity and weak mixing are exploited to obtain uniform rates of convergence of estimates to expected values. We extend the results to dynamic Markov games and to the case when the Markov decision process is partially observable and policies are nonstationary and have memory.

The problem of uniform convergence of the empirical average to the value function for discounted MDPs was studied in [20, 30] in a machine learning context. While [20] considered only finite state and action spaces, [30] obtains the conditions for uniform

convergence in terms of the simulation model rather than the geometric characteristics (such as covering numbers or the P-dimension) of the simulation function space as opposed to that of the more natural policy space. Large-deviations results for finite state and action spaces for the empirical state-action frequencies and general reward functions were obtained in [4, 24]. A different approach more akin to importance sampling is explored in [33].

While the problem of uniform estimation of the value function for discounted and average-reward partially observed MDPs is of interest in itself, it is also connected with the system identification problem [10, 48]. Also interesting and important for many applications are computationally tractable methods (such as through simulation) for approximating the optimal policy [31]. The simulation-based estimates, such as those proposed in this paper, have been used in a gradient-based method for finding Nash equilibrium policies in a pursuit-evasion game problem [38], though the theoretical understanding is far from complete. Other simulation-based methods for finding approximations to the optimal policy include [25], a likelihood-ratio type gradient estimation method for finite state spaces, and [6], which imposes certain differentiability and regularity assumptions on the derivatives of the policies with respect to the parameters.

The rest of the paper is organized as follows. Section 2 relates the work to the probably approximately correct (PAC) learning model and the system identification problem. Section 3 is a warm-up section presenting preliminaries and relating the covering numbers of the space of transition probabilities of the induced Markov chains and the policy space. Section 4 presents the estimation methodology using the "simple" simulation model and discusses its combinatorial complexity. Section 5 obtains uniform sample complexity results for estimation of values of discounted-reward MDPs. Section 6 considers average-reward MDPs. Section 7 provides the extension to partially observed MDPs with general policies. Some proofs are relegated to the appendices to maintain a smooth exposition.

2. Relation to PAC learning and system identification. System identification is studied in a general function learning setting in the PAC learning model [17, 42]. A fundamental relationship between system identification and empirical process theory was established by Vapnik and Chervonenkis in [47].

Consider a bounded real-valued measurable function $f \in \mathcal{F}$ over a set $\mathbb{X}$ with a probability measure P. We equip $\mathcal{F}$ with a pseudometric such as the one defined in (2.1) below. Unless necessary, we ignore all measurability issues throughout the paper. These have been discussed at length in [15, 34].

The goal is to estimate or "learn" f from independent samples $S = \{(x_1, f(x_1)), \ldots, (x_n, f(x_n))\}$. Say that $\mathcal{F}$ is *PAC-learnable* if there is an algorithm that maps S to $h_{n,f} \in \mathcal{F}$ such that for any $\epsilon > 0$, the probability that the empirical error

$$\operatorname{err}(f, h_{n,f}) := \int |f(x) - h_{n,f}(x)| P(dx) \tag{2.1}$$

is greater than ϵ goes to zero as $n \to \infty$. (Note that $h_{n,f}$ is a function of S.) In other words, for n large enough the probability that the error is larger than ϵ is smaller than some given $\delta > 0$.

The class of functions $\mathcal{F}$ has the *uniform convergence of empirical means (UCEM) property* if

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_P[f(X)] \right| \to 0$$

in probability. It is known that a class of bounded real-valued functions with the UCEM property is not only PAC-learnable but *PUAC (probably uniformly approximately correct)-learnable* [47], i.e.,

$$\lim_{n\to\infty} \mathrm{P}^n \left\{ \sup_{f\in\mathcal{F}} \mathrm{err}(f, h_{n,f}) > \epsilon \right\} = 0.$$

Thus if the mean value of each function in a family can be determined with small error and high probability, the function itself can be "identified" with small error and high probability [45, 46, 47, 49]. One such (minimum empirical risk) algorithm was discovered in [7] when the function class $\mathcal{F}$ satisfies a certain (finite covering number) condition.

The PAC learning model has been generalized to the case when the inputs are Markovian [1, 16], but it has not been extended to MDPs and games. We provide that extension in this paper.

3. Preliminaries. Consider an *MDP* M with countable state space $\mathbb{X}$ and action space $\mathbb{A}$, transition probability function $P_a(x, x')$, initial state distribution λ, and a measurable reward function $r(x)$ with values in $[0, R]$. The value function for a policy π is the expected discounted reward

$$V(\pi) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(x_t)\right],$$

where $0 < \gamma < 1$ is a discount factor and x_t is the state at time t under policy π. For the average-reward case the value function is

$$V(\pi) = \mathbb{E}\left[\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(x_t)\right].$$

Let Π_0 denote the space of all *stationary stochastic policies* $\{\pi(x, a) : a \in \mathbb{A}, x \in \mathbb{X}, \sum_a \pi(x, a) = 1\}$ and let $\Pi \subseteq \Pi_0$ be the subset of policies of interest. The MDP M under a fixed stationary policy π induces a Markov chain with transition probability function $P_\pi(x, x') = \sum_a P_a(x, x')\pi(x, a)$. The initial distribution on the Markov chains is λ, and we identify P_π with the Markov chain. Denote $\mathcal{P} := \{P_\pi : \pi \in \Pi\}$.

We seek conditions on the policy space Π such that a simulation-based estimate $\hat{V}(\pi)$ converges to the value function $V(\pi)$ in probability uniformly over all policies in Π. For this, as we will see in section 5, it is essential to understand the geometry of the space $\mathcal{P}$, and hence of Π. We do this by relating the *covering numbers* of Π with that of $\mathcal{P}$, which are then related to a space of (simulation) functions $\mathcal{F}$ that we define in section 4.

Let $\mathbb{X}$ be an arbitrary set and let λ be a probability measure on $\mathbb{X}$. Given a set $\mathcal{F}$ of real-valued functions on $\mathbb{X}$, ρ a metric on $\mathbb{R}$, let $d_{\rho(\lambda)}$ be the pseudometric on $\mathcal{F}$ with respect to measure λ,

$$d_{\rho(\lambda)}(f, g) = \int \rho(f(x), g(x))\lambda(dx).$$

A subset $\mathcal{G} \subseteq \mathcal{F}$ is an ϵ-*net* for $\mathcal{F}$ if $\forall f \in \mathcal{F}$, $\exists g \in \mathcal{G}$ with $d_{\rho(\lambda)}(f, g) < \epsilon$. The size of the minimal ϵ-net is the ϵ-*covering number*, denoted $N(\epsilon, \mathcal{F}, d_{\rho(\lambda)})$. The ϵ-*capacity* of $\mathcal{F}$ under the ρ metric is $C(\epsilon, \mathcal{F}, \rho) = \sup_\lambda N(\epsilon, \mathcal{F}, d_{\rho(\lambda)})$. Essentially, the ϵ-net can be

seen as a subset of functions that can ϵ-approximate any function in $\mathcal{F}$. The covering number is a measure of the richness of the function class. The richer it is, the more approximating functions we will need for a given measure of approximation ϵ. The capacity makes it independent of the underlying measure λ on $\mathbb{X}$. (See [21] for an elegant treatment of covering numbers.)

Let σ be a probability measure on $\mathbb{A}$. We now define the following L_1-pseudometric on Π:

$$d_{L_1(\sigma\times\lambda)}(\pi,\pi') := \sum_{a\in\mathbb{A}} \sigma(a) \sum_{x\in\mathbb{X}} \lambda(x)|\pi(x,a) - \pi'(x,a)|,$$

and the total variation pseudometric on $\mathcal{P}$,

$$d_{TV(\lambda)}(P,P') := \sum_{y\in\mathbb{X}} \left| \sum_{x\in\mathbb{X}} \lambda(x)(P(x,y) - P'(x,y)) \right|.$$

Note that covering numbers of function spaces can be defined for pseudometrics, and a metric structure is not necessary (see [17, 21, 49]).

Bounds on covering numbers are obtained in terms of various combinatorial dimensions.[1] Thus, we first relate the covering number of $\mathcal{P}$ with a combinatorial dimension of Π. Recall some measures of combinatorial dimension.

Let $\mathcal{F}$ be a set of binary-valued functions from $\mathbb{X}$ to $\{0,1\}$. Say that $\mathcal{F}$ *shatters* $\{x_1,\ldots,x_n\}$ if the set $\{(f(x_1),\ldots,f(x_n)), f\in\mathcal{F}\}$ has cardinality 2^n. The largest such n is the VC-dim($\mathcal{F}$). Intuitively, this means that the function class $\mathcal{F}$ can distinguish between a set of n points from the set $\mathbb{X}$.

Let $\mathcal{F}$ be a set of real-valued functions from $\mathbb{X}$ to $[0,1]$. Say that $\mathcal{F}$ *P-shatters* $\{x_1,\ldots,x_n\}$ if there exists a witness vector $c=(c_1,\ldots,c_n)$ such that the set $\{(\eta(f(x_1)-c_1),\ldots,\eta(f(x_n)-c_n)), f\in\mathcal{F}\}$ has cardinality 2^n; $\eta(\cdot)$ is the sign function. The largest such n is the P-dim($\mathcal{F}$). This is a generalization of VC-dim, and for $\{0,1\}$-valued functions, the two definitions are equivalent.

Other combinatorial dimensions such as the fat-shattering dimension introduced in [2] yield both an upper and lower bound on the covering numbers but in this paper we will use the P-dim. Results using fat-shattering dimension can be established similarly.

Given a policy space Π, let $\mathcal{P}$ denote the set of transition probabilities of the Markov chains it induces. We relate the covering numbers of the two spaces under the pseudometrics defined above.

LEMMA 3.1. *Suppose* P-dim(Π) $= d$. *Suppose there is a probability measure* λ *on* $\mathbb{X}$ *and a probability measure* σ *on* $\mathbb{A}$ *such that* $\pi(x,a)/\sigma(a) \le K\ \forall x\in\mathbb{X}, a\in\mathbb{A}, \pi\in\Pi$. *Then, for* $0<\epsilon<e/4\log_2 e$,

$$N(\epsilon,\mathcal{P},d_{TV(\lambda)}) \le N\left(\epsilon,\frac{\Pi}{\sigma},d_{L_1(\sigma\times\lambda)}\right) \le 2\left(\Phi\left(\frac{2eK}{\epsilon}\right)\right)^d,$$

where $\Phi(x) = x\log x$.

The proof is in the appendix.

We now give an example illustrating the intuition behind the concept of P-dim. It is similar in spirit to well-known results about finite-dimensional linear spaces. It

[1] These are different from the algebraic dimension. Examples of combinatorial dimensions are VC-dimension and P-dimension. A class of real-valued functions may have infinite algebraic dimension but finite P-dimension. See [3, 49] for more details.

shows that for an MDP with finite state and action spaces, the set of all stationary stochastic policies has a finite P-dimension equal to the number of free parameters needed to represent the set of policies being considered.

Example 3.2. Let $\mathbb{X} = \{1, \ldots, N\}$ and $\mathbb{A} = \{1, \ldots, M\}$. Then, P-dim $(\Pi_0) = N(M-1)$ for $M > 1$.

Proof. Consider the set of all stochastic policies:

$$\Pi_0 = \left\{ \pi : \mathbb{X} \times \mathbb{A} \to [0,1] \mid \sum_{a \in \mathbb{A}} \pi(x,a) = 1 \ \forall x \in \mathbb{X} \right\}.$$

Let $S = \{(x_1, a_1), \ldots, (x_N, a_1), \ldots, (x_1, a_{M-1}), \ldots, (x_N, a_{M-1})\}$. We can find $c_{11}, \ldots, c_{NM-1}$ such that the $N(M-1)$-dimensional vector

$$\begin{gathered}(\ \eta(\pi(x_1,a_1) - c_{11}), \ldots, \eta(\pi(x_1,a_{M-1}) - c_{1M-1}), \ldots \\ \vdots \quad \vdots \\ \ldots, \eta(\pi(x_N,a_1) - c_{N1}), \ldots, \eta(\pi(x_N,a_{M-1}) - c_{NM-1}) \ )\end{gathered}$$

yields all possible binary vectors as π runs over Π; $\eta(\cdot)$ is the sign function. Consider the first row. Note that the probabilities there together with $\pi(x_1, a_M)$ sum to 1. Choose all c_{1j} to be $1/M$. Then, we can get all possible binary vectors in the first row. Since the subsequent rows are independent, we can do the same for all of them. Thus, we can get all possible binary vectors of length $N(M-1)$. So Π_0 shatters S.

However, if we add another point, say (x_1, a_M), to S, the first row will sum to 1. In this case we cannot get all the 2^M possible binary vectors. Thus, the P-dimension of Π_0 is $N(M-1)$. □

4. The simulation model. We estimate the value $V(\pi)$ of policy $\pi \in \Pi$ from independent samples of the discounted rewards. The samples are generated by a simulation "engine" h. This is a deterministic function to which we feed a "noise" sequence $\omega = (\omega_1, \omega_2, \ldots)$ (with ω_i being i.i.d. from uniform distribution over $\Omega = [0,1]$) and an initial state x_0 (drawn from distribution λ). The engine h generates a sample trajectory with the same distribution as the Markov chain corresponding to π. The function $h : \mathbb{X} \times \mathbb{A} \times \Omega \to \mathbb{X}$ gives the next state x' given the current state x, action taken a, and noise ω_i. Several such sample trajectories are generated using i.i.d. noise sequences and initial states. Each sample trajectory yields an empirical total discounted reward. The estimate of $V(\pi)$, $\hat{V}(\pi)$ is the average of the empirical total discounted reward for the various sample trajectories.

Because simulation cannot be performed indefinitely, we stop the simulation at some time T, after which the contribution to the total discounted reward falls below $\epsilon/2$ for required estimation error bound ϵ. T is the $\epsilon/2$-*horizon time*.

Many simulation functions are possible. We will work with the following simple simulation model. For the rest of this paper, we consider the state space to be $\mathbb{X} = \mathbb{N}$.

DEFINITION 4.1 (simple simulation model). *The* simple *simulation model h for a given MDP is given by*

$$h(x, a, \omega) = \inf\{y \in \mathbb{X} : \omega \in [F_{a,x}(y-1), F_{a,x}(y))\},$$

in which $F_{a,x}(y) := \sum_{y' \le y} P_a(x, y')$ is the cummulative distribution function (c.d.f.) corresponding to the transition probability function $P_a(x, y)$.

Similarly, with a slight abuse of notation, we define the simple *simulation model h for the Markov chain P as*

$$h(x, P, \omega) = \inf\{y \in \mathbb{X} : \omega \in [F_{P,x}(y-1), F_{P,x}(y))\},$$

where $F_{P,x}(y) := \sum_{y' \le y} P(x, y')$ *is the c.d.f. corresponding to the transition probability function* $P(x, y)$.

This is the simplest method of simulation. For example, to simulate a probability distribution on a discrete state space, we partition the unit interval so that the first subinterval has length equal to the mass on the first state, the second subinterval has length equal to the mass on the second state, and so on. Perhaps surprisingly, there are other simulation functions h' that generate the same Markov chain, but which have a much larger complexity than h.

The sample trajectory $\{x_t\}$ for policy π is obtained by

$$x_{t+1} = f_{P_\pi}(x_t, \omega_{t+1}) = h(x_t, P_\pi, \omega_{t+1}),$$

in which P_π is the transition probability function of the Markov chain induced by π and $\omega_{t+1} \in \Omega$ is noise. The initial state x_0 is drawn according to the given initial state distribution λ. The function $f_{P_\pi} : \mathbb{X} \times \Omega \to \mathbb{X}$ is called the *simulation function* for the Markov chain transition probability function P_π. (The reader may note that the above definition is to ease understanding in this section. In the next section, we will redefine the domain and range of the simulation functions.) As before, $\mathcal{P} = \{P_\pi : \pi \in \Pi\}$. We denote by $\mathcal{F} = \{f_P : P \in \mathcal{P}\}$ the set of all simulation functions induced by $\mathcal{P}$.

To every $P \in \mathcal{P}$, there corresponds a function $f \in \mathcal{F}$. Observe that $f \in \mathcal{F}$ simulates $P \in \mathcal{P}$ given by

$$P(x, y) = \mu_0\{\omega : f(x, \omega) = y\},$$

where μ_0 is the Lebesgue measure on Ω. Unless specified otherwise, $\mathcal{F}$ will denote the set of simulation functions for the class $\mathcal{P}$ under the simple simulation model.

In the previous section, we related the covering numbers of policy space Π and $\mathcal{P}$. However, as we shall see in the next section, the convergence properties of our estimate of the value function really depend on the covering number of $\mathcal{F}$. Thus, we now show that the complexity of the space $\mathcal{F}$ is the same as that of $\mathcal{P}$ if $\mathcal{P}$ is convex.

The result is in the same spirit as Theorem 13.9 in [12] for finite-dimensional linear vector spaces. However, the setting here is different. We provide an independent proof.

LEMMA 4.2. *Suppose* $\mathcal{P}$ *is convex (being generated by a convex space of policies) with* P*-dimension* d. *Let* $\mathcal{F}$ *be the corresponding space of simple simulation functions induced by* $\mathcal{P}$. *Then,* P-dim $(\mathcal{F}) = d$. *Moreover, the algebraic dimension of* $\mathcal{P}$ *is also* d.

Proof. There is a one-to-one map between the space of simple simulation functions $\mathcal{F}$ and the space of c.d.f.'s $\tilde{\mathcal{F}}$ corresponding to $\mathcal{P}$. ($\tilde{\mathcal{F}} = \{\tilde{F} : \tilde{F}(x, y) = \sum_{y' \le y} P(x, y'), P \in \mathcal{P}\}$.) $\mathcal{F}$ and $\tilde{\mathcal{F}}$ have the same P-dimension because, for any $F \in \mathcal{F}$, $F(x, \omega) > y$ if and only if for the corresponding $\tilde{F} \in \tilde{\mathcal{F}}$, $\tilde{F}(x, y) < \omega$. Thus, $\tilde{\mathcal{F}}$ shatters $\{(x_1, y_1), \ldots, (x_d, y_d)\}$ with witness vector $(\omega_1, \ldots, \omega_d)$ if and only if $\mathcal{F}$ shatters $\{(x_1, \omega_1), \ldots, (x_d, \omega_d)\}$ with witness vector $(y_1, \ldots, y_d)$. So in the following discussion we treat them as the same space $\mathcal{F}$.

Because $\mathcal{P}$ has P-dimension d, there exists $S = \{(x_1, y_1), \ldots, (x_d, y_d)\}$ that is shattered by $\mathcal{P}$ with some witness vector $c = (c_1, \ldots, c_d)$. Consider the projection of the set $\mathcal{P}$ on the S coordinates: $\mathcal{P}|_S = \{(P(x_1, y_1), \ldots, P(x_d, y_d)) : P \in \mathcal{P}\}$. The definition of shattering implies that there is a d-dimensional hypercube contained in $\mathcal{P}|_S$ with center c. Also note that $\mathcal{P}|_S$ is convex and its algebraic dimension is d. To argue that the algebraic dimension of $\mathcal{P}$ cannot be $d + 1$, suppose that it is. Then it would contain $d+1$ coordinates such that the projection of $\mathcal{P}$ along those coordinates contains a hypercube of dimension $d+1$. Thus, $\mathcal{P}$ would shatter $d+1$ points with the

center of the hypercube being a witness vector. But that contradicts the assumption that the P-dimension of $\mathcal{P}$ is d.

Thus for convex spaces, the algebraic dimension and P-dimension are equal.

Next, $\mathcal{F}$ is obtained from $\mathcal{P}$ by an invertible linear transformation; hence its algebraic dimension is also d. Thus, it has d coordinates S such that the projected space $\mathcal{F}|_S$ has algebraic dimension d. Moreover, it contains a hypercube of dimension d. Hence, its P-dimension is at least d. Since the argument is reversible starting from space $\mathcal{F}$ to space $\mathcal{P}$, it implies P-dim $(\mathcal{P})$ = P-dim $(\mathcal{F})$. □

It may be noted that convexity is essential to the above argument. From several examples it appears that the result is not true without convexity. But we are unable to offer a concrete counterexample.

5. Discounted-reward MDPs. We now consider uniform value function estimation from simulation for discounted-reward MDPs.

For the rest of the paper, we redefine $\mathcal{F}$ to be a set of measurable functions from $Y := \mathbb{X} \times \Omega^\infty$ onto itself which simulate $\mathcal{P}$, the transition probabilities induced by Π under the simple simulation model. However, each function depends only on the first component of the sequence $\omega = (\omega_1, \omega_2, \ldots)$. Thus the results and discussion of the previous section hold. Let θ be the left-shift operator on Ω^∞, $\theta(\omega_1, \omega_2, \ldots) = (\omega_2, \omega_3, \ldots)$. For a policy π, our simulation system is

$$(x_{t+1}, \theta\omega) = f_{P_\pi}(x_t, \omega),$$

in which x_{t+1} is the next state starting from x_t and the simulator also outputs the shifted noise sequence $\theta\omega$. This definition of the simulation function is introduced to facilitate the iteration of simulation functions. Denote $\mathcal{F} := \{f_P : \mathbb{X} \times \Omega^\infty \to \mathbb{X} \times \Omega^\infty, P \in \mathcal{P}\}$ and $\mathcal{F}^2 := \{f \circ f : Y \to Y, f \in \mathcal{F}\}$ and $\mathcal{F}^t$ its generalization to t iterations. Similarly, we redefine the reward function as $r : \mathbb{X} \times \Omega^\infty \to [0, R]$.

The estimation procedure is the following. Obtain n initial states $x_0^1, \ldots, x_0^n$ drawn as i.i.d. according to λ, and n noise sequences $\omega^1, \ldots, \omega^n \in \Omega^\infty$ drawn according to μ. Denote the samples by $S = \{(x_0^1, \omega^1), \ldots, (x_0^n, \omega^n)\}$. Under the simple simulation model, $f_P(x, \omega) := (h(x, P, \omega_1), \theta\omega)$ and as before, $\mathcal{F} := \{f_P : P \in \mathcal{P}\}$. For a given initial state and noise sequence, the simulation function yields a reward sequence, the reward at time t given by $R_t(x_0, \omega) := r \circ f_P \circ \cdots \circ f_P(x_0, \omega)$, with f_P composed t times. The empirical total discounted reward for a given state sequence then is $\sum_{t=0}^\infty \gamma^t R_t(x_0, \omega)$. Our estimate of $V(\pi)$ from n simulations, each conducted for $\epsilon/2$-horizon time T, is $\hat{V}_n(\pi) := \frac{1}{n} \sum_{i=1}^n [\sum_{t=0}^T \gamma^t R_t^\pi(x_0^i, \omega^i)]$.

We first present a key technical result which relates the covering number of the iterated functions $\mathcal{F}^t$ under the ρ pseudometric with the covering number for $\mathcal{F}$ under the L_1 pseudometric, for which bounds are known in terms of the P-dim of $\mathcal{F}$.

Let μ be any probability measure on Ω^∞ and λ the initial distribution on $\mathbb{X}$. Denote the product measure on Y by $\mathbb{P} = \lambda \times \mu$, and on Y^n by $\mathbb{P}^n$. Define two pseudometrics on $\mathcal{F}$,

$$\rho_{\mathbb{P}}(f, g) = \sum_x \lambda(x) \mu\{\omega : f(x, \omega) \neq g(x, \omega)\}$$

and

$$d_{L_1(\mathbb{P})}(f, g) := \sum_x \lambda(x) \int |f(x, \omega) - g(x, \omega)| d\mu(\omega).$$

Here, we take $|f(x,\omega)-g(x,\omega)|$ to denote $|x'-x''|+||\theta\omega-\theta\omega||$, where $f(x,\omega)=(x',\theta\omega)$ and $g(x,\omega)=(x'',\theta\omega)$, and $||\cdot||$ is the l_1-norm on Ω^∞. Recall that $x',x''\in\mathbb{N}$. Then, $|f(x,\omega)-g(x,\omega)|=|x'-x''|$, the usual l_1 distance.

LEMMA 5.1. *Let λ be the initial distribution on $\mathbb{X}$ and let λ_f be the (one-step) distribution given by $\lambda_f(y)=\sum_x \lambda(x)\mu\{\omega : f(x,\omega)=(y,\theta\omega)\}$ for $f\in\mathcal{F}$. Suppose that*

$$K := \max\left\{\sup_{f\in\mathcal{F},y\in\mathbb{X}} \frac{\lambda_f(y)}{\lambda(y)}, 1\right\} < \infty. \tag{5.1}$$

Then,

$$\rho_{\mathbb{P}}(f^t,g^t) \le K^t d_{L_1(\mathbb{P})}(f,g)$$

and

$$N(\epsilon,\mathcal{F}^t,\rho_{\mathbb{P}}) \le N(\epsilon/K^t,\mathcal{F},d_{L_1(\mathbb{P})}).$$

The proof is in the appendix.

The condition of the lemma essentially means that under distribution λ the change in the probability mass on any state under any policy after one transition is bounded. It should be noted that for simulation, we can choose the initial state distribution and it should be such that $\lambda(y)>0\ \forall\ y$. Further, if $\lambda(y)=0$, the Markov chains are such that we must have $\lambda_f(y)=0$ as well, i.e., $\lambda_f\ll\lambda$. A particular case where this is satisfied is a set of positive recurrent Markov chains, say with the same invariant distribution π. If we choose $\lambda=\pi$, then $\lambda_f=\pi$ and the condition is trivially satisfied.

We now show that the estimate converges to the expected discounted reward uniformly over all policies in Π, and we also obtain the uniform rate of convergence.

THEOREM 5.2. *Let $(\mathbb{X},\Gamma,\lambda)$ be a measurable state space. Let $\mathbb{A}$ be the action space and r the $[0,R]$-valued reward function. Let $\Pi\subseteq\Pi_0$ be the space of stationary stochastic policies, $\mathcal{P}$ the space of Markov chain transition probabilities induced by Π, and $\mathcal{F}$ the space of simulation functions of $\mathcal{P}$ under the simple simulation model h. Suppose that* P-dim$(\mathcal{F})\le d$ *and the initial state distribution λ is such that $K:=\max\{\sup_{f\in\mathcal{F},x\in\mathbb{X}}\frac{\lambda_f(x)}{\lambda(x)},1\}$ is finite. Let $\hat{V}_n(\pi)$ be the estimate of $V(\pi)$ obtained by averaging the reward from n samples. Then, given any $\epsilon,\delta>0$,*

$$\mathbb{P}^n\left\{\sup_{\pi\in\Pi}|\hat{V}_n(\pi)-V(\pi)|>\epsilon\right\}<\delta$$

for

$$n \ge \frac{32R^2}{\alpha^2}\left(\log\frac{4}{\delta}+2d\left(\log\frac{32eR}{\alpha}+T\log K\right)\right). \tag{5.2}$$

T is the $\epsilon/2$-horizon time and $\alpha=\epsilon/2(T+1)$.

Proof. Fix a policy π. Let P be the induced Markov chain transition probability function simulated by the simple simulation function f_P. Let $R_t(x_0,\omega):=r\circ f_P\circ\cdots\circ f_P(x_0,\omega)$, with f_P composed t times, be the reward at time t, and denote $\mathcal{R}_t:=\{R_t : \mathbb{X}\times\Omega^\infty\to[0,R], P\in\mathcal{P}\}$. Let $V^T(\pi)$ be the expected discounted reward truncated up to T steps and $\hat{V}_n(\pi)=\frac{1}{n}\sum_{i=1}^n[\sum_{t=0}^T\gamma^t R_t^\pi(x_0^i,\omega^i)]$ its estimate from n finite time

simulations. Then,

$$\begin{aligned}|V(\pi)-\hat{V}_n(\pi)| &\le |V(\pi)-V^T(\pi)|+|V^T(\pi)-\hat{V}_n(\pi)|,\\ &\le |V^T(\pi)-\hat{V}_n^T(\pi)|+\frac{\epsilon}{2},\\ &\le \sum_{t=0}^{T}\left|\frac{1}{n}\sum_{i=1}^{n}[R_t^\pi(x_0^i,\omega^i)-\mathbb{E}(R_t^\pi)]\right|+\frac{\epsilon}{2}.\end{aligned}$$

Here, the expectation is with respect to the product measure $P_\pi^t\times\lambda\times\mu$. We show that with high probability, each term in the sum over t is bounded by $\alpha=\epsilon/2(T+1)$.

Note that

$$\int |r(f^t(x,\omega))-r(g^t(x,\omega))|d\mu(\omega)d\lambda(x)\le R\cdot\sum_x\lambda(x)\mu\{\omega: f^t(x,\omega)\ne g^t(x,\omega)\},$$

which as in Lemma 5.1 implies that

$$d_{L_1(\mathbb{P})}(r\circ f^t, r\circ g^t)\le R\cdot\rho_{\mathbb{P}}(f^t,g^t)\le R\cdot K^T d_{L_1(\mathbb{P})}(f,g).$$

Applying Theorem 3 from [17] with the "α" in the statement of that theorem set equal to $\epsilon/4R$ and $\nu=2R$, and using Lemma 5.1 and the inequality above, we get

$$\begin{aligned}&\mathbb{P}^n\left\{\sup_{R_t\in\mathcal{R}_t}\left|\frac{1}{n}\sum_{i=1}^{n}R_t(x_0^i,\omega^i)-\mathbb{E}(R_t)\right|>\alpha\right\}\\ &\le 2\mathbb{E}\left[\min\left(2N\left(\frac{\alpha}{16},\mathcal{R}_t|_S,d_{l_1}\right)e^{(-\frac{n\alpha^2}{32R^2})},1\right)\right]\\ &\le 4\sup_{\mathbb{P}}N\left(\frac{\alpha}{16},\mathcal{R}_t,d_{L_1(\mathbb{P})}\right)\exp\left(-\frac{n\alpha^2}{32R^2}\right)\\ &\le 4C\left(\frac{\alpha}{16},\mathcal{R}_t,d_{L_1}\right)\exp\left(-\frac{n\alpha^2}{32R^2}\right)\\ &\le 4\left(\frac{32eRK^T}{\alpha}\log\frac{32eRK^T}{\alpha}\right)^d\exp\left(-\frac{n\alpha^2}{32R^2}\right).\end{aligned}$$

This implies that the estimation error is bounded by α, with probability at least δ, if the number of samples is

$$n\ge\frac{32R^2}{\alpha^2}\left(\log\frac{4}{\delta}+2d\left(\log\frac{32eR}{\alpha}+T\log K\right)\right).\qquad\square$$

Remarks. 1. Theorem 5.2 implies that $\sup_{\pi\in\Pi}|\hat{V}_n(\pi)-V(\pi)|$ converges to zero in probability; hence the policy space Π is PAC-learnable. As in [30], the theorem assumes that the P-dimension of the $\mathcal{F}$ space is finite. Combined with Lemma 4.2, this gives the following corollary.

COROLLARY 5.3. *Under assumption* (5.1) *and if* $\mathcal{P}$ *is convex with* P-dim$(\mathcal{P})=d$, *result* (5.2) *of Theorem* 5.2 *holds.*

2. Our sample complexity is of the same order, in terms of δ,ϵ,T,R, and d, as the results of [30], but the two results are not directly comparable due to the different assumptions made. In fact, the major challenge in obtaining the uniform rate of convergence is relating the covering numbers and P-dimensions of the policy space Π

and $\mathcal{P}$ with the space $\mathcal{F}$. This is what we accomplished in this paper and what is missing in [30]. Also, unlike in [30], we do not require the simulation functions to be Lipschitz continuous. For a discrete state space, this is not a realistic assumption, as the following examples show.

(a) Consider the Markov chain on $\mathbb{N}$ such that the only transitions are from state 1 to state 2 with probability 1/2, to state 4 with probability $1/4, \ldots$, to state 2^k with probability $1/2^k$, etc. Let $\omega_l^k = \sum_{i=1}^{k+l-1} 2^{-i}$ and $\epsilon_l^k = 2^{-(k+l)}$ and define $\hat{\omega}^k = (\omega_1^k - \epsilon_1^k/2, \omega_2^k - \epsilon_2^k/2, \ldots)$ and $\check{\omega}^k = (\omega_1^k + \epsilon_1^k/2, \omega_2^k + \epsilon_2^k/2, \ldots)$. Thus, with noise sequence $\hat{\omega}^k$, the transition from state 1 is to state $\hat{x} = 2^k$, while with noise sequence $\check{\omega}^k$, the transition from state 1 is to state $\check{x} = 2^{k+1}$. Define the metric ρ on $\mathbb{X} \times \Omega^\infty$, $\rho((x_1, \hat{\omega}^k), (x_2, \check{\omega}^k)) = |x_1 - x_2| + |\hat{\omega}_1 - \check{\omega}_1|$. Then, it can be verified that

$$\rho(f(1, \hat{\omega}^k), f(1, \check{\omega}^k)) = |\hat{x} - \check{x}| + \epsilon_2^k \geq 2^k \ \forall k,$$

whereas Lipschitz continuity would require that it be less that $C2^{-(k+1)}$ for some positive constant C and every k. Thus, f is not Lipschitz continuous on Ω^∞.

(b) Consider the following Markov chain: state space $\mathbb{X} = \mathbb{N}$ again endowed with the same metric ρ as in the above example. Transitions are deterministic: transition from an even state n is to state $2n$, and from an odd state $n+1$ is to $3n$. Then, $\rho(f(n+1, \omega), f(n, \omega)) = n$ and so is not Lipschitz continuous on $\mathbb{X}$.

These examples demonstrate that Lipschitz continuity of the simulation functions on a discrete state space is not the right assumption to make.

3. *Markov games.* The generalization of the results of this section to discounted-reward Markov games is relatively straightforward. It is of considerable interest for many applications [38, 32]. Consider two players playing a Markov game with action spaces $\mathbb{A}_1$ and $\mathbb{A}_2$, state space $\mathbb{X}$, and transition function $P_{a_1,a_2}(x, y), a_1 \in \mathbb{A}_1, a_2 \in \mathbb{A}_2$. We consider only stationary policy spaces Π_1 and Π_2. The two reward functions r_1 and r_2 depend only on the state and have values in $[0, R]$. Denote the discounted-reward functions by $V_1(\pi_1, \pi_2)$ and $V_2(\pi_1, \pi_2)$ with discount factor $0 < \gamma < 1$. Denote the set of Markov chains induced by the policies in $\Pi_1 \times \Pi_2$ by $\mathcal{P} = \{P_{\pi_1,\pi_2} : (\pi_1, \pi_2) \in \Pi_1 \times \Pi_2)\}$ with $P_{\pi_1,\pi_2}(x, y) = \sum_{a_1,a_2} P_{a_1,a_2}(x, y)\pi_1(x, a_1)\pi_2(x, a_2)$.

Of interest is a uniform sample complexity bound such that the error in estimating both $V_1(\pi_1, \pi_2)$ and $V_2(\pi_1, \pi_2)$ is within ϵ with probability at least $1 - \delta$. This is now easily obtained by using Theorem 5.2 and bounding the maximum of the estimation error in V_1 and V_2. Such estimates may be used to compute the Nash equilibria [38].

A bound on the ϵ-covering number of $\mathcal{P}$ can be easily obtained by constructing $\epsilon/2$-covers for Π_1 and Π_2 under the L_1 metric, and then the ϵ cover for $\Pi_1 \times \Pi_2$ is obtained by taking the product of the two sets. The rest of the argument is the same as in Lemma 3.1.

4. We now show through two examples that the choice of simulation model to use for generating the sample trajectories affects whether we get uniform convergence or not.

Example 5.4. (i) Consider the three-state Markov chain in Figure 5.1 with the following transition matrix:

$$P = \begin{bmatrix} 0 & p & 1-p \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and with state 0 always the initial state. Let $\mathcal{P} = \{P \text{ as defined above with } 0 \leq p \leq 1\}$ denote the set of all such Markov chains as p varies. It is easy to check that $\mathcal{P}$ is

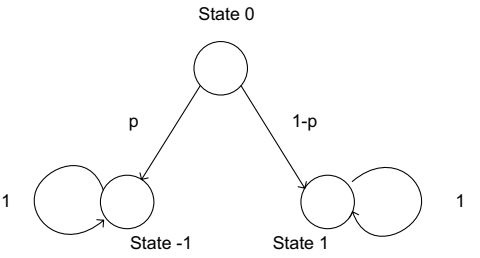


FIG. 5.1. *A three-state Markov chain with initial state* 0.

convex and has P-dimension 1. Let the reward in state 0 be 0, in state -1 be -1, and in state 1 be 1. This defines a Markov reward process, and for a given p we get the value $V_p = 1 - 2p$. Let $\mathcal{F}$ be the set of simple simulation functions that simulate the Markov chains in $\mathcal{P}$. By Lemma 4.2, P-dim($\mathcal{F}$) is 1 as well.

Now, consider the following set of simulations functions. Fix $p \in [0,1]$. Let $x_1, x_2, \ldots$ be rational numbers in $[0,1]$ such that

$$0 \le x_1 \le x_1 + p/2 \le x_2 \le x_2 + p/4 \le \cdots \le 1. \tag{5.3}$$

Generate a uniform $[0,1]$ random number ω. If it falls in $[x_1, x_1+p/2] \cup [x_2, x_2+p/4] \cup \cdots$, then the next state is -1; otherwise it is 1. This simulation function simulates a Markov chain with parameter p. Let $\tilde{\mathcal{F}}$ denote the set of all such simulation functions as p varies in $[0,1]$ and for every possible sequence of rational numbers satisfying (5.3). It is easy to check that P-dim($\tilde{\mathcal{F}}$) is not finite. Now, consider any $\omega_1, \omega_2, \ldots, \omega_n \in [0,1]$. Then, there exists a p and $x_1, x_2, \ldots$ such that $\omega_k \in [x_k, x_k + p/2^k]$, which means there is a simulation function which always goes to state -1 for this sequence of random numbers. Since n is arbitrary, this implies $\hat{V}_p = -1 \neq V_p$ for some p. Thus, in this case, even though the second set of simulation functions $\tilde{\mathcal{F}}$ simulates the same set of Markov chains $\mathcal{P}$, we do not get uniform convergence.

In the example above, there are many simulation functions in $\mathcal{F}$ that simulate the same Markov chain in $\mathcal{P}$. We now modify the example so that there is only one simulation function for each Markov chain.

(ii) Fix a positive integer m. Consider the $[0,1]$ interval divided into m^2 subintervals of length $1/m^2$. Pick any m of these subintervals, which can be done in $\binom{m^2}{m}$ ways. Arrange these choices in any order. Define

$$p_{m,k} = \frac{1}{m+1} + \frac{k}{\binom{m^2}{m} m(m+1)} \le \frac{1}{m}, \qquad k = 1, \ldots, \binom{m^2}{m}.$$

Scale the length of each of the m subintervals of the kth such choice to $p_{m,k}/m \le 1/m^2$, keeping the beginning point of the subintervals the same. Now, for every $p_{m,k}$ as defined above with $k = 1, \ldots, \binom{m^2}{m}$, and $m = 1, 2, \ldots$, define a simulation function $f_{m,k}$ such that for the Markov chain with parameter $p_{m,k}$, if the current state is zero, and the uniform $[0,1]$ random number falls in any of the subintervals, then the next state is -1; otherwise it is 1. All other Markov chains are simulated according to the simple simulation model. Thus, we get a new set of simulation functions $\bar{\mathcal{F}}$ with each Markov chain being simulated by only one simulation function. It is easy to check that P-dim($\bar{\mathcal{F}}$) is not finite. Further, given any $\omega_1, \omega_2, \ldots, \omega_n \in [0,1]$, we can always

pick an m large enough such that there is a k_m so that all of the ω_i lie outside the m subintervals, each of length $p_{m,k_m}/m$. Thus, the simulation goes to state -1 for each of the ω_i. Thus, $\hat{V}_{p_{m,k_m}} \neq V_{p_{m,k_m}}$ and we do not get uniform convergence.

These examples demonstrate that uniform convergence for MDPs depends not only on the geometry of the policy space but also on how the MDP is simulated to obtain the sample trajectories. The simple simulation model is in some sense the best model since it has the same complexity under convexity as the set of Markov chains it simulates (Lemma 4.2). However, these examples do not satisfy the conditions provided in Theorem 5.2. Note that those are sufficient conditions only.

An alternative simulation method for finite-dimensional convex spaces. An important special case is when the policy space is the convex hull of a finite number of policies, i.e., when all policies are (random) mixtures of finitely many policies. While the previous simulation method would still work, we present an alternative simulation method that exploits the convex structure.

The simulation method is the following. Suppose there are two policies π_i for $i = 0, 1$, each inducing a Markov chain P_i with simulation function f_i. Consider the mixture $\pi_w = w\pi_0 + (1-w)\pi_1$, so $P_w = wP_0 + (1-w)P_1$. For t steps, we have

$$P_w^t = w^t P_0^t + w^{t-1}(1-w)P_0^{t-1}P_1 + w^{t-1}(1-w)P_0^{t-2}P_1P_0 + \cdots + (1-w)^t P_1^t.$$

Note that P_0 and P_1 need not commute. To obtain $\hat{V}_n(\pi)$, we first determine the rewards at time t for $t = 0, \ldots, T$. To estimate the rewards, first draw the initial states $x_0^1, \ldots, x_0^n$ from λ, and the noise sequences $\omega^1, \ldots, \omega^n$ from μ. Then carry out 2^t simulations, one for each term in the sum of the equation above. For example, if $i_t \cdots i_1$ is the binary representation of k, then the contribution to reward from the kth term is determined by $r \circ f_{i_t} \circ \cdots \circ f_{i_1}(x_0^i, \omega^i)$. The estimate of the contribution to reward from the kth term, $\hat{R}_t^k$, is the mean over the n initial state and noise sequence pairs. The estimate of the reward at time t is

$$\hat{R}_t = w^t \hat{R}_t^0 + w^{t-1}(1-w)\hat{R}_t^1 + \cdots + (1-w)^t \hat{R}_t^{2^t-1}$$

and the value function estimate $\hat{V}_n(\pi)$ is $\sum_{t=0}^T \gamma^t \hat{R}_t$. This can be generalized to a policy space which is a convex hull of any finite number of policies.

THEOREM 5.5. *Let $\Pi = conv\{\pi_0, \ldots, \pi_{d-1}\}$, $\mathcal{P} = conv\{P_0, \ldots, P_{d-1}\}$ be the space of Markov chains induced by Π, and $\mathcal{F}$ be the space of simulation functions of $\mathcal{P}$ under the simple simulation model. Let $\hat{V}_n(\pi)$ be the estimate of $V(\pi)$ obtained from n samples. Then, given any $\epsilon, \delta > 0$*

$$\mathbb{P}^n \left\{ \sup_{\pi \in \Pi} |\hat{V}_n(\pi) - V(\pi)| > \epsilon \right\} < \delta$$

for $n \geq \frac{R^2}{2\alpha^2}\left(\log\frac{1}{\delta} + T\log 2d\right)$, where T is the $\epsilon/2$-horizon time and $\alpha = \epsilon/2(T+1)$.

Proof. Consider any $\pi \in \Pi$ and the corresponding $P \in \mathcal{P}$. Let $P = \sum_{i=0}^{d-1} a_i P_i$ with $a_i \geq 0$ and $\sum_i a_i = 1$. Thus, P^t can be written as

$$P^t = \sum_{k=0}^{d^t-1} w_k^t P_{i_t} \cdots P_{i_1}, \tag{5.4}$$

where $i_t \cdots i_1$ is a d-ary representation of k and the nonnegative weights w_k^t are such that $\sum_{k=0}^{d^t-1} w_k^t = 1$ and can be determined easily.

To obtain $\hat{V}_n(\pi)$, we need to simulate P for T steps as before. The t-step reward is determined by the state at time t, whose distribution is given by $\nu_t = P^t\lambda$. To estimate the reward, first draw the initial states $x_0^1, \ldots, x_0^n$ from λ, and the noise sequence $\omega^1, \ldots, \omega^n$ from μ. We carry out d^t simulations, one for each term in the sum of (5.4). Recall that $i_t \cdots i_1$ is a d-ary representation of k. Thus, to determine the contribution to the empirical reward at time t due to the kth term in (5.4), the state at time t is determined by $r \circ f_{i_t} \circ \cdots \circ f_{i_1}(x_0^i, \omega^i)$. Thus, an estimate of the expected reward at time t is

$$\hat{R}_t(w^t) = \sum_{k=0}^{d^t-1} w_k^t \left(\frac{1}{n} \sum_{i=1}^{n} r \circ f_{i_t} \circ \cdots \circ f_{i_1}(x_0^i, \omega^i) \right), \tag{5.5}$$

a consistent estimator of $\mathbb{E}_{P^t\lambda}[r(x)]$. $\hat{V}_n(\pi)$ is now given by $\sum_{t=0}^{T} \gamma^t \hat{R}_t(w^t(\pi))$, where $w^t(\pi)$ are the weights determined in 5.4 by policy π at time t. Denote $W_t := \{w^t(\pi) : \pi \in \Pi\}$.

Note that

$$\mathbb{E}_{P^t\lambda}[r(x)] = \sum_{k=0}^{d^t-1} w_k^t \left(\mathbb{E}_{P_{i_t} \cdots P_{i_1}\lambda}[r(x)] \right), \tag{5.6}$$

where the quantity in the brackets of (5.5) (denote it by $\hat{R}_t^k$) is an estimator for the quantity in the brackets of (5.6). Thus,

$$\begin{aligned} \mathbb{P}^n \left\{ \sup_{\pi \in \Pi} |\hat{V}_n(\pi) - V(\pi)| > \epsilon \right\} &\le \mathbb{P}^n \left\{ \sup_{w^t \in W_t} |\hat{R}_t(w^t) - \mathbb{E}_{P^t\lambda}[r(x)]| > \alpha \right\} \\ &\le d^T \max_k \mathbb{P}^n \{ |\hat{R}_t^k - \mathbb{E}_{P_{i_t} \cdots P_{i_1}\lambda}[r(x)]| > \alpha \} \\ &\le 2d^T \exp\left(-\frac{2n\alpha^2}{R^2} \right), \end{aligned}$$

where $\alpha = \epsilon/2(T+1)$ and the last inequality follows from Hoeffding's inequality. From the above, the sample complexity can be obtained as in the proof of Theorem 5.2. □

Remarks. This alternative simulation model has the same order of sample complexity as in the earlier case, but it has greater computational complexity since, for each chain, d^T simulations need to be carried out. However, if several Markov chains need to be simulated, as when performing an optimal policy search, the simulation is carried out only *once* for all the chains because the estimates for various mixtures are obtained by appropriately weighing the d^t estimates, $\hat{R}_t^k$.

Also, to obtain the estimates for $t \le T$, one need not repeat the simulation for each t. Instead, the T-step simulation suffices to yield the t-step estimate by simply ignoring the simulation beyond $t+1$. Thus, only d^T simulations are needed to obtain an estimate for a reward for any step $t \le T$ and any $P \in \mathcal{P}$.

6. Average-reward MDPs. Some MDP problems use the average-reward criterion. However, there are no published results on simulation-based uniform estimates of value functions of average-reward MDPs. We present such a result in this section.

Unlike the discounted-reward case, where we simulated several different sample paths, starting with different initial states, here the estimate is obtained from only one sample path. This is possible only when the policies we consider are such that

the Markov chains are stationary, ergodic, and weakly mixing. (The conditions under which a policy induces a stationary, ergodic Markov chain can be found in Chapter 5 of [9].) A related problem is addressed in [22], which uses Csiszer's concentration inequality to bound reward functionals for Doeblin chains. However, the bound is not uniform and distribution dependent.

Let λ_π denote the invariant measure of the Markov chain $\{X_k\}_{k=0}^\infty$ with transition probability function P_π, and Λ be the set of all such invariant measures. Let $\mathbb{P}$ denote the probability measure for the process. We assume that there is a Markov chain $\mathbb{P}_0$ with invariant measure (steady state distribution) λ_0 such that $\lambda_\pi \ll \lambda_0$, i.e., λ_π is absolutely continuous with respect to λ_0, meaning that $\lambda_\pi(A) = 0$ if $\lambda_0(A) = 0$ for any measurable set A [18]. We call such a chain a *reference* Markov chain. Let $\hbar_\pi(x) := \lambda_\pi(x)/\lambda_0(x)$ be the Radon–Nikodym derivative and assume that it is uniformly bounded, $\hbar_\pi(x) \le K \ \forall\ x$ and π. Let $\mathcal{H}$ be the set of all such Radon–Nikodym derivatives.

Our simulation methodology is to generate a sequence $\{x_k\}_{k=1}^n$ according to $\mathbb{P}_0$. We then multiply $r(x_t)$ by $\hbar_\pi(x_t)$ to obtain the tth-step reward for the Markov chain induced by policy π. The estimate of the value function is then obtained by taking an empirical average of the rewards, i.e.,

$$\hat{V}_n(\pi) = \frac{1}{n}\sum_{t=1}^{n} \tilde{r}_\pi(x_t),$$

in which $\tilde{r}_\pi(x) := r_\pi(x)\hbar_\pi(x)$. Let $\mathcal{R} = \{\tilde{r}_\pi : \pi \in \Pi\}$. Furthermore, in some problems, such as when the state space is multidimensional or complex, it may not be possible to integrate the reward function with respect to the stationary measure. In such cases, Monte Carlo type methods such as importance sampling are useful for estimating integrals. The method proposed here falls into such a category, and we present a uniform rate of convergence result for it. This approach is useful when it is difficult to sample from the stationary distributions of the Markov chains but easy to compute the derivative $\hbar_\pi(x)$.

We first state some definitions and results needed in the proof below. Let $\{X_n\}_{n=-\infty}^\infty$ be a process on the measurable space $(\mathbb{X}_{-\infty}^\infty, \mathcal{S}_{-\infty}^\infty)$. Consider a stationary, ergodic process with measure $\mathbb{P}$. Let P be the one-dimensional marginal, and $\bar{\mathbb{P}} = \prod_{-\infty}^\infty P$, the product measure under which the process is i.i.d.

Let $\mathbb{P}_{-\infty}^0$ and $\mathbb{P}_1^\infty$ be the semi-infinite marginals of $\mathbb{P}$. Define the *β-mixing coefficients* [14] as

$$\beta(k) := \sup\{|\mathbb{P}(A) - \mathbb{P}_{-\infty}^0 \times \mathbb{P}_1^\infty(A)| : A \in \sigma(X_{-\infty}^0, X_k^\infty)\}.$$

If $\beta(k) \to 0$ as $k \to \infty$, the process is said to be *β-mixing (or weakly mixing)*. From the definition of the β-mixing coefficients we get the following.

FACT 1 (see [29, 50]). *If $A \in \sigma(X_0, X_k, \ldots, X_{(m-1)k})$, $|\mathbb{P}(A) - \bar{\mathbb{P}}(A)| \le m\beta(k)$.*

We assume that the Markov chain $\mathbb{P}_0$ is β-mixing with mixing coefficients β_0. We also need the following generalization of Hoeffding's bound.

FACT 2 (McDiarmid–Azuma inequality [5, 27]). *Let $X_1, \ldots, X_n$ be i.i.d. drawn according to P and $g : X^n \to \mathbb{R}$. Suppose g has bounded differences, i.e.,*

$$|g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n)| \le c_i.$$

Then, $\forall \tau > 0$,

$$P^n\{g(X^n) - \mathbb{E}g(X^n) \ge \tau\} \le \exp\left(-\frac{2\tau^2}{\sum_{i=1}^n c_i^2}\right).$$

We now show that the estimation procedure enunciated above for the average-reward case produces estimates $\hat{V}(\pi)$ that converge uniformly over all policies to the true value $V(\pi)$. Moreover, we can obtain the rate of convergence, whose explicit form depends on the specific problem.

THEOREM 6.1. *Suppose the Markov chains induced by $\pi \in \Pi$ are stationary and ergodic. Assume there exists a Markov chain $\mathbb{P}_0$ with invariant measure λ_0 and mixing coefficient β_0 such that $\lambda_\pi \ll \lambda_0$ and the $\hbar_\pi$ are bounded by a constant K with* P-dim$(\mathcal{H}) \leq d$. *Denote by $\hat{V}_n(\pi)$ the estimate of $V(\pi)$ from n samples. Then, given any $\epsilon, \delta > 0$,*

$$\mathbb{P}_0 \left\{ \sup_{\pi \in \Pi} |\hat{V}_n(\pi) - V(\pi)| > \epsilon \right\} < \delta,$$

for n large enough such that $\gamma(m) + Rm\beta_0(k) + \tau(n) \leq \epsilon$, where $\tau(n) := (\frac{2R^2}{n} \log \frac{1}{\delta})^{1/2}$ and

$$\gamma(m) := \inf_{\alpha > 0} \left\{ \alpha + 8R \left(\frac{32eRK}{\alpha} \log \frac{32eRK}{\alpha} \right)^d \exp\left(-\frac{m\alpha^2}{32R^2} \right) \right\},$$

with $n = m_n k_n$ such that $m_n \to \infty$ and $k_n \to \infty$ as $n \to \infty$.

Proof. The idea of the proof is to use the β-mixing property and reduce the problem to one with i.i.d. samples. We then use techniques similar to those used in the proof of Theorem 5.2. The problem of iteration of simulation functions does not occur in this case, which makes the proof easier.

Let $x_1, x_2, \ldots, x_n$ be the state sequence generated according to $\mathbb{P}_0$ which can be done using the simple simulation model. Note that

$$E_0[r(x)\hbar_\pi(x)] = E_{\lambda_\pi}[r(x)],$$

in which E_{λ_π} and E_0 denote expectation taken with respect to the stationary measures λ_π and λ_0, respectively. Denote $\hat{\mathbb{E}}_n[\tilde{r}_\pi; x^n] := \frac{1}{n} \sum_{t=1}^n \tilde{r}_\pi(x_t)$ and observe that

$$\begin{aligned}
\mathbb{E}_0 \left[\sup_{\pi \in \Pi} |\hat{\mathbb{E}}_n[\tilde{r}_\pi; x^n] - E_0[\tilde{r}_\pi(x)]| \right] &\overset{(1)}{\leq} \frac{1}{k} \sum_{j=0}^{k-1} \mathbb{E}_0 \left[\sup_{\pi \in \Pi} \left| \frac{1}{m} \sum_{l=0}^{m-1} \tilde{r}_\pi(x_{lk+j}) - E_0[\tilde{r}_\pi(x)] \right| \right] \\
&\overset{(2)}{\leq} \alpha + R\mathbb{P}_0 \left\{ \sup_{\pi \in \Pi} \left| \frac{1}{m} \sum_{l=0}^{m-1} \tilde{r}_\pi(x_{lk}) - E_0[\tilde{r}_\pi(x)] \right| > \alpha \right\} \\
&\overset{(3)}{\leq} \alpha + Rm\beta_0(k) \\
&\quad + R\bar{\mathbb{P}}_0 \left\{ \sup_{\pi \in \Pi} \left| \frac{1}{m} \sum_{l=0}^{m-1} \tilde{r}_\pi(x_{lk}) - E_0[\tilde{r}_\pi(x)] \right| > \alpha \right\},
\end{aligned}$$

in which $\mathbb{E}_0$ is expectation with respect to $\mathbb{P}_0$, and $\bar{\mathbb{P}}_0$ is the i.i.d. product measure corresponding to $\mathbb{P}_0$. Inequality (1) follows by triangle inequality, (2) by stationarity and the fact that the reward function is bounded by R, and (3) by definition of the β-mixing coefficients.

CLAIM 1. *Suppose $\hbar_\pi(x) \leq K$ for all x and π and that* P-dim$(\mathcal{H}) \leq d$. *Then, $\forall \alpha > 0$*

$$C(\alpha, \mathcal{R}, d_{L_1}) \leq 2 \left(\frac{2eRK}{\alpha} \log \frac{2eRK}{\alpha} \right)^d.$$

Proof. Let λ_0 be a probability measure on $\mathbb{X}$. Observe that for $\tilde{r}_1, \tilde{r}_2 \in \mathcal{R}$,

$$\begin{aligned} d_{L_1(\lambda_0)}(\tilde{r}_1, \tilde{r}_2) &= \sum_x |\tilde{r}_1(x) - \tilde{r}_2(x)| \lambda_0(x) \\ &\le R \cdot \sum_x |\hbar_1(x) - \hbar_2(x)| \lambda_0(x) \\ &= R \cdot d_{L_1(\lambda_0)}(\hbar_1, \hbar_2). \end{aligned}$$

As argued for similar results earlier in the paper, this implies the desired conclusion. □

From Theorem 5.7 in [49], we then get

$$\begin{aligned} \bar{\mathbb{P}}_0 \left\{ \sup_{\pi \in \Pi} \left| \frac{1}{m} \sum_{l=0}^{m-1} \tilde{r}_\pi(x_{lk}) - E_0[\tilde{r}_\pi(x)] \right| > \alpha \right\} &\le 4C\left(\frac{\alpha}{16}, \mathcal{R}, d_{L_1}\right) \exp\left(-\frac{m\alpha^2}{32R^2}\right) \\ &\le 8\left(\frac{32eRK}{\alpha} \log \frac{32eRK}{\alpha}\right)^d \exp\left(-\frac{m\alpha^2}{32R^2}\right). \end{aligned}$$

Substituting above, we get

$$\mathbb{E}_0 \left[\sup_{\pi \in \Pi} |\hat{\mathbb{E}}_n[\tilde{r}_\pi; x^n] - E_0[\tilde{r}_\pi(x)]| \right] \le \gamma(m) + Rm\beta_0(k).$$

Now, defining $g(x^n)$ as the argument of $\mathbb{E}_0$ above and using the McDiarmid–Azuma inequality with $c_i = R/n$, we obtain that

$$\mathbb{P}_0 \left\{ \sup_{\pi \in \Pi} |\hat{\mathbb{E}}_n[\tilde{r}_\pi; x^n] - E_0[\tilde{r}_\pi(x)]| \ge \gamma(m) + Rm\beta_0(k) + \tau(n) \right\} < \delta,$$

where $\delta = \exp(-2n\tau^2(n)/R^2)$, and hence we get the desired result. □

Note that by assumption of the mixing property, $\beta_0(k_n) \to 0$, and for fixed ϵ and δ, $\gamma(m_n), \tau(n) \to 0$ as $n \to \infty$. The sample complexity result is implicit here since given functions γ, β_0, and τ, we can determine n, m_n, and k_n such that $n = m_n k_n$, $m_n \to \infty$, $k_n \to \infty$, and $\gamma(m_n) + Rm_n\beta_0(k_n) + \tau \le \epsilon$ for given ϵ and δ. The existence of m_n and k_n sequences such that $m_n\beta(k_n) \to 0$ is guaranteed by Lemma 3.1 in [49]. This implies $\delta \to 0$ as $n \to \infty$, and thus the policy space Π is PAC-learnable under the hypothesis of the theorem.

One of the assumptions we have made is that for policy π, the Radon–Nikodym derivative $\hbar_\pi$ is bounded by K. This essentially means that all the Markov chains are close to the *reference* Markov chain in the sense that the probability mass on any state does not differ by more than a multiplicative factor of K from that of $\mathbb{P}_0$. The assumption that $\mathcal{H}$ has finite P-dimension is less natural but essential to the argument. Let us now show the existence of reference Markov chains through an example.

Example 6.2. Consider any Markov chain with invariant measure λ_0. Consider any parametrized class of functions $\hbar_\pi : \mathbb{X} \to [0, 1]$, $\pi \in \Pi$ such that $\sum_x \lambda_0(x)\hbar_\pi(x) = 1$, and P-dim$(\{\hbar_\pi : \pi \in \Pi\}) = d$. Denote $\lambda_\pi(x) = \lambda_0(x)\hbar_\pi(x)$ and consider a set of Markov chains $\mathcal{P}$ with invariant measures λ_π. Then, clearly, λ_0 is an invariant measure of a *reference* Markov chain for the set $\mathcal{P}$ of Markov chains.

7. Partially observable MDPs with general policies. We now consider partially observed discounted-reward MDPs with general policies (nonstationary with

memory). The setup is as before, except that the policy depends on observations $y \in \mathbb{Y}$, governed by the (conditional) probability $\nu(y|x)$ of observing $y \in \mathbb{Y}$ when the state is $x \in \mathbb{X}$. Let h_t denote the history $(y_0, a_1, y_1, \ldots, a_t, y_t)$ of observations and actions up to time t.

The results of section 5 extend when the policies are nonstationary; however, there are many subtleties regarding the domain and range of simulation functions, and measures, and some details are different. Let $\mathcal{H}_t = \{h_t = (y_0, a_1, y_1, \ldots, a_t, y_t) : a_s \in \mathbb{A}, y_s \in \mathbb{Y}, 0 \le s \le t\}$. Let Π be the set of policies $\pi = (\pi_1, \pi_2, \ldots)$, with $\pi_t : \mathcal{H}_t \times \mathbb{A} \to [0,1]$ a probability measure on $\mathbb{A}$ conditioned on $h_t \in \mathcal{H}_t$. Let Π_t denote the set of all policies π_t at time t with $\pi \in \Pi$. This gives rise to a conditional state transition function $P_t(x, x'; h_t)$, the probability of transition from state x to x' given history h_t up to time t. Under π,

$$P_t(x, x'; h_t) = \sum_a P_a(x, x')\pi_t(h_t, a).$$

Let $\mathcal{P}_t$ denote the set of all P_{π_t} induced by the policies π_t with $\pi \in \Pi$. Then, defining the usual $d_{TV(\lambda)}$ metric on $\mathcal{P}_t$ and the usual L_1 metric on Π_t, we get the next result.

LEMMA 7.1. *Suppose Π_t and $\mathcal{P}_t$ are as defined above with* P-dim$(\Pi_t) = d$. *Assume λ and ρ are probability measures on $\mathbb{X}$ and $\mathcal{H}_t$, respectively, and σ a probability measure on $\mathbb{A}$ such that $\pi_t(h_t, a)/\sigma(a) \le K \ \forall h_t \in \mathcal{H}_t, a \in \mathbb{A}, \pi_t \in \Pi_t$. Then, for $0 < \epsilon < e/4 \log_2 e$,*

$$N(\epsilon, \mathcal{P}_t, d_{TV(\lambda\times\rho)}) \le N(\epsilon, \Pi_t/\sigma, d_{L_1(\sigma\times\rho)}) \le \left(\frac{2eK}{\epsilon} \log \frac{2eK}{\epsilon}\right)^d.$$

The proof can be found in the appendix.

Let $\mathcal{F}_t$ be the set of simulation functions of $\mathcal{P}_t$ under the simple simulation model. Thus, $f_t \in \mathcal{F}_t$ for $t \ge 2$ is defined on $f_t : \mathbb{X} \times \mathcal{H}_{t-1} \times \Omega^\infty \to \mathbb{X} \times \mathcal{H}_t \times \Omega^\infty$, while $f_1 \in \mathcal{F}_1$ shall be defined on $f_1 : \mathbb{X} \times \Omega^\infty \to \mathbb{X} \times \mathcal{H}_1 \times \Omega^\infty$. This is because at time $t = 1$, there is no history, and the state transition depends only on the initial state and the noise. For $t > 1$, the state transition depends on the history as well. Further, the function definitions have to be such that the composition $f_t \circ f_{t-1} \circ \cdots \circ f_1$ is well defined.

It is straightforward to verify that Lemma 4.2 extends to Lemma 7.2.

LEMMA 7.2. *Suppose $\mathcal{P}$ is convex (being generated by a convex space of policies Π). Let* P-dim$(\mathcal{P}_t) = d$. *Let $\mathcal{F}_t$ be the corresponding space of simple simulation functions induced by $\mathcal{P}_t$. Then,* P-dim$(\mathcal{F}_t) = d$.

By $\mathcal{F}^t$ we shall denote the set of functions $f^t = f_t \circ \cdots \circ f_1$, where $f_s \in \mathcal{F}_s$ and they arise from a common policy π. Note that $f^t : \mathbb{X} \times \Omega^\infty \to \mathcal{Z}_t \times \Omega^\infty$, where $\mathcal{Z}_t = \mathbb{X} \times \mathcal{H}_t$. We shall consider the following pseudometric on $\mathcal{F}_t$ with respect to a measure λ_t on $\mathcal{Z}_{t-1}$ for $t \ge 2$ and measure σ on Ω^∞,

$$\rho_t(f_t, g_t) := \sum_{z \in \mathcal{Z}_{t-1}} \lambda_t(z)\sigma\{\omega : f_t(z, \omega) \ne g_t(z, \omega)\}.$$

We shall take ρ_1 as the pseudometric on $\mathcal{F}^t$ with respect to the product measure $\lambda \times \sigma$. Let

$$\lambda_{f^t}(z) := \sum_{z \in \mathcal{Z}_t} \lambda(x)\sigma\{\omega : f^t(x, \omega) = (z, \theta^t\omega)\}$$

be a probability measure on $\mathcal{Z}_t$. We now state the extension of the technical lemma needed for the main theorem of this section.

LEMMA 7.3. *Let λ be a probability measure on $\mathbb{X}$ and λ_{f^t} be the probability measure on $\mathcal{Z}_t$ as defined above. Suppose that* P-dim $(\mathcal{F}_t) \le d$ *and there exists probability measures λ_t on $\mathcal{Z}_t$ such that $K := \max\{\sup_t \sup_{f^t \in \mathcal{F}^t, z \in \mathcal{Z}_t} \frac{\lambda_{f^t}(z)}{\lambda_{t+1}(z)}, 1\} < \infty$. Then, for $1 \le t \le T$,*

$$N(\epsilon, \mathcal{F}^t, \rho_1) \le N\left(\frac{\epsilon}{Kt}, \mathcal{F}_t, \rho_t\right) \cdots N\left(\frac{\epsilon}{Kt}, \mathcal{F}_1, \rho_1\right) \le \left(\frac{2eKt}{\epsilon} \log \frac{2eKt}{\epsilon}\right)^{dt}.$$

The proof can be found in the appendix. We now obtain our sample complexity result.

THEOREM 7.4. *Let $(\mathbb{X}, \Gamma, \lambda)$ be the measurable state space, $\mathbb{A}$ the action space, $\mathbb{Y}$ the observation space, $P_a(x, x')$ the state transition function, and $\nu(y|x)$ the conditional probability measure that determines the observations. Let $r(x)$ be the real-valued reward function bounded in $[0, R]$. Let Π be the set of stochastic policies (nonstationary and with memory in general), $\mathcal{P}_t$ be the set of state transition functions induced by Π_t, and $\mathcal{F}_t$ the set of simulation functions of $\mathcal{P}_t$ under the simple simulation model. Suppose that* P-dim $(\mathcal{P}_t) \le d$. *Let λ and σ be probability measures on $\mathbb{X}$ and $\mathbb{A}$, respectively, and λ_{t+1} a probability measure on $\mathcal{Z}_t$ such that $K := \max\{\sup_t \sup_{f^t \in \mathcal{F}^t, z \in \mathcal{Z}_t} \frac{\lambda_{f^t}(z)}{\lambda_{t+1}(z)}, 1\}$ is finite, where λ_{f^t} is as defined above. Let $\hat{V}_n(\pi)$ be the estimate of $V(\pi)$ obtained from n samples. Then, given any $\epsilon, \delta > 0$, and with probability at least $1 - \delta$,*

$$\sup_{\pi \in \Pi} |\hat{V}_n(\pi) - V(\pi)| < \epsilon$$

for $n \ge \frac{32R^2}{\alpha^2}\left(\log \frac{4}{\delta} + 2dT(\log \frac{32eR}{\alpha} + \log KT)\right)$, where T is the $\frac{\epsilon}{2}$-horizon time and $\alpha = \frac{\epsilon}{2}(T+1)$.

Remarks. A special case is when the policies are *stationary and memoryless*, i.e., $\pi_t = \pi_1 \; \forall\, t$, and $\pi_1 : \mathbb{Y} \times \mathbb{A} \to [0, 1]$ depends only on the current observation. Let Π_1 be the set of all π_1. Then, each policy $\pi \in \Pi$ induces a time-homogeneous Markov chain with probability transition function given by $P_\pi(x, x') = \sum_{a,y} P_a(x, x')\pi_1(y, a)\nu(y|x)$. Let $\mathcal{P}$ denote the set of all probability transition functions induced by Π. In general if Π is convex, $\mathcal{P}$ is convex. We will denote the set of simulation functions under the simple simulation model for $P \in \mathcal{P}$ by $\mathcal{F}$. Suppose that P-dim $(\mathcal{P}) = d$; then by Lemma 4.2, P-dim $(\mathcal{P})$ = P-dim $(\mathcal{F}) = d$. This implies that the sample complexity result of Theorem 5.2 for discounted-reward MDPs holds for the case when the state is partially observable and the policies are stationary and memoryless. Thus, for uniform convergence and estimation, partial observability of the state does not impose any extra cost in terms of sample complexity. Note that in the case of general polices, the sample complexity is $O(T \log T)$ times more.

8. Conclusions. The paper considers simulation-based value function estimation methods for Markov decision processes (MDPs). Uniform sample complexity results are presented for the discounted-reward case. The combinatorial complexity of the space of simulation functions under the proposed simple simulation model is shown to be the same as that of the underlying space of induced Markov chains when the latter is convex. Using ergodicity and weak mixing leads to similar uniform sample complexity result for the average-reward case, when a *reference* Markov chain exists.

Extensions of the results are obtained when the MDP is partially observable with general policies. Remarkably, the sample complexity results have the same order for both completely and partially observed MDPs when stationary and memoryless

policies are used. Sample results for discounted-reward Markov games can be deduced easily as well.

The results can be seen as an extension of the theory of PAC (probably approximately correct) learning for partially observable Markov decision processes (POMDPs) and games. PAC theory is related to the system identification problem.

One of the key contributions of this paper is the observation that how we simulate an MDP matters for obtaining uniform estimates. This is a new (and surprising) observation. Thus, the results of this paper can also be seen as the first steps towards developing an empirical process theory for MDPs. Such a theory would go a long way in establishing a theoretical foundation for computer simulation of complex engineering systems.

We have used Hoeffding's inequality in obtaining the rate of convergence for discounted-reward MDPs and used the McDiarmid–Azuma inequality for the average-reward MDPs, though more sophisticated and tighter inequalities of Talagrand [41] (see also Samson [37]) can be used as well. This would yield better results and is part of future work.

Appendix A. Proof of Lemma 3.1.

Proof. We first relate the $d_{TV(\lambda)}$ pseudometric on $\mathcal{P}$ with the $d_{L_1(\lambda\times\sigma)}$ pseudometric on Π. Pick any $\pi, \pi' \in \Pi$ and denote $P = P_\pi$ and $P' = P_{\pi'}$. Then,

$$\begin{aligned} d_{TV(\lambda)}(P,P') &= \sum_y \left| \sum_x \lambda(x) \sum_a P_a(x,y)(\pi(x,a) - \pi'(x,a)) \right| \\ &\le \sum_a \sum_x \sum_y \lambda(x) P_a(x,y) |\pi(x,a) - \pi'(x,a)| \\ &\le \sum_a \sigma(a) \sum_x \lambda(x) |\pi(x,a) - \pi'(x,a)| / \sigma(a) \\ &= d_{L_1(\sigma\times\lambda)}(\pi/\sigma, \pi'/\sigma). \end{aligned}$$

The second inequality above follows from the triangle inequality and the fact that the order of the sums over y, x, and a can be changed by Fubini's theorem [18] and noting that $\sum_y P_a(x,y) = 1$. Thus, if $\Pi'/\sigma \subseteq \Pi/\sigma$ is an ϵ-net for Π/σ, $\{P_\pi, \pi \in \Pi'\}$ is an ϵ-net for $\mathcal{P}$. Further, the spaces Π and Π/σ have the same P-dimension, as can be easily verified. The bound on the covering number is then given by a standard result (see Theorem 4.3 of [49]). □

Appendix B. Proof of Lemma 5.1.

Proof. Consider any $f, g \in \mathcal{F}$, and $x \in \mathcal{X}$. Then,

$$\begin{aligned} &\mu\{\omega : f^t(x,\omega) \ne g^t(x,\omega)\} \\ &= \mu\{\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) = g^{t-1}(x,\omega)\} \\ &\quad + \mu\{\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) \ne g^{t-1}(x,\omega)\} \\ &= \mu\{\cup_y (\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) = g^{t-1}(x,\omega) = (y,\theta^{t-1}\omega))\} \\ &\quad + \mu\{\cup_y (\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) \ne g^{t-1}(x,\omega), f^{t-1}(x,\omega) = (y,\theta^{t-1}\omega))\} \\ &\le \mu\{\cup_y (\omega : f(y,\theta^{t-1}\omega) \ne g(y,\theta^{t-1}\omega), f^{t-1}(x,\omega) = g^{t-1}(x,\omega) = (y,\theta^{t-1}\omega))\} \\ &\quad + \mu\{\cup_y (\omega : f^{t-1}(x,\omega) \ne g^{t-1}(x,\omega), f^{t-1}(x,\omega) = (y,\theta^{t-1}\omega))\} \\ &\le \sum_y \mu\{\omega : f(y,\theta^{t-1}\omega) \ne g(y,\theta^{t-1}\omega) | f^{t-1}(x,\omega) = (y,\theta^{t-1}\omega)\} \mu\{\omega : f^{t-1}(x,\omega) = y\} \\ &\quad + \mu\{\omega : f^{t-1}(x,\omega) \ne g^{t-1}(x,\omega)\}. \end{aligned}$$

It is easy to argue that $\lambda_{f^{t-1}}(y) \le K^{t-1}\lambda(y)$, where $\lambda_{f^{t-1}}(y) = \sum_x \lambda(x)\mu\{\omega : f^{t-1}(x,\omega) = (y,\theta^{t-1}\omega)\}$. Thus, multiplying both sides of the above sequence of inequalities and summing over x, and observing that

$$\mu\{\omega : f(y,\theta^{t-1}\omega) \ne g(y,\theta^{t-1}\omega)\} = \mu\{\omega' : f(y,\omega') \ne g(y,\omega')\},$$

we get that the first part of the right side is

$$\sum_y \lambda_{f^{t-1}}(y)\mu\{\omega' : f(y,\omega') \ne g(y,\omega')\} \le K^{t-1} \cdot \sum_y \lambda(y)\mu\{\omega' : f(y,\omega') \ne g(y,\omega')\}.$$

This implies that

$$\begin{aligned}\rho_{\mathbb{P}}(f^t, g^t) &\le K^{t-1}\rho_{\mathbb{P}}(f,g) + \rho_{\mathbb{P}}(f^{t-1}, g^{t-1}) \\ &\le (K^{t-1} + K^{t-2} + \cdots + 1)\rho_{\mathbb{P}}(f,g) \\ &\le K^t \rho_{\mathbb{P}}(f,g),\end{aligned}$$

where the second inequality is obtained by induction. Now,

$$\sum_x \lambda(x)\mu\{\omega : f(x,\omega) \ne g(x,\omega)\} \le \sum_x \lambda(x) \int |f(x,\omega) - g(x,\omega)| d\mu(\omega),$$

and thus $\rho_{\mathbb{P}}(f^t, g^t) \le K^t \cdot d_{L_1}(f,g)$, which proves the required assertion. □

Appendix C. Proof of Lemma 7.1.

Proof. Pick any $\pi_t, \pi'_t \in \Pi_t$, and denote $P = P_{\pi_t}$ and $P' = P_{\pi'_t}$. Then,

$$\begin{aligned}d_{TV(\lambda\times\rho)}(P,P') &= \sum_x \lambda(x) \sum_{x'\in\mathbb{X}} \left| \sum_{a\in\mathbb{A}, h_t\in\mathcal{H}_t} P_a(x,x')(\pi_t(h_t,a) - \pi'_t(h_t,a)) \right| \rho(h_t) \\ &\le \sum_{h_t} \rho(h_t) \sum_x \lambda(x) \sum_a \sum_{x'} P_a(x,x') |\pi_t(h_t,a) - \pi'_t(h_t,a)| \\ &\le \sum_{h_t} \rho(h_t) \sum_a \sigma(a) \left| \frac{\pi_t(y,a)}{\sigma(a)} - \frac{\pi'_t(y,a)}{\sigma(a)} \right| \\ &= d_{L_1(\sigma\times\rho)} \left(\frac{\pi_t}{\sigma}, \frac{\pi'_t}{\sigma} \right).\end{aligned}$$

The second inequality above follows by changing the order of the sums over a and x', noting that $\sum_{x'} P_a(x,x') = 1$ and denoting $\sum_x \lambda(x) = 1$. The rest of the argument is the same as in Lemma 3.1. □

Appendix D. Proof of Lemma 7.3. The proof of Lemma 7.3 is similar but the details are somewhat more involved.

Proof. Consider any $f^t, g^t \in \mathcal{F}^t$, and $x \in \mathcal{X}$. Then,

$$\begin{aligned}&\mu\{\omega : f^t(x,\omega) \ne g^t(x,\omega)\} \\ &\quad = \mu\{\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) = g^{t-1}(x,\omega)\} \\ &\qquad + \mu\{\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) \ne g^{t-1}(x,\omega)\} \\ &\quad = \mu\{\cup_{z\in\mathcal{Z}_{t-1}}(\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) = g^{t-1}(x,\omega) = (z,\theta^{t-1}\omega))\} \\ &\qquad + \mu\{\cup_{z\in\mathcal{Z}_{t-1}}(\omega : f^t(x,\omega) \ne g^t(x,\omega), f^{t-1}(x,\omega) \\ &\qquad\qquad \ne g^{t-1}(x,\omega), f^{t-1}(x,\omega) = (z,\theta^{t-1}\omega))\}\end{aligned}$$

$$\begin{aligned}
&\leq \mu\{\cup_{z\in\mathcal{Z}_{t-1}}(\omega : f^t(x,\omega) \neq g^t(x,\omega), f^{t-1}(x,\omega) = g^{t-1}(x,\omega) = (z,\theta^{t-1}\omega))\} \\
&\quad + \mu\{\cup_{z\in\mathcal{Z}_{t-1}}(\omega : f^{t-1}(x,\omega) \neq g^{t-1}(x,\omega), f^{t-1}(x,\omega) = (z,\theta^{t-1}\omega))\} \\
&\leq \sum_{z\in\mathcal{Z}_{t-1}} \mu\{\omega : f_t(z,\theta^{t-1}\omega) \neq g_t(z,\theta^{t-1}\omega) | f^{t-1}(x,\omega) \\
&\qquad\qquad = (z,\theta^{t-1}\omega)\}\mu\{\omega : f^{t-1}(x,\omega) = (z,\theta^{t-1}\omega)\} \\
&\quad + \mu\{\omega : f^{t-1}(x,\omega) \neq g^{t-1}(x,\omega)\}.
\end{aligned}$$

Multiplying both sides of the above sequence of inequalities and summing over x, and observing again that

$$\mu\{\omega : f_t(z,\theta^{t-1}\omega) \neq g_t(z,\theta^{t-1}\omega)\} = \mu\{\omega' : f_t(z,\omega') \neq g_t(z,\omega')\},$$

we get that the first part of the right side is

$$\sum_{z\in\mathcal{Z}_{t-1}} \lambda_{f^{t-1}}(z)\mu\{\omega' : f_t(z,\omega') \neq g_t(z,\omega')\} \leq K\cdot \sum_{z\in\mathcal{Z}_{t-1}} \lambda_t(z)\mu\{\omega' : f_t(z,\omega') \neq g_t(z,\omega')\}.$$

This by induction implies

$$\rho_1(f^t, g^t) \leq K(\rho_t(f_t, g_t) + \cdots + \rho_1(f_1, g_1)),$$

which implies the first inequality. For the second inequality, note that the ρ pseudometric and the L_1 pseudometric are related; thus,

$$\sum_z \lambda_t(z)\mu\{\omega : f_t(z,\omega) \neq g_t(z,\omega)\} \leq \sum_z \lambda_t(z) \int |f_t(z,\omega) - g_t(z,\omega)| d\mu(\omega),$$

which relates their covering numbers. Also, the covering number under the L_1 pseudometric can be bounded in terms of the P-dim of appropriate spaces. □

Acknowledgments. We thank Peter Bartlett, Tunc Simsek, and Antonis Dimakis for many helpful discussions. The counterexample 5.4(ii) was jointly developed with Antonis Dimakis. We also thank the anonymous referees for their comments which helped to greatly improve the quality of the paper.

REFERENCES

[1] D. Aldous and U. Vazirani, *A Markovian extension of Valiant's learning model*, Inform. Comput., 117 (1995), pp. 181–186.
[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, *Scale-sensitive dimension, uniform convergence and learnability*, J. ACM, 44 (1997), pp. 615–631.
[3] M. Anthony and P. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, UK, 1999.
[4] E. Altman and O. Zeitouni, *Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality*, Math. Oper. Res., 19 (1994), pp. 955–974.
[5] K. Azuma, *Weighted sums of certain dependent random variables*, Tôhoku Math. J. (2), 19 (1967), pp. 357–367.
[6] J. Baxtar and P. L. Bartlett, *Infinite-horizon policy-gradient estimation*, J. Artificial Intelligence Res., 15 (2001), pp. 319–350.
[7] G. M. Benedek and A. Itai, *Learning with respect to fixed distributions*, Theoret. Comput. Science, 86 (1991), pp. 377–390.
[8] V. Bentkus, *On Hoeffding's inequalities*, Ann. Probab., 32 (2004), pp. 1650–1673.
[9] V. Borkar, *Topics in Controlled of Markov Chains*, Pitman Res. Notes Math. 240, Longman Scientific Technical, Harlow, John Wiley, New York, 1991.

[10] M. C. Campi and P. R. Kumar, *Learning dynamical systems in a stationary environment*, Systems Control Lett., 34 (1998), pp. 125–132.
[11] H. Chernoff, *A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations*, Ann. Math. Statist., 23 (1952), pp. 493–507.
[12] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.
[13] P. Diaconis and D. Freedman, *Iterated random functions*, SIAM Rev., 41 (1999), pp. 45–76.
[14] P. Doukhan, *Mixing: Properties and Examples*, Lecture Notes in Statist., 85, Springer-Verlag, Berlin, 1994.
[15] R. Dudley, *Uniform Central Limit Theorems*, Cambridge University Press, Cambridge, UK, 1999.
[16] D. Gamarnik, *Extensions of the PAC framework to finite and countable Markov chains*, IEEE Trans. Inform. Theory, 49 (2003), pp. 338–345.
[17] D. Haussler, *Decision theoretic generalizations of the PAC model for neural nets and other learning applications*, Inform. Comput., 100 (1992), pp. 78–150.
[18] P. R. Halmos, *Measure Theory*, Springer-Verlag, Berlin, 1974.
[19] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 13–30.
[20] M. Kearns, Y. Mansour, and A. Y. Ng, *Approximate planning in large POMDPs via reusable trajectories*, in Advances in Neural Information Processing Systems 12, MIT Press, 2000.
[21] A. N. Kolmogorov and V. M. Tihomirov, *ϵ-entropy and ϵ-capacity of sets in functional space*, Amer. Math. Soc. Transl. Ser. 2, 17 (1961), pp. 277–364.
[22] I. Kontoyiannis, L. A. L. Montano, and S. P. Meyn, *Relative entropy and exponential deviation bounds for general Markov chains*, in Proceedings of the International Symposium on Information Theory, IEEE, 2005, pp. 1563–1567.
[23] M. Ledoux, *The Concentration of Measure Phenomenon*, Math. Surveys Monogr. 89, AMS, Providence, RI, 2001.
[24] S. Mannor and J. N. Tsitsiklis, *On the empirical state-action frequencies in Markov decision processes under general policies*, Math. Oper. Res., 30 (2005), pp. 545–561.
[25] P. Marbach and J. N. Tsitsiklis, *Approximate gradient methods in policy-space optimization of Markov reward processes*, Discrete Event Dyn. Sys., 13 (2003), pp. 111–148.
[26] K. Marton, *Measure concentration for Euclidean distance in the case of dependent random variables*, Ann. Probab., 32 (2004), pp. 2526–2544.
[27] C. McDiarmid, *On the method of bounded differences*, in Surveys in Combinatorics, Cambridge University Press, Cambridge, UK, 1989, pp. 148–188.
[28] P. Massart, *Some applications of concentration inequalities to statistics*, Ann. Fac. Sci. Toulouse Math. (6), 9 (2000), pp. 245–303.
[29] A. Nobel and A. Dembo, *A note on uniform laws of averages for dependent processes*, Statist. Probab. Lett., 17 (1993), pp. 169–172.
[30] A. Y. Ng and M. I. Jordan, *Pegasus: A policy search method for large MDPs and POMDPs*, in Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence, Kaufman, San Franciso, CA, 2000, pp. 405–414.
[31] C. H. Papadimitriou and J. N. Tsitsiklis, *The complexity of Markov decision processes*, Math. Oper. Res., 12 (1987), pp. 441–450.
[32] M. Pesendorfer and P. S. Dengler, *Identification and Estimation of Dynamic Games*, Working Paper 9726, National Bureau of Economic Research, Cambridge, MA, 2003.
[33] L. Peshkin and S. Mukherjee, *Bounds on sample size for policy evaluation in Markov environments*, in Computational Learning Theory, Lecture Notes in Comput. Sci. 2111, Springer, Berlin, 2001, pp. 616–629.
[34] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, Berlin, 1984.
[35] D. Pollard, *Empirical Process Theory and Applications*, Institute of Mathematical Statistics, Hayward, CA, 1990.
[36] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley and Sons, New York, 1994.
[37] P-M. Samson, *Concentration of measure inequalities for Markov chains for Φ-mixing processes*, Ann. Probab., 28 (2000), pp. 416–461.
[38] D. H. Shim, H. J. Kim, and S. Sastry, *Decentralized nonlinear model predictive control of multiple flying robots in dynamic environments*, in Proceedings of the 42nd Conference IEEE Conference on Decision and Control, IEEE, Los Alamitos, 2003, pp. 3621–3626.
[39] J. M. Steele, *Probability Theory and Combinatorial Optimization*, SIAM, Philadelphia, PA, 1997.
[40] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Haute Études, Sci. Publ. Math., 81 (1995), pp. 73–205.

[41] M. Talagrand, *A new look at independence*, Ann. Probab., 24 (1996), pp. 1–34.
[42] L. Valiant, *A theory of the learnable*, Commun. ACM, 27 (1984), pp. 1134–1142.
[43] S. Van de Geer, *On Hoeffding's inequality for dependent random variables*, in Empirical Process Techniques for Dependent Data, H. Dehling, T. Mikosch, and M. Sorensen, eds., Birkhäuser Boston, Boston, MA, 2002, pp. 161–170.
[44] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer-Verlag, Berlin, 1996.
[45] V. N. Vapnik and A. Ya. Chervonenkis, *On the uniform convergence of relative frequencies to their probabilities*, Theory Probab. Appl., 16 (1971), pp. 264–280.
[46] V. N. Vapnik and A. Ya. Chervonenkis, *Necessary and sufficient conditions for the uniform convergence of means to their expectations*, Theory Probab. Appl., 26 (1981), pp. 532–553.
[47] V. Vapnik and A. Chervonenkis, *The necessary and sufficient conditions for consistency in the empirical risk minimization method*, Pattern Recognition and Image Analysis, 1 (1991), pp. 283–305.
[48] M. Vidyasagar and R. L. Karandikar, *System identification: A learning theory approach*, in Proceedings of the 40th IEEE Conference on Decision and Control, IEEE, Los Alamitos, CA, 2001, pp. 2001–2006
[49] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks*, 2nd ed., Springer-Verlag London, LTD., London, 2003.
[50] B. Yu, *Rates of convergence of empirical processes for mixing sequences*, Ann. Probab., 22 (1994), pp. 94–116.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1657–1679

MAXIMIZING VISIBILITY IN NONCONVEX POLYGONS: NONSMOOTH ANALYSIS AND GRADIENT ALGORITHM DESIGN*

ANURAG GANGULI†, JORGE CORTÉS‡, AND FRANCESCO BULLO§

Abstract. This paper presents a motion control algorithm for a planar mobile observer such as a mobile robot equipped with an omnidirectional camera. We propose a nonsmooth gradient algorithm for the problem of maximizing the area of the region visible to the observer in a non-self-intersecting nonconvex polygon. First, we show that the visible area is almost everywhere a locally Lipschitz function of the observer location. Second, we provide a novel version of the LaSalle invariance principle for discontinuous vector fields and Lyapunov functions with a finite number of discontinuities. Finally, we establish the asymptotic convergence properties of the nonsmooth gradient algorithm and illustrate numerically its performance.

Key words. geometric optimization, dynamical systems, robotics, nonsmooth analysis, stability theory

AMS subject classifications. 78M50, 37N35, 68T40, 49J52, 34D20

DOI. 10.1137/050621918

1. Introduction.

Consider a single-point mobile robot in a planar nonconvex environment modeled as a non-self-intersecting polygon: how should the robot move in order to monotonically increase the area of its visible region (i.e., the region within its line of sight)? This problem is the subject of this paper. The following are the modeling assumptions in our method of approach. The dynamical model for the robot's motion is a first-order system of the form $\dot{p} = u$, where p refers to the position of the robot in the environment and u is the driving input. The robot is equipped with an omnidirectional line-of-sight range sensor; the range of the sensor is larger than the diameter of the environment. The robot does not know the entire environment nor its position in it, and its instantaneous motion depends only on what is within its line of sight (this assumption restricts our attention to memoryless feedback laws).

In broad terms, this work is related to numerous references on optimal sensor location and motion planning coming from the computational geometry, geometric optimization, and robotics literature. The problem we consider is akin to the next best view problems in robotics for two-dimensional map building. In these map-building problems the objective is to compute the next position of a robot in an environment with obstacles that maximizes the gain in visible area. A heuristic is proposed and simulated in [10] and in the early work [13]. Other relevant references can be found in computational geometry. For example, the classic art gallery problem is to find the smallest number of such guards necessary for each point of the environment to be visible to at least one guard; see [7, 1]. Also studied in computational geometry is the

*Received by the editors January 3, 2005; accepted for publication (in revised form) April 14, 2006; published electronically November 14, 2006. This material is based upon work supported in part by AFOSR through Award F49620-02-1-0325 and by ONR through YIP Award N00014-03-1-0512. A preliminary version of this manuscript appeared in *Proceedings of the American Control Conference*, Portland, OR, 2005.
http://www.siam.org/journals/sicon/45-5/62191.html

†Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 W. Main St., Urbana, IL 61801 (aganguli@uiuc.edu, http://motion.mee.ucsb/∼aganguli).

‡Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064 (jcortes@ucsc.edu, http://www.soe.ucsc.edu/∼jcortes).

§Department of Mechanical Engineering, University of California, Santa Barbara, CA 93106 (bullo@engineering.ucsb.edu, http://motion.mee.ucsb.edu).

problem of locating a guard in a non-self-intersecting polygon so as to maximize the visible area. This problem is still open to the best of our knowledge and is the subject of ongoing research; see [12, 16, 4]. Key differences exist between the computational geometric approach to this problem and our sensor-based feedback approach. In the computational geometric version, the data about the entire polygon is available a priori, the difficulties are of combinatorial nature, and the solutions can be thought of as open loop. In the problem of interest in this paper, we consider the feedback control problem for a mobile robot based upon only local knowledge of the environment and without recollection of past trajectories. The work that is perhaps closest in spirit to our work is the numerical approach proposed in [3].

A second set of relevant references are those on nonsmooth stability analysis. Indeed, our approach to maximizing visible area is to design a nonsmooth gradient flow. To define our proposed algorithm we rely on the notions of generalized gradient [5] and of Filippov solutions for differential inclusions [9]. To study our proposed algorithm we extend recent results on the stability and convergence properties of nonsmooth dynamical systems, as presented in [15, 2]. Finally, the present work has some interesting connections with the study of the behavior of certain territorial animals. A particularly relevant reference is the study of the effect of visibility on space use by red-capped cardinals [8]. These are birds that defend territories along shorelines of rivers and lakes and tend to spend the majority of their time near peninsulas (areas that offer greater amount of visibility of their respective territories) rather than bays.

The contributions of this paper are threefold. First, we prove some basic properties of the area visible from a point observer in a nonconvex polygon Q; see Figure 1.1. Namely, we show that the area of the visibility polygon, as a function of the observer position, is a locally Lipschitz function almost everywhere and that the finite point set of discontinuities consists of the reflex vertices of the polygon Q. Additionally, we compute the generalized gradient of the function and show that the function is not, in general, regular.

Second, we provide a generalized version of certain stability theorems for discontinuous vector fields available in the literature [15, 2]. Specifically, we provide a generalized nonsmooth LaSalle invariance principle for discontinuous vector fields, Filippov solutions, and Lyapunov functions that are locally Lipschitz almost every-

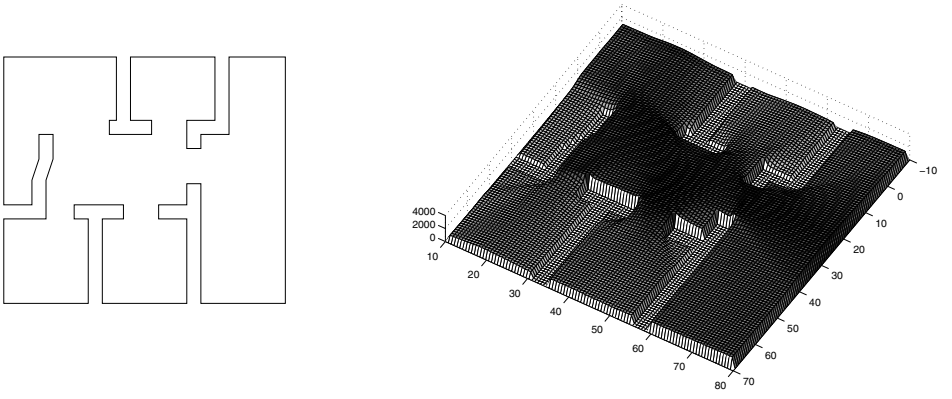


FIG. 1.1. *The image on the left shows a nonconvex polygonal environment shaped like a typical floor plan. The image on the right shows the variation of the visible area in the environment as a function of the position of a point observer.*

where (except for a finite set of discontinuities).

Third and last, we use these novel results to design a nonsmooth gradient algorithm that monotonically increases the area visible to a point observer. To the best of our knowledge, this is the first provably correct algorithm for this version of the next best view problem. We illustrate the performance of our algorithm via simulations for some interesting polygons.

Before proceeding with the technical content, we provide here a detailed comparison between our proposed local feedback method and a two-phase approach, where the guard first explores the entire environment and later finds the approximate location of the maximum. The approximate global maximum could be computed using, for example, the method in [4]. Let us refer to the latter approach as the *explore and optimize method*. Clearly in this latter method, the optimal location of the observer is a global maximum, whereas in the former it is a local maximum. However, our approach does not require any memory for the observer. In the explore and optimize method, the observer needs to remember the environment as it is being explored. As the size of the environment increases, so does the amount of memory and run time required by the observer. Another problem that arises in the exploration of unknown environments in the absence of accurate global positioning is that of *simultaneous location and mapping (SLAM).* To explain briefly, to build a map of the environment accurately, one needs an accurate estimate of the position of the observer. This is not available due to odometry errors and lack of global positioning ability. Therefore, to accurately localize the observer inside the environment, an accurate map of the environment is needed but which is again unavailable due to measurement sensor errors. Various approaches to solving the SLAM problem have been studied, but most of the accurate approaches again are intensive in terms of computation and memory. In contrast, the fact that our approach is local and requires no memory renders it more robust to errors in computation and sensor measurements. Another advantage of being memoryless is that it works for environments that may change quasi-statically, whereas any other method relying on memory is ineffective in this case. Finally, note that out local optimization approach might be used in conjunction with a global search method to find the globally optimum position. For example, after finding the position of a local optimum, the observer could perform a random walk in the environment. While executing the random walk, the observer must compare the magnitude of the visible area with the value at the previously discovered optimum. If the value is greater, find the local optimum using the approach in the paper. This procedure is repeated, and one can expect to find the global optimum after sufficient time has elapsed.

The paper is organized as follows. Section 2 contains the analysis of the smoothness properties and of the generalized gradient of the function of interest. Section 3 contains the novel results on nonsmooth stability analysis. Section 4 presents the nonsmooth gradient algorithm and the properties of the resulting closed-loop system. Finally, the simulations in section 5 illustrate the convergence properties of the algorithm.

2. The area visible from an observer. In this section we study the area of the region visible to a point observer equipped with an omnidirectional camera. We show that the visible area, as a function of the location of the observer, is locally Lipschitz, except at a finite point set. We prove that, for general nonconvex polygons, the function is not regular. We also provide expressions for the generalized gradient of the visible area function wherever it is locally Lipschitz. We have included the notions of locally Lipschitz functions and related concepts whenever they first appear

in the text of the paper.

Let us start by introducing the set of lines on the plane $\mathbb{R}^2$. For $(a,b,c) \in \mathbb{R}^3 \setminus \{(0,0,c) \in \mathbb{R}^3 \mid c \in \mathbb{R}\}$, define the equivalence class $[(a,b,c)]$ by

$$[(a,b,c)] = \{(a',b',c') \in \mathbb{R}^3 \mid (a,b,c) = \lambda(a',b',c'),\ \lambda \in \mathbb{R}\}.$$

The *set of lines on* $\mathbb{R}^2$ is defined as

$$\mathbb{L} = \{[(a,b,c)] \subset \mathbb{R}^3 \mid (a,b,c) \in \mathbb{R}^3,\ a^2 + b^2 \neq 0\}.$$

It is possible to show that $\mathbb{L}$ is a two-dimensional manifold, sometimes referred to as the affine Grassmannian of lines in $\mathbb{R}^2$; see [11].

Next, two useful functions are defined. Let $f_{\mathrm{pl}} : \mathbb{R}^2 \times \mathbb{R}^2 \setminus \{(p,p) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid p \in \mathbb{R}^2\} \to \mathbb{L}$ map two distinct points in $\mathbb{R}^2$ to the line passing through them. Given distinct $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^2$, we have that

$$f_{\mathrm{pl}}((x_1,y_1),(x_2,y_2)) = [(y_2 - y_1, x_1 - x_2, y_1x_2 - x_1y_2)].$$

If $l_1 \parallel l_2$ denotes that two lines $l_1, l_2 \in \mathbb{L}$ are parallel, then let $f_{\mathrm{lp}} : \mathbb{L}^2 \setminus \{(l_1,l_2) \in \mathbb{L}^2 \mid l_1 \parallel l_2\} \to \mathbb{R}^2$ map two nonparallel lines to their unique intersection point. Given two lines $[(a_1,b_1,c_1)]$ and $[(a_2,b_2,c_2)]$, we have that

$$f_{\mathrm{lp}}([(a_1,b_1,c_1)],[(a_2,b_2,c_2)]) = \left(\frac{b_2c_1 - b_1c_2}{a_2b_1 - a_1b_2}, \frac{a_1c_2 - a_2c_1}{a_2b_1 - a_1b_2}\right).$$

Note that the maps f_{pl} and f_{lp} are class C^ω; i.e., they are analytic over their domains.

Now, let us turn our attention to the polygonal environment. Let Q be a non-self-intersecting polygon, possibly nonconvex. A polygon is said to be non-self-intersecting if the only points in the plane belonging to two polygon edges are the polygon vertices. Such a polygon has a well-defined interior and exterior. Note that a non-self-intersecting polygon can contain holes. Let $\mathring{Q}$ and ∂Q denote the interior and the boundary of Q, respectively. Let $\mathrm{Ve}(Q) = (v_1,\dots,v_n)$ be the list of vertices of Q ordered counterclockwise. The *interior angle of a vertex* v of Q is the angle formed inside Q by the two edges of the boundary of Q incident at v. The point $v \in \mathrm{Ve}(Q)$ is a *reflex vertex* if its interior angle is strictly greater than π. Let $\mathrm{Ve}_r(Q)$ be the list of reflex vertices of Q. If S is a finite set, then let $|S|$ denote its cardinality.

A point $q \in Q$ is *visible from* $p \in Q$ if the segment between q and p is contained in Q. The *visibility polygon* $S(p) \subset Q$ from a point $p \in Q$ is the set of points in Q visible from p. It is convenient to think of $p \mapsto S(p)$ as a map from Q to the set of polygons contained in Q. It must be noted that the visibility polygon is not necessarily a non-self-intersecting polygon.

DEFINITION 2.1. *Let v be a reflex vertex of Q, and let $w \in \mathrm{Ve}(Q)$ be visible from v. The* (v,w)-generalized inflection segment $I(v,w)$ *is the set*

$$I(v,w) = \{q \in S(v) \mid q = \lambda v + (1-\lambda)w,\ \lambda \geq 1\}.$$

A reflex vertex v of Q is an anchor of $p \in Q$ *if it is visible from p and if $\{q \in S(v) \mid q = \lambda v + (1-\lambda)p, \lambda > 1\}$ is not empty.*

In other words, a reflex vertex is an anchor of p if it occludes a portion of the environment from p. Figure 2.1 illustrates the various notions defined above. Given a point q and a line l, let $\mathrm{dist}(q,l)$ denote the distance between them.

Note that any generalized inflection segment $I(v,w)$ splits the polygonal environment into two smaller polygons. The vertex w is visible from any point on $I(v,w)$

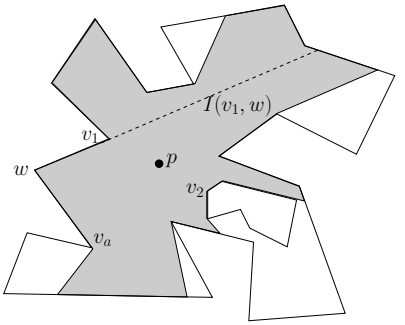


FIG. 2.1. *Reflex vertices v_1 and v_2, a generalized inflection segment $I(v_1, w)$, an anchor v_a of p, and the visibility polygon (shaded region) from p. Note that the polygonal environment has a hole.*

and from the interior of only *one* of the two smaller polygons. Intuitively, it then follows that if p belongs to the interior of a polygon and does not lie on a generalized inflection segment, then in a neighborhood of p the number of vertices of the visibility polygon does not change and their positions vary smoothly as a function of p. This is described formally in the following theorem.

THEOREM 2.2. *Let $\{I_\alpha\}_{\alpha\in\mathcal{A}}$ be the set of generalized inflection segments of Q, and let P be a connected component of $Q \setminus \bigcup_{\alpha\in\mathcal{A}} I_\alpha$. For all $p \in P$, the visibility polygon $S(p)$ is non-self-intersecting and has a constant number of vertices, say $\mathrm{Ve}(S(p)) = \{u_1(p), \ldots, u_k(p)\}$. For all $i \in \{1, \ldots, k\}$, the map $P \ni p \mapsto u_i(p)$ is C^ω and*

$$\mathrm{d}u_i(p) = \begin{cases} 0, & u_i(p) \in \mathrm{Ve}(Q), \\ \dfrac{\mathrm{dist}(v_a, l)}{(\mathrm{dist}(p, l) - \mathrm{dist}(v_a, l))^2 \sqrt{a^2 + b^2}} \begin{bmatrix} -b \\ a \end{bmatrix} \begin{bmatrix} y - y_a \\ x_a - x \end{bmatrix}^T, & u_i(p) = f_{\mathrm{lp}}(f_{\mathrm{pl}}(v_a, p), l), \end{cases}$$

where $v_a = (x_a, y_a)$ is an anchor of p and where $l = [(a, b, c)]$ is a line defined by an edge of Q.

Proof. The first part of the proof is by contradiction. Let $|\mathrm{Ve}(S(p'))| > |\mathrm{Ve}(S(p))|$ for some point $p' \in P$. This means that at least one additional vertex is visible from p' that was occluded by an anchor of p. Two cases may arise. First, when the additional vertex belongs to $\mathrm{Ve}(Q)$, then by our definition, p and p' must lie on opposite sides of a generalized inflection segment. This is a contradiction. Second, if the additional vertex does not belong to $\mathrm{Ve}(Q)$, it must be the projection of a reflex vertex (acting as an anchor). Here again two cases may arise: (1) the reflex vertex is visible from p, and (2) it is not. The first case is possible only if the reflex vertex is visible but does not act as an anchor. So, positive lengths of both sides adjoining the reflex vertex must also be visible from p and at least one of the sides is completely not visible from p', since there is a projection. This means that p and p' lie on opposite sides of a generalized inflection segment generated by the reflex vertex and one of its adjacent vertices. This is a contradiction. The second case is possible if the reflex vertex in question is occluded by another reflex vertex. But this means that p and p' lie on opposite sides of the generalized inflection segment from the reflex vertex to the anchor occluding the reflex vertex; again this is a contradiction. If, on the other hand, $|\mathrm{Ve}(S(p'))| < |\mathrm{Ve}(S(p))|$, then the above arguments hold by interchanging p and p'.

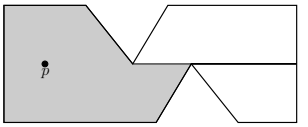


FIG. 2.2. *The visibility polygon of the point represented by p. Note that there exists a ray emanating from p which intersects the environment at three points, and hence the corresponding visibility polygon is self-intersecting.*

Hence, p and p' lie on opposite sides of a generalized inflection segment, which is a contradiction. This completes the proof that $|\operatorname{Ve}(S(p'))|$ is constant for all $p' \in P$.

Let $p \in P$. Since the visibility polygon $S(p)$ is star-shaped and since any ray emanating from p can intersect Q at most at two distinct points, then $S(p)$ is non-self-intersecting. (Indeed, if the ray emanating from p intersects the environment at three or more points inside $S(p)$, then p must belong to a generalized inflection segment. See Figure 2.2.)

Regarding the second statement, it is clear that if $u_i(p)$ is a vertex of Q, then it is independent of p. Instead, if $u_i(p) \notin \operatorname{Ve}(Q)$, then

$$u_i(p) = f_{\mathrm{lp}}(f_{\mathrm{pl}}((x,y),(x_a,y_a)),\ell),$$

where $p = (x,y)$, $v_a = (x_a, y_a)$ is an anchor of p, and ℓ is the line, determined by an edge of Q, that identifies u_i. Now, $p \in P$ implies $p \neq v_a$. It follows that $f_{\mathrm{pl}}(p, v_a)$ is C^ω for all $p \in P$. Also, from the definition of $u_i(p)$, it is clear that $f_{\mathrm{pl}}(p, v_a) \nparallel \ell$. Therefore, for all $p \in P$, $f_{\mathrm{lp}}(f_{\mathrm{pl}}(p, v_a), \ell)$ is C^ω; this implies that $p \mapsto u_i(p)$ is also C^ω. The formula for the derivative can be verified directly. □

Next, the area of a visibility polygon as a function of the observer location is studied; see Figure 1.1. Recall that the area of a non-self-intersecting polygon Q with counterclockwise-ordered vertices $\operatorname{Ve}(Q) = ((x_1,y_1),\dots,(x_n,y_n))$ is given by

$$A(Q) = \frac{1}{2}\sum_{i=1}^{n} x_i(y_{i-1} - y_{i+1}),$$

where $(x_0,y_0) = (x_n,y_n)$ and $(x_{n+1},y_{n+1}) = (x_1,y_1)$. As in the previous theorem, let $\{I_\alpha\}_{\alpha\in\mathcal{A}}$ be the set of generalized inflection segments of Q, and let P be a connected component of $Q \setminus \bigcup_{\alpha\in\mathcal{A}} I_\alpha$. Next, if $p \in P$, the visibility polygon from p has a constant number of vertices, say $k = |\operatorname{Ve}(S(p))|$, is non-self-intersecting, and satisfies $A\circ S(p) = \sum_{i=1}^{k} x_i(y_{i-1} - y_{i+1})$, where $\operatorname{Ve}(S(p)) = (u_1,\dots,u_k)$ are ordered counterclockwise, $u_i(p) = (x_i, y_i)$, $u_0 = u_k$, and $u_{k+1} = u_1$. Therefore, $P \ni p \mapsto A\circ S(p)$ is also C^ω and

$$\mathrm{d}(A\circ S)(p) = \sum_{i=1}^{k} \frac{\partial A(u_1,\dots,u_k)}{\partial u_i}\mathrm{d}u_i(p). \tag{2.1}$$

Remark 2.3. For any $u_i(p) \notin \operatorname{Ve}(Q)$, we have

$$\frac{\partial(A\circ S)}{\partial u_i}\mathrm{d}u_i(p) = \frac{\operatorname{dist}(v_a,l)}{2}\frac{\operatorname{dist}(u_{i+1},l) - \operatorname{dist}(u_{i-1},l)}{(\operatorname{dist}(p,l) - \operatorname{dist}(v_a,l))^2}\begin{bmatrix} y - y_a \\ x_a - x \end{bmatrix}^T. \tag{2.2}$$

Note here that $\frac{\partial(A\circ S)}{\partial u_i}\mathrm{d}u_i(p)$ is perpendicular to $p - v_a$.

To illustrate (2.1) and (2.2), it is convenient to introduce the *versor* operator defined by $\operatorname{vers}(X) = X/\|X\|$ if $X \in \mathbb{R}^2 \setminus \{0\}$ and by $\operatorname{vers}(0) = 0$. We depict the normalized gradient $\operatorname{vers}(\mathrm{d}(A\circ S))$ of the visible area function in Figure 2.3.

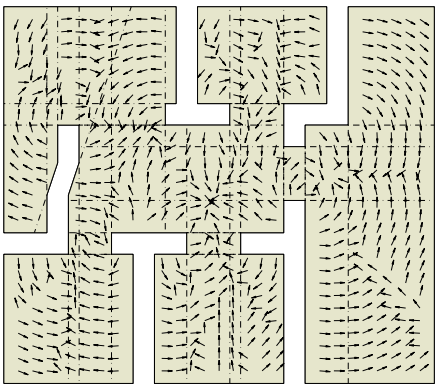


FIG. 2.3. *Normalized gradient of the visible area function over the nonconvex polygon depicted in Figure* 1.1. *The dashed lines represent some of the generalized inflection segments.*

We will now characterize the smoothness properties of the map $A \circ S$ over a polygon Q excluding the set of reflex vertices. Before that, we present the following notion.

DEFINITION 2.4. *A function* $f : \mathbb{R}^N \to \mathbb{R}$ *is said to be* locally Lipschitz near $x \in \mathbb{R}^N$ *if there exist positive constants* L_x *and* ϵ *such that* $|f(y) - f(y')| \leq L_x \|y - y'\|$ *for all* $y, y' \in B(x, \epsilon)$, *where* $B(x, \epsilon)$ *is an* N*-dimensional open ball of radius* ϵ *and centered at* x.

Note that continuously differentiable functions at x are locally Lipschitz near x.

THEOREM 2.5. *The map* $A \circ S$ *restricted to* $Q \setminus \mathrm{Ve}_r(Q)$ *is locally Lipschitz.*

Proof. By Theorem 2.2, it suffices to consider points lying on generalized inflection segments. Let p belong to multiple, say m, generalized inflection segments $\{I_\alpha\}_{\alpha \in \{1,\ldots,m\}}$. Let ϵ be small enough such that no generalized inflection segments intersect $B(p, \epsilon)$ other than $\{I_\alpha\}_{\alpha \in \{1,\ldots,m\}}$. For $\alpha \in \{1, \ldots, m\}$, let v_{k_α} be the anchor determining the generalized inflection segment I_α. Without loss of generality, it can be assumed that no anchor is visible from p other than $v_{k_1}, \ldots, v_{k_m}$. For $\alpha \in \{1, \ldots, m\}$, lines $l_\alpha \perp f_{\mathrm{pl}}(p, v_{k_\alpha})$ can be constructed with the property that $l_\alpha \cap Q = \emptyset$ and the vector $v_{k_\alpha} - p$ points toward l_α. Let h_α be the line parallel to l_α, tangent to $B(\epsilon, p)$, and intersecting the segment from p to v_{k_α}. Let p' and p'' belong to $B(p, \epsilon) \cap (Q \setminus \mathrm{Ve}_r(Q))$. Next, let $q'_\alpha = f_{\mathrm{lp}}(f_{\mathrm{pl}}(p', v_{k_\alpha}), l_\alpha)$ and $q''_\alpha = f_{\mathrm{lp}}(f_{\mathrm{pl}}(p'', v_{k_\alpha}), l_\alpha)$; see Figure 2.4. Let v'_α and v''_α be the intersections between h_α and the lines $f_{\mathrm{pl}}(p', v_{k_\alpha})$ and $f_{\mathrm{pl}}(p'', v_{k_\alpha})$, respectively.

Now, $|A(v_{k_\alpha}, q'_\alpha, q''_\alpha)| = \frac{1}{2}\|q'_\alpha - q''_\alpha\| \operatorname{dist}(v_{k_\alpha}, l_\alpha)$. But from Figure 2.4, it is easy to see that $\|q'_\alpha - q''_\alpha\| = \frac{\operatorname{dist}(v_{k_\alpha}, l_\alpha)}{\|v_{k_\alpha} - p\| - \epsilon} \|v'_\alpha - v''_\alpha\|$ and that $\|v'_\alpha - v''_\alpha\| < \|p' - p''\|$. For $K_\alpha(p) = \frac{1}{2} \frac{\operatorname{dist}(v_{k_\alpha}, l_\alpha)^2}{\|v_{k_\alpha} - p\| - \epsilon}$, the following is true:

$$\begin{aligned} |A(S(p')) - A(S(p''))| &\leq \sum_{\alpha=1}^{m} |A(v_{k_\alpha}, q'_\alpha, q''_\alpha)| \\ &\leq \sum_{\alpha=1}^{m} K_\alpha(p) \|p' - p''\|. \end{aligned}$$

This fact is illustrated by Figure 2.5. This completes the proof that $Q \setminus \mathrm{Ve}_r(Q) \ni p \mapsto A \circ S(p)$ is locally Lipschitz. ∎

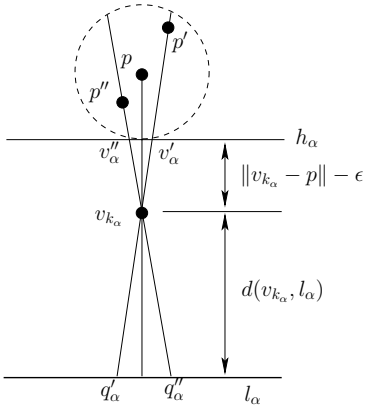


FIG. 2.4. *Definition of the lines l_α, h_α and the points $q'_\alpha, q''_\alpha, v'_\alpha, v''_\alpha$.*

It is clear that the map $A \circ S$ is not differentiable everywhere. However, other notions of derivatives might still be defined for it. The usual *right directional derivative* and the *generalized directional derivative* of any function f at x in the direction of $v \in \mathbb{R}^N$ are defined, respectively, as

$$f'(x;v) = \lim_{t \to 0^+} \frac{f(x+tv) - f(x)}{t}, \qquad f^o(x;v) = \limsup_{\substack{y \to x \\ t \to 0^+}} \frac{f(y+tv) - f(y)}{t}.$$

For a locally Lipschitz function, the limit in the definition of $f'(x;v)$ does not always exist, whereas the limit in $f^o(x;v)$ is always well defined. Also, from Rademacher's theorem [5], we know that locally Lipschitz functions are continuously differentiable almost everywhere (in the sense of Lebesgue measure). If Ω_f denotes the set of points

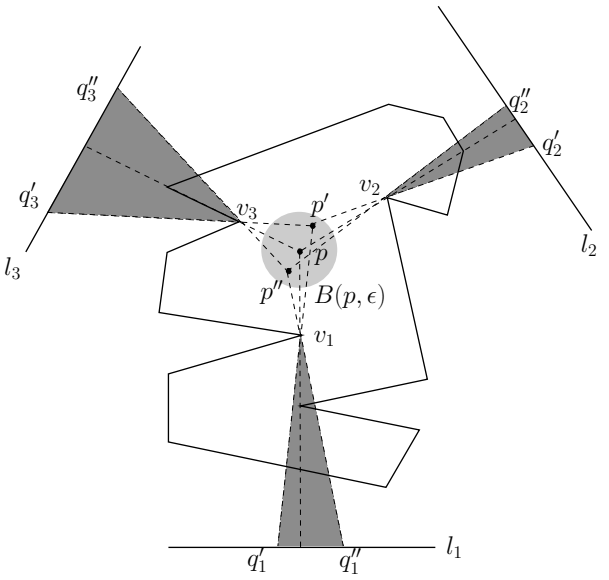


FIG. 2.5. *Upper bounds on the change in area. Here $m = 3$.*

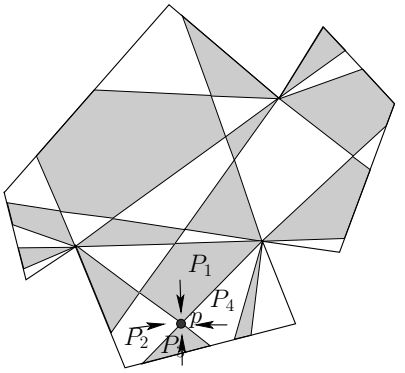


FIG. 2.6. *Partition of Q. The generalized gradient of the area function at p is the convex hull of the gradient of four functions $A_1, \ldots, A_4$ at p.*

in $\mathbb{R}^N$ at which f fails to be differentiable, and S denotes any other set of measure zero, the *generalized gradient* of f is defined by

$$\partial f(x) = \operatorname{co}\left\{ \lim_{i \to +\infty} \mathrm{d}f(x_i) \mid x_i \to x\,,\ x_i \notin S \cup \Omega_f \right\}.$$

Note that this definition coincides with $\mathrm{d}f(x)$ if f is continuously differentiable at x. The generalized gradient and the generalized directional derivative (cf. Proposition 2.1.2 in [5]) are related by $f^o(x; v) = \max\{\zeta \cdot v \mid \zeta \in \partial f(x)\}$ for each $v \in \mathbb{R}^N$.

To obtain the expression for the generalized gradient of $A \circ S$, the polygon Q is partitioned as follows.

LEMMA 2.6. *Let $\{I_\alpha\}_{\alpha \in \mathcal{A}}$ be the set of generalized inflection segments of Q. There exists a unique partition $\{\overline{P}_\beta\}_{\beta \in \mathcal{B}}$ of Q, where P_β is a connected component of $Q \setminus \bigcup_{\alpha \in A} I_\alpha$ and $\overline{P}_\beta$ denotes its closure.*

Figure 2.6 illustrates this partition for the given nonconvex polygon. For $\beta \in \mathcal{B}$, define $A_\beta : \overline{P}_\beta \to \mathbb{R}_+$ by

$$A_\beta(p) = A \circ S(p), \quad \text{for } p \in P_\beta,$$

and by continuity on the boundary of P_β. It turns out that the maps A_β, $\beta \in \mathcal{B}$, are continuously differentiable[1] on $\overline{P}_\beta$. Equation (2.1) gives the value of the gradient for $p \in P_\beta$. However, in general, for $p \in \overline{P}_{\beta_1} \bigcap \cdots \bigcap \overline{P}_{\beta_m} \setminus \mathrm{Ve}_r(Q)$, based on Theorem 2.5 and Lemma 2.6, we can write

$$\partial(A \circ S)(p) = \operatorname{co}\left\{ \mathrm{d}A_{\beta_1}(p), \ldots, \mathrm{d}A_{\beta_m}(p) \right\}. \tag{2.3}$$

This completes our study of the generalized gradient of the locally Lipschitz function $A \circ S$. Apart from differentiability, another smoothness property of $A \circ S$ that we will characterize is the one of regularity. As will be clear later, this will be crucial in constructing a set-valued estimate of the rate at which $A \circ S$ changes as the observer moves. We define regularity first and then show how $A \circ S$ is not regular in many interesting situations.

[1] A function is continuously differentiable on a closed set if (1) it is continuously differentiable on the interior, and (2) the limit of the derivative at a point in the boundary does not depend on the direction from which the point is approached.

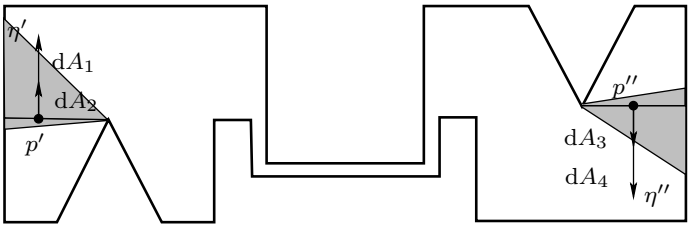


FIG. 2.7. *Example polygon for which $A \circ S$ and $-A \circ S$ restricted to $Q \setminus \mathrm{Ve}_r(Q)$ are not regular. Note here that $\mathrm{d}A_1$ and $\mathrm{d}A_2$ are not perfectly aligned with η'. Also, $\mathrm{d}A_3$ and $\mathrm{d}A_4$ are not perfectly aligned with η''.*

DEFINITION 2.7. *A function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be* regular at $x \in \mathbb{R}^N$ *if for all $v \in \mathbb{R}^N$, $f'(x; v)$ exists and $f^o(x; v) = f'(x; v)$.*

Again, a continuously differentiable function at x is regular at x. Also, a locally Lipschitz function at x which is convex is also regular (cf. Proposition 2.3.6 in [5]).

LEMMA 2.8. *There exists a nonconvex polygon Q such that the maps $A \circ S$ and $-A \circ S$ restricted to $Q \setminus \mathrm{Ve}_r(Q)$ are not regular.*

Proof. We present an example to justify the above statement. In Figure 2.7, $\partial(A \circ S)(p') = \mathrm{co}\{\mathrm{d}A_1, \mathrm{d}A_2\}$, where $\|\mathrm{d}A_1\| \gg \|\mathrm{d}A_2\|$. Take a vector η' perpendicular to the generalized inflection segment to which p' belongs (see Figure 2.7). It is clear that $(A \circ S)'(p; \eta') = \mathrm{d}A_2 \cdot \eta'$. However, $(A \circ S)^0(p'; \eta') = \max\{\zeta \cdot \eta' | \zeta \in \partial(A \circ S)(p')\} = \mathrm{d}A_1 \cdot \eta' > \mathrm{d}A_2 \cdot \eta'$. Again, in Figure 2.7, $\partial(-A \circ S)(p'') = \mathrm{co}\{-\mathrm{d}A_3, -\mathrm{d}A_4\}$, where $\| - \mathrm{d}A_4\| \gg \| - \mathrm{d}A_3\|$. Take a vector η'' perpendicular to the generalized inflection segment to which p'' belongs (see Figure 2.7). It is clear that $-(A \circ S)'(p''; \eta'') = -\mathrm{d}A_4 \cdot \eta''$. However, $(A \circ S)^0(p''; \eta'') = \max\{\zeta \cdot \eta'' | \zeta \in \partial(A \circ S)(p'')\} = -\mathrm{d}A_3 \cdot \eta'' > -\mathrm{d}A_4 \cdot \eta''$. ∎

3. An invariance principle in nonsmooth stability analysis. This section presents results on stability analysis for discontinuous vector fields via nonsmooth Lyapunov functions. The results extend the work in [2] and will be useful in the next control design section; see also [6]. We refer the reader to [9] and to Appendix A for some useful nonsmooth analysis concepts that we have not included in the main body of the paper.

In what follows we shall study differential equations of the form

$$\dot{x}(t) = X(x(t)), \tag{3.1}$$

where $X : \mathbb{R}^N \to \mathbb{R}^N$ is a measurable and essentially locally bounded, possibly discontinuous vector field. We understand the solution of this equation in the Filippov sense following [9]. For each $x \in \mathbb{R}^N$, consider the set

$$K[X](x) = \bigcap_{\delta > 0} \bigcap_{\mu(S)=0} \mathrm{co}\{X(B(x, \delta) \setminus S)\},$$

where μ denotes the usual Lebesgue measure in $\mathbb{R}^N$. Alternatively, one can show [14] that there exists a set S_X of measure zero such that

$$K[X](x) = \mathrm{co}\left\{ \lim_{i \to +\infty} X(x_i) \mid x_i \to x\,,\ x_i \notin S \cup S_X \right\},$$

where S is any set of measure zero. A Filippov solution (see Appendix A) of (3.1) on an interval $[t_0, t_1] \subset \mathbb{R}$ is defined as a solution of the differential inclusion

$$\dot{x} \in K[X](x)\,. \tag{3.2}$$

Since the set-valued map $K[X] : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ is upper semicontinuous with nonempty, compact, convex values and locally bounded (cf. [9]), the existence of Filippov solutions of (3.1) is guaranteed by Lemma A.3. A set M is *weakly invariant* (respectively, *strongly invariant*) for (3.1) if for each $x_0 \in M$, M contains a maximal solution (respectively, all maximal solutions) of (3.1).

We now introduce another useful tool. Given a locally Lipschitz function $f : \mathbb{R}^N \to \mathbb{R}$, the *set-valued Lie derivative of f with respect to X* at x is defined as

$$\widetilde{\mathcal{L}}_X f(x) = \{a \in \mathbb{R} \mid \exists v \in K[X](x) \text{ such that } \zeta \cdot v = a \ \ \forall \zeta \in \partial f(x)\}\ .$$

For each $x \in \mathbb{R}^N$, $\widetilde{\mathcal{L}}_X f(x)$ is a closed and bounded interval in $\mathbb{R}$, possibly empty. If f is continuously differentiable at x, then $\widetilde{\mathcal{L}}_X f(x) = \{\mathrm{d}f \cdot v \mid v \in K[X](x)\}$. If, in addition, X is continuous at x, then $\widetilde{\mathcal{L}}_X f(x)$ corresponds to the singleton $\{\mathcal{L}_X f(x)\}$, the usual Lie derivative of f in the direction of X at x.

We are now ready to state the first result in this section.

LEMMA 3.1. *Let $X : \mathbb{R}^N \to \mathbb{R}^N$ be measurable and essentially locally bounded, and let $f : \mathbb{R}^N \to \mathbb{R}$ be locally Lipschitz. Let $\gamma : [t_0, t_1] \to \mathbb{R}^N$ be a Filippov solution of X such that $f(\gamma(t))$ is regular for almost all $t \in [t_0, t_1]$. Then*

(i) $\frac{\mathrm{d}}{\mathrm{d}t}(f(\gamma(t)))$ *exists for almost all $t \in [t_0, t_1]$, and*

(ii) $\frac{\mathrm{d}}{\mathrm{d}t}(f(\gamma(t))) \in \widetilde{\mathcal{L}}_X f(\gamma(t))$ *for almost all $t \in [t_0, t_1]$.*

Proof. The result is an immediate consequence of Lemma 1 in [2]. □

The following result is a generalization of the classic LaSalle invariance principle for smooth vector fields and smooth Lyapunov functions to the setting of discontinuous vector fields and nonsmooth Lyapunov functions.

THEOREM 3.2 (LaSalle invariance principle). *Let $X : \mathbb{R}^N \to \mathbb{R}^N$ be measurable and essentially locally bounded, and let $S \subset \mathbb{R}^N$ be compact and strongly invariant for X. Let $C \subset S$ consist of a finite number of points, and let $f : S \to \mathbb{R}$ be locally Lipschitz on $S \setminus C$ and bounded from below on S. Assume the following properties hold:*

(A1) *if $x \in S \setminus C$, then either $\max \widetilde{\mathcal{L}}_X f(x) \le 0$ or $\widetilde{\mathcal{L}}_X f(x) = \emptyset$;*

(A2) *if $x \in C$ and if γ is a Filippov solution of X with $\gamma(0) = x$, then $\lim_{t \to 0^-} f(\gamma(t)) \ge \lim_{t \to 0^+} f(\gamma(t))$; and*

(A3) *if $\gamma : \overline{\mathbb{R}}_+ \to S$ is a Filippov solution of X, then $f \circ \gamma$ is regular almost everywhere.*

Define $Z_{X,f} = \{x \in S \setminus C \mid 0 \in \widetilde{\mathcal{L}}_X f(x)\}$, and let M be the largest weakly invariant set contained in $(\overline{Z}_{X,f} \cup C)$. Then the following statements hold:

(i) *if $\gamma : \overline{\mathbb{R}}_+ \to S$ is a Filippov solution of X, then $f \circ \gamma$ is monotonically nonincreasing;*

(ii) *each Filippov solution of X with initial condition in S approaches M as $t \to +\infty$; and*

(iii) *if M consists of a finite number of points, then each Filippov solution of X with initial condition in S converges to a point of M as $t \to +\infty$.*

Proof. Fact (i) is a consequence of Assumptions (A1), (A2), and (A3) and of Lemma 3.1.

In what follows we shall require the following notion. Given a curve $\gamma : \mathbb{R}_+ \to \mathbb{R}^N$, the positive limit set of γ, denoted by $\Omega(\gamma)$, is the set of $y \in \mathbb{R}^N$ for which there

exists a sequence $\{t_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ such that $t_k < t_{k+1}$, for $k \in \mathbb{N}$, $\lim_{k\to+\infty} t_k = +\infty$, and $\lim_{k\to+\infty} \gamma(t_k) = y$. For $x \in S$, let γ_1 be a Filippov solution of X with $\gamma_1(0) = x$, and let $\Omega(\gamma_1)$ be the limit set of γ_1. Under this setting, $\Omega(\gamma_1)$ is nonempty, bounded, connected, and weakly invariant; see [9]. Furthermore, $\Omega(\gamma_1) \subset S$ because S is strongly invariant and closed.

To prove fact (ii), it suffices to show that $\Omega(\gamma_1) \subset \overline{Z}_{X,f} \cup C$. Trivially, $\Omega(\gamma_1) \cap C \subset C$. Let $y \in \Omega(\gamma_1)\backslash C$ so that f is locally Lipschitz at y. There exists a sequence $\{t_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to+\infty} \gamma_1(t_k) = y$. Because $f \circ \gamma_1$ is monotonically nonincreasing and f is bounded from below, $\lim_{t\to+\infty} f(\gamma_1(t))$ exists and is equal to, say, $a \in \mathbb{R}$. Now, by continuity of f, $a = \lim_{k\to+\infty} f \circ \gamma_1(t_k) = f(y)$. This proves that $f(y) = a$ for all $y \in \Omega(\gamma_1) \setminus C$. At this point we distinguish two cases. First, assume that y is an isolated point in $\Omega(\gamma_1)$. Then, clearly, there exists a Filippov solution of X, say γ_2, such that $\gamma_2(t) = y$ for all $t \geq 0$. Hence $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_2(t)) = 0$, and, by Lemma 3.1, $0 \in \widetilde{\mathcal{L}}_X f(\gamma_2(t))$, or in other words $y \in Z_{X,f}$. Second, assume that y is not isolated in $\Omega(\gamma_1)$, and let γ_2 be a Filippov solution of X with $\gamma_2(0) = y$. Since f is continuous at y and $\Omega(\gamma_1)$ contains a finite number of points of discontinuity of f, there exists $\delta > 0$ such that $f(y') = a$ for all $y' \in B(y, \delta) \cap \Omega(\gamma_1)$. Therefore, there exists $t' > 0$ such that $f(\gamma_2(t)) = a$ for all $t \in [0, t']$. Hence, we have $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma_2(t)) = 0$ for all $t \in [0, t']$. It follows from Lemma 3.1 that for all $t \in [0, t']$, we have $0 \in \widetilde{\mathcal{L}}_X f(\gamma_2(t))$, or in other words $\gamma_2(t) \in Z_{X,f}$. By continuity of γ_2 at $t = 0$, we have that $\gamma_2(0) = y \in \overline{Z}_{X,f}$. Since $\Omega(\gamma_1)$ is weakly invariant, we have $\Omega(\gamma_1) \subset M$, and hence γ_2 approaches M.

We now prove fact (iii). If M consists of a finite number of points, and since $\Omega(\gamma_1) \subset M$ is connected, $\Omega(\gamma_1)$ is a point. Hence, by the argument in the preceding paragraph, each Filippov solution of X approaches a point of M. In other words, it converges to a point of M. □

COROLLARY 3.3. *The LaSalle invariance principle is valid under the following relaxed assumption:*

(A3) *if $\gamma : \overline{\mathbb{R}}_+ \to S$ is a Filippov solution of X, then almost everywhere either $f \circ \gamma$ or $-f \circ \gamma$ is regular.*

Proof. The proof is a consequence of the fact that $\frac{\mathrm{d}}{\mathrm{d}t}(f(\gamma(t)))$ exists and belongs to $\widetilde{\mathcal{L}}_X f(\gamma(t))$ if and only if $\frac{\mathrm{d}}{\mathrm{d}t}(-f(\gamma(t)))$ exists and belongs to $\widetilde{\mathcal{L}}_X(-f)(\gamma(t))$. Thus result (ii) of Lemma 3.1 still holds and the proof of the LaSalle invariance principle remains unchanged. □

4. Maximizing the area visible from a mobile observer. In this section we build on the analysis results obtained thus far to design an algorithm that maximizes the area visible to a mobile observer. We aim to reach local maxima of the visible area $A \circ S$ by designing some appropriate form of a gradient flow for the discontinuous function $A \circ S$. We now present an *introductory and incomplete* version of the algorithm: the objective is to steer the mobile observer along a path for which the visible area is guaranteed to be nondecreasing (see Algorithm 1 on next page).

The remainder of this section is dedicated to formalizing this loose description.

4.1. A modified gradient vector field. Before describing the algorithm to maximize the area visible to the mobile observer, we introduce the following useful notions. Given a non-self-intersecting polygon Q with $\mathrm{Ve}(Q) = (v_1, \ldots, v_n)$ and $\epsilon > 0$, define the following quantities:

(i) let the *ϵ-expansion of Q* be $Q^\epsilon = \{p \mid ||p - q|| \leq \epsilon \text{ for some } q \in Q\}$;

(ii) for $i \in \{1, \ldots, n\}$, let P_i^ϵ be the open set delimited by the edge $\overline{v_i v_{i+1}}$, the bisectors of the external angles at v_i and v_{i+1}, and the boundary of Q^ϵ;

ALGORITHM 1

Name:	Increase visible area for Q
Goal:	Maximize the area visible to a mobile observer
Assumption:	Generalized inflection segments of Q do not intersect. Initial position does not belong to a generalized inflection segment.

Let $p(t)$ denote the observer position at time t inside the nonconvex polygon Q. The observer performs the following tasks at each time instant:

compute visibility polygon $S(p(t)) \subset Q$,
if $p(t)$ does not belong to any generalized inflection segment or to the boundary of Q, **then**
move along the versor of the gradient $\mathrm{d}(A \circ S)$
else if $p(t)$ belongs to a generalized inflection segment but not to the boundary of Q, **then**
depending on the generalized gradient $\partial(A \circ S)$, either slide along the segment or leave the segment in an appropriate direction
else if $p(t)$ belongs to the boundary of Q but not to a reflex vertex, **then**
depending on the projection of $\partial(A \circ S)$ along the boundary, either slide along the boundary or move in an appropriate direction toward the interior of Q
else
either follow a direction of ascent of $A \circ S$ or stop
end if

(iii) for ϵ small enough and for any point p in Q^ϵ, let $\mathrm{prj}_Q(p)$ be uniquely equal to $\arg\min\{||p' - p|| \mid p' \in \partial Q\}$; and
(iv) for $p \in \cup_{i\in\{1,\dots,n\}} P_i^\epsilon$, let the *outward normal* $n(\mathrm{prj}_Q(p))$ be the unit vector directed from $\mathrm{prj}_Q(p)$ to p.

We illustrate these notions in Figure 4.1. Note that $\mathrm{prj}_Q(p)$ can never be a reflex vertex. We can now define a vector field on Q^ϵ as follows:

$$X_Q(p) = \begin{cases} \mathrm{vers}(\mathrm{d}(A \circ S)(p)) & \text{if } p \in \mathring{Q} \setminus \{I_\alpha\}_{\alpha\in\mathcal{A}}, \\ -n(\mathrm{prj}_Q(p)) & \text{if } p \in P_i^\epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

(Recall that the versor operator is defined by $\mathrm{vers}(Y) = Y/\|Y\|$ if $Y \in \mathbb{R}^2 \setminus \{0\}$ and by $\mathrm{vers}(0) = 0$.) Note that X_Q is well defined because at $p \in \mathring{Q} \setminus \{I_\alpha\}_{\alpha\in\mathcal{A}}$ the function $A \circ S$ is analytic. Clearly, X_Q is not continuous on Q^ϵ. However, the set of points where it is discontinuous is of measure zero. Almost everywhere in the interior of Q, the vector field X_Q is equal to the normalized gradient of $A \circ S$, as depicted in Figure 2.3.

Remark 4.1. An important observation in this setting is that at all points p where $A \circ S$ is locally Lipschitz, we have $K[\mathrm{d}(A \circ S)](p) = \partial(A \circ S)(p)$. In such a case it is also true that for all $\eta \in \partial(A \circ S)(p)$, there exists at least one $\delta > 0$ such that $\delta\eta \in K[X_Q](p)$ and vice versa.

We now present the differential equation describing the motion of the observer:

$$\dot{p}(t) = X_Q(p(t)). \tag{4.1}$$

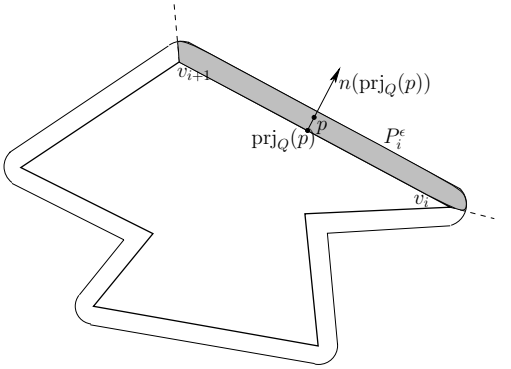


FIG. 4.1. *The ϵ-expansion Q^ϵ of the non-self-intersecting polygon Q, an open set P_i^ϵ, and the corresponding outward normal $n(\mathrm{prj}_Q(p))$.*

A Filippov solution of (4.1) on an interval $[t_0, t_1] \subset \mathbb{R}$ is defined as a solution of the differential inclusion

$$\dot{p}(t) \in K[X_Q](p(t)), \tag{4.2}$$

where $K[X_Q]$ is the usual Filippov differential inclusion associated with X_Q; see Appendix A. Since X_Q is measurable and bounded, the existence of a Filippov solution is guaranteed. We study uniqueness and completeness of Filippov solutions in the following lemma.

LEMMA 4.2. *The following statements hold true:*

(i) *there exists a non-self-intersecting polygon Q for which the corresponding vector field X_Q admits multiple Filippov solutions starting from the same initial condition; and*

(ii) *any non-self-intersecting polygon Q is a strongly invariant set for the corresponding vector field X_Q and, therefore, any Filippov solution is defined over $\overline{\mathbb{R}}_+$.*

Proof. We present an example to justify the statement (i). In Figure 4.2, at the point p_0 on the generalized inflection segment, both directions η_1 and η_2 belong to $\partial(A \circ S)(p_0)$. Three distinct Filippov solutions of (4.1) exist. Two of the solutions start from p_0 along the two directions η_1 and η_2, while the third solution is $p(t) = p_0$ for all $t \geq 0$. Statement (ii) is a consequence of the definition of X_Q on P_i^ϵ for $i \in \{1, \ldots, n\}$. ∎

We now claim that any solution of the differential inclusion (4.2) has the property that the visible area increases monotonically. To prove these desirable properties, we first present the following results in nonsmooth analysis.

4.2. Properties of solutions and convergence analysis. To prove the convergence properties of the solution of (4.2) using the results presented in section 3, we must first define a suitable Lyapunov function. Intuitively, since our objective is

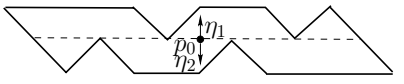


FIG. 4.2. *Three Filippov solutions exist starting from the point p_0.*

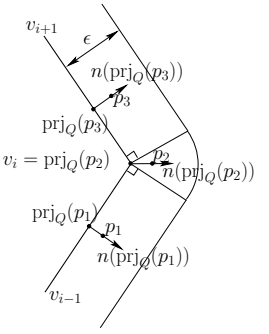


FIG. 4.3. *Extending the function $A \circ S$ to A_Q^ϵ. Note the direction of $n(\mathrm{prj}_Q(p_i))$ at all points p_i.*

to maximize the visible area, our Lyapunov function should be closely related to it. For $\epsilon > 0$, we now define the *extended area function* A_Q^ϵ at all points $p \in Q \bigcup \{\cup_i P_i^\epsilon\}$. The extended function coincides with the original function on the interior and on the boundary of Q and is defined appropriately outside:

$$A_Q^\epsilon(p) = \begin{cases} A \circ S(p) & \text{if } p \in Q, \\ A \circ S(\mathrm{prj}_Q(p)) - ||p - \mathrm{prj}_Q(p)|| & \text{if } p \in \cup_i P_i^\epsilon. \end{cases}$$

For all $p \in \partial Q \setminus \mathrm{Ve}\, Q$, A_Q^ϵ satisfies (see Figure 4.3)

$$A_Q^\epsilon{}'(p; n(\mathrm{prj}_Q(p))) = -1.$$

Remark 4.3. The extended area function A_Q^ϵ is locally Lipschitz on $(Q \setminus \mathrm{Ve}_r(Q)) \bigcup \{\cup_i P_i^\epsilon\}$ and analytic almost everywhere on $Q \bigcup \{\cup_i P_i^\epsilon\}$.

The following theorem characterizes the regularity of the map $p \mapsto -A_Q^\epsilon(p)$ along a Filippov solution of X_Q. This is important to prove that the area of the visibility polygon is nondecreasing along any Filippov solution of X_Q

THEOREM 4.4. *Let $G(Q)$ be the subset of Q where both maps $p \mapsto -A_Q^\epsilon(p)$ and $p \mapsto A_Q^\epsilon(p)$ are not regular. Then any Filippov solution $\gamma : \overline{\mathbb{R}}_+ \to Q$ of X_Q has the property that $\gamma(t) \notin G(Q)$ for almost all $t \in \overline{\mathbb{R}}_+$ unless γ reaches a critical point of $K[X_Q]$.*

Proof. Note that $G(Q)$ is a subset of $\cup_{\alpha \in \mathcal{A}} I_\alpha$. This is a consequence of Theorem 2.2 and the fact that functions are regular at points of differentiability. Given a generalized inflection segment I_α, let l_α be the line extending I_α, and let t_α be one of the two unit tangent vectors to I_α. A Filippov solution γ of X_Q slides along I_α starting from $p_0 \in I_\alpha$ only if $\partial A_Q^\epsilon(p_0)$ contains either t_α or $-t_\alpha$. It then suffices to show that if $\partial A_Q^\epsilon(p_0)$ contains t_α or $-t_\alpha$, then either A_Q^ϵ or $-A_Q^\epsilon$ is regular at p_0. Let us also assume that p_0 does not belong to any other generalized inflection segment. If this were not the case, then either p_0 is a critical point or the Filippov solution does not belong to the point of intersection for almost all $t \in \overline{\mathbb{R}}_+$.

Let l_α divide $\mathbb{R}^2$ into two open half planes H_1 and H_2. There exists $\delta > 0$ such that A_Q^ϵ is analytic on $H_i \cap B(p_0, \delta)$, $i \in \{1, 2\}$; see Figure 4.4. On l_α, we have $(A_Q^\epsilon)_1 = (A_Q^\epsilon)_2$, where $(A_Q^\epsilon)_i$ is the function A_Q^ϵ restricted to H_i. Let $p' \in B(p_0, \delta)$ and, without loss of generality, let $p' \in H_2$. Let n be the normal to I_α at p_0 pointing away from p'. Note that in terms of the notation introduced in section 4.1, $n = -n(\mathrm{prj}_{l_\alpha}(p'))$, where $\mathrm{prj}_{l_\alpha}(p') = \arg\min\{||p' - p|| \mid p \in l_\alpha\}$. Now, $(A_Q^\epsilon)_1$ can be extended to $H_2 \cap B(p_0, \delta)$ by analyticity. Likewise, $(A_Q^\epsilon)_2$ can be extended to $H_1 \cap B(p_0, \delta)$. Since the functions

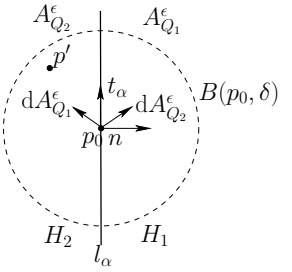


FIG. 4.4. *The point p_0 lies on the generalized inflection segment l. H_1 and H_2 are half planes on either side of l. n and t are normal and parallel to l, respectively. The other arrows indicate the directions of $\mathrm{d}A^\epsilon{}_Q(p)$ on either side of l.*

$(A^\epsilon_Q)_i$, $i \in \{1, 2\}$, are analytic, they can be written as the expansions of their Taylor series:

$$(A^\epsilon_Q)_i(p') = (A^\epsilon_Q)_i(p_0) + \mathrm{d}((A^\epsilon_Q)_i)(p_0) \cdot (p' - p_0) + O(\|p' - p_0\|^2).$$

It follows from the above set of equations that

$$(A^\epsilon_Q)_2(p') - (A^\epsilon_Q)_1(p') = \big(\mathrm{d}((A^\epsilon_Q)_2) - \mathrm{d}((A^\epsilon_Q)_1)\big) \cdot (p' - p_0) + O(\|p' - p_0\|^2).$$

Note that n is the same for all $p' \in H_2$. Now, $p' - p_0 = -c_1 n + c_2 t_\alpha$ such that $c_1 \geq 0$. Also, $\mathrm{d}(A^\epsilon_Q)_1(p_0) \cdot t_\alpha = \mathrm{d}(A^\epsilon_Q)_2(p_0) \cdot t_\alpha$, since $(A^\epsilon_Q)_1(p) = (A^\epsilon_Q)_2(p)$ for $p \in I_\alpha$. Therefore,

$$(A^\epsilon_Q)_2(p') - (A^\epsilon_Q)_1(p') = c_1 \left(\mathrm{d}(A^\epsilon_Q)_1(p_0) \cdot n - \mathrm{d}(A^\epsilon_Q)_2(p_0) \cdot n\right) + O(\|p' - p_0\|^2).$$

Now, either t_α or $-t_\alpha$ belongs to $\partial A^\epsilon_Q(p_0) = \mathrm{co}\{\mathrm{d}(A^\epsilon_Q)_1, \mathrm{d}(A^\epsilon_Q)_2\}$ if and only if the product of $\mathrm{d}(A^\epsilon_Q)_1(p_0) \cdot n$ and $\mathrm{d}(A^\epsilon_Q)_2(p_0) \cdot n$ is less than or equal to zero (see Figure 4.4). If $\mathrm{d}(A^\epsilon_Q)_1(p_0) \cdot n = 0$ and $\mathrm{d}(A^\epsilon_Q)_2(p_0) \cdot n = 0$, then clearly A^ϵ_Q is C^1 at p_0 and hence regular. Otherwise, let us assume, without loss of generality, that $\mathrm{d}(A^\epsilon_Q)_1(p_0) \cdot n - \mathrm{d}(A^\epsilon_Q)_2(p_0) \cdot n < 0$. Therefore, there exists $\eta_2 > 0$ such that $(A^\epsilon_Q)_2(p') - (A^\epsilon_Q)_1(p') \leq 0$ for $p' \in H_2 \cap B(p_0, \eta_2)$. Similarly, there exists $\eta_1 > 0$ such that for $p' \in H_1 \cap B(p_0, \eta_1)$, we have $(A^\epsilon_Q)_1(p') - (A^\epsilon_Q)_2(p') \leq 0$. Thus, there exists a neighborhood around p_0 where $A^\epsilon_Q(p) = \min\{(A^\epsilon_Q)_1(p), (A^\epsilon_Q)_2(p)\}$ or $-A^\epsilon_Q(p) = \max\{-(A^\epsilon_Q)_1(p), -(A^\epsilon_Q)_2(p)\}$. Since $(A^\epsilon_Q)_i$, $i \in \{1, 2\}$, are smooth functions, it follows from Proposition 2.3.12 in [5] that $-A^\epsilon_Q$ is regular at p_0. On the other hand, if we assume that $\mathrm{d}(A^\epsilon_Q)_1(p_0) \cdot n - \mathrm{d}(A^\epsilon_Q)_2(p_0) \cdot n > 0$, then we get that A^ϵ_Q is regular at p_0. □

In the following theorem, the functions A^ϵ_Q and $-A^\epsilon_Q$ are used as candidate Lyapunov functions to show the convergence properties of Filippov solutions of X_Q.

THEOREM 4.5. *Any Filippov solution $\gamma : \overline{\mathbb{R}}_+ \to Q$ of X_Q has the following properties:*

(i) *$t \mapsto A \circ S(\gamma(t))$ is continuous and monotonically nondecreasing; and*

(ii) *γ approaches the set of critical points of $K[X_Q]$.*

Proof. Let us start by showing that if γ is a Filippov solution of X_Q, then $A \circ S \circ \gamma$ is continuous. The reader is referred to Figure 4.5 for an introduction of notations used. Let X^r_Q and X^θ_Q be the components of X_Q parallel and perpendicular to $p - v$, respectively. Similarly, let $\mathrm{d}(A \circ S(p))^r$ and $\mathrm{d}(A \circ S(p))^\theta$ be the components of $\mathrm{d}(A \circ S(p))$ parallel and perpendicular to $p - v$, respectively. Note

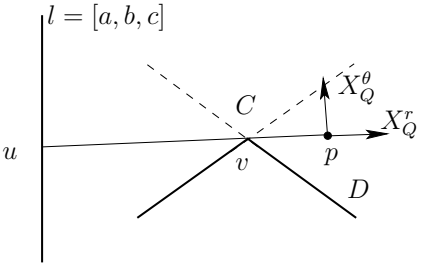


FIG. 4.5. *Illustration of various notions used in Theorem* 4.5. *The dashed lines represent generalized inflection segments generated by the reflex vertex* v *and vertices adjacent to it. These divide the region around* v *that is inside* Q *into three subregions* C, D, *and* E. $u \in \mathrm{Ve}(S(p))$ *lies on the line* l. *The generalized inflection segments including the vertex* v *are assumed to belong to region* C. *Note that* $D \cap C = \emptyset$.

that if $\|\mathrm{d}(A \circ S(p))\| \neq 0$, then $\|X_Q^r\| = \frac{\|\mathrm{d}(A \circ S(p))^r\|}{(\|\mathrm{d}(A \circ S(p))^r\|^2 + \|\mathrm{d}(A \circ S(p))^\theta\|^2)^{1/2}}$ and $\|X_Q^\theta\| = \frac{\|\mathrm{d}(A \circ S(p))^\theta\|}{(\|\mathrm{d}(A \circ S(p))^r\|^2 + \|\mathrm{d}(A \circ S(p))^\theta\|^2)^{1/2}}$. Let $\epsilon > 0$ be such that $\|\mathrm{d}(A \circ S(p))\| \neq 0$ for all $p \in B(v, \epsilon) \cap D$. For now, let us also assume that $\{\cup_{\alpha \in \mathcal{A}} I_\alpha\} \cap B(v, \epsilon) \cap D = \emptyset$. We now claim that in $B(v, \epsilon) \cap D$, $\mathrm{d}(A \circ S(p))^\theta = \Omega(1/\|p - v\|)$ and $\mathrm{d}(A \circ S(p))^r = O(1)$. In other words there exist constants $k_\theta > 0$ and $k_r > 0$ such that $\|\mathrm{d}(A \circ S(p))^\theta\| \geq \frac{k_\theta}{\|p - v\|}$ and $\|\mathrm{d}(A \circ S(p))^r\| \leq k_r$. Notice that $\mathrm{d}(A \circ S(p)) = \mathrm{d}(A \circ S(p))^r + \mathrm{d}(A \circ S(p))^\theta = \sum_i \frac{\partial(A \circ S)}{\partial u_i} \mathrm{d}u_i(p)$. Let $u_1 = u$. From (2.2), it is clear that $\frac{\partial(A \circ S)}{\partial u} \mathrm{d}u(p)$ is perpendicular to $p - v$ and hence contributes only to $\mathrm{d}(A \circ S(p))^\theta$. Also $\| \sum_{i \geq 2} \frac{\partial(A \circ S)}{\partial u_i} \mathrm{d}u_i(p)\|$ is bounded for all $p \in B(v, \epsilon) \cap D$. Therefore, $\mathrm{d}(A \circ S(p))^\theta = \frac{\partial(A \circ S)}{\partial u} \mathrm{d}u(p) + \Omega(1) = \Omega(\| \frac{\partial(A \circ S)}{\partial u} \mathrm{d}u(p)\|)$ and $\mathrm{d}(A \circ S(p))^r = O(1)$. Again from (2.2), we have

$$\left\| \frac{\partial(A \circ S)}{\partial u} \mathrm{d}u(p) \right\| = \frac{\mathrm{dist}(v, l)}{2} \frac{\|p - v\| |\,\mathrm{dist}(u_2, l) - \mathrm{dist}(u_n, l)|}{(\mathrm{dist}(p, l) - \mathrm{dist}(v, l))^2}.$$

Now, $|\,\mathrm{dist}(p, l) - \mathrm{dist}(v, l)| \leq \|p - v\|$. Therefore,

$$\left\| \frac{\partial(A \circ S)}{\partial u} \mathrm{d}u(p) \right\| = \Omega\left(\frac{|\,\mathrm{dist}(u_2, l) - \mathrm{dist}(u_n, l)|}{\|p - v\|} \right).$$

Since p does not lie on a generalized inflection segment, either $u_n = v$ or $u_2 = v$. Without loss of generality, let $u_n = v$. Since u belongs to l, clearly u_2 must belong to l. Hence $|\,\mathrm{dist}(u_2, l) - \mathrm{dist}(u_n, l)| = \mathrm{dist}(v, l)$ and is a constant for all $p \in B(v, \epsilon) \cap D$. Thus

$$\left\| \frac{\partial(A \circ S)}{\partial u} \mathrm{d}u(p) \right\| = \Omega\left(\frac{1}{\|p - v\|} \right).$$

Hence, $\mathrm{d}(A \circ S(p))^\theta = \Omega(\frac{1}{\|p - v\|})$. Therefore $\frac{\|\mathrm{d}(A \circ S(p))^\theta\|}{\|\mathrm{d}(A \circ S(p))^r\|} \geq \frac{k_\theta}{k_r \|p - v\|}$. It follows that

$$\|X_Q^r\| = \frac{1}{(1 + \frac{\|\mathrm{d}(A \circ S(p))^\theta\|^2}{\|\mathrm{d}(A \circ S(p))^r\|^2})^{1/2}} \leq \frac{1}{(1 + \frac{k_\theta^2}{k_r^2 \|p - v\|^2})^{1/2}} = \frac{k_r \|p - v\|}{(k_\theta^2 + k_r^2 \|p - v\|^2)^{1/2}}$$
$$\leq \frac{k_r \|p - v\|}{k_\theta}.$$

Note that a convex combination of finitely many X_Q^r will also admit a similar inequality, and so the assumption that $\{\cup_{\alpha\in\mathcal{A}} I_\alpha\}\cap B(v,\epsilon)\cap D = \emptyset$ is not limiting. Now let $\gamma(t)$ be a solution of X_Q such that $\gamma(0) = v$. Let T be any time such that $\|\gamma(T) - v\| = R$ and for all $t \in [0,T]$, $\gamma(t) \in B(v,\epsilon)\cap D$ and $X_Q^r(\gamma(t))$ is directed away from v. Then, clearly, $R = \int_0^T X_Q^r \mathrm{d}t \le R\frac{k_r}{k_\theta}T$. In other words the time T taken for a trajectory to travel any distance R is greater than $\frac{k_\theta}{k_r}$. This is a contradiction. Therefore, our assumption that for all $t \in [0,T]$, $\gamma(t) \in B(v,\epsilon)\cap D$ is false. So, the trajectory must belong to C for some finite time interval contained in $[0,T]$. We can choose R as small as possible, and this implies that there exists a finite time interval $[0,T_C]$ for which $\gamma(t) \in C$. It follows trivially that $t \mapsto A\circ S(\gamma(t))$ is right continuous at t, where $\gamma(t) = v$. We can prove similarly that $t \mapsto A\circ S(\gamma(t))$ is left continuous at t, where $\gamma(t) = v$ by considering the vector field $-X_Q$ in place of X_Q. This completes the proof that $t \mapsto A\circ S(\gamma(t))$ is continuous.

Next, we show that Assumptions (A1), (A2), and (A3) in Theorem 3.2 hold. Let $p \in Q \setminus \mathrm{Ve}_r(Q)$, and take $a \in \widetilde{\mathcal{L}}_{X_Q}(-A_Q^\epsilon)(p)$. By definition, there exists $k \in K[X_Q](p)$ such that $a = k\cdot\zeta$ for all $\zeta \in -\partial A_Q^\epsilon(p)$. In particular, it is true for $\zeta = -\delta k$, for some $\delta > 0$; see Remark 4.1. Therefore, $a = -\delta\|k\|^2 \le 0$. This proves that either $\max \widetilde{\mathcal{L}}_{X_Q}(-A_Q^\epsilon)(p) \le 0$ or $\widetilde{\mathcal{L}}_{X_Q}(-A_Q^\epsilon)(p) = \emptyset$; i.e., Assumption (A1) is satisfied. Assumption (A2) is a consequence of the continuity of $A\circ S\circ\gamma$. Finally, Assumption (A3) is a consequence of Theorem 4.4. Applying now Theorem 3.2 and its corollary, we conclude that fact (i) holds. Moreover, we also deduce that any Filippov solution of X_Q converges to the largest weakly invariant set M contained in $\overline{Z}_{X_Q,-A_Q^\epsilon} \cup \mathrm{Ve}_r(Q)$.

To prove fact (ii), let us show that $M = \{p \in Q \mid 0 \in K[X_Q](p)\}\cap(\overline{Z}_{X_Q,-A_Q^\epsilon} \cup \mathrm{Ve}_r(Q))$. Based on Theorem 4.4, Theorem 3.2, and Corollary 3.3, it suffices to show that M is contained in $\{p \in Q \mid 0 \in K[X_Q](p)\}$. Let us note that the set $\{p \in Q \mid 0 \in K[X_Q](p)\}$ is weakly invariant and can be established to be closed following the same reasoning as in Proposition 2.1.1 in [6]. Let $x \in Z_{X_Q,-A_Q^\epsilon}$. Then, $0 \in \widetilde{\mathcal{L}}_{X_Q}(-A_Q^\epsilon)(x)$; i.e., there exists $k \in K[X_Q](x)$ such that $\zeta\cdot k = 0$ for all $\zeta \in -\partial A_Q^\epsilon(x)$. But, $k \in K[X_Q](x)$ implies that there exists $\delta > 0$ such that $\delta k \in -\partial A_Q^\epsilon(x)$; see Remark 4.1. Thus, for $\zeta = \delta k$, we get $\delta\|k\|^2 = 0$, that is, $0 \in K[X_Q](x)$. This shows that $Z_{X_Q,-A_Q^\epsilon} \subset \{p \in Q \mid 0 \in K[X_Q](x)\}$. Next, let $x \in \mathrm{Ve}_r(Q)\cap M$. If the set $\{x\}$ is weakly invariant, then by definition $0 \in K[X_Q](x)$. If, on the other hand, x is not isolated in M, then there exists a sequence of points $\{x_m\}_{m\in\mathbb{N}}$ converging to x such that $x_m \in Z_{X_Q,-A_Q^\epsilon}$ or, alternatively, $0 \in K[X_Q](x_m)$. Because $\{p \in Q \mid 0 \in K[X_Q](p)\}$ is closed, it follows that $0 \in K[X_Q](x)$. Thus we proved that any weakly invariant set contained in $Z_{X_Q,-A_Q^\epsilon} \cup \mathrm{Ve}_r(Q)$ is a subset of $\{p \in Q \mid 0 \in K[X_Q](p)\}$. Again, as in Proposition 2.1.1 in [6], it can be shown that $Z_{X_Q,-A_Q^\epsilon}$ is a closed set, and hence the claim that $M \subset \{p \in Q \mid 0 \in K[X_Q](p)\}$ follows. □

Theorem 4.5 implies that the single observer converges to a critical point of $A\circ S$ or to a reflex vertex of Q. However, as shown in Figure 5.2, the presence of noise or computational inaccuracies actually works to drive the observer away from a reflex vertex that is not a local maximum. This will also be true for other critical points that are not local maxima.

5. Simulation results. To conduct experiments, a simulation environment has been developed in MATLAB. There are two levels of the code. The lower level consists of a library containing routines to answer queries such as whether two points in

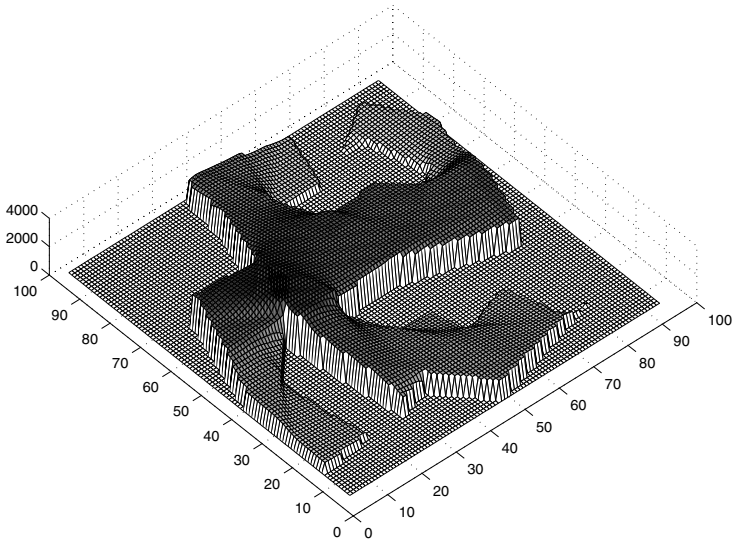


FIG. 5.1. *Example of visible area function over a typical nonconvex polygon.*

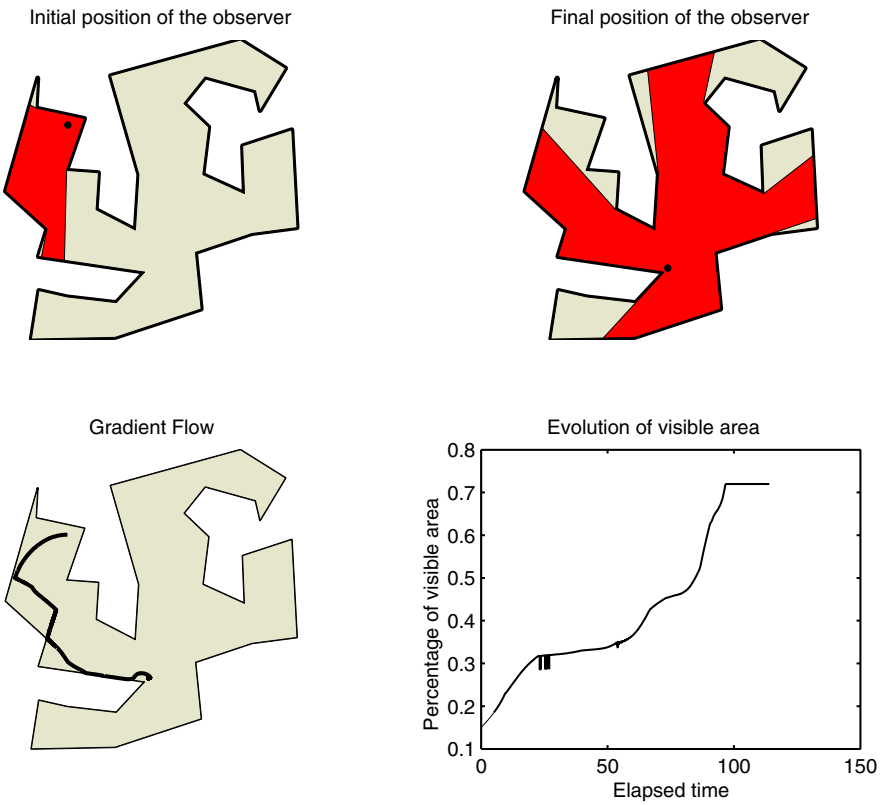


FIG. 5.2. *Simulation results of the gradient algorithm for the nonconvex polygon depicted in Figure* 5.1. *The observer arrives, in finite time, at a local maximum. Note here that the observer visits a reflex vertex at some point in its trajectory but comes out of it due to computational inaccuracies because it is not a local maximum.*

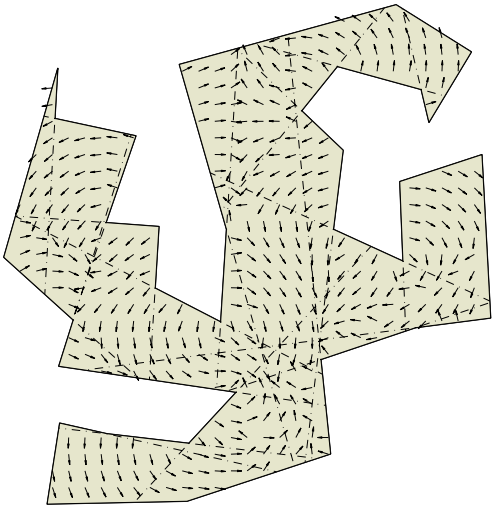


FIG. 5.3. *Example of vector field over the nonconvex polygon in Figure* 5.1.

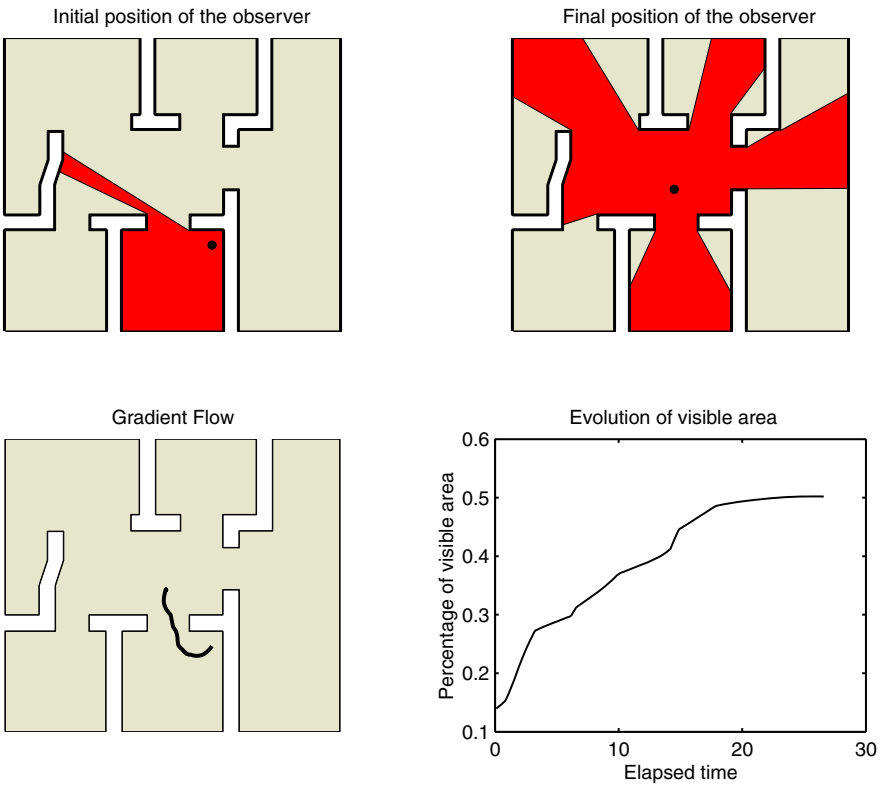


FIG. 5.4. *Simulation results of the gradient algorithm for the nonconvex polygon in Figure* 1.1. *The observer arrives, in finite time, at a local maximum.*

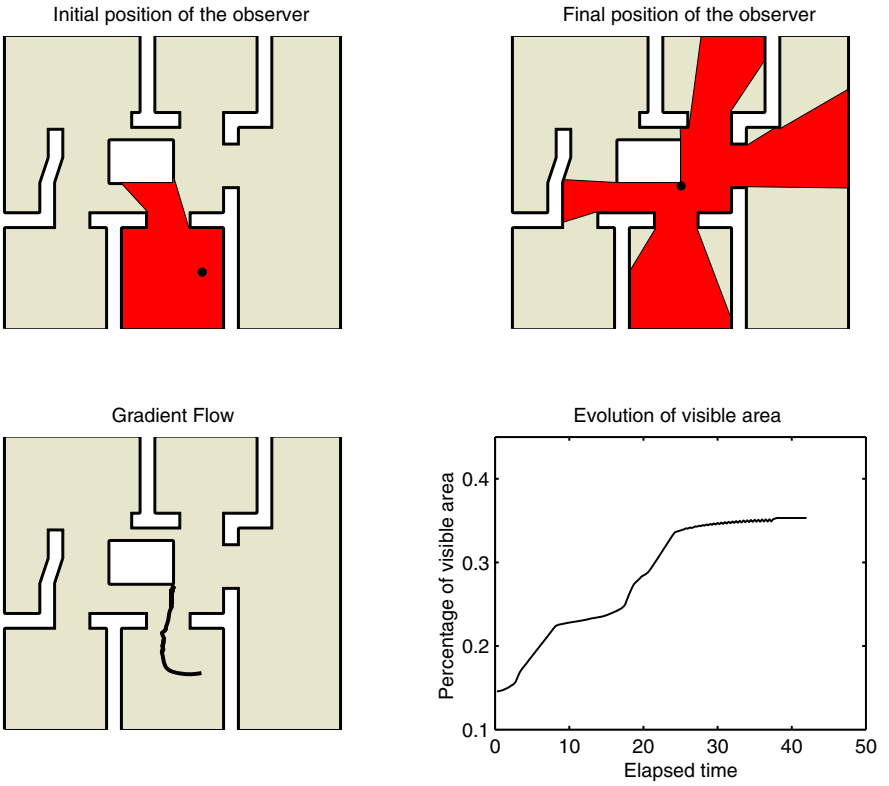


FIG. 5.5. *Simulation results of the gradient algorithm for an observer in a nonconvex environment with a hole. The observer arrives, in finite time, at a reflex vertex.*

a two-dimensional polygonal environment are visible to each other. The higher level utilizes these routines and consists of two major portions. In the first, the vertices of the visibility polygon are obtained by means of an $O(n^2)$ algorithm, where n is the number of vertices of the polygonal environment. These are then sorted in counterclockwise order to compute the visibility polygon. The second consists of the controller which decides the direction and the step size of the observer motion at each time instant. The main task of the controller is the calculation of the generalized gradient of the visible area function which is a natural outcome of (2.1) and (2.3). *Such a framework gives the flexibility to easily implement other visibility-based algorithms for single or multiple observers in a polygonal environment. This can be done by extracting the appropriate information using the low level functions and implementing the desired controller.*

Figure 5.1 illustrates a typical nonconvex polygon; see Figure 5.3 for the phase portrait of the vector field X_Q for the same polygon. Figures 5.2 and 5.4 illustrate the performance of the gradient algorithm in equation (4.2) for the environments in Figures 5.1 and 1.1, respectively. Computational inaccuracies in the implementation of the algorithm to calculate the visibility polygon have been noticed in some configurations; see the plot of the evolution of visible area with time in Figure 5.2. Simulation results for an observer in a polygonal environment containing a hole is shown in Figure 5.5. Our experiments suggest that the observer reaches a local maximum of the visible area in finite time; however, this can be shown not to be true in general.

6. Conclusions. This paper introduces a gradient-based algorithm to optimally locate a mobile observer in a nonconvex environment. We have presented nonsmooth analysis and control design results. The simulation results illustrate that, in the presence of noise, the observer reaches a local maximum of the visible area. In a "highly nonconvex" environment, a single observer may not be able to see a large fraction of the environment. In such a case, a team of observers can be deployed to achieve the same task. We therefore plan to investigate this same visibility objective for teams of observers. Other directions of future research include practical robotic implementation issues as well as other combined mobility and visibility problems.

Appendix A. Nonsmooth analysis and discontinuous vector fields. In this appendix we review some basic facts and standard notation from nonsmooth analysis [5].

Given a locally Lipschitz function $f : \mathbb{R}^N \to \mathbb{R}$, a point $x \in \mathbb{R}^N$ which verifies that $0 \in \partial f(x)$ is called a *critical point of* f. The extrema of Lipschitz functions are characterized by the following result.

PROPOSITION A.1. *Let f be a locally Lipschitz function at $x \in \mathbb{R}^N$. If f attains a local minimum or maximum at x, then $0 \in \partial f(x)$; i.e., x is a critical point.*

Let $\mathrm{Ln} : 2^{\mathbb{R}^N} \to 2^{\mathbb{R}^N}$ be the set-valued map that associates with each closed subset S of $\mathbb{R}^N$ the set of its least-norm elements $\mathrm{Ln}(S)$. For a locally Lipschitz function f, we consider the *generalized gradient vector field* $\mathrm{Ln}(\partial f) : \mathbb{R}^N \to \mathbb{R}^N$ given by $x \mapsto \mathrm{Ln}(\partial f)(x) = \mathrm{Ln}(\partial f(x))$.

THEOREM A.2. *Let f be a locally Lipschitz function at x. Assume that $0 \notin \partial f(x)$. Then, there exists $T > 0$ such that $f(x - t\,\mathrm{Ln}(\partial f)(x)) \leq f(x) - \frac{t}{2}\|\mathrm{Ln}(\partial f)(x)\|^2$, $0 < t < T$. The vector $-\mathrm{Ln}(\partial f)(x)$ is called a* direction of descent.

For differential equations with discontinuous right-hand sides we understand the solutions in terms of differential inclusions following [9]. Let $F : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ be a set-valued map. Consider the differential inclusion

$$\dot{x} \in F(x)\,. \tag{A.1}$$

A solution to this equation on an interval $[t_0, t_1] \subset \mathbb{R}$ is defined as an absolutely continuous function $x : [t_0, t_1] \to \mathbb{R}^N$ such that $\dot{x}(t) \in F(x(t))$ for almost all $t \in [t_0, t_1]$. Given $x_0 \in \mathbb{R}^N$, the existence of at least a solution with initial condition x_0 is guaranteed by the following lemma.

LEMMA A.3. *Let the map F be upper semicontinuous with nonempty, compact, and convex values. Then, given $x_0 \in \mathbb{R}^N$, there exists at least a solution of* (A.1) *with initial condition x_0.*

Acknowledgments. We would like to express our thanks to the referees and the associate editor handling this paper for their thoughtful comments that helped us improve the presentation of the paper.

REFERENCES

[1] P. K. AGARWAL AND M. SHARIR, *Efficient algorithms for geometric optimization*, ACM Comput. Surveys, 30 (1998), pp. 412–458.

[2] A. BACCIOTTI AND F. CERAGIOLI, *Stability and stabilization of discontinuous systems and nonsmooth Lyapunov functions*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 361–376.

[3] L.-T. CHENG AND Y.-H. TSAI, *Visibility optimization using variational approaches*, Commun. Math. Sci., 3 (2005), pp. 425–451.

[4] O. CHEONG, A. EFRAT, AND S. HAR-PELED, *On finding a guard that sees most and a shop that sells most*, in Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, 2004, pp. 1091–1100.

[5] F. H. Clarke, *Optimization and nonsmooth analysis*, Canad. Math. Soc. Ser. Monogr. Adv. Texts, John Wiley & Sons, New York, 1983.

[6] J. Cortés and F. Bullo, *Coordination and geometric optimization via distributed dynamical systems*, SIAM J. Control Optim., 44 (2005), pp. 1543–1574.

[7] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, 2nd ed., Springer-Verlag, New York, 2000.

[8] P. K. Eason and J. A. Stamps, *The effect of visibility on space use by territorial red-capped cardinals*, Behaviour, 138 (2001), pp. 19–30.

[9] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Sides*, Math. Appl. 18, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[10] H. H. González-Baños and J.-C. Latombe, *Navigation strategies for exploring indoor environments*, Internat. J. Robotics Res., 21 (2002), pp. 829–848.

[11] J. E. Goodman and R. Pollack, *Foundations of a theory of convexity on affine Grassmann manifolds*, Mathematika, 42 (1995), pp. 305–328.

[12] D. S. Hochbaum and A. Pathria, *Analysis of the greedy approach in problems of maximum k-coverage*, Naval Res. Logist., 45 (1998), pp. 615–627.

[13] K. Kakusho, T. Kitahashi, K. Kondo, and J.-C. Latombe, *Continuous purposive sensing for 2D map building*, in Proceedings of the IEEE International Conference of Systems, Man and Cybernetics, Vancouver, BC, Canada, 1995, pp. 1472–1477.

[14] B. Paden and S. S. Sastry, *A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators*, IEEE Trans. Circuits and Systems, 34 (1987), pp. 73–82.

[15] D. Shevitz and B. Paden, *Lyapunov stability theory of nonsmooth systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1910–1914.

[16] P. Valtr, *Guarding galleries where no point sees a small area*, Israel J. Math., 104 (1998), pp. 1–16.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1680–1696

STABILITY ROBUSTNESS OF RETARDED LTI SYSTEMS WITH SINGLE DELAY AND EXHAUSTIVE DETERMINATION OF THEIR IMAGINARY SPECTRA*

RIFAT SIPAHI† AND NEJAT OLGAC‡

Abstract. In this paper we consider the stability robustness of the general class of vector LTI (linear time invariant) equations with a single delay, $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{x}(t-\tau)$, $\mathbf{x} \in \mathbf{R}^n$. The robustness is against the uncertain, but constant delay, $\tau \in \mathbf{R}^+$. We first present a set of novel propositions and state that the solution must start from the complete knowledge of imaginary spectra of the system, and the corresponding delays. The propositions claim that such spectra form a set of manageably small number of members, and this number is upper bounded by n^2 regardless of the composition of $\mathbf{A}$ and $\mathbf{B}$ matrices. They also claim that the infinite-dimensional system at hand has an outstanding discipline regarding these imaginary spectra. This discipline invites the recently developed concept called the cluster treatment of characteristic roots (CTCR). The CTCR procedure requires a complete and precise determination of the imaginary spectra of the system. There are many procedures in the literature to achieve this. They are, in fact, some variations of the five main methods of different levels of precision and complexity. There is, however, no study known to the authors for presenting a comparison among these methods. This paper addresses this need. We first offer an overview of each of the five methods and then compare their numerical performances over an example case study.

Key words. linear time-delayed dynamics, quasi polynomial, imaginary spectra detection, robust stability

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/050633238

1. Problem statement and an explicit function for stability.

The stability of linear time invariant retarded time delayed systems (LTI–TDS) has been a very active research topic during the past several decades [1, 2, 3, 4, 5]. Numerous contributions by renowned investigators can be found in the literature on the subject. Although at present the focus of attention in the time delayed systems (TDS) community is directed toward much more complex dynamics (such as parametric uncertainties, robustness, time-varying time delays, nonlinear TDS), the LTI–TDS has an undisputed knowledge base which offers plenty of insight into some realistic problems. Furthermore it is the authors' belief that the LTI–TDS field still remains rich with challenging and unsolved problems. Some existing methods, for instance, present new knowledge, which have not been recognized until recently [4, 6, 7, 26]. Some others suggest variations on the earlier techniques—to overcome some subtle and hidden impracticalities—mainly from a numerical deployment point of view [8].

The general dynamics in question is

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{x}(t-\tau), \tag{1.1}$$

*Received by the editors June 7, 2005; accepted for publication (in revised form) April 18, 2006; published electronically November 16, 2006. This work was partially supported by research funds from the DOE (DE-FG02-04ER25656) and the NSF (CMS-0439980, CMS-0539980, DMI 0522910).
http://www.siam.org/journals/sicon/45-5/63323.html

†Mechanical Engineering Department, University of Connecticut, Storrs, CT 06269-3139; currently at Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115 (rifat@coe.neu.edu).

‡Mechanical Engineering Department, University of Connecticut, Storrs, CT 06269-3139 (olgac@engr.uconn.edu).

where $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{A}$ and $\mathbf{B} \in \mathbf{R}^{n\times n}$ are known matrices with ranks n and p ($\leq n$), respectively, and $\tau \in \mathbf{R}^+$, which is the only free parameter in (1.1). The question is to determine the stability outlook of the system in the semi-infinite τ domain. The characteristic equation of the system is

$$CE(s,\tau) = \det(s\,\mathbf{I} - \mathbf{A} - \mathbf{B}\,e^{-\tau s}) = 0, \tag{1.2}$$

and it contains time delays of a commensurate nature with degree up to p; i.e., there exist exponential terms $e^{-k\tau s}$, $k = 0, 1, \ldots, p$, in (1.2), where $p = \text{rank}(\mathbf{B}) \leq n$. The system is infinite dimensional and as such possesses infinitely many characteristic roots, also known as the spectrum $\sigma(\tau)$ of the system. The question of stability for systems of this class translates into some conditions on τ to guarantee that all the infinitely many characteristic roots lie on the stable left half of the complex plane ($\mathbf{C}^-$). In pursuit of this, we present a remark and two relevant propositions next, partially following [6, 15] which contain the highlights of the CTCR paradigm.

Remark 1. If a delay, τ_0, results in a spectrum, which contains an imaginary root, i.e., $s = \omega i \in \sigma(\tau)$, there exist infinitely many delays

$$\tau_k = \tau_0 \mp \frac{2\pi}{\omega}k, \quad k = 0, 1, 2, \ldots, \tag{1.3}$$

which impart the same $s = \omega i \in \sigma(\tau_k)$. In short, the intersection of these infinitely many spectra contains a pair of imaginary roots,

$$(\mp\omega i) \in \bigcap_{k=0}^{\infty} \sigma(\tau_k). \tag{1.4}$$

Notice that these τ_k values corresponding to a single $s = \omega i$ are separated by $2\pi/\omega$. We name the smallest positive of them the *kernel delay value*, τ_{ker}. τ_{ker} begets all the other delays as per (1.3), and we call them the *offspring delay values*. Briefly, the set of infinitely many delay values

$$\left\{\tau_{ker} + \frac{2\pi}{\omega}k,\ k = 0, 1, 2, \ldots\right\} \text{ all result in } s = \omega i \in \sigma\left(\tau_{ker} + \frac{2\pi}{\omega}k\right). \tag{1.5}$$

PROPOSITION 1. *For the entire $\tau \in \mathbf{R}^+$ domain,*

(i) *there is a manageably small number, say m, of imaginary roots, $s = \omega i$, of* (1.1)*;*

(ii) *this number is upper bounded by n^2, i.e., $m \leq n^2$, regardless of the composition of the constant matrices $\mathbf{A}$ and $\mathbf{B}$.*

We will present the proof of Proposition 1 later in the text. We wish to state here, however, that this proposition brings an extraordinary mathematical confinement to the otherwise vastly distributed delay values creating a spectrum, which contains an imaginary root, i.e., $\{\tau > 0 \mid \sigma(\tau) \cap \mathbf{C}^0 \neq 0\}$, where $\mathbf{C}^0$ represents the imaginary axis.

PROPOSITION 2. *Assume that all m of the imaginary roots $\{\omega\} \in \mathbf{R}^{m+}$ and all corresponding* kernel delay values *$\{\tau_{ker}\} \in \mathbf{R}^{m+}$ are known. The* root tendency*, which is defined as*

$$RT|_k = sgn\left[\text{Re}\left(\left.\frac{\partial s}{\partial \tau}\right|_{\substack{s=\omega i \\ \tau = \tau_{ker} + \frac{2\pi}{\omega}j}}\right)\right], \quad j = 0, 1, 2, \ldots, \tag{1.6}$$

is invariant with respect to the counter j.

Proof of Proposition 2. The characteristic equation of (1.2) can be rewritten as

$$CE(s,\tau)=\sum_{k=0}^{p}a_k(s)e^{-k\tau s}=0, \tag{1.7}$$

where $p=\text{rank}(\mathbf{B})$, $p\le n$, is the degree of commensuracy in the dynamics and $a_k(s)$ are polynomials of degree $n-k$. A variational form of this equation is

$$\sum_{k=0}^{p}([a_k'(s)-a_k(s)k\tau]e^{-k\tau s}ds-a_k(s)kse^{-k\tau s}d\tau)=0, \tag{1.8}$$

where $a_k'(s)=da_k(s)/ds$. The root sensitivities at a point $s=\omega i$ corresponding to the delays given in (1.5) are derived from (1.8) as

$$\left.\frac{ds}{d\tau}\right|_{\substack{s=\omega i\\ \tau=\tau_{ker}+\frac{2\pi}{\omega}j}}=\left.\frac{\sum_{k=0}^{p}a_k(s)kse^{-k\tau s}}{\sum_{k=0}^{p}[a_k'(s)-a_k(s)k\tau]e^{-k\tau s}}\right|_{\substack{s=\omega i\\ \tau=\tau_{ker}+\frac{2\pi}{\omega}j}}. \tag{1.9}$$

It is trivial to show that

$$sgn\left[\text{Re}\left(\left.\frac{ds}{d\tau}\right|_{\substack{s=\omega i\\ \tau=\tau_{ker}+\frac{2\pi}{\omega}j}}\right)\right]=sgn[H(s)|_{s=\omega i}], \tag{1.10}$$

where

$$H(s)=\text{Im}\left(\left.\frac{\sum_{k=0}^{p}a_k'(s)e^{-k\tau s}}{\sum_{k=0}^{p}a_k(s)ke^{-k\tau s}}\right|_{\substack{s=\omega i\\ \tau=\tau_{ker}+\frac{2\pi}{\omega}j}}\right).$$

To obtain (1.10), one can divide the numerator and the denominator of (1.9) by $\sum_{k=0}^{n}a_k(s)kse^{-k\tau s}$ and study the real part of the new expression. Expression (1.10) is obviously independent of the counter j. We call this feature the *root tendency invariance property.* □

Proposition 2 simply implies that the root transition at ωi is either destabilizing ($RT=+1$) or stabilizing ($RT=-1$) for a given τ_{ker} and its offspring $\tau_{ker}+\frac{2\pi}{\omega}j$, $j=0,1,2,\ldots$, which render this particular imaginary root.

Using Propositions 1 and 2, one can create an explicit function of τ for the number of unstable roots (i.e., the number of roots in the set $\sigma(\tau)\cap\mathbf{C}^+$) as follows:

$$NU(\tau)=NU(0)+\sum_{k=1}^{m}\left[\Gamma\left(\frac{\tau-\tau_{ker}}{\Delta\tau_k}\right)\cdot U(\tau,\tau_{ker})\cdot RT_k\right]_k, \tag{1.11}$$

where $NU(0)$ is the number of unstable roots when $\tau=0$, $U(\tau,\tau_{ker})=$ step function in τ with the step taking place at τ_{ker}:

$$U(\tau,\tau_{ker})=\begin{cases}0, & 0<\tau<\tau_{ker},\\ 1 \quad \text{for} & \tau\ge\tau_{ker},\ \omega=0,\\ 2, & \tau\ge\tau_{ker},\ \omega\neq 0.\end{cases} \tag{1.12}$$

$\Gamma(x)=$ ceiling function of x, and Γ returns the smallest integer greater than or equal to x. The $[\,]_k$ notation in (1.11) is simply used to re-emphasize that τ_{ker} needs to be changed by the counter k. The $NU(\tau)$ expression requires only knowledge of the following:

(i) NU(0);
(ii) τ_{ker}, $k = 1, \ldots, m$, the smallest τ values corresponding to each ω (m of them) as per Proposition 1;
(iii) $\Delta\tau_k = 2\pi/\omega_k$, $k = 1, \ldots, m$;
(iv) $RT|_k$, $k = 1, \ldots, m$, the invariant root tendencies, as per Proposition 2.

The τ intervals, where $NU = 0$, render stable behavior for the system. $NU > 0$, on the other hand, would mean instability. Since the *kernel delay set* and the corresponding m crossing roots are all known, expression (1.11) exhaustively defines all the stable regions of the system in the delay space, $\tau \in \mathbf{R}^+$.

It is obvious from (1.11) that the stability robustness declaration is possible only if and only if the m members of the $\{\tau_{ker}\}$ kernel delay set and the corresponding m crossing roots $s = \omega i$ are exhaustively and accurately determined. We devote the rest of the paper to reviewing the five distinctive methodologies to achieve this. We also include their comparisons based on accuracy and efficiency. These methodologies are

(a) the Schur–Cohn method (Hermite matrix formation) [3, 11];
(b) elimination of transcendental terms [5];
(c) the matrix pencil–Kronecker sum method [3, 12];
(d) the Kronecker multiplication/elementary transformation method [13];
(e) Rekasius substitution [14].

2. Brief review of the methodologies. In this section we revisit the five main methodologies mentioned to prepare for the comparative work. Let us take the expanded form of (1.2) and explain each method based on that.

(a) Schur–Cohn criterion as per [3, 11]. The formation starts with rewriting (1.7), multiplying it by $e^{k\tau s}$, $k = 0, 1, \ldots, p-1$. This generates p equations in terms of $e^{k\tau s}$, $k = -p, \ldots, -1, 0, 1, \ldots, p-2, p-1$, which are $2p$ linearly independent terms. Next let us consider the companion equation, $CE(-s, \tau) = 0$, which is also satisfied for $s = \omega i$ due to the fact that the imaginary characteristic root ωi always appears as a complex conjugate pair,

$$\overline{CE}(s, \tau) = CE(-s, \tau) = \sum_{k=0}^{p} a_k(s) e^{-k\tau s} = \sum_{k=0}^{p} \overline{a}_k(s) e^{k\tau s}, \tag{2.1}$$

where $\overline{f}(s) = f(-s)$ is indeed the conjugate operation when $s = \omega i$. We then multiply $\overline{CE}(s, \tau)$ with $e^{-k\tau s}$, $k = 1, 2, \ldots, p$, generating additional p equations in terms of the same $2p$ linearly independent terms $e^{k\tau s}$, $k = -p, -p+1, \ldots, p-1$. Both of these sets of p equations can be combined into a single matrix equation as

(2.2)

$$\begin{bmatrix}
0 & 0 & 0 & \ldots & a_0 & a_1 & a_2 & \ldots & \ldots & a_p \\
0 & 0 & \ldots & a_0 & a_1 & a_2 & \ldots & & a_p & 0 \\
\vdots & & \ddots & \ddots & \vdots & \vdots & & \ddots & & \vdots \\
\vdots & \ddots & \ddots & & \vdots & \vdots & \ddots & & & \vdots \\
a_0 & a_1 & a_2 & & a_{p-1} & a_p & \ldots & \ldots & \ldots & 0 \\
0 & 0 & & 0 & \overline{a}_p & \overline{a}_{p-1} & \overline{a}_{p-2} & \ldots & \ldots & \overline{a}_0 \\
0 & 0 & \ldots & \overline{a}_p & \overline{a}_{p-1} & \overline{a}_{p-2} & \ldots & & \overline{a}_0 & 0 \\
\vdots & & \ddots & \ddots & \vdots & \vdots & \ddots & & \vdots & \vdots \\
\vdots & \ddots & \ddots & & \vdots & \vdots & & & \ddots & \vdots \\
\overline{a}_p & \overline{a}_{p-1} & \ldots & \ldots & \overline{a}_1 & \overline{a}_0 & 0 & \ldots & \ldots & 0
\end{bmatrix}
\begin{bmatrix}
e_{p-1} \\ e_{p-2} \\ \vdots \\ \vdots \\ \vdots \\ e_1 \\ e_0 \\ \vdots \\ \vdots \\ e_{-p}
\end{bmatrix}
= \mathbf{A}_1 \mathbf{E}_1 = \mathbf{0},$$

where e_k represents $e^{k\tau s}$ as shorthand notation to prevent cluttering the equation, and $\mathbf{A}_1$, $\mathbf{E}_1$ are evidently the $(2p \times 2p)$ matrix and the $(2p \times 1)$ vector, respectively. If one rewrites this equation by rearranging the exponential terms, it can be cast in the form

(2.3)
$$\begin{bmatrix} a_0 & 0 & 0 & \dots & 0 & a_p & a_{p-1} & \dots & a_2 & a_1 \\ a_1 & a_0 & 0 & \dots & 0 & 0 & a_p & a_{p-1} & \dots & a_2 \\ \vdots & & \ddots & \ddots & \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \vdots & \vdots & \ddots & & & \vdots \\ a_{p-1} & a_{p-2} & \dots & \dots & a_0 & 0 & 0 & \dots & 0 & a_p \\ \overline{a}_p & 0 & \dots & \dots & 0 & \overline{a}_0 & \overline{a}_1 & \dots & \dots & \overline{a}_{p-1} \\ \overline{a}_{p-1} & \overline{a}_p & \dots & \dots & \dots & 0 & \overline{a}_0 & \dots & \dots & \overline{a}_{p-2} \\ \vdots & & \ddots & \ddots & \vdots & \vdots & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & & \vdots & \vdots & & & \ddots & \vdots \\ \overline{a}_1 & \overline{a}_2 & \dots & \dots & \overline{a}_p & 0 & \dots & \dots & 0 & \overline{a}_0 \end{bmatrix} \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ \vdots \\ e_{p-1} \\ e_{-p} \\ e_{-p+1} \\ \vdots \\ \vdots \\ e_{-1} \end{bmatrix} = \mathbf{A}_2\mathbf{E}_2 = \mathbf{0}.$$

Obviously for a nontrivial solution of $\mathbf{E}_2$, the $\mathbf{A}_2$ matrix must be singular:

$$\det \mathbf{A}_2(s) = 0. \tag{2.4}$$

This matrix, $\mathbf{A}_2$, is known to be the Schur–Cohn matrix. Notice the favorable fragmentation of $\mathbf{A}_2$ into four $p \times p$ segments in the rearranged form as

$$\mathbf{A}_2 = \begin{bmatrix} \Lambda_1 & \Lambda_2 \\ \Lambda_2^H & \Lambda_1^H \end{bmatrix}, \tag{2.5}$$

where Λ^H implies the Hermitian of Λ, and (2.5) presents a compact form adopted by [3, 11].[1] Λ_1 and Λ_2 are evidently matrices from (2.3).

This method suggests that if (1.7) has any imaginary root pair $s = \mp\omega i$, it should also satisfy (2.4). Consequently, the question of finding all the imaginary roots of (1.2) reduces to finding the imaginary roots of (2.4), which is a polynomial of s with degree $2np$. Notice that the original system (1.1) is a retarded time delayed system. Thus the nondelayed term of the characteristic equation (1.7), $a_0(s)$, is an nth degree polynomial of s. So the problem is cast into determining the purely imaginary roots of the $2np$ degree polynomial equation (2.4), which can produce maximum np pairs of imaginary roots. For cases when $p = \operatorname{rank}(\mathbf{B}) = n$, the supremum of this number becomes n^2.

Evaluation of the $\det \mathbf{A}_2(s)$, however, needs a symbolic operation, while $2p$ terms are multiplied and added $2p$ times for expanding the determinant. Each of these $2p$ multiplications would create some round-off errors, eventually resulting in a polynomial of s with erroneous coefficients. This operation ultimately yields poor precision in determining the desired imaginary roots.

An alternative to this symbolic evaluation of a determinant is to determine the eigenvalues of a constant matrix (see Theorem 2.1 of [3]). Namely,

$$\det[\omega \mathbf{I} - \mathbf{P}] \equiv \det(\mathbf{P}_n^{-1}) \det(\mathbf{A}_2)|_{s=\omega i}, \tag{2.6}$$

[1] We wish to make a remark on the formulation error of equations (3.79) and (3.80) in [11], which need to be corrected according to (2.3) above.

where

$$\mathbf{P} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{0} & \mathbf{I} \\ -\mathbf{P}_n^{-1}\mathbf{P}_0 & -\mathbf{P}_n^{-1}\mathbf{P}_1 & \dots & \dots & \dots & -\mathbf{P}_n^{-1}\mathbf{P}_{n-1} \end{bmatrix}, \tag{2.7}$$

$$\mathbf{P}_k = \begin{pmatrix} i^k\mathbf{T}_k & i^k\mathbf{H}_k \\ (-i)^k\mathbf{H}_k^T & (-i)^k\mathbf{T}_k^T \end{pmatrix}, \quad k = 0, 1, 2, \dots, n,$$

$$\mathbf{H}_k = \begin{bmatrix} a_{pk} & a_{p-1,k} & \dots & a_{1k} \\ 0 & \ddots & \ddots & a_{2k} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{pk} \end{bmatrix}, \quad \mathbf{T}_k = \begin{bmatrix} a_{0k} & 0 & \dots & 0 \\ a_{1k} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{p-1,k} & \dots & a_{1k} & a_{0k} \end{bmatrix} \tag{2.8}$$

with $k = 0, 1, \dots, n-1$ and $a_j(s) = \sum_{k=0}^{p} a_{jk}s^k$ which are the terms defined in (1.7). Notice that (2.6) indicates that the imaginary roots of $\det \mathbf{A}_2 = 0$ are identical to the real eigenvalues of $\mathbf{P}$, which is a constant matrix. So the numerical procedure is now converted into a simpler and more precise one, i.e., "real eigenvalue" determination of a constant matrix.

(b) Elimination of transcendental terms (as introduced by [5] and utilized in [16, 17, 18, 19, 20, 21]). This procedure follows a starting premise similar to that in the Schur–Cohn methodology in (a). If $CE(s, \tau)$ of (1.7) has an imaginary root, then, correspondingly, $\overline{CE}(s, \tau)$ of (2.1) should have the same root. Multiplying (2.1) by $e^{-p\tau s}$, we obtain

$$e^{-p\tau s}\overline{CE}(s, \tau) = \sum_{k=0}^{p} \overline{a}_k(s)e^{(k-p)\tau s} = 0. \tag{2.9}$$

One can then eliminate the highest commensuracy term (i.e., $e^{-p\tau s}$) between (1.7) and (2.9), yielding a new equation

$$CE_1(s, \tau) = \sum_{k=0}^{p-1} a_k^{(1)}e^{-k\tau s} = 0, \tag{2.10}$$

which is of commensuracy degree $p - 1$. If one repeats this procedure of eliminating the highest degree commensuracy terms p times successively, one arrives at

$$CE_p(s) = a_0^{(p)}(s) = 0, \tag{2.11}$$

an algebraic characteristic equation in which no transcendentality remains. One can show that $a_0^{(p)}(s)$ is a polynomial of degree $n2^p$, of which purely imaginary roots are in question. Notice that due to the successive substitution of "s" with "$-s$" during the manipulations, the imaginary roots of the original characteristic equation CE are preserved, although the degree of the s terms in polynomials $CE_i(s)$ continuously increases. Ultimately there remains only $n2^p$ finite roots of (2.11) instead of the

infinitely many roots of the original equation (1.7). It is guaranteed that only the imaginary roots of these two equations are identical. Therefore searching for the imaginary roots of (2.11) is a sufficient procedure for the mission. The practical usage of this analytically elegant procedure in the literature is very limited [16, 18, 20, 21] because of the round-off errors it invites during the successive evaluation of $CE_p(s)$, albeit in alphanumeric form.

One can further prove that the $CE_p(s)$ and $\det \mathbf{A}_2(s)$ in (2.4) have a common factor, which is the complete polynomial of $\det \mathbf{A}_2(s)$, because they are obtained based on the same fundamental premise of $s \to -s$ substitution, and they represent the same system of (1.1). When numerically executed, however, this factorization disappears due to successive round-off errors. Ultimately one finds two sets of roots, which may be close to one another but not identical.

Clearly, for $n = p = 1, 2$ the degrees of the polynomials of (2.4) and (2.11) are identical, and they are equal to 2. For these cases, the equations are indeed identical. For $n = p > 2$, which implies the case of full rank $\mathbf{B}$ matrix ($p = n$), $n2^n > 2n^2$, and clearly the procedure in (a) is a much more favorable proposition for determining the purely imaginary roots. Notice that the $n2^n - 2n^2$ excess roots may also contain some false imaginary roots, which should not appear at all. We suppress the proofs of these statements, but we will revisit them for example case studies later.

(c) Matrix pencil–Kronecker sum method introduced in [3, 12]. This procedure departs from (1.2), which is rewritten as

$$\det[s\,\mathbf{I} - (\mathbf{A} + \mathbf{B}z)] = 0, \quad z = e^{-\tau s}. \tag{2.12}$$

Using the argument that if $s = \omega i$ is a root of (2.12), we see that so is $s = -\omega i$ when z is replaced with $1/z$. One can say that the eigenvalues of $\mathbf{A} + \mathbf{B}z$ and $\mathbf{A} + \mathbf{B}z^{-1}$ must be $s = \mp\omega i$, which can also be expressed using the property of the Kronecker sum of matrices (see the appendix for a definition). A commonly known property of this operation is that the eigenvalues of the Kronecker sum of two matrices are equal to the pairwise sum of the individual eigenvalues of the matrices [22, 23]. That is, at least one of the eigenvalues of the following matrix has to be zero:

$$(\mathbf{A} + \mathbf{B}z) \oplus (\mathbf{A} + \mathbf{B}z^{-1}), \tag{2.13}$$

where $\oplus$ represents the Kronecker summation. In other words,

$$\det[(\mathbf{A} + \mathbf{B}z) \oplus (\mathbf{A} + \mathbf{B}z^{-1})] = 0. \tag{2.14}$$

This equation gives rise to a polynomial in z of degree $2n^2$ for $n = p$. One needs to solve the $2n^2$ roots of (2.14) and determine those which have the unity magnitude $|z| = 1$. Only those roots represent $s = \omega i$ eigenvalues as per (2.12). For the roots which satisfy this condition, one next solves the imaginary roots $s = \omega i$ from (2.12). Notice that by substituting z as a complex number in (2.12), one obtains a polynomial with complex but constant coefficients. Therefore, most of the neat features of ordinary polynomials with constant coefficients disappear. For instance, there is no guarantee of the complex conjugate feature of the roots. Therefore to decide whether an imaginary root $s = \omega i + \epsilon$, $\epsilon << 1$, is really an imaginary root, except that it is displaced infinitesimally due to numerical/computational error, is not a trivial task. This particular point alone, from the numerical deployment point of view, suggests a drawback.

Similarly to the methodology explained for (a), one can convert the symbolic determinant evaluation of (2.14) into an equivalent eigenvalue determination of a

constant matrix; see Theorem 3.1 of [3] or [24]. In this new form, a generalized eigenvalue problem appears as

$$\det[z\mathbf{U}-\mathbf{V}] = z^{n^2}\det[(\mathbf{A}+\mathbf{B}z)\oplus(\mathbf{A}+\mathbf{B}z^{-1})] = 0, \tag{2.15}$$

where

$$\mathbf{U} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2 \end{bmatrix}_{2n^2\times 2n^2}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{B}_0 & -\mathbf{B}_1 \end{bmatrix}_{2n^2\times 2n^2},$$

and $\mathbf{B}_0 = \mathbf{I}\otimes\mathbf{B}^T$, $\mathbf{B}_1 = \mathbf{A}\otimes\mathbf{A}^T$, $\mathbf{B}_2 = \mathbf{B}\otimes\mathbf{I}$, all of which have dimensions of $(n^2\times n^2)$. Again the generalized eigenvalue operation is numerically a much more reliable and efficient operation than evaluating the roots of the determinant in (2.14). We now wish to utilize these findings to prove Proposition 1.

Proof of Proposition 1. It is clear that (2.15) results in a polynomial of z with degree $2n^2$. Consequently, there can be at most n^2 pairs of complex conjugate unitary solutions for z, and that is the upper bound of the number of root crossings claimed in Proposition 1. Similar arguments can be made following the discussions in section 2(a) immediately after (2.5). The maximum possible number of imaginary eigenvalues for $\mathbf{A}_2$ in (2.3) is n^2 (when $p = n$). □

(d) Kronecker multiplication/elementary transformation method [13]. Before we proceed with this method, we wish to define a critical elementary vectorization transformation, $\xi : \mathbf{C}^{n\times n} \to \mathbf{C}^{n^2\times 1}$ [22, 23]. The ξ operation converts a matrix $\mathbf{M}^{(n\times n)} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n]^T_{n\times n}$ into $(\xi\mathbf{M})^{(n^2\times 1)} = [\mathbf{m}_1^T, \mathbf{m}_2^T, \ldots, \mathbf{m}_n^T]^T_{1\times n^2}$ and the multiplication of three $n\times n$ matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ into a Kronecker product of dimension $n^2\times n^2$ [22, 23], which is given as

$$\xi(\mathbf{P}_1\mathbf{P}_2\mathbf{P}_3) = (\mathbf{P}_1\otimes\mathbf{P}_3^T)\xi\mathbf{P}_2, \tag{2.16}$$

where $(\bullet)^T$ denotes the transpose of $(\bullet)$. The aim is to form a $\mathbf{P}_1\mathbf{P}_2\mathbf{P}_3$ product, which will then be mapped into the right-hand side of (2.16). This mapping, as explained below, brings convenience to solving the pure imaginary roots of dynamics (1.1).

The procedure departs from (1.1), for which a solution of the form $\mathbf{x}(t) = e^{st}\mathbf{v}$ is suggested, where $(s, \mathbf{v}^{(n\times 1)})$ is an eigenvalue-eigenvector pair (both of which are complex in general). Differentiating this expression and substituting in (1.1), one gets

$$(s\mathbf{I}-\mathbf{A}-\mathbf{B}e^{-\tau s})\mathbf{v} = 0, \tag{2.17}$$

which can be rewritten as

$$(s\mathbf{I}-\mathbf{A})\mathbf{v} = e^{-\tau s}\mathbf{B}\mathbf{v}. \tag{2.18}$$

If $s = \omega i$ is a root of (2.18), then so is its conjugate $s = -\omega i$. We can express this by conjugating the complex equation in (2.18),

$$\mathbf{v}^*(-s\mathbf{I}-\mathbf{A}^T) = e^{\tau s}\mathbf{v}^*\mathbf{B}^T, \tag{2.19}$$

where $(\bullet)^*$ denotes the conjugate transpose of $(\bullet)$. One can now multiply (2.18) and (2.19) side by side to get

$$(s\,\mathbf{I}-\mathbf{A})\;\mathbf{V}\;(s\,\mathbf{I}+\mathbf{A}^T) \;=\; -\mathbf{B}\,\mathbf{V}\,\mathbf{B}^T \quad \text{with } \mathbf{V}^{(n\times n)} \;=\; \mathbf{v}\mathbf{v}^*. \tag{2.20}$$

Equation (2.20) is exactly in the form to be transformed by using ξ as defined in (2.16). It returns

$$\begin{aligned} &\{(s\,\mathbf{I}-\mathbf{A})\otimes(s\,\mathbf{I}+\mathbf{A})+\mathbf{B}\otimes\mathbf{B}\}\,\xi\mathbf{V}=0\\ &\Rightarrow\ \boldsymbol{\lambda}(s)\,\xi\mathbf{V}=0, \end{aligned} \tag{2.21}$$

where $\boldsymbol{\lambda}(s)$ is evidently the matrix operation. For nontrivial solutions of $\xi\mathbf{V}\neq\mathbf{0}$, the only way to satisfy (2.21) is to set

$$\det\boldsymbol{\lambda}(s)=0. \tag{2.22}$$

We can conclude that the desired imaginary roots are determined by solving the roots of a $2n^2$ degree polynomial (2.22). However, one should take notice that this root-finding algorithm for higher dimensions ($2n^2>10$) becomes numerically unreliable due to repeated round-off errors in the determinant expansion procedure unless the operation is performed using a very large number of significant digits. In that case, however, excessive computational cost will appear. In order to circumvent this difficulty using lower precision calculations, one can expand (2.21) using Kronecker product identities as defined in [22, 23]. The outcome of this is a matrix polynomial as follows:

$$\boldsymbol{\lambda}(s)=\mathbf{G}_0\,s^2+\mathbf{G}_1\,s+\mathbf{G}_2, \tag{2.23}$$

where $\mathbf{G}_0=\mathbf{I}\otimes\mathbf{I}$, $\mathbf{G}_1=\mathbf{I}\otimes\mathbf{A}-\mathbf{A}\otimes\mathbf{I}$, $\mathbf{G}_2=\mathbf{B}\otimes\mathbf{B}-\mathbf{A}\otimes\mathbf{A}$. Then this matrix polynomial can be linearized [24] by the fact that $\mathbf{G}_0$ is an invertible matrix. The linearized form of (2.23) is expressed as

$$\mathbf{F}=\begin{bmatrix}\mathbf{0}&\mathbf{T}_0\\ -\mathbf{T}_2&-\mathbf{T}_1\end{bmatrix}\quad\text{with }\mathbf{T}_0=\mathbf{I}\otimes\mathbf{I},\ \mathbf{T}_1=\mathbf{G}_0^{-1}\mathbf{G}_1,\ \mathbf{T}_2=\mathbf{G}_0^{-1}\mathbf{G}_2, \tag{2.24}$$

where the zeros of $\det(\boldsymbol{\lambda}(s))=0$ are the eigenvalues of $\mathbf{F}$, i.e., $\det(\boldsymbol{\lambda}(s))=\det(s\mathbf{I}-\mathbf{F})=0$. The imaginary roots of dynamics (1.1) have to be among the eigenvalues of $\mathbf{F}$, which can be computed by many practical and efficient routines such as the *eig*($\mathbf{F}$) subroutine of MATLAB or *eigenvalues*($\mathbf{F}$) subroutine of Maple. Eventually, one can arrive at a numerically more reliable set of imaginary roots of dynamics (1.1), as demonstrated in the example section.

(e) Rekasius substitution as introduced in [14] and utilized by [6, 8, 25]. This critical procedure is an exact substitution of transcendental terms in (1.7) with

$$e^{-\tau s}=\frac{1-Ts}{1+Ts}\quad\text{when } s=\omega\,i\ \textit{only}, \tag{2.25}$$

where $T\in\Re$ and

$$\tau=\frac{2}{\omega}\left[\tan^{-1}(\omega T)\mp\ell\pi\right],\quad \ell=0,1,\ldots. \tag{2.26}$$

This exact substitution creates a new characteristic equation,

$$CE(s,T)=\sum_{k=0}^{p}a_k(s)\left(\frac{1-Ts}{1+Ts}\right)^k=0. \tag{2.27}$$

Multiplying (2.27) by $(1+T\,s)^p$, one obtains

$$\sum_{k=0}^{p} a_k(s)\,(1+T\,s)^{p-k}(1-T\,s)^k \;=\; 0. \tag{2.28}$$

Considering that $a_k(s)$ are ordinary polynomials, (2.28) is nothing other than a polynomial in s with parameterized coefficients in T. Since the system in (1.1) is of retarded type, the highest degree term of s is n and it is in $a_0(s)$. Equation (2.28), therefore, contains a polynomial of s with degree $n+p$. The problem is to determine *all* $T \in \mathbf{R}$ values, which cause imaginary roots of $s = \omega\, i$ for this equation. This can be done by forming the Routh array of (2.28) and setting the only term in the s^1 row to zero [9, 10, 26]. It can be shown that this polynomial is of degree $n\,p$ in T, of which only the real roots are searched. Once these roots are determined, the corresponding crossing frequencies ($s = \omega\, i$) can be found using the auxiliary equation, which is formed by the s^2 row of the Routh array [9, 26]. Notice that the s^2 row has two terms, which are functions of T. They must agree in sign for those T values to yield imaginary roots. Final results are exhaustive in detecting all the imaginary characteristic roots we set out to solve, and their number is upper bounded by n^2 (again in support of Proposition 1).

We wish to discuss two limiting cases here. The first has to do with the extrema of the T domain, namely $T = \mp\infty$. It is easy to check whether these unbounded T values are of interest to us or not. That is, if $T \to \mp\infty$ and there is a finite imaginary root s, the corresponding $e^{-\tau s} \to -1$ (from (2.25)) leaves $CE\,(s\,,\,\tau\,)$ as a simple polynomial of s. The roots of this polynomial can be found easily. If any one of these roots is purely imaginary, that root becomes part of the solution of interest.

The second limiting case is for the imaginary root(s) appearing at the origin, $s = \omega\, i$ with $\omega = 0$, which makes $e^{-\tau s} = 1$ as per (2.25). Consequently one needs to check if

$$\sum_{k=0}^{p} a_k(0) = 0 \tag{2.29}$$

holds from (1.7). If it does, there should be at least one stationary root at $s = 0$, which remains there for all $\tau \in \mathbf{R}^+$. It is easy to determine the multiplicity of this root for some τ values simply by checking if the successive derivatives of (1.7) with respect to s are also zero for the same τ values. For instance, if $s = 0$ is a double root, the second root may cross over the imaginary axis, altering the stability of the system, as the stationary one remains fixed at the origin.

Both of these limiting cases can appear, regardless of the methods (a)–(e). That is, $e^{-\tau s} \to \pm 1$ can occur as a property of the system at hand.

3. A numerical case study and comparative observations. We now take an example case study to display a comparison among the five methodologies we discussed above. Consider the numerical example in [6] which has $p = \mathrm{rank}(\mathbf{B}) = 3 = n$:

$$\mathbf{A} = \begin{pmatrix} -1 & 13.5 & -1 \\ -3 & -1 & -2 \\ -2 & -1 & -4 \end{pmatrix}, \qquad \mathbf{B} = \begin{bmatrix} -5.9 & 7.1 & -70.3 \\ 2 & -1 & 5 \\ 2 & 0 & 6 \end{bmatrix}. \tag{3.1}$$

The respective characteristic equation is

$$\begin{aligned} CE(s,\ \tau) \ &= \ s^3 + 6\,s^2 + 45.5\,s + 111 \ + \ (0.9\,s^2 - 116.8\,s - 22.1)\,e^{-\tau s} \\ &\quad + (90.9\,s - 185.1)\,e^{-2\tau s} + \ 119.4\,e^{-3\tau s} = 0, \end{aligned} \tag{3.2}$$

in which $a_k(s)$, $k = 0, \ldots, 3$, expressions are readily identified as represented in (1.7). For this system the following exhaustive list of (τ_{ker}, ω) is given in [6]. We cross-check this list, using a sufficiently high number of digits in performing the methodologies (a)–(e), and obtain Table 3.1.

TABLE 3.1
The fundamental delays and the imaginary spectra.

τ_{ker}	ω
7.208	0.8404
0.8725	2.1109
0.1859	2.9123
0.1624	3.0351
0.2219	15.5032

Here the notation $(\tau_{ker,j}, \omega_j)$ implies the minimum positive delay, $\tau_{ker,j}$, defined by (1.3) and the corresponding root at $\omega_j\, i$. In (3.2) we now start utilizing the five methodologies as described earlier. The main point of comparison is to be able to precisely declare the complete Table 3.1. Computational efficiency may also be considered as a basis for comparison, and we will address this point later in the text. In order to conserve space, the numerical results are given in truncated forms at the fourth decimal except where necessary for the arguments.

(a) Schur–Cohn procedure. $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ matrices of (2.5) are readily formed using the terms of $a_0(s), \ldots, a_3(s)$. Notice that the degree of $\det(\mathbf{A}_2)$ of (2.4) is $2np = 18$. Thus the mission is reduced to determining the purely imaginary roots of

$$\begin{aligned} \det(A_2) \ &= \ s^{18} + \ 10^{-8}\ s^{17} + 165.81\ \ s^{16} + 0.3\ 10^{-4}\, s^{15} - 18681.53\, s^{14} \\ &+ 0.022\ s^{13} - 248693.82\ \ s^{12} + \ 4.5\ s^{11} + 35141.36\ \ s^{10} + 330.83\ s^9 \\ &+ 0.41\ 10^{10}\ s^8 + 8671.27\ s^7 + \ 0.94\ \ 10^{11}\ \ s^6 + \ 72932.09\ s^5 + 0.71\ \ 10^{12}\ s^4 \\ &+ 148284.13\ s^3 + 0.19\ \ 10^{13} s^2 + 29249.72\ s\ \ + 0.1\ 10^{13} = 0 \end{aligned} \tag{3.3}$$

which displays two major obstacles as follows:

(i) Equation (3.3) is expected to have only even powers of s. Due to the accumulated numerical error in the symbolic manipulations, one cannot achieve this even with 60-digit precision (for this exercise). The trial-based determination of the significant digits needed is an important hindrance. For comparison purposes, we consider the numerical errors at the level of up to 1/4 of significant digits as acceptable accuracy. For a 20-digit operation, any error of 10^{-15} or less is considered to be zero.

(ii) Evaluation of the imaginary roots of (3.3) results in the form of $s = \omega\, i + \varepsilon$, $\varepsilon << 1$, lending themselves to another question for the number of significant digits needed. For instance, using 20 *digits* of precision in Maple, (3.3) gives the following apparently imaginary roots:

$$\begin{aligned} &0.21\ 10^{-7} \mp 0.8404i, \quad 0.28\ 10^{-6} \mp 2.1109i, \quad -0.55\ 10^{-8} \mp 15.5032i, \\ &-0.13\ 10^{-5} \mp 2.9123i, \quad 0.69\ 10^{-6} \mp 3.0351i. \end{aligned}$$

As can be seen, the numerical errors in real parts (which are supposed to be zero) are at least of the order of 10 magnitudes larger than the computational precision, and they all violate the 1/4-digit rule. This point makes it necessary to adjust the accuracy in calculations.

An alternative but more reliable procedure along this line was described earlier over (2.6)–(2.8). Accordingly, the real eigenvalues of a matrix are evaluated instead of the roots of a polynomial. Using this procedure, one obtains the five crossing frequencies (which correspond to the real eigenvalues of $\mathbf{P}$) as $0.8404 - 0.44\ 10^{-17}\, i$, $2.1109{+}0.66\ 10^{-16}\, i$, $2.9123{+}0.63\ 10^{-15}\, i$, $3.0351{+}0.68\ 10^{-15}\, i$, $15.5032{+}0.72\ 10^{-17}\, i$, still with 20-digit precision. These numbers are all acceptably close to the desired results under the proposed 1/4-digit rule.

(b) Elimination of transcendental terms. Starting from the characteristic equation (3.2) and following the three steps as described in section 2(b), one can sequentially eliminate $e^{-3\,\tau\, s}$, $e^{-2\,\tau\, s}$, and $e^{-\tau\, s}$. Notice that these steps preserve the purely imaginary roots of the characteristic equation. The final form should contain only even powers of s:

$$\begin{aligned} CE_p(s) &= s^{24} + 220.81\ s^{22} - 8823.73\ s^{20} - 0.12\ 10^{7}\ s^{18} - 0.27\ 10^{8}\ s^{16} \\ &+ 0.39\ 10^{10}\ s^{14} + 0.32\ 10^{12}\ s^{12} + 0.89\ 10^{13}\ s^{10} + 0.12\ 10^{15}\ s^{8} \\ &+ 0.81\ 10^{15}\ s^{6} + 0.28\ 10^{16}\ s^{4} + 0.44\ 10^{16}\ s^{2} + 0.2\ 10^{16} = 0. \end{aligned} \tag{3.4}$$

The imaginary roots of the original system have to be among the roots of this polynomial. We again use 20 significant digits of precision in Maple during the symbolic manipulations. The same level of precision is used in the following steps of root finding. (Cautionary note: Even for the case of up to 18 digits of precision in Maple, one can still observe odd powered terms appearing in (3.4).)

Equation (3.4) results in 24 symmetric roots with respect to the origin, of which only the purely imaginary ones are important. They are

$$\begin{aligned} &\mp 5.9977i, \quad \mp 3.9460i, \quad \mp 1.8587i, \quad \mp 0.8404i, \\ &\mp 2.1109i, \quad \mp 2.9123i, \quad \mp 3.0351i, \quad \mp 15.5032i. \end{aligned} \tag{3.5}$$

It is easy to demonstrate that the first three roots on the first row are faulty findings and do not represent true crossings. To see that, one can substitute them into the original characteristic equation (3.2) just to show that they do not satisfy this equation. On the other hand, the remaining five roots do. They also yield $|\, e^{-\tau\, \omega\, i}\, | = 1$ from which one can determine the respective time delays.[2]

One can also observe that the difference between the degrees of (3.3) and (3.4) is 6. In fact the relation between the two polynomials is described as

$$CE_p(s,\ \tau)\ =\ P_1(s)\ \det(\mathbf{A}_2), \tag{3.6}$$

where the degree $(P_1) = 6$, and its roots result in the three false pairs of imaginary roots for the specific numerical example, as mentioned above.

[2]Incidentally, the procedure described in [5, eq. (48)] yields incorrect time delays due to the accumulation of numerical error. Primarily an ill conditioning occurs due to some coefficients in (2.11) being 15–16 orders of magnitude apart from one another.

(c) Matrix pencil–Kronecker sum application. Deploying (2.14) we obtain

$$
\begin{aligned}
&\det\left[(\mathbf{A}+\mathbf{B}\,z)\oplus(\mathbf{A}+\mathbf{B}\,z^{-1})\right] \;=\; (0.17\;10^{7}\;z^{18}-141137.96\;z^{17}\\
&+0.23\;10^{8}\;z^{16}-0.17\;10^{9}\;z^{15}+0.52\;10^{9}\;z^{14}-0.81\;10^{9}\;z^{13}+0.62\;10^{9}\;z^{12}\\
&-0.32\;10^{8}\;z^{11}-0.43\;10^{9}\;z^{10}+0.56\;10^{9}\;z^{9}\;-0.43\;10^{9}\;z^{8}-0.32\;10^{8}\;z^{7}\\
&+0.62\;10^{9}\;z^{6}-0.81\;10^{9}\;z^{5}+0.52\;10^{9}\;z^{4}-0.17\;10^{9}\;z^{3}+0.22\;10^{8}\;z^{2}\\
&-141137.96\;z+0.17\;10^{7})/z^{9}=0.
\end{aligned}
\tag{3.7}
$$

We are seeking the roots of (3.7) with magnitude equal to 1. There is a major difficulty, specifically in deciding the tolerance level of $|\,z\,|=1$. This difficulty disappears, however, when higher precision (above 20 digits) is used. For example, with 20 digits of precision, the errors in the magnitudes of the five roots become precisely zero. They are listed in Table 3.2, together with their corresponding five imaginary roots, which are obtained by replacing $z=e^{-\tau\,\omega\,i}$, $|\,z\,|=1$, in (3.2) and solving for ω.

TABLE 3.2
The tolerances in evaluating z's and the imaginary spectra.

$\lvert 1-z\rvert$	Imaginary spectra
0	$-0.57\;10^{-19}+0.8404i$
0	$0.47\;10^{-20}+2.1109i$
0	$-0.73\;10^{-19}+2.9123i$
0	$0.5\;10^{-19}+3.0352i$
0	$-10^{-21}+15.5032i$

Notice that the cumulative effects of numerical error in imaginary spectra are all acceptable in 1/4-digit sense. However, with up to 13 digits of precision, one observes relatively large accumulated errors (violating the 1/4-digit rule) in the computation of imaginary roots. Therefore if a root of z displays $|\,z\,|-1=\varepsilon$, it is quite difficult to assess whether we should take it as an indicator of crossing or not. Increasing the digits higher than 14 (obviously including 20) this concern disappears for this case study.

When the more reliable procedure is followed, as mentioned in section 2(c), $\mathbf{U}$ and $\mathbf{V}$ matrices are trivially obtained from their definitions. The generalized eigenvalues of the $(\mathbf{U},\mathbf{V})$ pair, which have unitary magnitudes (obtained with 20-digit precision again) are listed in Table 3.3.

TABLE 3.3
The tolerances in evaluating z's and the corresponding imaginary spectra.

$\lvert 1-z\rvert$	Imaginary spectra
$-0.14\;10^{-17}$	$-0.25\;10^{-17}\mp 0.8404i$
$-0.34\;10^{-17}$	$0.27\;10^{-17}\mp 2.1109i$
$0.12\;10^{-16}$	$0.18\;10^{-16}\mp 2.9123i$
$-0.78\;10^{-17}$	$0.31\;10^{-16}\mp 3.0351i$
$0.4\;10^{-18}$	$-0.95\;10^{-18}\mp 15.5032i$

Interestingly, on this implementation one can use 9-digit precision and can still obtain acceptable errors looking at the real parts of the imaginary roots, as shown in Table 3.4. This table indicates that the generalized eigenvalue computation can be

performed with a much smaller number of digits, such as 9, without sacrificing the accuracy of the roots.

TABLE 3.4
The tolerances in evaluating z's and the imaginary spectra.

$\lvert 1-z\rvert$	Imaginary spectra
0	$-0.15\ 10^{-7} \mp 0.8404i$
-10^{-8}	$0.66\ 10^{-8} \mp 2.1109i$
-10^{-8}	$0.4\ 10^{-7} \mp 2.9123i$
$-0.18\ 10^{-7}$	$-0.42\ 10^{-7} \mp 3.0351i$
0	$0.27\ 10^{-7} \mp 15.5032i$

(d) Kronecker multiplication/elementary transformation method. Following the procedure given in section 2(d), one obtains (2.22), again with 20-digit precision, as

$$\begin{aligned}\det\ \boldsymbol{\lambda}(s)\ =\ & s^{18} + 0.3\ 10^{-9}\ s^{17} + 165.81\ s^{16}\ + 0.99\ 10^{-6}\ s^{15} - 18681.53\ s^{14}\\ & + 0.0007\ s^{13} -\ 248693.48\ s^{12}\ + 0.16\ s^{11}\ + 35183.49\ s^{10} +\ 12.97\ s^{9}\\ & + 0.41\ 10^{10}\ s^{8} + 370.2\ s^{7} + 0.94\ 10^{11}\ s^{6} + 3348.51\ s^{5}\ + 0.71\ 10^{12}\ s^{4}\\ & + 7209.8\ s^{3} + 0.19\ 10^{13}\ s^{2} + 1480\ s + 0.1\ 10^{13} = 0.\end{aligned} \tag{3.8}$$

Purely imaginary roots of this equation are the expected crossing frequencies as given in Table 3.5. With this precision level, the accumulated errors in the real parts are relatively large (in the 1/4-digit sense) in magnitude as can be seen in this table. As mentioned in section 2(d), in order to improve the accuracy of the imaginary roots of (1.1), we pursue computing the eigenvalues of matrix $\mathbf{F}$, eig($\mathbf{F}$). For this computation, 10- and 20-digit precision are used separately. The results are given in Table 3.6, with the error terms in the real parts being acceptably small. Thus we conclude that this method can produce the imaginary roots accurately even with 10-digit operations.

TABLE 3.5

Imaginary spectra
$0.1\ 10^{-8} \mp 0.8404i$
$0.13\ 10^{-7} \mp 2.1109i$
$-0.82\ 10^{-7} \mp 2.9123i$
$0.53\ 10^{-7} \mp 3.0351i$
$-0.1\ 10^{-9} \mp 15.5032i$

TABLE 3.6
Imaginary spectra.

10-digit precision	20-digit precision
$0.8847\ 10^{-15} \mp 0.8404i$	$0.88\ 10^{-18} \mp 0.8404i$
$0.111\ 10^{-13} \mp 2.1109i$	$-0.1\ 10^{-18} \mp 2.1109i$
$-0.1856\ 10^{-12} \mp 2.9123i$	$0.19\ 10^{-16} \mp 2.9123i$
$0.172\ 10^{-12} \mp 3.0351i$	$-0.19\ 10^{-16} \mp 3.0352i$
$-0.355\ 10^{-14} \mp 15.5032i$	$0.66\ 10^{-17} \mp 15.5032i$

(e) Rekasius substitution. In this procedure, at least 13 digits of precision are required in order to obtain the crossing frequencies as listed in the beginning of this section. The characteristic equation of (3.2) is rewritten after the Rekasius transformation as

$$\begin{aligned} CE(s,T) \;=\; & T^3 s^6 \;+\; (5.1\,T^3 + 3\,T^2)\,s^5 + (253.2\,T^3 + 17.1\,T^2 + 3\,T)\,s^4 \\ & + (-171.4\,T^3 + 162.4\,T^2 + 18.9\,T + 1)\,s^3 \\ & + (898.4\,T^2 - 71.2\,T + 6.9)\,s^2 \;+ (137.8\,T + 19.6)\,s \;+ 23.2 = 0 \end{aligned} \tag{3.9}$$

for which we need all $T \in \Re$ values yielding the imaginary roots. The numerator of the only term of the s^1 row of the respective Routh array in (3.9) is set to zero for this:

$$\begin{aligned} & T^3\,(0.4004\;10^7\,T^9 \;-\; 0.5418\;10^6\;T^8 \;-\; 0.1060\;10^7\;T^7 \;-\; 0.7869\;10^5\;T^6 \\ & \quad - 0.1501\;10^5\;T^5 \;+\; 0.1216\;10^4\;T^4 \;+\; 0.4011\;10^3\;T^3 \\ & \quad - 0.1025\;10^2\,T^2 \;+\; 0.1197\;T \;-\; 0.1120) \;=\; 0. \end{aligned} \tag{3.10}$$

Notice that we are seeking the real roots of (3.10), and three of them are at $T = 0$. The remaining roots, i.e., 9 $(= n^2)$ of them, have to be solved. Out of these nine roots only five happen to be real and all five satisfy the sign agreement condition in the s^2 row of the Routh array (see the note in section 2(e)). They concur with the results of [6] which uses a slightly different procedure to arrive at these T values. An auxiliary equation is formed by the terms of the s^2 row generating the crossing frequencies, ω. The T versus ω correspondence is shown in (3.11):

$$[T] = \begin{bmatrix} -0.4269 \\ -0.1332 \\ 0.0828 \\ 0.0952 \\ 0.6232 \end{bmatrix} \rightarrow [\omega] = \begin{bmatrix} 15.5032 \\ 0.8404 \\ 3.0351 \\ 2.9123 \\ 2.1109 \end{bmatrix}. \tag{3.11}$$

Note that by using T and ω, one can also find the respective delays τ as per (2.26).

4. Comparative comments and conclusions. The five procedures described above reduce the problem at hand to the following solutions:

(a) Schur–Cohn: Roots of the $2n^2$ degree complex polynomial are solved for purely imaginary roots.
(b) Elimination of transcendental terms: The $n\,2^n$ degree real polynomial is solved for imaginary roots.
(c) Kronecker sum: The $2n^2$ degree polynomial is solved for complex roots with $|\,z\,| = 1$.
(d) Kronecker multiplication: The $2n^2$ degree polynomial is solved for imaginary roots.
(e) Rekasius: The n^2 degree polynomial is solved for real roots ($T \in \mathbf{R}$).

From the perspective of developmental steps, the least involved path appears to be in (e). Nevertheless, utilizing the appropriate matrix operations, methods (a), (c), and (d) prove to be equally potent, producing the desired results. In fact for the example case, method (d) generates the results with the smallest number of significant digits among the five. Method (b) is the one that requires special attention, avoiding the false solutions. As the dimension n increases, method (b) becomes more problematic to process, leaving the other four methods as plausible paths to follow.

We wish to mention two attractive features of method (e): First, the degree of the polynomial in question is considerably smaller than all the other four. Second, we seek the real roots (T) only, not the complex ones. This is a great relief from a numerical perspective, as complex (particularly purely imaginary) roots are very hard to detect when numerical errors creep in. A similar hardship appears when the test of $|z| = 1$ is performed in method (c).

We should also make note of the fact that methods (b) and (e) have a symbolic segment in the deployment, as the remaining three methods are numerical. Upon the request of one of the reviewers, we performed a study of efficiency on all of the methods (i.e., the CPU times consumed) including the symbolic as well as numerical operations. These CPU costs did not show noteworthy variations from one method to another. For instance, the tabulations given in section 3 for all five methods were obtained within 1 sec. on a PC with an Intel Centrino 1.6 MHz processor with 512 MB RAM. It is clear that this efficiency measure is not as important as the accuracy of the findings unless the operation is repeated a large number of times for some parametric studies. But this subject is outside the scope of the present paper.

It is also noteworthy that the authors are engaged in expanding the procedures described here to systems with multiple independent delays. In fact, [10] is a document which utilizes procedure (e) for systems with two independent delays. Several other studies on the deployment of procedure (c) for multiple delay cases are also in progress.

Appendix. Let $\mathbf{A} = [a_{ij}] \in \mathbf{R}^{k \times \ell}$, $\mathbf{B} = [b_{ij}] \in \mathbf{R}^{m \times n}$. Then the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \otimes \mathbf{B}$, is defined as follows:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{n1}\mathbf{B} & \cdots & a_{nn}\mathbf{B} \end{bmatrix} \in \mathbf{R}^{km \times \ell n}. \tag{A.1}$$

If $k = \ell$ and $m = n$, then the Kronecker sum of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{A} \oplus \mathbf{B}$, is defined by

$$\mathbf{A} \oplus \mathbf{B} = \mathbf{A} \otimes \mathbf{I}_m + \mathbf{I}_k \otimes \mathbf{B} \in \mathbf{R}^{km \times km}. \tag{A.2}$$

Acknowledgment. The authors wish to thank James Louisell for valuable contributions in the section related to Kronecker multiplication.

REFERENCES

[1] J. K. Hale and S. M. Verduyn Lunel, *An Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[2] K. L. Cooke and P. van den Driessche, *On zeros of some transcendental equations*, Funkcial. Ekvac., 29 (1986), pp. 77–90.

[3] J. Chen, G. Gu, and C. N. Nett, *A new method for computing delay margins for stability of linear delay systems*, Systems Control Lett., 26 (1995), pp. 107–117.

[4] S.-I. Niculescu, *Delay Effects on Stability*, Springer-Verlag London, Ltd., London, 2001.

[5] K. E. Walton and J. E. Marshall, *Direct method for TDS stability analysis*, Proc. IEE-D, 134 (1987), pp. 101–107.

[6] N. Olgac and R. Sipahi, *An exact method for the stability analysis of time delayed LTI systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 793–797.

[7] K. Gu and S.-I. Niculescu, *Further remarks on additional dynamics in various model transformations of linear delay systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 497–500.

[8] A. Thowsen, *The Routh-Hurwitz method for stability determination of linear differential-difference systems*, Internat. J. Control, 33 (1981), pp. 991–995.

[9] N. Olgac and R. Sipahi, *The cluster treatment of characteristic roots and the neutral type time-delayed systems*, ASME J. Dynam. Systems Measurement Control, 127 (2005), pp. 88–97.

[10] R. Sipahi and N. Olgac, *Complete stability robustness of third-order LTI multiple time-delay systems*, Automatica J. IFAC, 41 (2005), pp. 1413–1422.

[11] S. Barnett, *Polynomials and Linear Control Systems*, Marcel Dekker, New York, 1983.

[12] J.-H. Su, *The asymptotic stability of linear autonomous systems with commensurate time delays*, IEEE Trans. Automat. Control, 40 (1995), pp. 1114–1117.

[13] J. Louisell, *A matrix method for determining the imaginary axis eigenvalues of a delay system*, IEEE Trans. Automat. Control, 46 (2001), pp. 2008–2012.

[14] Z. V. Rekasius, *A stability test for systems with delays*, in Proceedings of the Joint Automatic Control Conference, 1980, Paper No. TP9-A.

[15] R. Sipahi and N. Olgac, *Active vibration suppression with time delayed feedback*, J. Vibration and Acoustics, 125 (2003), pp. 384–388.

[16] I. Tuzcu and N. Ahmadian, *Delay-independent stability of uncertain control systems*, J. Vibration and Acoustics, 124 (2002), pp. 277–283.

[17] D. Filipovic and N. Olgac, *Delayed resonator with speed feedback-design and performance analysis*, Mechatronics, 12 (2002), pp. 393–413.

[18] N. Jalili and N. Olgac, *Multiple delayed resonator vibration absorbers for multi-degree-of-freedom mechanical structures*, J. Sound Vibration, 223 (1999), pp. 567–585.

[19] L. Naimark and E. Zeheb, *All constant gain stabilizing controllers for an interval delay system with uncertain parameters*, Automatica J. IFAC, 33 (1997), pp. 1669–1675.

[20] E. Tissir and A. Hmamed, *Further results on stability of* $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}(t-\tau)$, Automatica J. IFAC, 32 (1996), pp. 1723–1726.

[21] S. Arunsawatwong, *Stability of retarded delay differential systems*, Internat. J. Control, 65 (1996), pp. 347–364.

[22] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 772–781.

[23] L. Qiu and E. J. Davison, *The stability robustness determination of state space models with real unstructured perturbations*, Math. Control Signals Systems, 4 (1991), pp. 247–267.

[24] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, London, 1982.

[25] D. Hertz, E. I. Jury, and E. Zeheb, *Simplified analytic stability test for systems with commensurate time delays*, Proc. IEE-D, 131 (1984), pp. 52–56.

[26] N. Olgac and R. Sipahi, *A practical method for analyzing the stability of neutral type LTI-time delayed systems*, Automatica J. IFAC, 40 (2004), pp. 847–853.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1697–1713

NONPOSITIVE CURVATURE AND PARETO OPTIMAL COORDINATION OF ROBOTS*

ROBERT GHRIST† AND STEVEN M. LAVALLE‡

Abstract. Given a collection of robots sharing a common environment, assume that each possesses a graph (a one-dimensional complex also known as a *roadmap*) approximating its configuration space and, furthermore, that each robot wishes to travel to a goal while optimizing elapsed time. We consider vector-valued (or *Pareto*) optima for collision-free coordination on the product of these roadmaps with collision-type obstacles. Such optima are by no means unique: in fact, continua of Pareto optimal coordinations are possible. We prove a finite bound on the number of optimal coordinations in the physically relevant case where all obstacles are *cylindrical* (i.e., defined by pairwise collisions). The proofs rely crucially on perspectives from geometric group theory and $CAT(0)$ *geometry*. In particular, the finiteness bound depends on the fact that the associated *coordination space* is devoid of positive curvature. We also demonstrate that the finiteness bound holds for systems with moving obstacles following known trajectories.

Key words. $CAT(0)$ *geometry*, Pareto optimality, configuration space, nonpositive curvature

AMS subject classifications. Primary, 58E17; Secondary, 49J27, 49K27

DOI. 10.1137/040609860

1. Introduction.

1.1. Motivation.

In numerous settings, the coordination of multiple robots remains a basic and challenging research issue. Autonomous guided vehicles (AGVs) are used in a wide variety of industrial settings for problems such as material handling, palletizing, paper roll handling, assembly, conveying, and people moving. Typically, AGVs reliably traverse a fixed roadmap of paths in a complicated factory environment. Although the paths avoid collisions with obstacles in the workspace, the efforts of numerous AGVs may have to be coordinated in a way that avoids collisions between AGVs while at the same time maximizing productivity.

If we wish to coordinate the motions of N robots in a common environment, what is an appropriate notion of optimality? Minimizing the average time robots take to reach their goal? Minimizing the time that the last robot takes? Such approaches are common (e.g., [16, 20, 29]) and may be appropriate in some cases; however, it is important to recognize that scalarization of a vector of N criteria occurs in this process. Each robot has its own cost function, e.g., elapsed time. These N criteria are then converted—often in an arbitrary manner—into a single criterion to be optimized.

In this paper, we investigate the optimization problem for multiple robot coordination without scalarizing the vector-valued cost function. This centers on the notion of *Pareto optimality* [24, 26], a concept which is widely used in mathematical economics to model individual consumers striving to optimize distinct economic goals. This brand of optimization is richer in the sense that no choice of scalarization is

*Received by the editors June 10, 2004; accepted for publication (in revised form) June 17, 2006; published electronically November 16, 2006. This research was supported by DARPA HR0011-05-1-0008.

http://www.siam.org/journals/sicon/45-5/60986.html

†Department of Mathematics and Coordinate Science Laboratory, University of Illinois, Urbana, IL 61801 (ghrist@math.uiuc.edu). This author's work was supported by NSF PECASE DMS-0337713.

‡Department of Computer Science, University of Illinois, Urbana, IL 61801 (lavalle@cs.uiuc.edu).

involved. In particular the following hold:

1. Any optimum of *any* scalarization of the vector-valued cost function (which is monotone increasing in the components) is in fact one of the Pareto optima; see Lemma 2.6.
2. For settings in which priorities change over time, knowledge of the Pareto optima allows for easy adaptation. This is particularly relevant in automation settings where priority changes arise from, say, changing needs for parts delivery or output.
3. By filtering out all of the motion plans that are not worth considering and presenting the user with a small set of the best alternatives, additional criteria, such as priority or the amount of sacrifice one robot makes, can be applied to automatically select a particular motion plan.

Given the desire to filter the space of all possible coordination schemes to a small set of best cases independent of biases on the robots, we are certainly most interested in the cases where this collection of optima is *finite*. A finiteness criterion is the principal result of this paper.

1.2. Contributions. In section 2 we will phrase the coordination problem as an optimal path-planning problem on a *roadmap coordination space* $\mathcal{X}$: a product of finite metric graphs $\{\Gamma_i\}_1^N$ with obstacle sets removed. The optimization criteria are distinct for each robot. Thus, the appropriate category of optimal paths is that of the so-called *Pareto optimal* solutions. Roughly speaking, a path between fixed endpoints is *locally Pareto optimal* if and only if deforming it to any path which decreases one of the goal times increases some other goal time. A path is *globally Pareto optimal* if and only if any other path which decreases one goal time increases some other. Paths are *Pareto equivalent* if and only if they are connected by a 1-parameter family of paths that preserve all goal times; this equivalence generates Pareto optimal *classes*. (See section 2.2 for precise definitions.)

Depending on the coordination space, nonequivalent Pareto optimal classes may be isolated or may appear in uncountable continuous families. Examples appear in section 3. Most desirable for classification and cataloging purposes are those coordination spaces which admit a finite number of globally Pareto optimal classes.

The principal optimization result of this paper is that for coordination spaces which possess "cylindrical" obstacles—those defined by pairwise collisions between robots—there is a complete classification of locally Pareto optimal classes and a finite bound on the global Pareto optima.

Main Theorem. *Let $\mathcal{X}$ be a cylindrical coordination space. The the following hold:*

1. *Fixing endpoints, the locally Pareto optimal path classes are in bijective correspondence with homotopy classes of paths.*
2. *There is a finite bound on the number of globally Pareto optimal path classes.*

The first statement implies that the set of Pareto optimal paths within a fixed homotopy classes of paths is connected in this space of paths. The second statement means that only a finite number of these local optima are in fact globally Pareto optimal. Proofs appear in section 5.

Furthermore, we extend the finiteness bound to a class of systems involving moving obstacles; see section 7. Remarkably, our techniques easily adapt to problems for which obstacles of varying shape, speed, and trajectory can interfere with the robot motion.

The proof of the Main Theorem uses in an essential manner certain techniques

from geometric group theory, namely, the geometry of $CAT(0)$ *spaces* (see, e.g., [8]) and certain *normal forms* for groups acting on $CAT(0)$ cubical complexes (as in [21]). These tools are applicable to the proof of the Main Theorem thanks to the following lemma.

MAIN LEMMA. *Let $\mathcal{X}$ be a cylindrical coordination space. Then $\mathcal{X}$ is locally $CAT(0)$.*

This lemma is a broad generalization of the result by Abrams [1] that configuration spaces of graphs are locally $CAT(0)$ in the product metric.

Though it falls outside the scope of this paper, the $CAT(0)$ techniques we develop here are directly applicable to the problem of *computing* normal forms of Pareto optimal paths. See [14] for an exact algorithm for computing Pareto optimal paths in a cylindrical coordination space.

1.3. History. The problem of coordinating robots along fixed roadmaps can be considered as a special case of general motion planning for multiple robots. Previous approaches to multiple-robot motion planning are often categorized as *centralized* or *decoupled*. A centralized approach typically constructs a path in a composite configuration space, which is derived from the Cartesian product of the configuration spaces of the individual robots (see, e.g., [5, 6, 27]). A decoupled approach typically generates paths for each robot independently and then considers the interactions between the robots (see, e.g., [9, 12, 22]). The results of this paper are for decoupled coordination.

The approach in [12] prioritizes the robots and defines a sequence of planning problems for which each problem involves moving one robot while those with higher priority are considered as predictable, moving obstacles. This involves the construction of two-dimensional path-time space [17] over which the velocity of the robot is tuned to avoid collisions with the moving obstacles.

In [3, 7, 11, 25, 22, 28] robot paths are independently determined, and a coordination diagram is used to plan a collision-free trajectory along the paths. The approaches in [3, 25] additionally consider dynamics. In [18, 30], an independent roadmap is computed for each robot, and coordination occurs on the Cartesian product of the roadmap path domains. The suitability of one approach over the other is usually determined by the trade-off between computational complexity associated with a given problem and the amount of completeness that is lost. In some applications, such as the coordination of automated guided vehicles, the roadmap might represent all allowable mobility for each robot.

2. Definitions.

2.1. Coordination spaces. This paper concerns the coordination of N robots, each having a roadmap Γ_i (a graph within the configuration space of the ith robot) precomputed independently of the other robots. All graphs are assumed finite and possessed of an intrinsic metric (inherited from the configuration space in which it resides).

DEFINITION 2.1. *A* roadmap coordination space *of graphs $\{\Gamma_i\}_1^N$ is any space of the form*

$$\mathcal{X} = \left(\prod_{i=1}^{N} \Gamma_i \right) - \mathcal{O}, \tag{2.1}$$

where $\mathcal{O}$ denotes an (open) obstacle set.

Assumption 2.2. For the remainder of the paper, all coordination spaces are assumed to have obstacle sets which are sufficiently nice. Namely, the boundary

of the obstacle set, $\partial\mathcal{O}$, is piecewise real-analytic and *collared*: there exists some embedding of $\partial\mathcal{O} \times [0,1] \hookrightarrow \mathcal{X}$ which takes $\partial\mathcal{O} \times \{0\} \to \partial\mathcal{O}$ via the identity map.

This assumption rules out obstacle sets $\mathcal{O}$ for which the system "locks up" in a singular configuration. The piecewise smoothness condition for $\partial\mathcal{O}$ is imposed to avoid discussions of *chattering* and nonrectifiable paths.

Example 2.3. A special class of a roadmap coordination space arises when all of the graphs Γ_i are identical and the obstacle set is an open neighborhood of the collision set $\{x_i = x_j : \text{for some } x_i \in \Gamma_i\,;\, x_j \in \Gamma_j\,;\, i \neq j\}$. In this case, one can consider the workspace to be the graph itself, and the roadmap coordination space is precisely the configuration space of N labeled points on the graph.

In practice, a large class of coordination spaces arise by assigning as illegal states $\mathcal{O}$ those configurations for which there is a collision between robots in the common workspace. For such a system, the obstacle set $\mathcal{O}$ has a very particular "cylindrical" structure since all collisions are determined by pairwise data.

DEFINITION 2.4. *A coordination space $\mathcal{X}$ is said to be* cylindrical *if $\mathcal{O}$ is of the form*

$$\mathcal{O} = \bigcup_{i<j} \left\{ (x_k)_1^N \in \prod_{k=1}^{N} \Gamma_i \,:\, (x_i, x_j) \in \Delta_{i,j} \right\} \tag{2.2}$$

for some (open) sets $\Delta_{i,j} \subset \Gamma_i \times \Gamma_j$, where $1 \leq i < j \leq N$.

That is to say, a cylindrical coordination space is one for which illegal states are determined by pairwise configurations. If two robots have collided, it makes no difference what the positions or configurations of the remaining robot are—this state still counts as an illegal "collision" state.

Configuration spaces of graphs are examples of cylindrical coordination spaces, since the obstacle set $\mathcal{O}$ is the union of (neighborhoods of) sets of the form $\{x_i = x_j : i \neq j\}$. Noncylindrical coordination spaces arise when the legality of a potential collision between two robots is dependent upon the configuration of a third, e.g., hallways in the workspace through which two but not three robots can squeeze.

2.2. Pareto optimality. Pareto optimization refers to optimization of vector-valued functions [23]. In the context of robotics applications, Pareto optimization arises when distinct robots possess distinct goals and cost functions for evaluating performance. Each robot wishes to optimize its cost function independently of the others.

Mathematically, this is characterized as follows. Given a parameterized path $\gamma : [0,T] \to \mathcal{X}$ in a coordination space, each robot executes the projected path $\gamma_i = \text{PROJ}_i \circ \gamma$, where PROJ_i denotes projection onto the ith factor. Given cost functions $\{\tau_i\}$, $i = 1, \ldots, N$, the cost vector for γ is the vector $\tau(\gamma) = (\tau_i(\gamma_i))_i^N$. The case in which τ_i measures *elapsed time* from start to goal is an important and characteristic example, though more general cost functions are allowed.

Assumption 2.5. We restrict our attention to parameterized paths which are *admissible*, meaning that velocity vectors have all components bounded above by one. There are no restrictions on acceleration: only velocity is bounded. It will be assumed for simplicity that the cost functions τ_i agree with elapsed time. To change the velocity bound or the cost function, one can perform a simple rescaling of the geometry of the coordination space.[1]

[1]The assumption about potentially unbounded acceleration is not so easily skirted. The results of this paper are unknown in that case.

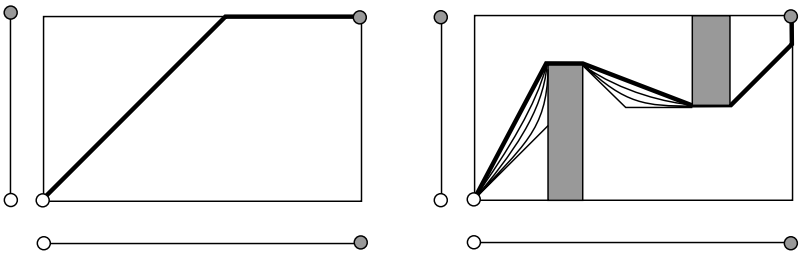


FIG. 1. *Left: the unique locally and globally Pareto optimal path between corners of a rectangle with regards to elapsed time; right: an envelope of Pareto optimal paths weaving through obstacles forms a single equivalence class.*

A path $\gamma : [0, T] \to \mathcal{X}$ is *Pareto optimal* if and only if $\tau(\gamma)$ is minimal with respect to the partial order on vectors:

$$\tau(\gamma) \leq \tau(\gamma') \Leftrightarrow \tau_i(\gamma_i) \leq \tau_i(\gamma'_i) \quad \forall\, i = 1, \ldots, N. \tag{2.3}$$

The local Pareto optima compose the set of all optima for all monotone scalarizations (such as, e.g., average time-to-goal and nonlinear generalizations thereof) as stated in the following theorem.

LEMMA 2.6. *For any scalarization $f : \mathbb{R}^n \to \mathbb{R}$ with $\partial f / \partial x_i > 0$, all minima of $f \circ \tau$ are global Pareto optima.*

Proof. Given any minimizer for $f \circ \tau$ which is not globally Pareto optimal, there exists a path which decreases some τ_i without increasing any of the others; by monotonicity, this decreases the f-value of the path, which is a contradiction. □

In some cases, locally Pareto optimal paths are unique. In the rectangular coordination space of Figure 1 (left), there is a unique locally (and globally) Pareto optimal path. The diagonal of the rectangle is not locally Pareto optimal unless the rectangle is a square.

Locally Pareto optimal paths are usually not unique. Two paths γ and γ' are *Pareto equivalent* if and only if they are homotopic through locally Pareto optimal paths which are equal in the partial order; i.e., $\tau(\gamma) = \tau(\gamma')$. Figure 1 (right) illustrates a single locally Pareto optimal class with many representatives. We show in section 3 that certain roadmap coordination spaces admit a continuum of locally/globally Pareto optimal *classes*.

2.3. A little topology. Experts can safely bypass this brief primer of the topological concepts used throughout the paper. See, e.g., [15] for a comprehensive treatment.

Given a topological space X, a *path* is a (continuous) map $\alpha : [0, 1] \to X$. Two paths α and α' sharing identical endpoints are said to be *homotopic* if there exists a continuous 1-parameter family of paths $\{\alpha_t\}_0^1$ satisfying (1) $\alpha_0 = \alpha$; (2) $\alpha_1 = \alpha'$; and (3) α_t has the same endpoints for all t.

Topologists keep track of different homotopy classes of paths by computing a more general object: the fundamental group. Given a fixed basepoint $x_0 \in X$, the *fundamental group* $\pi_1(X, x_0)$ is defined to be the set of homotopy equivalence classes of paths in X which have endpoints at x_0. The group operation is concatenation of paths: $\alpha * \beta$ means "do α then do β." This is well defined (since all endpoints are x_0) and yields a group with identity element the *contractible* loops in x. For path-connected spaces, $\pi_1(X, x_0)$ is independent of the basepoint x_0 (up to isomorphism of

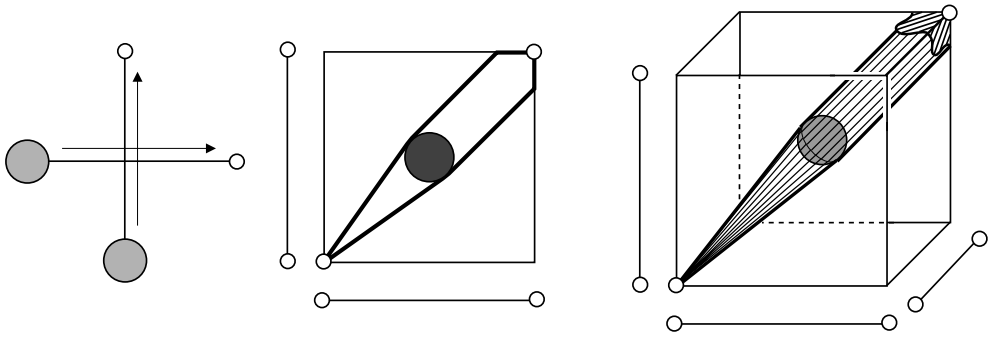


FIG. 2. *The case of two disc-like robots translating along edges (left) yields a two-dimensional coordination space with a disc obstacle possessing two globally Pareto optimal classes (center). Three robots with a spherical obstacle possess a continuum of globally Pareto optimal classes (right).*

groups), and one may unambiguously denote the fundamental group $\pi_1(X)$. A *simply connected* space X is one which is connected and has $\pi_1(X)$ a trivial group.

Any space X which has a reasonably nice local structure (e.g., a cell complex) and a nontrivial fundamental group can be "unwrapped" to a larger simply connected space which is locally indistinguishable from X. The *universal cover* of X, $\tilde{X}$, is defined to be a simply connected space along with a projection map $p : \tilde{X} \to X$ which is locally a homeomorphism. For example, the universal cover of the circle S^1 is $\mathbb{R}$ with the projection $p : x \mapsto x \mod 1$.

3. Examples. We illustrate a few examples of simple coordination spaces and globally Pareto optimal path classes.

Example 3.1. Consider the case where $N = 2$, $\Gamma_1 = \Gamma_2 = [-2, 2]$, and $\mathcal{O} = \{(x, y) : x^2 + y^2 < 1\}$, which corresponds to a pair of identical disc-shaped robots sliding along interval roadmaps which intersect in the workspace at right angles. There are exactly two globally Pareto optimal classes of paths from $(-2, -2)$ to $(2, 2)$, as illustrated in Figure 2 (left). The difference between these two paths lies in which robot decides to pause in order to allow the other to pass through the intersection. Note also that this is a cylindrical coordination space and hence satisfies the Main Theorem above.

Example 3.2. We modify the previous example by letting $N = 3$ and choosing $\mathcal{O}$ to be a round ball of radius 1 at the origin. By the symmetry of $\mathcal{X}$ about the diagonal of the cube, it is clear that there is a circle's worth of paths which begin at $(-2, -2)$, trace a straight line which is tangent to $\mathcal{O}$, and then exit this sphere tangentially with slope 1. The projection of this family of paths to the first two coordinates includes as special cases the distinct Pareto optima of Example 3.1, as well as a continuum of paths whose goal times continuously interpolate the two. Hence, there is a continuum of globally Pareto optimal classes.

Note that the symmetry in these problems is not crucial: the number of globally Pareto optima in both examples is robust to small perturbations in $\mathcal{O}$. Note as well that the obstacle set in Example 3.2 is not generated by pairwise collisions of the robots; the obstacle set is determined by the positions of all three robots.

Example 3.3. Coordination problems without the fixed path constraint can have continua of optima more easily than in the fixed path case. Consider the problem of translating two unit squares in the plane in such a way as to exchange their positions.

Assume that the squares are centered at $(-1, 0)$ and $(1, 0)$. It is straightforward to exhibit a continuous family of distinct globally Pareto optimal classes. Extrema of this family consist of those coordinations for which one square translates exclusively in the horizontal direction, while the other square moves vertically in order to get out of the way. The intermediate coordinations involve one square translating up by an amount $0 < h < 1$ and the other translating down by a total amount $1 - h$. This example works with arbitrary translations or with translations restricted to coordinate axis directions.

4. Discretization. Coordination spaces come equipped with a natural geometry as a subset of a (flat) product of metric graphs. For the remainder of this paper, we consider the *path metric*—the distance between points is defined in terms of the length of a shortest path (geodesic) between the points. Our tools for dealing with coordination spaces stem from knowledge of the geometry of the path metric.

4.1. Nonpositive curvature. Let (X, d) denote a complete metric space for which local geodesics exist: between any two sufficiently close points $p, q \in X$, there exists a path in X whose length is equal to $d(p, q)$. The notion of curvature bounds for such spaces is classical and can be found in the work of, e.g., Alexandrov (see [8] for references and a comprehensive introduction). We repeat that it is not necessary for X to be a manifold, much less smooth. In this "synthetic" approach to geometry, one interrogates a space with geodesic triangles.

For each triple of points $p, q, r \in X$, draw the triangle in X with geodesics $\overline{pq}$, $\overline{qr}$, and $\overline{rp}$. Let p', q', and r' denote three points in the Euclidean plane $\mathbb{E}^2$ forming a *comparison triangle* whose edges have length $d(p, q)$, $d(q, r)$, and $d(r, p)$, respectively. Choose a pair on points s and t on the edges of the geodesic triangle in X; then, consider the corresponding points s' and t' in $\mathbb{E}^2$. The $CAT(0)$ *inequality*[2] for X is the following: for every p, q, r in X, and for every s, t, one has $d(s, t) \leq \|s' - t'\|$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{E}^2$.

DEFINITION 4.1. *A complete geodesic space X is $CAT(0)$ if the $CAT(0)$ inequality holds for each geodesic triangle in X. A geodesic space X is* nonpositively curved *or* NPC *if it is locally $CAT(0)$, that is, if the $CAT(0)$ inequality holds for sufficiently small geodesic triangles in X.*

Otherwise said, X has nowhere positive curvature if small geodesic triangles in X are "no fatter" than similar triangles in the Euclidean plane: see Figure 3. By assigning the appropriate definition of angles in a geodesic metric space [8], this condition translates into saying that the sum of the angles of a geodesic triangle in an NPC space is no larger than π. It is a standard fact that X is $CAT(0)$ if and only if it is NPC and simply connected [8].

The existence of *some* metric of nonpositive curvature on a space X implies several strong results about the topology and geometry of X. For example [8]: (1) the universal cover of X is contractible; (2) the fundamental group $\pi_1(X)$ has no torsional elements; and, most importantly for our purposes, (3) geodesics on X are unique up to homotopy, fixing the endpoints.

The problem of determining when a space is $CAT(0)$ or NPC is surprisingly easy for piecewise-Euclidean cube complexes, that is, spaces built from Euclidean cubes glued together along faces. Since the individual cubes are flat, any nonzero curvature, if it exists in the complex, must be focused along the faces and detectable at the

[2]The term "CAT" is Gromov's, used in deference to the work of Cartan, Alexandrov, and Toponogov. Alternately, one can take this as *comparare ab triangulos* ("to compare by triangles").

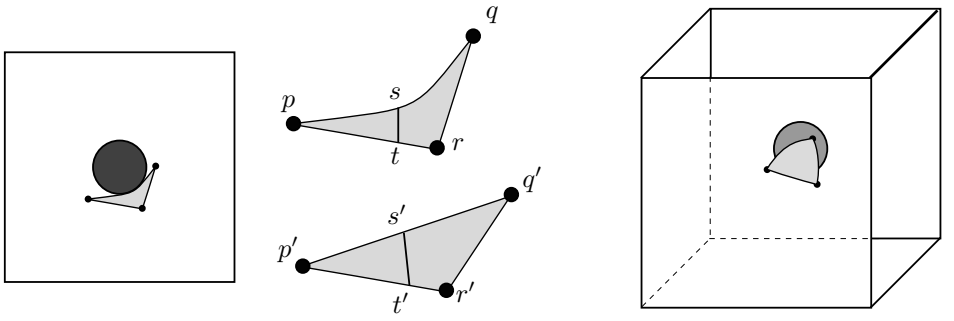


FIG. 3. *The coordination space of Example* 3.1 *is NPC (left) since triangles are "skinnier" than their Euclidean counterparts (center), whereas that of Example* 3.2 *is not NPC (right).*

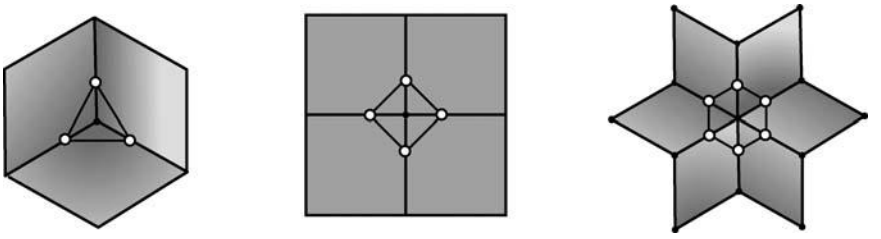


FIG. 4. *Vertices detect curvature for a cube complex: from left to right, positive, zero, and negative curvature in two-dimensional cube complexes. The link condition holds for the latter two of these complexes.*

vertices of the cube complex. Intuitively, positive curvature arises whenever there is a convex "corner" in the complex.

More specifically, let v denote a vertex of a cube complex X. The number and types of cubes in X which have a corner at v are encoded in a combinatorial object called the *link* of v; see, e.g., [8] for details. The following result of Gromov (which we have translated into simpler language for clarity) characterizes nonpositive curvature in terms of this structure and is therefore called the *link condition*.

THEOREM 4.2 (Gromov's link condition). *A piecewise Euclidean cube complex is NPC if and only if the following holds for each vertex v. Given any set of $K \geq 3$ edges of X incident to v, any pair of which are part of a square in X, all K edges are part of a K-dimensional cube with corner at v.*

This is illustrated for several two-dimensional complexes in Figure 4. The intuition behind this theorem is that positive curvature in a cube complex is entirely determined by the presence of a "sharp" corner.

4.2. Discretized coordination spaces. The key step in our analysis of cylindrical coordination spaces is a spatial discretization of them into cubical complexes. Discretized configuration spaces have recently been used in other robotics contexts [1, 2, 13].

Consider a coordination space $\mathcal{X}$ for roadmaps $\{\Gamma_i\}$. For simplicity, assume that all the edge lengths of the graphs Γ_i are rationally related. Because of this, one may choose a length $\delta > 0$ so that all edge lengths are integer multiples of δ. Denote by $\Gamma_i^{(n)}$ the graph obtained from Γ_i by subdividing it into edges of length $2^{-n}\delta$.

DEFINITION 4.3. *For each $n \geq 0$, denote by $\mathcal{X}^{(n)} \subset \Gamma_1^{(n)} \times \cdots \times \Gamma_N^{(n)}$ the maximal closed subcomplex of the product cubical complex which does not intersect the obstacle set $\mathcal{O}$. This cubical complex is called the* stage-n discretized coordination space.

The cubical nature of these spaces forms a convenient structure with which to manipulate paths. Our strategy is to approximate locally Pareto optimal paths with paths which follow along the diagonals of cubes in $\mathcal{X}^{(n)}$ for n sufficiently large. Working with these *cube paths* will allow us to skirt such technical issues as regularity of paths. The following result is crucial.

THEOREM 4.4. *Cylindrical coordination spaces are NPC.*

Proof. By assumption, all coordination spaces in this paper are complete geodesic metric spaces under the path metric. We begin by showing that $\mathcal{X}^{(n)}$ converges to $\mathcal{X}$ in the Hausdorff sense; i.e., given any point $p \in \mathcal{X}$ and an $\epsilon > 0$, there exists an $M > 0$ such that an ϵ-neighborhood of p intersects $\mathcal{X}^{(n)}$ for all $n > M$. This is certainly true for p in the interior of $\mathcal{X}$, since p is bounded away from $\mathcal{O}$. Since $\mathcal{X}$ is collared, points in the boundary of $\mathcal{X}$ are limits of points in the interior, and convergence follows.

It is known that a complete NPC space which is the Hausdorff limit of NPC spaces is itself NPC (see [8, Lemma II.3.10(1)], which is stated for $CAT(0)$ spaces; apply this to the universal covers of $\mathcal{X}$ and $\mathcal{X}^{(n)}$ to obtain the result). Hence, it suffices to show that $\mathcal{X}^{(n)}$ is NPC.

Since these are cubical complexes, Gromov's link condition is necessary and sufficient to show NPC. Therefore, let v be any vertex of $\mathcal{X}^{(n)}$. Assume that there is a collection of $K \geq 3$ edges $\{e_i\}_1^K$ incident to v that pairwise bound squares in $\mathcal{X}^{(n)}$. Each e_i corresponds to an edge of $\Gamma_{\alpha(i)}$ for some indexing α satisfying $\alpha(i) \neq \alpha(j)$ for all $i \neq j$. The square bound by edges e_j and e_k corresponds to a square in

$$\Gamma_{\alpha(j)} \times \Gamma_{\alpha(k)} - \Delta_{\alpha(j),\alpha(k)}$$

for each j and k. By the cylindrical structure of the holes in $\mathcal{X}$, the product of the edges $\{e_i\}$ is a K-dimensional cube in $\prod_1^m \Gamma_{\alpha(i)}$ which does not intersect $\Delta_{\alpha(j),\alpha(k)}$ for any pair j, k. This therefore defines a K-dimensional cube in $\mathcal{X}^{(n)}$ spanning the edges. The link condition is satisfied. □

Theorem 4.4 generalizes a number of known results about NPC configuration spaces, especially the results of [1] about configuration spaces of points on graphs, and more general "colored" configuration spaces of graphs.[3]

4.3. Cube paths. The spatial discretization of coordination spaces extends to a discrete version of paths. The following notion of discrete paths is crucial to future proofs.

DEFINITION 4.5. *Let X denote a cubical complex. A* cube path *from vertices v_0 to v_N in X is an ordered sequence of closed cubes $\mathcal{C} = \{C_i\}_1^N$ of X which satisfy* (1) *$C_i \cap C_{i+1} = \{v_i\}$, a vertex; and* (2) *$C_i$ is the smallest cube of X containing v_{i-1} and v_i. A cube path is said to be* normal *if in addition for all i,* (3) *$C_{i+1} \cap \mathsf{St}(C_i) = v_i$, where $\mathsf{St}(C_i)$ is the* star *of C_i (all cubes in X, including C_i, which have C_i as a face).*

To each cube path is associated a piecewise linear (PL) path in X given by the chain of straight line segments from v_i to v_{i+1} for $i = 0, \ldots, N-1$: see Figure 5. Roughly speaking, a normal cube path is one which uses the highest dimensional cubes as early as possible in the path.[4] This hints at the intuition of normal paths as

[3]This proof is simpler than that of Abrams in that it requires no direction computation of the homotopy groups of the Hausdorff limit.

[4]These paths are called "normal" in [21] since they are used to generate normal forms for languages of fundamental groups acting on $CAT(0)$ cube complexes.

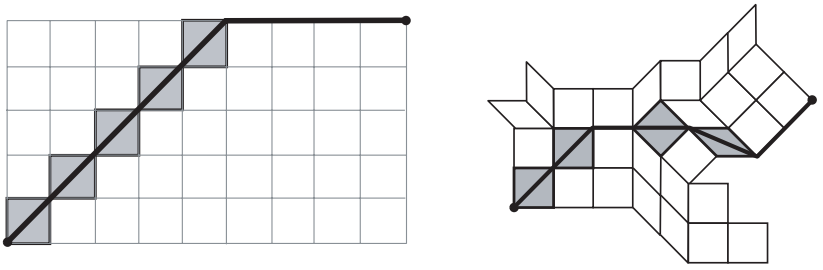


FIG. 5. *Examples of normal cube paths (with the diagonals highlighted).*

locally Pareto optimal, an intuition which we demonstrate is entirely justified. The key fact about normal cube paths is that, like geodesics on an NPC space, they are unique up to homotopy.

COROLLARY 4.6. *Any cube path between vertices of $\mathcal{X}^{(n)}$ is homotopic to a unique normal cube path.*

Proof. This is true for any NPC cube complex via Proposition 3.3 of [21]. By Theorem 4.4, this holds for $\mathcal{X}^{(n)}$. □

5. Topological bounds on Pareto optima. In this section we demonstrate a uniqueness result for locally Pareto optimal classes. The strategy of our proof is to show that any locally Pareto optimal path class in $\mathcal{X}$ contains a representative which is a limit of normal cube paths in $\mathcal{X}^{(n)}$. For the remainder of the paper, $\mathcal{X}$ will denote a cylindrical roadmap coordination space.

The geometry on $\mathcal{X}$ is that defined by the path metric: the metric distance between points of $\mathcal{X}$ is defined to be the length of a geodesic path between the points. Unless otherwise noted, the default notion of length is the standard length obtained by integrating the ℓ^2 norm of the velocity vectors over the path. Other notions of length are useful. Given a velocity vector, the ℓ^∞ norm of that vector is the usual definition: in this case, the maximal speed of the AGVs is traveling at that instant. The ℓ^∞ length of a path in $\mathcal{X}$ is the integral of the ℓ^∞ norm of the velocity vectors over the path. This ℓ^∞ length measures the total elapsed time that the AGVs take to execute the coordination implied by the path. As such, ℓ^∞ geodesics between points correspond to "fastest" coordinations.

LEMMA 5.1. *Any ℓ^∞ geodesic path on $\mathcal{X}$ can be C^0- and ℓ^∞-approximated by a cube path in $\mathcal{X}^{(n)}$ for n sufficiently large.*

Proof. Given an ℓ^∞ geodesic α, those portions of the curve which lie near $\partial\mathcal{O}$ may not be in $\mathcal{X}^{(n)}$. However, given the embedded collar $\Phi : \partial\mathcal{O} \times [0,1] \hookrightarrow \mathcal{X}$ from Assumption 2.2, the curve α can be pushed into the interior of $\mathcal{X}$ by increasing the second coordinate of Φ. This yields a curve β which avoids $\partial\mathcal{O}$, is freely homotopic to α, and is both C^0 and ℓ^∞-close.

We next approximate β by a cube path. Since the image of β is compact, there exists an $\epsilon > 0$ so that a tube of radius ϵ about the image of β does not intersect $\partial\mathcal{O}$. Choose n so that the grid size $\delta = 2^{-n}$ is less than ϵ. Modify β so that its endpoints are at vertices of $\mathcal{X}^{(n)}$. Consider the vertex $v^0 = \beta(0)$ and the ℓ^∞ balls $B_r = B_r^\infty(v^0)$ of radius r about v^0 in $\mathcal{X}^{(n)}$. Let $w^1 = \beta \cap \partial B_\delta$ and $w^2 = \beta \cap \partial B_{2\delta}$.

Let $F = B_\delta(v^0) \cap B_\delta(w^2)$. Clearly, $w^1 \in F$, and for any $x \in F$ it holds that the PL path $v^0 \to x \to w^2$ is of the same ℓ^∞ length as the restriction of β to this neighborhood. We claim that F contains a vertex of the grid structure on $\mathcal{X}^{(n)}$. To see this, note that F is a subset of some face of B_δ defined by the set of all points

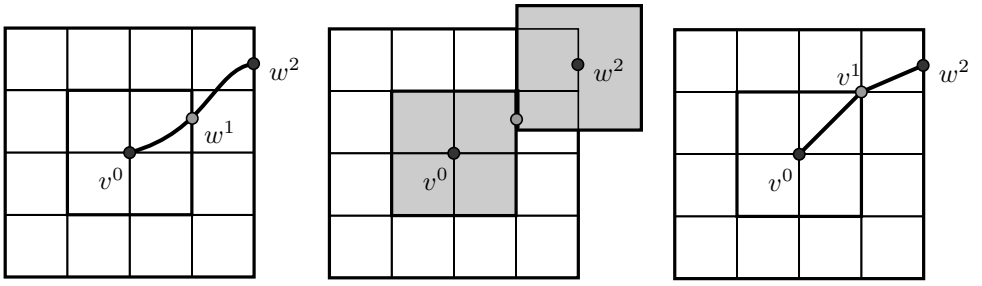


FIG. 6. *Approximation of a path by a cube path.*

$x = (x_i)$ such that

$$x_i \in [v_i^0 - \delta, v_i^0 + \delta] \cap [w_i^2 - \delta, w_i^2 + \delta], \qquad i = 1, \ldots, N. \tag{5.1}$$

Since these intervals have the same length, it follows that either $v_i^0 - \delta$ or $v_i^0 + \delta$ lies within the intersection for each i. Therefore, F contains a gridpoint v^1 of $\mathcal{X}^{(n)}$, and we may replace the segment of the path β with a PL path from v^0 to v^1 to w^2: see Figure 6. Repeat the argument inductively, beginning at the vertex v^1 and considering the ℓ^∞ balls about v^1 of radius δ and 2δ.

Marching along yields a PL path passing through vertices of $\mathcal{X}^{(n)}$ which are sequentially of ℓ^∞ distance δ. This is therefore a cube path in $\mathcal{X}^{(n)}$. This completes the proof of the lemma, as all modifications are homotopies within a neighborhood of the initial path α. □

LEMMA 5.2. *Let α be a locally Pareto optimal path on $\mathcal{X}$. One can find homotopic cube paths with endpoints close to those of α such that all the goal times approximate those of α.*

Proof. Any locally Pareto optimal path is a concatenation of paths which go from the attainment of the ith goal to the $(i+1)$st goal. Each of these subpaths is an ℓ^∞ geodesic whose ℓ^∞ length contributes to the elapsed time. Applying Lemma 5.1 inductively to these subpaths gives a path which approximates all the elapsed times. □

A simple algorithm, `LeftGreedy`, converts a cube path $\mathcal{C}$ into a normal cube path by sweeping along $\mathcal{C}$ from left to right, comparing the star of each cube with its neighbor to the right (line 3) and shifting the common directions to the left cube (lines 4–5).

ALGORITHM **1.** `LeftGreedy`($\mathcal{C}$).

Require: $\mathcal{C} = \{C_i\}_1^m$ is an ℓ^∞ geodesic cube path
1: **while** $\mathcal{C}$ is not normal **do**
2: **for** $i = 1, \ldots, m$ **do**
3: $Z \leftarrow C_{i+1} \cap \mathsf{St}(C_i)$
4: $C_{i+1} \leftarrow C_{i+1}/Z$
5: $C_i \leftarrow C_i \times Z$
6: **end for**
7: **end while**

LEMMA 5.3. *Given a locally Pareto optimal cube path $\mathcal{C}$,* `LeftGreedy`($\mathcal{C}$) *is a normal cube path which is Pareto equivalent to $\mathcal{C}$.*

Proof. Since common directions are shifted to the left, each sweep through $\mathcal{C}$ for which $C_{i+1} \cap \mathsf{St}(C_i) \neq \{v_i\}$ for some i strictly decreases the positive integer-valued function $\sum_i i \cdot \dim(C_i)$; hence, convergence.

To argue that `LeftGreedy` can be executed through Pareto equivalence, note first that the shifting process of lines 4–5 results in a homotopic path, since all modifications take place within $\mathsf{St}(C_i) \cup C_{i+1}$. If the jth coordinate x_j reaches its goal at vertex v_i of $\mathcal{C}$, then all subsequent vertices v_k $(k \geq i)$ lie within the hyperplane of $\mathcal{X}$ having the same jth coordinate. Hence, in line 3, the common factor Z is trivial in the j^{th} coordinate for v_i and beyond. Thus, the time-to-goal for x_j is unchanged by steps 4–5. □

We now may prove the Main Theorem.

THEOREM 5.4. *Locally Pareto optimal path classes on $\mathcal{X}$ are in bijective correspondence with homotopy classes of paths (fixing the endpoints).*

Proof. Existence of a locally Pareto optimum does not rely on the cylindrical nature of $\mathcal{X}$, but rather on the fact that it is compact and piecewise real-analytic. Via Lemma 2.6 applied to each homotopy class, the existence of a locally Pareto optimal path is guaranteed by the existence of a path which minimizes some monotone scalarization of the cost function, e.g., average time-to-goal. For fixed endpoints, this scalar function is positive and bounded below over the space of piecewise real-analytic paths in a given homotopy class.

The classification theorem relies on the cylindrical nature of $\mathcal{X}$. Assume there are two locally Pareto optimal paths, α and α', which are homotopic. Approximate these paths by cube paths $\mathcal{C}$ and $\mathcal{C}'$ on $\mathcal{X}^{(n)}$ for some large n via Lemma 5.2. The approximations $\mathcal{C}$ and $\mathcal{C}'$ are homotopic in $\mathcal{X}$ to α and α', respectively: we claim that $\mathcal{C}$ and $\mathcal{C}'$ are homotopic in $\mathcal{X}^{(n)}$. Since α and α' are homotopic in $\mathcal{X}$, there exists a map of a closed disc into $\mathcal{X}$ whose boundary is sent to $\alpha * (\alpha')^{-1}$ (that is, the loop obtained by concatenating loop α with the reverse of the loop α'). The collar Φ from Assumption 2.2 which is used to push α and α' into the interior of $\mathcal{X}^{(n)}$ also pushes this spanning disc into the interior. Thus, the paths $\mathcal{C}$ and $\mathcal{C}'$ are homotopic within $\mathcal{X}^{(n)}$.

Feed these cube paths into the algorithm `LeftGreedy`. Via Lemma 5.3, the outputs are normal cube paths which are Pareto equivalent to $\mathcal{C}$ and $\mathcal{C}'$, respectively. By Corollary 4.6, $\mathcal{C}$ and $\mathcal{C}'$ are the same path and hence have the same output times. Therefore, α and α' have output times which are arbitrarily close. □

This implies that the locally Pareto optimal classes are "discrete" and cannot arise in continua as in the noncylindrical case. It does not automatically follow that there is a bound on the number of globally Pareto optimal paths, since any cylindrical coordination space $\mathcal{X}$ which is not simply connected must have a countably infinite number of homotopy classes of paths. However, since the obstacle set is collared, the fundamental group of $\mathcal{X}$ is finitely generated, and, apart from this finite collection of generators, all other paths "wrap around" a loop multiple times. One rightly suspects that such repetition makes everyone's arrival times later and thus cannot be a globally Pareto optimal solution. We make this intuition rigorous below.

THEOREM 5.5. *The number of globally Pareto optimal paths between fixed endpoints on $\mathcal{X}$ is finite.*

Proof. Fix endpoints and consider $\{\gamma_i\}_1^\infty$ an infinite collection of nonequivalent globally Pareto optimal paths. By Theorem 5.4, all these are in different homotopy classes. The endpoints of these paths in the universal cover must be an unbounded discrete set, whose ℓ^∞ lengths tend uniformly to infinity, after taking a subsequence and reordering.

Denote by $\{T_i^j\}_{i=1,\dots,\infty}^{j=1,\dots,N}$ the $j = 1, \dots, N$ goal times of the path γ_i. By the pigeonhole principle, one may assume (after taking a subsequence) that for some k, $T_i^{k+1} \geq T_i^k + i$ for all i. By compactness of $\mathcal{X}$, the set of points $\{\gamma_i(T_i^k)\}$ has an accumulation point x_k. Likewise, the set of points $\{\gamma_i(T_i^{k+1})\}$ has an accumulation point x_{k+1}. For all i large, a small perturbation to the path γ_i after T_i^k in a neighborhood of x_k and x_{k+1} reduces T_i^j by more than $i-1$ for each $j > k$, without increasing any other goal times. This contradicts the assumption of globally Pareto optimal paths. □

COROLLARY 5.6. *There exists a finite bound on the number of globally Pareto optimal classes on $\mathcal{X}$ between fixed endpoints which is independent of the endpoints.*

Proof. For each fixed pair of endpoints in $\mathcal{X} \times \mathcal{X}$, there is a finite number of globally Pareto optimal classes via Theorem 5.5. The bound extends to small open neighborhoods in $\mathcal{X} \times \mathcal{X}$ via the same argument. Thanks to the compactness of $\mathcal{X}$, this open cover restricts to a finite subcover. □

This completes the proof of the Main Theorem.

6. Computing locally Pareto optima. The proof of the Main Theorem reveals a technique for computing "canonical" representatives for the locally Pareto optimal classes. One observes from the proof that passing to a sufficiently fine cubical discretization of the coordination space and choosing a normal cube path between endpoints yields an approximation to a locally Pareto optimal path. Normal cube paths are not difficult to compute, given the cubical approximation—one simply cuts across as many diagonals as early as possible in the path.

However, this procedure is not optimal as a means of generating representatives of a Pareto equivalence class, since it involves a potentially large and expensive discretization step and yields only an approximation to the desired optimum. A less expensive approach that yields the true locally Pareto optimal representative path comes from deriving a "continuous" version of a normal cube path which does not rely on having a cubical structure at all. These paths, called *left-greedy normal* paths, are well defined, unique up to homotopy, and locally Pareto optimal.

A companion paper [14] to this one gives proper definitions of left-greedy normal paths, along with algorithms for their computation, complexity bounds, and implementation in the context of coordinating several irregular-shaped robots translated along planar tracks. It is an interesting feature that the algorithms for computing these locally Pareto optimal paths rely crucially on the nonpositive curvature of the underlying coordination space.

7. Moving obstacles. Assume that, as before, we have N robots which traverse roadmaps $\{\Gamma_i\}_1^N$. If, as will typically occur, the obstacle set $\mathcal{O}$ is defined by pairwise collisions, then the resulting coordination space $\mathcal{X}$ will be cylindrical and the finiteness bounds apply.

Assume now that there are additional objects which, like the robots, move through the shared environment but, unlike the robots, are not controllable. These objects move along predetermined trajectories in the workspace. We make no assumptions about their speed, shape, or motion.[5] We assume *only* that the trajectory through the workspace is a fixed function of time. Such a system gives rise to a roadmap coordination system with moving obstacles.

We consider the problem of determining locally Pareto optimal coordinations of the N robots through the workspace, where all admissible paths are those which avoid

[5]Indeed, the moving obstacles may, during their trajectory, change shape, speed, or even connectivity.

collision with the moving obstacles. Our strategy is to augment the coordination space by adding one dimension for time, similar to the methods used in [17, 25].

THEOREM 7.1. *Any roadmap coordination system with moving obstacles possesses a finite number of globally Pareto optimal classes of paths.*

Proof. Let $\mathcal{X}$ denote the coordination space of the N robots obtained by ignoring the additional obstacles. We will build an augmented coordination space of the form $\overline{\mathcal{X}} \subset [0,T] \times \left(\prod_{k=1}^{N} \Gamma_k\right)$. Encode the positions of the moving objects as a time-parameterized subset of the workspace (where, without loss of generality, $0 \leq t \leq T$ for some T). For each $1 \leq i \leq N$, denote by $\overline{\mathcal{O}}_i \subset [0,T] \times \Gamma_i$ the set of points of the form (t, x_i), where the ith robot in configuration x_i is in collision with the moving obstacles' positions at time t. We define $\overline{\mathcal{X}}$ via the obstacle set

$$\overline{\mathcal{O}} = ([0,T] \times \mathcal{O}) \bigcup_i \left\{ (t, x_1, \ldots, x_N) \in [0,T] \times \prod_{k=1}^{N} \Gamma_k \,:\, (t, x_i) \in \mathcal{O}_i \right\}. \tag{7.1}$$

Claim 1. The augmented coordination space $\overline{\mathcal{X}}$ is cylindrical.

To prove this, note that $\overline{\mathcal{X}}$ is obtained from $[0,T] \times \prod_i \Gamma_i$ by removing sets defined either from $\mathcal{O}$ or from $\mathcal{O}_i$. These are all pairwise defined cylindrical subsets.

This completes Claim 1.

We now apply the same proof to $\overline{\mathcal{X}}$ with two key modifications. The first is that the time direction $[0,T]$ in $\overline{\mathcal{X}}$ is rigid: all admissible paths in $\overline{\mathcal{X}}$ must travel with unit speed in the t-direction. The second is that the endpoint is variable: paths start at the point $(0,p)$ and end at (T',q) for some fixed $p, q \in \mathcal{X}$ and for some variable $T' \leq T$.

For the moment, assume that T' is fixed.

Claim 2:. The number of admissible globally Pareto optimal paths from $(0,p)$ to (T',q) is finite.

To prove this, one shows that the proofs of Theorems 5.4 and 5.5 apply to admissible paths (those with unit speed in the t-direction). Not all homotopy classes of $\overline{\mathcal{X}}$ are represented, but only those which have unit speed in the t-direction. The crucial observation for the analogue of Theorem 5.4 is that Algorithm 1 applied to a path with unit speed in the t-direction preserves this property. The analogue of Theorem 5.5 in this setting is unchanged.

This completes Claim 2.

In the general case, the endpoint has t-coordinate T'. For each such T' there is a finite set of globally Pareto optimal paths in $\overline{\mathcal{X}}$.

Claim 3. The projections of these globally Pareto optimal paths in $\overline{\mathcal{X}}$ to $\mathcal{X}$ are independent of T' for T' sufficiently large.

To prove this, assume that $\mathcal{C}'$ is a normal cube path from $(0,p)$ to (T',q) in $\overline{\mathcal{X}}$. Let $\mathcal{C}$ denote the cube path obtained by appending to $\mathcal{C}'$ the straight line segment from (T',q) to (T,q). Claim 3 follows if $\mathcal{C}$ is also a normal cube path.

Assume that $\mathcal{C}$ is not normal. Then, since both pieces of $\mathcal{C}$ are normal ($\mathcal{C}'$ and the segment $(T',q) \to (T,q)$), the normal condition must fail precisely at the vertex (T',q). This would imply that the star of the last cube in $\mathcal{C}'$ intersects the segment $(T',q) \to (T,q)$; hence, the last cube of $\mathcal{C}'$ had no component in the t-direction. However, this is a contradiction, since $\mathcal{C}'$ is an admissible path with constant speed 1 in the t-direction.

This completes Claim 3 and the theorem. □

The problem of computing locally Pareto optimal paths in this case is no different from that in the stationary obstacle case of the previous sections. The motion of

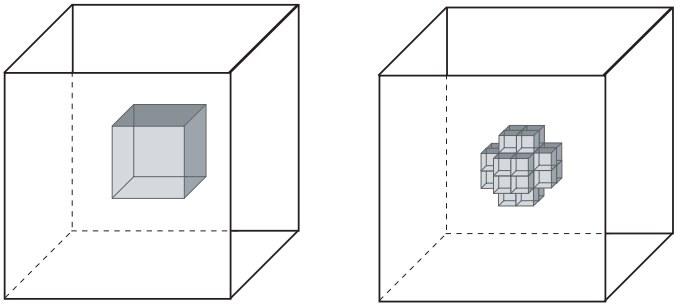


FIG. 7. *As one approximates the obstacle of Example* 3.2 *by finer and finer cubes, the number of normal cube paths from one corner to its antipode grows exponentially.*

the obstacles is incorporated in the geometry of the coordination space in the time-dimension.

8. On positive curvature. We claim that the crucial reason for the finiteness bound is that the cylindrical coordination spaces can be approximated by cubical complexes without any vertices of positive curvature. For more general coordination spaces, all the steps of the proof of the finiteness bound hold *except* the uniqueness of normal cube paths up to homotopy.

For example, if one approximates the space of Example 3.2 by discretizations $\mathcal{X}^{(n)}$, then, as before, there are a finite number of normal cube paths. For example, in $\mathcal{X}^{(2)}$, $\mathcal{O}$ is approximated by a cube and there are exactly three normal cube paths. However, they are not unique up to homotopy, and the number of such paths grows exponentially in the discretization step n, giving rise to the continuum of optima in the limit; see Figure 7.

The positive curvature in Example 3.2 is transparent: the obstacle set $\mathcal{O}$ clearly has a positively curved boundary. But in Example 3.3, the positive curvature is more subtle and arises from the fact that the obstacle set $\mathcal{O}$ is not axis-aligned. Consider first the case in which the squares have full mobility in all directions. We discretize this $\mathcal{X}$ to a cube complex by letting edges correspond to sliding a square in either the horizontal or vertical direction. Higher dimensional cubes correspond to slides which "commute" (they are compatible without causing a collision between the two squares).

We claim that positive curvature is created in the discretization. It suffices to demonstrate the failure of the link condition, namely, to show that there exists a triple of edges which pairwise commute but which together yield an illegal state. This is illustrated in Figure 8. Each pair of the three moves exhibited there commutes (with one pair generating a diagonal slide), but all three performed simultaneously generates a collision.

There is one last type of failure on which we remark. If we modify Example 3.3 so that squares are constrained to slide solely in the horizontal and vertical directions, then, indeed, the discretized cube complex now has no positive curvature, since sliding the same square horizontally and vertically no longer constitutes commutative moves. Because of this, however, there are nontrivial loops in $\pi_1(\mathcal{X})$ generated by moving a single square north-east-south-west in turn. There is one such loop for each vertex in $\mathcal{X}^{(n)}$, and the length of the loop is four times the discretized edge length. As $n \to \infty$, the number of such loops becomes arbitrarily large while their lengths become arbitrarily small: there is a "froth" of π_1. Even though normal cube paths (and hence

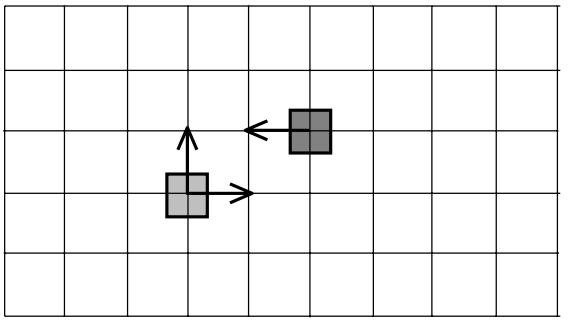


FIG. 8. *The link condition fails for cubical approximations to the coordination space of Example* 3.3*: the three translations shown can be performed simultaneously pairwise, but performing all three at once causes a collision.*

locally Pareto optimal representatives) are unique up to homotopy, the number of such homotopy classes blows up exponentially. In this case, the argument involving π_1 being finitely generated no longer holds, and, indeed, there are an uncountably infinite number of globally Pareto optimal paths in the limit.

9. Conclusion. The techniques we have introduced here—$CAT(0)$ geometry and nonpositive curvature—are very classical concepts. Yet their applications outside of geometry and algebra are in the very beginning stages, and we are hopeful that other problems in robotics and optimization will benefit from these perspectives. We end by noting that the problem of computing global Pareto optima is not dissimilar to that of computing Euclidean geodesics. It is well known that computing shortest paths in three dimensions is NP-hard in general [10] (see, e.g., [19] for an up-to-date account of hardness results for geodesics). It is worth noting that every example surveyed in [19] which has an NP-hard geodesic problem is not a $CAT(0)$ space. It would be a good omen for the applicability of $CAT(0)$ techniques if the geodesic problem simplified in the case of $CAT(0)$ spaces in general. In a parallel application, pursuit-evasion problems on domains of dimension greater than two depend crucially on the geometry and topology of the domain. For certain pursuit games, the existence of a successful pursuit strategy can be guaranteed on a $CAT(0)$ space [4].

REFERENCES

[1] A. ABRAMS, *Configuration Spaces and Braid Groups of Graphs*, Ph.D. thesis, University of California at Berkeley, Berkeley, CA, 2000.

[2] A. ABRAMS AND R. GHRIST, *State complexes for metamorphic robot systems*, Int. J. Robot. Res., 23 (7/8), pp. 809–824.

[3] S. AKELLA AND S. HUTCHINSON, *Coordinating the motions of multiple robots with specified trajectories*, in Proceedings of the IEEE International Conference on Robotics and Automation, 2002, pp. 624–631.

[4] S. ALEXANDER, R. BISHOP, AND R. GHRIST, *Pursuit and evasion in arbitrary dimensions*, in Proceedings of the Robotics: Science and Systems Conference, to appear.

[5] M. D. ARDEMA AND J. M. SKOWRONSKI, *Dynamic game applied to coordination control of two arm robotic system*, in Differential Games—Developments in Modelling and Computation, R. P. Hämäläinen and H. K. Ehtamo, eds., Springer-Verlag, Berlin, 1991, pp. 118–130.

[6] J. BARRAQUAND AND J.-C. LATOMBE, *Robot motion planning: A distributed representation approach*, Int. J. Robot. Res., 10 (1991), pp. 628–649.

[7] Z. BIEN AND J. LEE, *A minimum-time trajectory planning method for two robots*, IEEE Trans. Robot. Automat., 8 (1992), pp. 414–418.

[8] M. BRIDSON AND A. HAEFLIGER, *Metric Spaces of Nonpositive Curvature*, Springer-Verlag, Berlin, 1999.

[9] S. J. Buckley, *Fast motion planning for multiple moving robots*, in Proceedings of the IEEE International Conference on Robotics and Automation, 1989, pp. 322–326.
[10] J. Canny and J. Reif, *Lower bounds for shortest path and related problems*, in Proceedings of the 28th Annual IEEE Symposium on Foundations of Computer Science, 1987, pp. 49–60.
[11] C. Chang, M. J. Chung, and B. H. Lee, *Collision avoidance of two robot manipulators by minimum delay time*, IEEE Trans. Syst. Man Cybern., 24 (1994), pp. 517–522.
[12] M. Erdmann and T. Lozano-Perez, *On multiple moving objects*, in Proceedings of the IEEE International Conference on Robotics and Automation, 1986, pp. 1419–1424.
[13] R. Ghrist, *Shape complexes for metamorphic robot systems*, in Algorithmic Foundations of Robotics V, STAR 7, Springer-Verlag, Berlin, Heidelberg, 2004, pp. 185–201.
[14] R. Ghrist, J. O'Kane, and S. M. LaValle, *Pareto optimal coordination on roadmaps*, Int. J. Robot. Res., 12 (2005), pp. 997–1010.
[15] A. Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, UK, 2001.
[16] H. Hu, M. Brady, and P. Probert, *Coping with uncertainty in control and planning for a mobile robot*, in Proceedings of the IEEE/RSJ International Workshop on Intelligent Robots and Systems, Osaka, Japan, 1991, pp. 1025–1030.
[17] K. Kant and S. W. Zucker, *Toward efficient trajectory planning: The path-velocity decomposition*, Int. J. Robot. Res., 5 (1986), pp. 72–89.
[18] S. M. LaValle and S. A. Hutchinson, *Path selection and coordination of multiple robots via Nash equilibria*, in Proceedings of the IEEE International Conference on Robotics and Automation, 1994, pp. 1847–1852.
[19] J. S. B. Mitchell and M. Sharir, *New results on shortest paths in three dimensions*, in Proceedings of the 20th Annual ACM Symposium on Computational Geometry, 2004, pp. 124–133.
[20] J. Miura and Y. Shirai, *Planning of vision and motion for a mobile robot using a probabilistic model of uncertainty*, in Proceedings of the IEEE/RSJ International Workshop on Intelligent Robots and Systems, Osaka, Japan, 1991, pp. 403–408.
[21] G. A. Niblo and L. D. Reeves, *The geometry of cube complexes and the complexity of their fundamental groups*, Topology, 37 (1998), pp. 621–633.
[22] P. A. O'Donnell and T. Lozano-Pérez, *Deadlock-free and collision-free coordination of two robot manipulators*, in Proceedings of the IEEE International Conference on Robotics and Automation, 1989, pp. 484–489.
[23] V. Pareto, *Cours d'Économie Politique*, Rouge, Lausanne, Switzerland, 1896.
[24] L. E. Parker, *Cooperative motion control for multi-target observation*, in Proceedings of the IEEE/RSJ International Workshop on Intelligent Robots and Systems, 1998, pp. 1591–1598.
[25] J. Peng and S. Akella, *Coordinating multiple robots with kinodynamic constraints along specified paths*, in Algorithmic Foundations of Robotics V, STAR 7, Springer-Verlag, Berlin, Heidelberg, 2004, pp. 221–237.
[26] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization*, Academic Press, New York, 1985.
[27] J. T. Schwartz and M. Sharir, *On the piano movers' problem:* III. *Coordinating the motion of several independent bodies*, Int. J. Robot. Res., 2 (1983), pp. 97–140.
[28] T. Simeon, S. Leroy, and J.-P. Laumond, *Path coordination for multiple mobile robots: A resolution complete algorithm*, IEEE Trans. Robot. Automat., 18, 2002.
[29] S.-H. Suh and K. G. Shin, *A variational dynamic programming approach to robot-path planning with a distance-safety criterion*, IEEE Trans. Robot. Automat., 4 (1988), pp. 334–349.
[30] P. Svestka and M. H. Overmars, *Coordinated motion planning for multiple car-like robots using probabilistic roadmaps*, in Proceedings of the IEEE International Conference on Robotics and Automation, 1995, pp. 1631–1636.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1714–1735

FEEDBACK AND FEEDFORWARD OUTPUT REGULATION OF BOUNDED UNIFORMLY CONTINUOUS SIGNALS FOR INFINITE-DIMENSIONAL SYSTEMS*

EERO IMMONEN† AND SEPPO POHJOLAINEN†

Abstract. This article presents a state space output regulation theory for linear infinite-dimensional systems (with bounded control, observation, and feedthrough operators) in which the reference and disturbance signals are bounded and uniformly continuous. Output regulation problems are posed and solved for both feedforward and error feedback controllers. Furthermore, under certain assumptions necessary conditions are obtained for those controllers which solve these output regulation problems. One consequence of these necessary conditions is a generalization of the well-known internal model principle. A key feature in our theory is that the exogenous system generating the reference and disturbance signals is infinite-dimensional.

Key words. output regulation, infinite-dimensional systems, infinite-dimensional exogenous systems, bounded uniformly continuous signals, strong stability, regulator equations

AMS subject classifications. 93B50, 93B52, 93D99, 37N35, 35R99

DOI. 10.1137/050623000

1. Introduction.

The term output regulation is usually associated with the stabilization of a dynamical system and asymptotic tracking of a given class of reference signals under a given class of disturbances. Controllers which achieve output regulation are often either of feedforward or feedback type. For finite-dimensional linear systems and simple reference/disturbance signals generated by systems of linear ordinary differential equations, such output regulation problems were studied intensively in the 1970s. A complete solution now exists for both feedback and feedforward controllers, e.g., in the work of Francis, Wonham, and Davison [8, 10, 11, 26]. In addition to this, the necessary structure of a feedback controller which solves the output regulation problem is well known; it is formulated as the internal model principle—one of the cornerstones in classical automatic control theory. The principle asserts, roughly, that output regulation can be achieved only if the control system encapsulates a model of the dynamical structure of the reference/disturbance signals [10, 11].

The work of Francis and Wonham initiated what is nowadays known as geometric output regulation theory. This terminology stems from the fact that they studied output regulation problems in geometric terms, such as subspace inclusions. During the past two decades several authors have extended this theory to infinite-dimensional systems. In the early 1980s Schumacher [21] constructed finite-dimensional controllers for infinite-dimensional plants in which the system operator has compact resolvent and a complete set of generalized eigenfunctions. His solution of the output regulation problem is also expressed in geometric terms (cf. Theorem 3.1 in [21]). Several years later Byrnes et al. [3] generalized Francis's theory [10] for infinite-dimensional systems in such a way that the geometric conditions were replaced by the so-called regulator equations. These equations (which are in fact also present in the finite-dimensional case [10] and in a certain form in [21]) express the geometric conditions in an alge-

*Received by the editors January 20, 2005; accepted for publication (in revised form) August 4, 2006; published electronically November 16, 2006.
http://www.siam.org/journals/sicon/45-5/62300.html
†Institute of Mathematics, Tampere University of Technology, PL 553, 33101 Tampere, Finland (Eero.Immonen@tut.fi, seppo.pohjolainen@tut.fi).

braic way, hence simplifying solution of the output regulation problem. Byrnes et al. [3] showed that solvability of the feedforward regulation problem, solvability of the error feedback regulation problem, and solvability of the regulator equations are all equivalent to each other provided that the plant and the finite-dimensional exogenous signal generator have sufficient stabilizability properties.

In the above approaches towards solving output regulation problems the class of admissible signals is rather small: The reference/disturbance signals are assumed to be generated by finite-dimensional exogenous systems, and they can be, e.g., ramps, constant signals, or sinusoids. Recently the feedforward output regulation theory of Byrnes et al. [3] was generalized for infinite-dimensional exogenous systems in [4, 5, 14, 15, 16]. In [4, 5] the one-dimensional wave equation is generating reference signals. In the present paper, as well as in [14], the exogenous system is *purpose-built* so that any bounded and uniformly continuous reference functions can be treated. This class of reference signals is essentially larger than that resulting from the use of a finite-dimensional exosystem; for example, it is possible to consider the asymptotic tracking of arbitrary sufficiently smooth periodic reference signals, as was done in [15, 16].

While in [14, 15, 16] only the use of feedforward control was considered, in this article we also extend the error feedback regulation theory of [3] to allow for any bounded uniformly continuous reference signals. The generalization is, however, not at all straightforward. In the case of a finite-dimensional exogenous system [3] it is reasonable to require the closed loop semigroup to be exponentially stable. However, in our setting this is not so; one may be required to stabilize an operator having an unbounded spectrum on the imaginary axis (see subsection 2.4). Since the error feedback regulation problem is solved by considering it as a feedforward control problem for the extended system, we are forced to solve both the feedforward (see section 3) and the error feedback (see section 4) regulation problems under the weaker assumption of strong closed loop stability. It turns out that, even under this weaker stability assumption, the regulator equations provide a solution to the output regulation problems. Yet their solvability is no longer necessary for output regulation unless certain additional conditions hold (see section 6).

The main results of this article are Theorems 3.1, 4.1, and 7.2:

- Theorem 3.1 provides a solution to the feedforward regulation problem. It shows that if the pair (A, B) in the plant is strongly stabilizable, then the feedforward control problem can be solved if the regulator equations (3.1) can be solved. In particular, this result generalizes Theorem IV.1 in [3], Theorem 3.1 in [16], and Corollary 3.3 in [14] for plants which are not necessarily exponentially stabilizable.
- Theorem 4.1 provides a solution to the error feedback regulation problem. It shows that if the closed loop system operator consisting of the plant and the controller generates a strongly stable semigroup, then output regulation occurs if two coupled pairs of regulator equations are satisfied—one for the plant (see (3.1)) and one for the controller (see (4.1)). A key feature in this result is that the controller structure is fixed, but the parameters can be freely chosen as long as the conditions are met. This result seems to be new even for finite-dimensional plants, and in particular its Corollary 4.2 generalizes Theorem IV.2 in [3], where the parameters of the controller are chosen to yield an observer.
- Theorem 7.2 is a direct generalization of the internal model principle in [10] for infinite-dimensional plants and bounded uniformly continuous ref-

erence/disturbance signals. Roughly stated, this result asserts that under certain auxiliary conditions the dynamics of the exogenous system must be embedded in the dynamics of the controller in order to obtain asymptotic tracking and closed loop stability. In the present form Theorem 7.2 seems to be new even for finite-dimensional plants, because the proof is new also for finite-dimensional exosystems (it utilizes semigroup methods, while matrix methods are used in [10]) and because the result also applies for general bounded and uniformly continuous reference/disturbance signals.

We conclude this introductory section with some remarks concerning the relation of our work to the so-called repetitive control scheme. Repetitive control is a frequency domain–based error feedback control technique introduced by Hara et al. [12] for the asymptotic tracking of periodic signals for finite-dimensional systems. Consequently the results of this paper partially overlap those in the repetitive control literature (see, e.g., [12, 27, 28] and the references therein). However, this overlap is not very large, as there are also some crucial differences between the present paper and the repetitive control literature. The most notable differences are the following: We work in the state space domain, we also allow for infinite-dimensional plants (this is not the case, e.g., in [12]), we consider both feedforward and error feedback control, and, finally, we also treat exogenous signals which are not periodic. Moreover, while in our case there is freedom in the choice of the parameters of the dynamic controller, the structure of a repetitive controller is fixed as it incorporates the internal model $\frac{1}{1-e^{-sp}}$ of p-periodic signals in the feedback loop. Yamamoto and Hara [28] have actually shown that it is also necessary to use this internal model if asymptotic tracking is to be achieved. Unfortunately, it turns out that the closed loop in a repetitive control system is notoriously difficult to stabilize, because this internal model is difficult to stabilize. Consequently in engineering applications one often resorts to the use of ad hoc design techniques to overcome this stabilizability problem; see, e.g., [25]. A similar stabilizability problem also arises in our work (section 5), although in our view the problem is not as severe in this paper as it is in the repetitive control literature. This is because the parameters of the controller are not fixed and because we only require strong stability of the semigroup generated by the closed loop system operator; for example, Yamamoto and Hara [28] require exponential stability (i.e., internal stability of the closed loop), which is difficult to achieve in practice. However, they could also conclude exponential decay of the tracking error.

1.1. Notation. For Banach spaces E and F, $\mathcal{L}(E,F)$ denotes the space of bounded linear operators $E \to F$ (all Banach spaces in this article are complex). The resolvent set of a closed linear operator $A : E \to F$ is denoted by $\rho(A)$, whereas $\sigma(A)$ denotes its spectrum. The approximate point spectrum of A is denoted by $\sigma_A(A)$ [9]. $R(\lambda, A)$ denotes (whenever it exists) the resolvent operator $(\lambda I - A)^{-1}$. If $\widetilde{E}$ is a subspace of E, then $A|_{\widetilde{E}}$ denotes the restriction of A to $\widetilde{E}$. The range of A is denoted by $\mathrm{ran}(A)$. A strongly continuous (or C_0) semigroup $T_A(t)$ in E is strongly stable if $\lim_{t\to\infty}\|T_A(t)x\| = 0$ for each $x \in E$. $T_A(t)$ is weakly stable if $\lim_{t\to\infty} f(T_A(t)x) = 0$ for each $x \in E$ and every functional $f \in \mathcal{L}(E,\mathbb{C})$. If A generates a C_0-semigroup in E and $B \in \mathcal{L}(F,E)$, then the pair (A,B) is strongly stabilizable if there exists $K \in \mathcal{L}(E,F)$ such that $A+BK$ generates a strongly stable semigroup in E. If $C \in \mathcal{L}(E,F)$, then the pair (A,C) is approximately observable if $\{\, x \in E \mid CT_A(t)x = 0 \text{ for all } t \geq 0 \,\} = \{0\}$. $\Re(z)$ denotes the real part of a complex number z and $i\mathbb{R}$ denotes the imaginary axis. $BUC(\mathbb{R},E)$ denotes the Banach space (with respect to sup-norm) of bounded uniformly continuous functions $f : \mathbb{R} \to E$.

The Carleman transform of $f \in BUC(\mathbb{R}, E)$ is defined by

$$\widetilde{f}(\lambda) = \begin{cases} \int_0^\infty e^{-\lambda t} f(t)dt, & \Re(\lambda) > 0, \\ -\int_{-\infty}^0 e^{-\lambda t} f(t)dt, & \Re(\lambda) < 0. \end{cases} \tag{1.1}$$

A point $\lambda_0 \in \mathbb{R}$ is called a regular point of f if $\widetilde{f}$ can be continued analytically into a neighborhood of $i\lambda_0$. The complement in $\mathbb{R}$ of the set of regular points is called the Carleman spectrum of f and it is denoted by $sp(f)$ [1, 24]. A function is almost periodic if it can be uniformly approximated by trigonometric polynomials [1].

2. Formulation of the output regulation problems. In this section we formulate the feedforward and error feedback output regulation problems. We begin by defining the plant and the exogenous system that generates the reference and disturbance signals.

2.1. The plant. We consider a plant described by the following (possibly infinite-dimensional) control system:

$$\dot{z}(t) = Az(t) + Bu(t) + \mathcal{U}_{dist}(t), \quad z(0) \in Z, \quad t \geq 0, \tag{2.1a}$$

$$y(t) = Cz(t) + Du(t), \quad t \geq 0. \tag{2.1b}$$

Here A generates a C_0-semigroup $T_A(t)$, $t \geq 0$, in a complex Banach space Z. The continuous input $u : \mathbb{R}_+ \to H$ and continuous output $y : \mathbb{R}_+ \to H$ take values in a complex Banach space H. The control operator $B \in \mathcal{L}(H, Z)$, the observation operator $C \in \mathcal{L}(Z, H)$, and the feedthrough operator $D \in \mathcal{L}(H)$ (bounded linear operators are denoted by $\mathcal{L}$). The continuous function $\mathcal{U}_{dist}$ is a disturbance (to be defined shortly). Equation (2.1a) is to be considered in the mild sense.

2.2. The exogenous system. Let $\mathcal{H}$ be a Banach space which is continuously embedded in $BUC(\mathbb{R}, H)$, the space of bounded uniformly continuous H-valued functions on $\mathbb{R}$ endowed with the sup-norm ($\mathcal{H}$ need not have the sup-norm). In symbols, let $\mathcal{H} \hookrightarrow BUC(\mathbb{R}, H)$. We let $T_S(t)$ denote the shift C_0-group in $BUC(\mathbb{R}, H)$ defined as $T_S(t)f = f(\cdot + t)$; its infinitesimal generator is $S = \frac{d}{dx}$ with the domain of definition $\mathcal{D}(S) = \{ f \in BUC(\mathbb{R}, H) \mid Sf \in BUC(\mathbb{R}, H) \} \subset BUC(\mathbb{R}, H)$. If $\mathcal{H}$ is invariant for $T_S(t)$ and the restrictions $T_S(t)|_{\mathcal{H}}$ to $\mathcal{H}$ constitute an isometric C_0-group, we denote this by $\mathcal{H} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$. In this case the generator of $T_S(t)|_{\mathcal{H}}$ is denoted by $S|_{\mathcal{H}}$.

Let $\mathcal{H} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$, and let Q denote the point evaluation at the origin in $\mathcal{H}$, i.e., $Qf = f(0)$ for $f \in \mathcal{H}$. Clearly Q is linear and $\|Qf\|_H = \|f(0)\|_H \leq \|f\|_\infty \leq c\|f\|_{\mathcal{H}}$ for some $c \geq 0$ (because $\mathcal{H} \hookrightarrow BUC(\mathbb{R}, H)$) and so $Q \in \mathcal{L}(\mathcal{H}, H)$. Moreover, $QT_S(t)|_{\mathcal{H}} f = f(x + t)|_{x=0} = f(t)$ for every $f \in \mathcal{H}$ and $t \in \mathbb{R}$. This leads us to the following definition.

DEFINITION 2.1 (the exogenous system). *Let $\mathcal{H} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$ and let $P \in \mathcal{L}(\mathcal{H}, Z)$ be a known disturbance operator. The exogenous system generating reference signals y_{ref} and disturbance signals $\mathcal{U}_{dist}$ is defined on the state space $\mathcal{H}$ as*

$$\dot{w}(t) = S|_{\mathcal{H}} w(t), \quad w(0) = w_0 \in \mathcal{H}, \tag{2.2a}$$

$$y_{ref}(t) = Qw(t), \quad t \in \mathbb{R}, \tag{2.2b}$$

$$\mathcal{U}_{dist}(t) = Pw(t), \quad t \in \mathbb{R}, \tag{2.2c}$$

where (2.2a) *is to be considered in the mild sense.*

By the above discussion it is easy to see that every reference signal y_{ref} in $\mathcal{H}$ can be obtained from (2.2b) by choosing $w(0) = y_{ref}$. Although we do not have such

precise knowledge of the disturbance signals $\mathcal{U}_{dist}$, the relation (2.2c) implies that the dynamical behavior of the disturbances must be similar to that of the reference signals: For example, if $\mathcal{H}$ consists of only p-periodic functions, then also every disturbance is a p-periodic function.

2.3. The feedforward regulation problem (FRP).

DEFINITION 2.2 (FRP). *The task in the feedforward regulation problem is to find operators $K \in \mathcal{L}(Z,H)$ and $L \in \mathcal{L}(\mathcal{H},H)$ having the following properties:*

1. *The pair (A,B) is strongly stabilizable using K, i.e., $A+BK$ generates a strongly stable C_0-semigroup T_{A+BK} on Z.*
2. *As the control law $u(t) = Kz(t) + Lw(t)$ is applied, in the extended system on $Z \times \mathcal{H}$ given by*

$$\dot{z}(t) = (A+BK)z(t) + (BL+P)w(t), \quad t \geq 0, \tag{2.3a}$$

$$\dot{w}(t) = S|_{\mathcal{H}} w(t) \tag{2.3b}$$

the tracking error $e(t) = y(t) - y_{ref}(t) = (C+DK)z(t) + (DL-Q)w(t) \to 0$ as $t \to \infty$ regardless of the initial conditions $z(0) \in Z$ and $w(0) \in \mathcal{H}$.

We hasten to emphasize that in the formulation of the FRP above, we do not require exponential stabilizability of the pair (A,B), as in [3, 16]. The concept of strong stability is more general than that of exponential stability, and it may be the more realistic one in applications. For example, for many partial differential equations it is known that solutions are stable, but no uniform decay rate exists [19]. Furthermore, that we do not require exponential stabilizability of (A,B) turns out to be important in the solution of the error feedback regulation problem described in the next subsection.

2.4. The error feedback regulation problem (EFRP). In practice the state $z(t)$ of the plant, which is used in the FRP, is usually not directly available for measurement. In this subsection we pose an output regulation problem which relies only on information on the outputs of the systems.

DEFINITION 2.3 (EFRP). *The task is to find an error feedback controller of the form*

$$\dot{x}(t) = Fx(t) + Ge(t), \quad x(0) \in X, \quad t \geq 0, \tag{2.4a}$$

$$u(t) = Jx(t) \tag{2.4b}$$

on some Banach state space X where F generates a C_0-semigroup, $G \in \mathcal{L}(H,X)$, and $J \in \mathcal{L}(X,H)$. The controller must satisfy the following requirements:

1. *In the closed loop system for $t \geq 0$,*

$$\dot{z}(t) = Az(t) + BJx(t) + Pw(t), \tag{2.5a}$$

$$\dot{x}(t) = GCz(t) + (F+GDJ)x(t) - GQw(t), \tag{2.5b}$$

$$\dot{w}(t) = S|_{\mathcal{H}} w(t), \tag{2.5c}$$

$$e(t) = Cz(t) + DJx(t) - Qw(t), \tag{2.5d}$$

the semigroup $T_{\mathcal{A}}(t)$ generated by $\mathcal{A} = \left(\begin{smallmatrix} A & BJ \\ GC & F+GDJ \end{smallmatrix}\right)$ in $Z \times X$ is strongly stable.

2. *The tracking error $e(t) \to 0$ as $t \to \infty$ for any initial conditions $z(0) \in Z$, $x(0) \in X$, and $w(0) \in \mathcal{H}$.*

Remark 2.4. Byrnes et al. [3] considered a similar error feedback regulation problem for finite-dimensional exosystems. However, they require exponential stability of $T_{\mathcal{A}}(t)$ and zero feedthrough, i.e., $D = 0$. It turns out that in our case exponential stability of $T_{\mathcal{A}}(t)$ is an unrealistic requirement if we plan to use the controller proposed by Byrnes et al. [3]. This is because their operator F (and hence also the closed loop system operator $\mathcal{A}$) contains a copy of $S|_{\mathcal{H}}$ according to the internal model principle. But it is well known that if $\mathcal{H}$ is infinite-dimensional, then $S|_{\mathcal{H}} + \Delta$ does not generate an exponentially stable C_0-semigroup for any compact operator $\Delta \in \mathcal{L}(\mathcal{H})$ (cf. Corollary 3.58 in [18]). Fortunately, although by the above we cannot in general require exponential stability of $T_{\mathcal{A}}(t)$ if $\mathcal{H}$ is infinite-dimensional, it is quite often possible to achieve strong stability of this closed loop semigroup even for infinite-dimensional spaces $\mathcal{H}$.

3. Sufficient conditions for the solvability of the FRP. In this section we show that if the so-called regulator equations can be solved, and if the pair (A, B) can be strongly stabilized, then the FRP can be solved. It is known that if (A, B) can actually be exponentially stabilized, then also a converse result is true, namely, that in this case exponentially fast output regulation implies solvability of the regulator equations (see [14]). However, in the case that (A, B) is only strongly stabilizable, solvability of the FRP alone does not imply solvability of the regulator equations; necessary conditions for the solvability of the regulator equations and the FRP are presented in section 6.

THEOREM 3.1. *Assume that the pair (A, B) is strongly stabilizable using $K \in \mathcal{L}(H, Z)$. If there exist $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ and $\Gamma \in \mathcal{L}(\mathcal{H}, H)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(A)$ and the regulator equations*

$$A\Pi + B\Gamma + P = \Pi S|_{\mathcal{H}} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{3.1a}$$

$$C\Pi + D\Gamma = Q \quad \text{in } \mathcal{H} \tag{3.1b}$$

are satisfied, then the control law $u(t) = Kz(t) + (\Gamma - K\Pi)w(t)$ solves the FRP.

Proof. Since by assumption $A + BK$ generates the strongly stable C_0-semigroup $T_{A+BK}(t)$, we only need to verify condition 2 in Definition 2.2. Let $L = \Gamma - K\Pi \in \mathcal{L}(\mathcal{H}, H)$. Since $\Pi S|_{\mathcal{H}} = (A + BK)\Pi + BL + P$ in $\mathcal{D}(S|_{\mathcal{H}})$, it is easy to see that

$$\begin{aligned}
&\int_0^t T_{A+BK}(t-\tau)(BL+P)T_S(\tau)|_{\mathcal{H}} w\,d\tau \\
&\quad = \int_0^t T_{A+BK}(t-\tau)\big[\Pi S|_{\mathcal{H}} - (A+BK)\Pi\big]T_S(\tau)|_{\mathcal{H}} w\,d\tau && (3.2)\\
&\quad = \int_0^t \frac{d}{d\tau} T_{A+BK}(t-\tau)\Pi T_S(\tau)|_{\mathcal{H}} w\,d\tau && (3.3)\\
&\quad = \Pi T_S(t)|_{\mathcal{H}} w - T_{A+BK}(t)\Pi w && (3.4)
\end{aligned}$$

for every $w \in \mathcal{D}(S|_{\mathcal{H}})$ and $t \geq 0$. By suitable density arguments it is in fact true that

$$\int_0^t T_{A+BK}(t-\tau)(BL+P)T_S(\tau)|_{\mathcal{H}} w\,d\tau = \Pi T_S(t)|_{\mathcal{H}} w - T_{A+BK}(t)\Pi w \tag{3.5}$$

for every $w \in \mathcal{H}$ and $t \geq 0$.

Consider then the composite operator $\mathcal{A}$ on the extended state space $Z \times \mathcal{H}$ (see (2.3)) defined as

$$\mathcal{A} = \begin{pmatrix} A + BK & BL + P \\ 0 & S|_{\mathcal{H}} \end{pmatrix}. \tag{3.6}$$

Since $A+BK$ generates the C_0-semigroup $T_{A+BK}(t)$ on Z and $S|_{\mathcal{H}}$ generates the C_0-(semi)group $T_S(t)|_{\mathcal{H}}$ on $\mathcal{H}$, it is clear that $\mathcal{A}$ generates a C_0-semigroup $T_{\mathcal{A}}(t)$ on $Z\times\mathcal{H}$, because $BL+P\in\mathcal{L}(\mathcal{H},Z)$ (see also [6, Lemma 3.2.2]). An easy calculation reveals that this semigroup is given by

$$T_{\mathcal{A}}(t)=\begin{pmatrix} T_{A+BK}(t) & \int_0^t T_{A+BK}(t-\tau)(BL+P)T_S(\tau)|_{\mathcal{H}}\,d\tau \\ 0 & T_S(t)|_{\mathcal{H}}\end{pmatrix} \tag{3.7}$$

$$=\begin{pmatrix} T_{A+BK}(t) & \Pi T_S(t)|_{\mathcal{H}}-T_{A+BK}(t)\Pi \\ 0 & T_S(t)|_{\mathcal{H}}\end{pmatrix}. \tag{3.8}$$

Let $z(0)=z_0\in Z$ and $w(0)=w_0\in\mathcal{H}$ be arbitrary. Then

$$T_{\mathcal{A}}(t)\left(\begin{smallmatrix} z_0\\ w_0\end{smallmatrix}\right)=\begin{pmatrix} T_{A+BK}(t)(z_0-\Pi w_0)+\Pi T_S(t)|_{\mathcal{H}}w_0\\ T_S(t)|_{\mathcal{H}}w_0\end{pmatrix}. \tag{3.9}$$

Since by (3.1b) we have $(C+DK)\Pi+DL-Q=C\Pi+D\Gamma-Q=0$, the explicit expression for the norm of the tracking error $e(t)$ is as follows:

$$\|e(t)\|=\|(C+DK)T_{A+BK}(t)(z_0-\Pi w_0)+(C\Pi+D\Gamma-Q)T_S(t)|_{\mathcal{H}}w_0\| \tag{3.10}$$

$$=\|(C+DK)T_{A+BK}(t)(z_0-\Pi w_0)\|\quad\forall t\geq 0 \tag{3.11}$$

Since $T_{A+BK}(t)$ is strongly stable and $C+DK\in\mathcal{L}(Z,H)$, we have that $e(t)\to 0$ as $t\to\infty$. This shows that the tracking condition of Definition 2.2 is also satisfied. The proof is then complete. □

Remark 3.2. In the case of a finite-dimensional space $H=\mathbb{C}^M$, $M\in\mathbb{N}$, it is sufficient in Theorem 3.1 for asymptotic tracking to occur that the pair (A,B) is merely weakly stabilizable. In this case $C+DK\in\mathcal{L}(Z,\mathbb{C}^M)$ and so $P_{e_k}(C+DK)\in\mathcal{L}(Z,\mathbb{C})$ for all $1\leq k\leq M$. Here $(e_j)_{1\leq j\leq N}$ is the canonical orthonormal basis of $\mathbb{C}^M$ and $P_{e_k}=\langle\cdot,e_k\rangle\in\mathcal{L}(\mathbb{C}^M,\mathbb{C})$. Hence for every $z\in Z$ and for every $1\leq k\leq M$ we have $\lim_{t\to\infty}P_{e_k}(C+DK)T_{A+BK}(t)z=0$ (see (3.11)). Thus also $\lim_{t\to\infty}(C+DK)T_{A+BK}(t)z=0$.

4. Sufficient conditions for the solvability of the EFRP. In this section we show that the regulator equations can be used to solve the EFRP too. In addition, it turns out to be possible to obtain a rather complete description of suitable error feedback controllers (2.4) in terms of solutions to another regulator equations—the so-called regulator equations for the error feedback controller—which have the same form as (3.1). We point out that to our knowledge no such description has been available for infinite-dimensional plants: Byrnes et al. [3], for example, give *one possible choice* for the stabilizing and regulating dynamic controller (2.4).

Our main result in this section is Theorem 4.1 below. Notable in its proof is the use of a marvelous argument due to Francis (see also Byrnes et al. [3]): The error feedback control problem is interpreted as a feedforward control problem for the extended system, and this problem is then solved in terms of the FRP.

THEOREM 4.1. *Assume that there exist* $\Pi\in\mathcal{L}(\mathcal{H},Z)$ *and* $\Gamma\in\mathcal{L}(\mathcal{H},H)$ *such that* $\Pi(\mathcal{D}(S|_{\mathcal{H}}))\subset\mathcal{D}(A)$ *and the regulator equations* (3.1) *are satisfied. If the parameters* F, G, *and* J *of the controller* (2.4) *can be chosen such that*

1. *the operator* $\mathcal{A}=\left(\begin{smallmatrix} A & BJ\\ GC & F+GDJ\end{smallmatrix}\right)$ *generates a strongly stable* C_0*-semigroup* $T_{\mathcal{A}}(t)$ *on* $Z\times X$*; and*
2. *there exists* $\Lambda\in\mathcal{L}(\mathcal{H},X)$ *such that* $\Lambda(\mathcal{D}(S|_{\mathcal{H}}))\subset\mathcal{D}(F)$ *and the regulator equations for the error feedback controller,*

$$F\Lambda=\Lambda S|_{\mathcal{H}}\quad\text{in }\mathcal{D}(S|_{\mathcal{H}}), \tag{4.1a}$$

$$J\Lambda = \Gamma \quad \text{in } \mathcal{H}, \tag{4.1b}$$

are satisfied, then with this triplet (F, G, J) the EFRP is solvable.

Proof. Let $\Theta(t) = \binom{z(t)}{x(t)} \in Z \times X$ and define

$$\mathcal{A} = \begin{pmatrix} A & BJ \\ GC & F + GDJ \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathcal{P} = \begin{pmatrix} P \\ -GQ \end{pmatrix}, \quad \mathcal{C} = \begin{pmatrix} C & DJ \end{pmatrix}, \quad \mathcal{D} = 0. \tag{4.2}$$

Then consider the system

$$\dot{\Theta}(t) = \mathcal{A}\Theta(t) + \mathcal{B}u(t) + \mathcal{P}w(t), \quad \Theta(0) \in Z \times X, \quad t \geq 0, \tag{4.3a}$$

$$y(t) = \mathcal{C}\Theta(t) + \mathcal{D}u(t), \quad t \geq 0, \tag{4.3b}$$

with $\dot{w}(t) = S|_{\mathcal{H}} w(t)$. Since the regulator equations (3.1) and the regulator equations (4.1) for the error feedback controller are satisfied, we have $\Pi S|_{\mathcal{H}} = A\Pi + BJ\Lambda + P$ and $\Lambda S|_{\mathcal{H}} = F\Lambda = GC\Pi + (F + GDJ)\Lambda - GQ$ in $\mathcal{D}(S|_{\mathcal{H}})$. Hence

$$\begin{pmatrix} \Pi \\ \Lambda \end{pmatrix} S|_{\mathcal{H}} = \begin{pmatrix} A & BJ \\ GC & F + GDJ \end{pmatrix} \begin{pmatrix} \Pi \\ \Lambda \end{pmatrix} + \begin{pmatrix} P \\ -GQ \end{pmatrix} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{4.4}$$

$$Q = \begin{pmatrix} C & DJ \end{pmatrix} \begin{pmatrix} \Pi \\ \Lambda \end{pmatrix} \quad \text{in } \mathcal{H}, \tag{4.5}$$

which shows that the regulator equations (3.1) for the system (4.3) have a solution. Consequently, by Theorem 3.1 and the strong stability assumption, for the system (4.3) the FRP is solvable for the control law $u(t) = \mathcal{K}\Theta(t) + \mathcal{L}w(t) \equiv 0$ (recall that $\mathcal{B} = 0$ and $\mathcal{D} = 0$). This implies that in (2.5d) the tracking error $e(t) = Cz(t) + DJx(t) - Qw(t) = (\mathcal{C} + \mathcal{D}\mathcal{K})\Theta(t) + (\mathcal{D}\mathcal{L} - Q)w(t) \to 0$ as $t \to \infty$ regardless of the initial conditions $z(0)$, $x(0)$, and $w(0)$. ∎

The following corollary generalizes heorem IV.2 in [3]. It presents one possible choice for the triplet (F, G, J) in the solution of EFRP under two stability assumptions and the assumption $D = 0$.

COROLLARY 4.2. *Let $X = Z \times \mathcal{H}$. Assume that $D = 0$ and that*

1. *there exist $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ and $\Gamma \in \mathcal{L}(\mathcal{H}, H)$ which solve the regulator equations (3.1);*
2. *the pair (A, B) is exponentially stabilizable using $K \in \mathcal{L}(Z, H)$; and*
3. *there exist $G_1 \in \mathcal{L}(H, Z)$ and $G_2 \in \mathcal{L}(H, \mathcal{H})$ such that the operator $\mathcal{A}_s = \begin{pmatrix} A - G_1 C & P + G_1 Q \\ -G_2 C & S|_{\mathcal{H}} + G_2 Q \end{pmatrix}$ generates a strongly stable C_0-semigroup $T_{\mathcal{A}_s}(t)$ in X.*

Let

$$F = \begin{pmatrix} A + BK - G_1 C & P + B(\Gamma - K\Pi) + G_1 Q \\ -G_2 C & S|_{\mathcal{H}} + G_2 Q \end{pmatrix}, \quad J = \begin{pmatrix} K & \Gamma - K\Pi \end{pmatrix}, \quad \text{and} \quad G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}. \tag{4.6}$$

Then with this choice of (F, G, J) the EFRP is solvable.

Proof. Let $\Lambda = \binom{\Pi}{I} \in \mathcal{L}(\mathcal{H}, Z \times \mathcal{H})$. Then $J\Lambda = K\Pi + \Gamma - K\Pi = \Gamma$ and it is elementary to verify that $\Lambda S|_{\mathcal{H}} = F\Lambda$ in $\mathcal{D}(S|_{\mathcal{H}})$. Hence Λ solves the regulator equations (4.1) for the error feedback controller (4.6).

Consider the operator

$$\mathcal{A} = \begin{pmatrix} A & BJ \\ GC & F \end{pmatrix} = \begin{pmatrix} A & BK & B(\Gamma - K\Pi) \\ G_1 C & A + BK - G_1 C & P + B(\Gamma - K\Pi) + G_1 Q \\ G_2 C & -G_2 C & S|_{\mathcal{H}} + G_2 Q \end{pmatrix}. \tag{4.7}$$

If we can establish that $\mathcal{A}$ generates a strongly stable C_0-semigroup $T_{\mathcal{A}}(t)$, then the error feedback controller (4.6) solves the EFRP by Theorem 4.1.

Applying a similarity transform U given as

$$U = \begin{pmatrix} I & 0 & 0 \\ I & -I & 0 \\ 0 & 0 & -I \end{pmatrix} \tag{4.8}$$

on $Z \times Z \times \mathcal{H}$ to $\mathcal{A}$ we obtain the operator $\widetilde{\mathcal{A}} = U\mathcal{A}U$ having the expression

$$\widetilde{\mathcal{A}} = \begin{pmatrix} A+BK & -BK & -B(\Gamma - K\Pi) \\ 0 & A - G_1C & P + G_1Q \\ 0 & -G_2C & S|_{\mathcal{H}} + G_2Q \end{pmatrix} := \begin{pmatrix} A+BK & \Delta \\ 0 & \mathcal{A}_s \end{pmatrix}. \tag{4.9}$$

By our assumption $\mathcal{A}_s$ generates a strongly stable C_0-semigroup in X and $A + BK$ generates an exponentially stable C_0-semigroup in Z. Clearly the C_0-semigroup generated by $\widetilde{\mathcal{A}}$ on $Z \times X$ is given by

$$T_{\widetilde{\mathcal{A}}}(t) = \begin{pmatrix} T_{A+BK}(t) & \int_0^t T_{A+BK}(t-s)\Delta T_{\mathcal{A}_s}(s)ds \\ 0 & T_{\mathcal{A}_s}(t) \end{pmatrix}. \tag{4.10}$$

Consequently $T_{\widetilde{\mathcal{A}}}(t)$ is strongly stable if

$$\lim_{t\to\infty} \int_0^t T_{A+BK}(t-s)\Delta T_{\mathcal{A}_s}(s)xds = 0 \quad \forall x \in X. \tag{4.11}$$

But (4.11) holds by Proposition 5.6.1 in [1]. This proves that also $T_{\mathcal{A}}(t)$ is strongly stable. □

5. When does $\mathcal{A}_s$ generate a strongly stable semigroup? In this section we discuss the strong detectability of the pair $(\mathcal{A}_0, \mathcal{C})$ where $\mathcal{A}_0 = \left(\begin{smallmatrix} A & P \\ 0 & S|_{\mathcal{H}} \end{smallmatrix}\right)$ and $\mathcal{C} = \begin{pmatrix} C & -Q \end{pmatrix}$. In other words, we want to find $\mathcal{G} \in \mathcal{L}(H, Z\times\mathcal{H})$ such that $\mathcal{A}_s = \mathcal{A}_0 + \mathcal{G}\mathcal{C}$ generates a strongly stable C_0-semigroup $T_{\mathcal{A}_s}(t)$. This is assumption 3 in Corollary 4.2 and consequently the question deserves some attention.

In the finite-dimensional case it is well known that the pair $(\mathcal{A}_0, \mathcal{C})$ is strongly (and hence also exponentially) detectable if the pair $(\mathcal{A}_0, \mathcal{C})$ is observable. However, the infinite-dimensional setting is much more complicated [6]. In the case of an infinite-dimensional plant and a finite-dimensional exogenous system Byrnes et al. [3] state that undetectability of the pair $(\mathcal{A}_0, \mathcal{C})$ indicates a redundancy in the exogenous system, which can be removed by eliminating superfluous exosystem variables. Based on this remark they assume that the pair $(\mathcal{A}_0, \mathcal{C})$ is exponentially detectable. However, if the exogenous system is not finite-dimensional, then it is unreasonable to assume that the pair $(\mathcal{A}_0, \mathcal{C})$ would be exponentially detectable. In fact, if $\mathcal{H}$ is infinite-dimensional, then by Corollary 3.58 in [18] the operator $S|_{\mathcal{H}} + \Delta$ does not generate an exponentially stable semigroup for any compact $\Delta \in \mathcal{L}(\mathcal{H})$; hence it is not realistic to assume exponential detectability of $(\mathcal{A}_0, \mathcal{C})$ if $\mathcal{H}$ is infinite-dimensional.

In the following two subsections we discuss the strong detectability of the pair $(\mathcal{A}_0, \mathcal{C})$. First, a method based on the use of a Riccati equation is presented. Thereafter we propose a direct approach in which we assume the pair (A, C) in the plant to be exponentially detectable.

5.1. Strong stability via a Riccati equation. Algebraic Riccati equations provide a useful method for the study of strong detectability of the pair $(\mathcal{A}_0, \mathcal{C})$. Using this approach, in Proposition 5.1 we present sufficient conditions for the existence of a strongly stabilizing output injection operator; the operator itself can be found by solving a suitable Riccati equation. Recall in that if H is a Hilbert space, then $U \in \mathcal{L}(H)$ is coercive provided $\langle Uh, h\rangle \geq \epsilon \|h\|^2$ for some $\epsilon > 0$ and all $h \in H$. In the following $\mathcal{A}_0^*$ denotes the adjoint operator of $\mathcal{A}_0$.

PROPOSITION 5.1. *Assume the following:*

1. *H and $Z \times \mathcal{H}$ are (separable) Hilbert spaces.*
2. *$\mathcal{A}_0$ generates a contraction C_0-semigroup in $Z \times \mathcal{H}$.*
3. *The pair $(\mathcal{A}_0^*, \mathcal{C})$ is approximately observable.*
4. *$\mathcal{A}_0$ has compact resolvent.*

Then for any coercive operators $U = U^* \in \mathcal{L}(H)$ and $R = R^* \in \mathcal{L}(H)$ the Riccati equation

$$\mathcal{A}_0 \Delta z + \Delta \mathcal{A}_0^* z - \Delta \mathcal{C}^* R^{-1} \mathcal{C} \Delta z + \mathcal{C}^* U^{-1} \mathcal{C} z = 0 \quad \forall z \in \mathcal{D}(\mathcal{A}_0) \tag{5.1}$$

has a unique self-adjoint solution $\Delta \in \mathcal{L}(Z \times \mathcal{H})$ such that $\mathcal{A}_s = \mathcal{A}_0 - \Delta^* \mathcal{C}^* (R^{-1})^* \mathcal{C}$ generates a strongly stable C_0-semigroup.

Proof. It is evident that since $\mathcal{A}_0$ has compact resolvent, so does $\mathcal{A}_0^*$ because for real $\alpha \in \rho(\mathcal{A}_0)$ we have $R(\alpha, \mathcal{A}_0^*) = R(\alpha, \mathcal{A}_0)^*$ (cf. Lemma A.3.65 in [6], Theorem 7.3 in [22], and equality (A.3.15) in [6]). Furthermore, $\mathcal{A}_0^*$ also generates a contraction semigroup on $Z \times \mathcal{H}$ because $\|T_{\mathcal{A}_0}(t)\| = \|T_{\mathcal{A}_0^*}(t)\| \leq 1$ for each $t \geq 0$. By Theorem 4 in [7], the algebraic Riccati equation (3.1) has a unique self-adjoint solution $\Delta \in \mathcal{L}(Z \times \mathcal{H})$. By Corollary 5 in [7] the operator $\mathcal{A}_s^* = \mathcal{A}_0^* - \mathcal{C}^* R^{-1} \mathcal{C} \Delta$ generates a strongly stable C_0-semigroup. Consequently its adjoint $\mathcal{A}_s = \mathcal{A}_0 - \Delta \mathcal{C}^* (R^{-1})^* \mathcal{C}$ generates a weakly stable C_0-semigroup. But since $\mathcal{A}_0$ has compact resolvent, the boundedly perturbed operator $\mathcal{A}_s$ also has compact resolvent (cf. [9, p. 159]). This implies that the semigroup generated by $\mathcal{A}_s$ is in fact strongly stable (see, e.g., [7, Lemma 1] or [18, Proposition 3.21]). ∎

If the assumptions of Proposition 5.1 are met, we may choose as the strongly stabilizing output injection $\mathcal{G} = -\Delta \mathcal{C}^* (R^{-1})^*$.

Although neither exponential stabilizability nor detectability is required of the plant, the assumptions of Proposition 5.1 may seem somewhat restrictive. However, assumption 1 can be met if H and Z are Hilbert spaces, and if one considers the asymptotic tracking of periodic reference signals; the state space $\mathcal{H}$ can then be chosen as some Sobolev-type space of periodic functions (cf. [16]). Assumptions 2 and 3 are of a technical nature, but they are easy to verify if, say, there are no disturbances. Observe that $S|_{\mathcal{H}}$ always generates a contraction C_0-group. On the other hand, $\mathcal{A}_0$ has compact resolvent if $S|_{\mathcal{H}}$ and A have compact resolvents. In the case of periodic reference signals (i.e., whenever $T_S(t)|_{\mathcal{H}}$ is periodic), $S|_{\mathcal{H}}$ often has compact resolvent [9]. Moreover many differential operators A have compact resolvents [9]. Hence assumptions 2–4 of Proposition 5.1 can be met in many important special cases.

5.2. A direct method for exponentially detectable pairs (A, C). The difficulty in obtaining stability of $T_{\mathcal{A}_s}(t)$ lies in the fact that $\mathcal{A}_0$ contains a copy of $S|_{\mathcal{H}}$ which is hard to stabilize. We next show that this unfortunate state of matters can be, up to a certain point, compensated by good stabilizability properties of the plant.

THEOREM 5.2. *Let $G_1 \in \mathcal{L}(H,Z)$ be such that $A+G_1C$ generates an exponentially stable C_0-semigroup $T_{A+G_1C}(t)$. Define the bounded linear operator $Y : \mathcal{H} \to Z$ as*

$$Yw = \int_0^\infty T_{A+G_1C}(t)(G_1Q-P)T_S(-t)|_{\mathcal{H}} w dt \quad \forall w \in \mathcal{H}. \tag{5.2}$$

If $S|_{\mathcal{H}} - G_2(CY+Q)$ generates a strongly stable C_0-semigroup, then so does $\mathcal{A}_s = \mathcal{A} + \mathcal{GC}$. Here

$$\mathcal{G} = \left(\begin{smallmatrix} G_1 - YG_2 \\ G_2 \end{smallmatrix}\right). \tag{5.3}$$

Proof. By the results of [23] we know that $YS|_{\mathcal{H}} = (A+G_1C)Y + G_1Q - P$ in $\mathcal{D}(S|_{\mathcal{H}})$. It is then easy to see that

$$\begin{pmatrix} I & Y \\ 0 & I \end{pmatrix} \left[\begin{pmatrix} A & P \\ 0 & S|_{\mathcal{H}} \end{pmatrix} + \begin{pmatrix} G_1 - YG_2 \\ G_2 \end{pmatrix} \begin{pmatrix} C & -Q \end{pmatrix} \right] \begin{pmatrix} I & -Y \\ 0 & I \end{pmatrix} \tag{5.4}$$

$$= \begin{pmatrix} A+G_1C & 0 \\ G_2C & S|_{\mathcal{H}} - G_2(CY+Q) \end{pmatrix} = \mathcal{A}^s \tag{5.5}$$

so that $\mathcal{A}_s$ is similar to the operator $\mathcal{A}^s$ in (5.5). The semigroup $T_{\mathcal{A}^s}(t)$ generated by $\mathcal{A}^s$ is given by

$$T_{\mathcal{A}^s}(t) = \begin{pmatrix} T_{A+G_1C}(t) & 0 \\ \int_0^t T_{S|_{\mathcal{H}}-G_2(CY+Q)}(t-s)G_2CT_{A+G_1C}(s)ds & T_{S|_{\mathcal{H}}-G_2(CY+Q)}(t) \end{pmatrix}, \quad t \geq 0, \tag{5.6}$$

where $T_{S|_{\mathcal{H}}-G_2(CY+Q)}(t)$ is the strongly stable C_0-semigroup generated by $S|_{\mathcal{H}} - G_2(CY+Q)$. Since by our assumption $T_{A+G_1C}(t)$ is exponentially stable, we only need to show that

$$\lim_{t\to\infty} \int_0^t T_{S|_{\mathcal{H}}-G_2(CY+Q)}(t-s)G_2CT_{A+G_1C}(s)wds = 0 \quad \forall w \in \mathcal{H} \tag{5.7}$$

to ensure that $\mathcal{A}^s$ (and hence also $\mathcal{A}_s$) generates a strongly stable C_0-semigroup. However, the fact that (5.7) holds for each $w \in \mathcal{H}$ follows immediately from [1, Proposition 5.6.4.]. ∎

Remark 5.3. The result of Theorem 5.2 could also be formulated as follows. Assume that for some $G_1 \in \mathcal{L}(H,Z), G_2 \in \mathcal{L}(H,\mathcal{H})$, and $R \in \mathcal{L}(\mathcal{H},H)$ the operator $A+G_1C$ generates an exponentially stable C_0-semigroup on Z and that $S|_{\mathcal{H}} - G_2R$ generates a strongly stable C_0-semigroup on $\mathcal{H}$. If the operator equations

$$YS|_{\mathcal{H}} = AY + G_1R - P \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{5.8a}$$

$$R = CY + Q \quad \text{in } \mathcal{H} \tag{5.8b}$$

have a solution Y, then $\mathcal{A}_s = \mathcal{A} + \mathcal{GC}$ generates a strongly stable C_0-semigroup on $Z \times \mathcal{H}$ if we utilize the operator $\mathcal{G}$ of (5.3). Observe that now $(A+G_1C)Y + G_1Q - P = AY + G_1R - P = YS|_{\mathcal{H}}$ in $\mathcal{D}(S|_{\mathcal{H}})$, so that the solution operator Y (if it exists) is given by (5.2).

6. Characterizations for the solvability of the regulator equations and the regulation problems. In the results of sections 3 and 4 there are two key ingredients in the solution of the output regulation problems: strong stability of an appropriate semigroup and solvability of the regulator equations. In particular, if

strong stability of the semigroup has been established, then the output regulation problem is solvable provided that certain regulator equations are solvable. In this section we discuss a converse question: Is it necessary to solve the regulator equations in order to solve the corresponding output regulation problem? An answer to this question also provides a deeper insight into what happens in the closed loop control system after the application of a control law involving solutions of the regulator equations. For example, it is well known [3, 16] that the regulator equations express the existence of a subspace in the extended system on which the actual and reference outputs coincide and which is invariant for the extended semigroup.

For exponentially stabilizable pairs (A, B) and finite-dimensional exogenous systems it has been shown in [3] that solvability of the FRP is equivalent to solvability of the regulator equations (3.1). This result was generalized for the infinite-dimensional exogenous systems (2.1) in [14]. We shall see in this section that in the case of strongly stabilizable pairs (A, B) things get slightly more complicated. In order to study the problem, we need some preliminary results and definitions. An important role in this section will be played by closed, translation invariant, operator invariant spaces (CTO spaces) [24]. We will show that if $\mathcal{H}$ is a CTO space which does not contain functions vanishing at $+\infty$, then solvability of the output regulation problems FRP and EFRP imply solvability of the regulator equations under certain additional conditions *not including* that of exponential stabilizability.

DEFINITION 6.1. *A subspace $\mathcal{H} \subset BUC(\mathbb{R}, H)$ is CTO if*

1. *the space $\mathcal{H}$ is closed with respect to* sup*-norm and translation invariant (i.e., invariant for $T_S(t)$);*
2. *the space $\mathcal{H}$ is operator invariant, i.e., for every $f \in \mathcal{H}$ and every $R \in \mathcal{L}(\mathcal{H}, H)$ the function $g : t \to RT_S(t)|_{\mathcal{H}} f$ is in $\mathcal{H}$.*

Example 6.2. The space $\mathcal{H} = P_p(\mathbb{R}, H) = \{ f \in BUC(\mathbb{R}, H) \mid f(x + p) = f(x)$ for all $x \in \mathbb{R} \}$ of p-periodic H-valued bounded uniformly continuous functions is evidently a CTO space.

Example 6.3. The space $\mathcal{H} = AP(\mathbb{R}, H) \subset BUC(\mathbb{R}, H)$ of H-valued almost periodic functions is a CTO space. In fact, clearly $AP(\mathbb{R}, H)$ is closed and translation invariant [1, 13]. Moreover, by definition, any almost periodic function $f \in AP(\mathbb{R}, H)$ can be approximated in the sup-norm by trigonometric polynomials $\sum_{n \in I} f_n e^{i\omega_n \cdot}$ (where I is a finite set of indices). Since $\|RT_S(t)|_{\mathcal{H}}\| \leq \|R\|$ for all $t \in \mathbb{R}$ and since $RT_S(t)|_{\mathcal{H}} \sum_{n \in I} f_n e^{i\omega_n \cdot}$ is again a trigonometric polynomial, it is evident that $RT_S(t)|_{\mathcal{H}} f$ can be uniformly approximated by trigonometric polynomials whenever the same is true for f. In conclusion, $AP(\mathbb{R}, H)$ is a CTO space.

Example 6.4. An even less trivial example of a CTO space is $\mathcal{H} = \{ f \in BUC(\mathbb{R}, H) \mid \mathrm{sp}(f) \subset \Lambda \}$, where $sp(f)$ denotes the Carleman spectrum of f [13, 24] and $\Lambda \subset \mathbb{R}$ is a closed set. We refer the reader to [24, 13] for more details (see in particular [24, Example 1]).

In what follows we derive necessary conditions for output regulation in the case that the reference signals are in some fixed CTO space $\mathcal{H}$. The following concept of regular admissibility of an operator will be crucial.

DEFINITION 6.5. *An operator $\Delta \in \mathcal{L}(\mathcal{H}, Z)$ is said to be regularly admissible for the semigroup $T_A(t)$ if there exists $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(A)$ and such that for every $w \in \mathcal{H}$ the function $t \to z(t) = \Pi T_S(t)|_{\mathcal{H}} w$ is a mild solution of the differential equation $\dot{z}(t) = Az(t) + \Delta T_S(t)|_{\mathcal{H}} w$ on the whole real line, i.e., z satisfies*

$$z(t) = T_A(t - s)z(s) + \int_s^t T_A(t - \tau)\Delta T_S(\tau)|_{\mathcal{H}} w d\tau \quad \forall t \geq s. \tag{6.1}$$

Remark 6.6. In (6.1) t and s need not be nonnegative. Moreover, $z(t) = \Pi T_S(t)|_{\mathcal{H}} w$, with $w \in \mathcal{H}$, need not be the *only* mild solution of the differential equation $\dot{z}(t) = Az(t) + \Delta T_S(t)|_{\mathcal{H}} w$ on the real line.

Remark 6.7. The significance of the concept of regular admissibility in our analysis is, roughly stated, the following: If an operator $\Delta \in \mathcal{L}(\mathcal{H}, Z)$ is regularly admissible for $T_A(t)$, then the dynamical behavior of some mild solution (namely, $\Pi T_S(t)|_{\mathcal{H}} w$) of the differential equation $\dot{z}(t) = Az(t) + \Delta T_S(t)|_{\mathcal{H}} w$ on the whole real line is solely determined by the dynamical behavior of the exosystem (2.2). For example, if $\mathcal{H} = P_p(\mathbb{R}, H)$ as in Example 6.2 and if $\Delta \in \mathcal{L}(\mathcal{H}, Z)$ is regularly admissible for $T_A(t)$, then the above differential equation has at least one p-periodic mild solution on the whole real line. This feature will be utilized in the subsequent results.

Lemma 6.8 below is one of the main results in this section. Observe that we do not need exponential stability of $T_A(t)$ in it.

LEMMA 6.8. *The operator equation* $\Pi S|_{\mathcal{H}} = A\Pi + \Delta$ *in* $\mathcal{D}(S|_{\mathcal{H}})$ *has a solution* $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ *if and only if* $\Delta \in \mathcal{L}(\mathcal{H}, Z)$ *is regularly admissible for* $T_A(t)$.

Proof. Let $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(A)$ and $\Pi S|_{\mathcal{H}} = A\Pi + \Delta$ in $\mathcal{D}(S|_{\mathcal{H}})$. Let $w \in \mathcal{H}$ be arbitrary and set $z(t) = \Pi T_S(t)|_{\mathcal{H}} w$. Then it is easy to see that

$$\int_s^t T_A(t-\tau)\Delta T_S(\tau)|_{\mathcal{H}} w d\tau = \Pi T_S(t)|_{\mathcal{H}} w - T_A(t-s)\Pi T_S(s)|_{\mathcal{H}} w \tag{6.2}$$

for every $t \geq s$. Hence for every $t \geq s$, we have

$$z(t) = T_A(t-s)\Pi T_S(s)|_{\mathcal{H}} w + \Pi T_S(t)|_{\mathcal{H}} w - T_A(t-s)\Pi T_S(s)|_{\mathcal{H}} w \tag{6.3}$$

$$= T_A(t-s)z(s) + \int_s^t T_A(t-\tau)\Delta T_S(\tau)|_{\mathcal{H}} w d\tau \tag{6.4}$$

so that $z(t)$ is a mild solution of the differential equation $\dot{z}(t) = Az(t) + \Delta T_S(t)|_{\mathcal{H}} w$ on the whole real line. In other words, Δ is regularly admissible for $T_A(t)$.

Conversely, suppose that there exists $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(A)$ and such that for every $w \in \mathcal{H}$ the function $t \to z(t) = \Pi T_S(t)|_{\mathcal{H}} w$ is a mild solution of the differential equation $\dot{z}(t) = Az(t) + \Delta T_S(t)|_{\mathcal{H}} w$ on the whole real line. Let $w \in \mathcal{D}(S|_{\mathcal{H}})$. Then since $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(A)$ and since $\mathcal{D}(S|_{\mathcal{H}})$ is invariant for $T_S(t)|_{\mathcal{H}}$, the function $t \to \Delta T_S(t)|_{\mathcal{H}} w$ is continuously differentiable, and we can differentiate both sides of the identity

$$\Pi T_S(t)|_{\mathcal{H}} w = T_A(t)\Pi w + \int_0^t T_A(t-\tau)\Delta T_S(\tau) w d\tau \quad \forall t \geq 0 \tag{6.5}$$

(cf. Proposition 1.3.6 in [1]) and set $t = 0$ to obtain $\Pi S|_{\mathcal{H}} = A\Pi + \Delta$ in $\mathcal{D}(S|_{\mathcal{H}})$. The proof is complete. □

Remark 6.9. If $T_A(t)$ is exponentially stable, then by Corollary 8 in [23] the operator equation $\Pi S|_{\mathcal{H}} = A\Pi + \Delta$ in $\mathcal{D}(S|_{\mathcal{H}})$ has a unique solution $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ for every $\Delta \in \mathcal{L}(\mathcal{H}, Z)$. Consequently, *every* $\Delta \in \mathcal{L}(\mathcal{H}, Z)$ is regularly admissible for an exponentially stable C_0-semigroup $T_A(t)$.

Lemma 6.8 shows that if $T_{A+BK}(t)$ is strongly stable and if we can solve the regulator equations (3.1), then we in fact obtain more than just the existence of a solution to the FRP: The operator $BL + P = B(\Gamma - K\Pi) + P$ is regularly admissible for $T_{A+BK}(t)$. In the case of p-periodic reference signals (with $\mathcal{H}$ as in Example 6.2) this implies that for every $w(0) \in \mathcal{H}$ equation (2.3a) has a p-periodic mild solution $z(t)$ in the sense of (6.1).

We now come to the other main results of this section. They provide—via Lemma 6.8—characterizations for the solvability of the output regulation problems FRP and EFRP. In the following we set $C_0^+(\mathbb{R},H)=\{\,f\in BUC(\mathbb{R},H)\mid f(t)\to 0 \text{ as } t\to+\infty\,\}$. The next theorem generalizes the results of [3, 14, 16] to the case in which the pair (A,B) is not necessarily exponentially stabilizable.

THEOREM 6.10. *Assume that $\mathcal{H}$ is a CTO space such that $\mathcal{H}\cap C_0^+(\mathbb{R},H)=\{0\}$. Let the pair (A,B) be strongly stabilizable using $K\in\mathcal{L}(Z,H)$. Then the FRP is solvable using the control law $u(t)=Kz(t)+Lw(t)$, where $L\in\mathcal{L}(\mathcal{H},H)$ and $BL+P$ is regularly admissible for $T_{A+BK}(t)$, if and only if there exists $\Pi\in\mathcal{L}(\mathcal{H},Z)$ and $\Gamma\in\mathcal{L}(\mathcal{H},H)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}}))\subset\mathcal{D}(A)$, $L=\Gamma-K\Pi$ and the regulator equations* (3.1) *are satisfied.*

Proof. If $\Pi S|_{\mathcal{H}}=A\Pi+B\Gamma+P$ in $\mathcal{D}(S|_{\mathcal{H}})$ and if $L=\Gamma-K\Pi$, then also $\Pi S|_{\mathcal{H}}=(A+BK)\Pi+BL+P$ in $\mathcal{D}(S|_{\mathcal{H}})$. Theorem 3.1 and Lemma 6.8 now directly yield the "if" part of the result, so we turn our attention to the "only if" part.

Now, Lemma 6.8 implies that whenever $BL+P$ is regularly admissible for $T_{A+BK}(t)$, there exists $\Pi\in\mathcal{L}(\mathcal{H},Z)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}}))\subset\mathcal{D}(A+BK)=\mathcal{D}(A)$ and $\Pi S|_{\mathcal{H}}=(A+BK)\Pi+BL+P$ in $\mathcal{D}(S|_{\mathcal{H}})$. Let $\Gamma=L+K\Pi\in\mathcal{L}(\mathcal{H},H)$. Then Π and Γ solve the first regulator equation (3.1a) and $L=\Gamma-K\Pi$.

We next show that also the second regulator equation (3.1b) is satisfied. Observe that since $\Pi S|_{\mathcal{H}}=(A+BK)\Pi+BL+P$ in $\mathcal{D}(S|_{\mathcal{H}})$, we have for every $w\in\mathcal{H}$ and every $t\geq 0$ that

$$\int_0^t T_{A+BK}(t-\tau)(BL+P)T_S(\tau)|_{\mathcal{H}}w\,d\tau=\Pi T_S(t)|_{\mathcal{H}}w-T_{A+BK}(t)\Pi w. \tag{6.6}$$

Consider then the composite operator $\mathcal{A}$ on the extended state space $Z\times\mathcal{H}$ (see (2.3)) defined as

$$\mathcal{A}=\begin{pmatrix}A+BK & BL+P\\ 0 & S|_{\mathcal{H}}\end{pmatrix}. \tag{6.7}$$

An easy calculation reveals that $\mathcal{A}$ generates the C_0-semigroup given as

$$T_{\mathcal{A}}(t)=\begin{pmatrix}T_{A+BK}(t) & \Pi T_S(t)|_{\mathcal{H}}-T_{A+BK}(t)\Pi\\ 0 & T_S(t)|_{\mathcal{H}}\end{pmatrix}. \tag{6.8}$$

Since the FRP is solved, for arbitrary $w(0)=w\in\mathcal{H}$ and $z(0)=\Pi w\in Z$ we have $\left(\begin{smallmatrix}z(t)\\ w(t)\end{smallmatrix}\right)=T_{\mathcal{A}}(t)\left(\begin{smallmatrix}\Pi w\\ w\end{smallmatrix}\right)=\left(\begin{smallmatrix}\Pi T_S(t)|_{\mathcal{H}}w\\ T_S(t)|_{\mathcal{H}}w\end{smallmatrix}\right)$ and the corresponding tracking error satisfies

$$e(t)=\big((C+DK)\Pi+DL-Q\big)T_S(t)|_{\mathcal{H}}w=(C\Pi+D\Gamma-Q)T_S(t)|_{\mathcal{H}}w\to 0 \text{ as } t\to\infty, \tag{6.9}$$

which implies that this tracking error function $e\in C_0^+(\mathbb{R},H)$. Furthermore, since $\mathcal{H}$ is CTO and $(C\Pi+D\Gamma-Q)\in\mathcal{L}(\mathcal{H},H)$, the function $t\to(C\Pi+D\Gamma-Q)T_S(t)|_{\mathcal{H}}w$ is also in $\mathcal{H}$. In other words, $e\in C_0^+(\mathbb{R},H)\cap\mathcal{H}=\{0\}$. Consequently $e(0)=(C\Pi+D\Gamma-Q)w=0$, and so also the second regulator equation (3.1b) is satisfied. ∎

Remark 6.11. Our assumption that $\mathcal{H}\cap C_0^+(\mathbb{R},H)=\{0\}$ corresponds to the assumption in [3] that the (finite-dimensional) exogenous system is neutrally stable. In [3] neutral stability means that the origin is Lyapunov stable forward and backward in time; it implies that the spectrum of the exosystem's generator is located on the

imaginary axis and that this matrix cannot have nontrivial Jordan blocks [3]. This, on the other hand, implies that no reference function can vanish at $+\infty$, which is precisely our assumption $\mathcal{H}\cap C_0^+(\mathbb{R},H)=\{0\}$. This assumption is very mild, because such functions could be asymptotically tracked using a control law $u(t)=Kz(t)$ where $K\in\mathcal{L}(Z,H)$ stabilizes the pair (A,B) strongly.

Theorem 3.1, Remark 6.9, and Theorem 6.10 immediately yield the following corollary; it was also proved in [14], with the assumption that $\mathcal{H}$ is a CTO space replaced by the requirement for exponentially fast tracking/rejection of the exogenous signals.

COROLLARY 6.12. *Assume that $\mathcal{H}$ is a CTO space such that $\mathcal{H}\cap C_0^+(\mathbb{R},H)=\{0\}$. Let the pair (A,B) be exponentially stabilizable using $K\in\mathcal{L}(Z,H)$. Then the FRP is solvable using the control law $u(t)=Kz(t)+Lw(t)$, where $L\in\mathcal{L}(\mathcal{H},H)$, if and only if there exists $\Pi\in\mathcal{L}(\mathcal{H},Z)$ and $\Gamma\in\mathcal{L}(\mathcal{H},H)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}}))\subset\mathcal{D}(A)$, $L=\Gamma-K\Pi$ and the regulator equations* (3.1) *are satisfied.*

For the solution of the EFRP we next obtain similar results. The following theorem generalizes the results in [3] to the case in which the closed loop system is not necessarily exponentially stable, the controller's parameters F,G, and J are not fixed, and the exogenous system may be infinite-dimensional.

THEOREM 6.13. *Let $\mathcal{H}$ be a CTO space such that $\mathcal{H}\cap C_0^+(\mathbb{R},H)=\{0\}$. Assume that the triplet (F,G,J) has been chosen such that $\mathcal{A}=\left(\begin{smallmatrix}A & BJ\\ GC & F+GDJ\end{smallmatrix}\right)$ generates a strongly stable C_0-semigroup on $Z\times\mathcal{H}$. Then the EFRP is solvable with this choice of (F,G,J) and the operator $\mathcal{P}=\left(\begin{smallmatrix}P\\ -GQ\end{smallmatrix}\right)$ is regularly admissible for the semigroup $T_{\mathcal{A}}(t)$ if and only if the regulator equations* (3.1) *and regulator equations* (4.1) *for the error feedback controller are satisfied for some $\Pi\in\mathcal{L}(\mathcal{H},Z)$, $\Gamma\in\mathcal{L}(\mathcal{H},H)$, and $\Lambda\in\mathcal{L}(\mathcal{H},X)$.*

Proof. If there exist Π,Γ, and Λ such that (3.1) and (4.1) are satisfied, then by Theorem 4.1 the triplet (F,G,J) solves the EFRP. Moreover, (4.4) and Lemma 6.8 show that $\mathcal{P}$ is regularly admissible for $T_{\mathcal{A}}(t)$. Therefore, we have proved the "if" part of the result, and we can turn our attention to the "only if" part.

Assume that the EFRP is solvable using a controller of the form (2.4) and that $\mathcal{P}$ is regularly admissible for $T_{\mathcal{A}}(t)$. Let $\Theta(t)=\left(\begin{smallmatrix}z(t)\\ x(t)\end{smallmatrix}\right)\in Z\times X$ and consider the system (4.3) with notation as in (4.2). Let $\mathcal{K}\in\mathcal{L}(Z\times X,H)$ and $\mathcal{L}\in\mathcal{L}(\mathcal{H},H)$ be arbitrary. Then a control law $u(t)=\mathcal{K}\Theta(t)+\mathcal{L}w(t)$ solves the FRP for the system (4.3) and $\mathcal{P}=\mathcal{B}\mathcal{L}+\mathcal{P}$ is regularly admissible for the semigroup generated by $\mathcal{A}+\mathcal{B}\mathcal{K}=\mathcal{A}$. Hence by Theorem 6.10 there exist $\left(\begin{smallmatrix}\Pi_1\\ \Pi_2\end{smallmatrix}\right)\in\mathcal{L}(\mathcal{H},Z\times X)$ and $\Gamma_0\in\mathcal{L}(\mathcal{H},H)$ such that $\left(\begin{smallmatrix}\Pi_1\\ \Pi_2\end{smallmatrix}\right)(\mathcal{D}(S|_{\mathcal{H}}))\subset\mathcal{D}(\mathcal{A})$ and

$$\begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}S|_{\mathcal{H}}=\begin{pmatrix}A & BJ\\ GC & F+GDJ\end{pmatrix}\begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}+\begin{pmatrix}0\\ 0\end{pmatrix}\Gamma_0+\begin{pmatrix}P\\ -GQ\end{pmatrix} \tag{6.10}$$

$$=\begin{pmatrix}A & BJ\\ GC & F+GDJ\end{pmatrix}\begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}+\begin{pmatrix}P\\ -GQ\end{pmatrix}\quad\text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{6.11}$$

$$Q=\begin{pmatrix}C & DJ\end{pmatrix}\begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}+0\Gamma_0=\begin{pmatrix}C & DJ\end{pmatrix}\begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}\quad\text{in } \mathcal{H}. \tag{6.12}$$

Letting $\Pi=\Pi_1$ and $\Gamma=J\Pi_2$, we immediately see that the regulator equations (3.1) are satisfied. Moreover, for $\Lambda=\Pi_2$ the regulator equations (4.1) for the error feedback controller are also satisfied. The proof is complete. ∎

The following corollary is evident by Theorem 4.1, Remark 6.9, and Theorem 6.13.

COROLLARY 6.14. *Let $\mathcal{H}$ be a CTO space such that $\mathcal{H} \cap C_0^+(\mathbb{R}, H) = \{0\}$. Assume that the triplet (F, G, J) has been chosen such that $\mathcal{A} = \left(\begin{smallmatrix} A & BJ \\ GC & F+GDJ \end{smallmatrix}\right)$ generates an exponentially stable C_0-semigroup on $Z \times \mathcal{H}$. Then the EFRP is solvable with this choice of (F, G, J) if and only if the regulator equations* (3.1) *and regulator equations* (4.1) *for the error feedback controller are satisfied for some $\Pi \in \mathcal{L}(\mathcal{H}, Z)$, $\Gamma \in \mathcal{L}(\mathcal{H}, H)$ and $\Lambda \in \mathcal{L}(\mathcal{H}, X)$.*

7. The internal model principle. One of the cornerstones in the classical automatic control theory of finite-dimensional systems is the internal model principle of Francis and Wonham [10, 11]. The principle asserts that closed loop stability and asymptotic tracking can be achieved only if the control system encapsulates a suitably reduplicated model of the dynamic structure of the exogenous signals which the regulator is required to process. Using frequency domain methods Yamamoto and Hara [28] generalized this principle for infinite-dimensional systems having a pseudorational impulse response (e.g., delay-differential systems). In the dynamic compensator of Corollary 4.2 the effect of this principle can be seen in the $(2, 2)$-block of F in (4.6): the internal model is just $S|_{\mathcal{H}}$.

In Theorem 7.2 below we show, using state space methods, that the existence of an internal model in the controller is—under some additional conditions—necessary for the solvability of the EFRP. In particular, we formalize the above idea that F must contain a copy of $S|_{\mathcal{H}}$. In order to do that, we need the following assumption.

ASSUMPTION 7.1. *The pair $\left(\left(\begin{smallmatrix} A & P \\ 0 & S|_{\mathcal{H}} \end{smallmatrix}\right), \left(\begin{smallmatrix} C & -Q \end{smallmatrix}\right)\right)$ is approximately observable.*

In finite dimensions Assumption 7.1 implies that the pair $\left(\left(\begin{smallmatrix} A & P \\ 0 & S|_{\mathcal{H}} \end{smallmatrix}\right), \left(\begin{smallmatrix} C & -Q \end{smallmatrix}\right)\right)$ is exponentially detectable. This was also assumed in the result of Francis [10], which Theorem 7.2 generalizes.

THEOREM 7.2. *Let Assumption* 7.1 *hold and let $\mathcal{H}$ be a CTO space such that $\mathcal{H} \cap C_0^+(\mathbb{R}, H) = \{0\}$. Assume that the EFRP is solvable for some triplet (F, G, J) and that $\mathcal{P} = \left(\begin{smallmatrix} P \\ -GQ \end{smallmatrix}\right)$ is regularly admissible for $\mathcal{A} = \left(\begin{smallmatrix} A & BJ \\ GC & F+GDJ \end{smallmatrix}\right)$. Then there exists a subspace $X_0 \subset X$ which is invariant for the semigroup $T_F(t)$ generated by F on X, and there exists a linear bijection $M : \mathcal{H} \to X_0$ such that $M \in \mathcal{L}(\mathcal{H}, X)$ and $T_S(t)|_{\mathcal{H}} = M^{-1} T_F(t) M$.*

Proof. By Theorem 6.13, there exist bounded linear operators Π, Γ, and Λ such that

$$\Pi S|_{\mathcal{H}} = A\Pi + BJ\Lambda + P \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{7.1a}$$

$$\Lambda S|_{\mathcal{H}} = F\Lambda \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{7.1b}$$

$$Q = C\Pi + DJ\Lambda \quad \text{in } \mathcal{H}. \tag{7.1c}$$

Since $\Lambda(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(F)$, from (7.1b) we see that for every $w \in \mathcal{D}(S|_{\mathcal{H}})$ and every $t \geq 0$

(7.2)

$$\Lambda T_S(t)|_{\mathcal{H}} w - T_F(t)\Lambda w = \Big|_0^t T_F(t-\tau)\Lambda T_S(\tau)|_{\mathcal{H}} w = \int_0^t \frac{d}{d\tau} T_F(t-\tau)\Lambda T_S(\tau)|_{\mathcal{H}} w \, d\tau$$

$$= \int_0^t T_F(t-\tau)[\Lambda S|_{\mathcal{H}} - F\Lambda] T_S(\tau)|_{\mathcal{H}} w \, d\tau = 0. \tag{7.3}$$

A suitable density argument shows that $\Lambda T_S(t)|_{\mathcal{H}} w - T_F(t)\Lambda w = 0$ for every $w \in \mathcal{H}$ and every $t \geq 0$.

We now show that Λ is injective. If $\Lambda w = 0$ for some $w \neq 0$, then by the above $BJ\Lambda T_S(t)|_{\mathcal{H}} w = 0$ and $DJ\Lambda T_S(t)|_{\mathcal{H}} w = 0$ for every $t \geq 0$. Consider the semigroup

$$T(t) = \begin{pmatrix} T_A(t) & \int_0^t T_A(t-s) P T_S(s)|_{\mathcal{H}} ds \\ 0 & T_S(t)|_{\mathcal{H}} \end{pmatrix} \tag{7.4}$$

generated by $\left(\begin{smallmatrix} A & P \\ 0 & S|_{\mathcal{H}} \end{smallmatrix}\right)$ on $Z \times \mathcal{H}$. We immediately see by (7.1a) and (7.1c) that

$$T(t)\left(\begin{smallmatrix} \Pi w \\ w \end{smallmatrix}\right) = \begin{pmatrix} T_A(t)\Pi w + \int_0^t T_A(t-s)(BJ\Lambda + P) T_S(s)|_{\mathcal{H}} w ds \\ T_S(t)|_{\mathcal{H}} w \end{pmatrix} \tag{7.5}$$

$$= \begin{pmatrix} T_A(t)\Pi w + \Pi T_S(t)|_{\mathcal{H}} w - T_A(t)\Pi w \\ T_S(t)|_{\mathcal{H}} w \end{pmatrix} \tag{7.6}$$

$$= \begin{pmatrix} \Pi T_S(t)|_{\mathcal{H}} w \\ T_S(t)|_{\mathcal{H}} w \end{pmatrix}. \tag{7.7}$$

Hence $\begin{pmatrix} C & -Q \end{pmatrix} T(t) \left(\begin{smallmatrix} \Pi w \\ w \end{smallmatrix}\right) = (C\Pi + DJ\Lambda - Q) T_S(t)|_{\mathcal{H}} w = 0$ for every $t \geq 0$. This violates the approximate observability assumption. Hence Λ is injective.

By the fact that $\Lambda T_S(t)|_{\mathcal{H}} = T_F(t)\Lambda$ in $\mathcal{H}$, it is evident that $X_0 = \operatorname{ran}(\Lambda)$ is a $T_F(t)$-invariant subspace of X. Consequently $M = \Lambda$ is a bounded linear bijection $\mathcal{H} \to X_0$ and $T_S(t)|_{\mathcal{H}} = M^{-1} T_F(t) M$, as was claimed. ∎

Using Remark 6.9, we obtain the following corollary which can be applied whenever $\mathcal{A}$ generates an exponentially stable semigroup—especially in the setting of Byrnes et al. [3], where $\mathcal{H}$ is finite-dimensional.

COROLLARY 7.3. *Let Assumption 7.1 hold and let $\mathcal{H}$ be a CTO space such $\mathcal{H} \cap C_0^+(\mathbb{R}, H) = \{0\}$. Assume that the EFRP is solvable for some triplet (F, G, J) such that $\mathcal{A} = \left(\begin{smallmatrix} A & BJ \\ GC & F + GDJ \end{smallmatrix}\right)$ generates an exponentially stable C_0-semigroup. Then there exists a subspace $X_0 \subset X$ which is invariant for the semigroup $T_F(t)$ generated by F on X, and there exists a linear bijection $M : \mathcal{H} \to X_0$ such that $M \in \mathcal{L}(\mathcal{H}, X)$ and $T_S(t)|_{\mathcal{H}} = M^{-1} T_F(t) M$.*

8. Examples. In this section we shall present three examples to illustrate the theory. Each of these examples elucidates some novel feature of the present paper, especially in relation to [3, 14, 15, 16]. The first example concerns the design of a feedforward controller for the asymptotic tracking of periodic signals for a plant in which the semigroup is not exponentially stable (or exponentially stabilizable as in [3, 14, 15, 16]) but only strongly stable. The second example concerns the design of a dynamic error feedback controller for the asymptotic tracking of periodic signals for a delay-differential equation; the same example was studied in [16] in the context of an open loop (i.e., feedforward) control. The last example of this section illustrates, for the case of a finite-dimensional plant, that the output regulation theory developed in this article is not confined to just periodic signals as in [15, 16], but the reference signals can also have uncountable Carleman spectrum (as opposed to [3, 15, 16]).

Example 8.1. The following is a model for a weakly damped string of unit length, with clamped ends [19, Example 1.1.1].

$$\frac{\partial^2 v(x,t)}{\partial t^2} + M \frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v(x,t)}{\partial x^2} \quad \text{for } t \geq 0 \text{ and } 0 \leq x \leq 1, \tag{8.1a}$$

$$v(0,t) = v(1,t) = 0 \quad \text{for } t \geq 0. \tag{8.1b}$$

Here M is a damping operator which will be defined shortly. We define the operator U by $Uv = -\frac{\partial^2 v}{\partial x^2}$ with $\mathcal{D}(U) = \{\, v \in H^2(0,1) \mid v(0) = v(1) = 0 \,\}$, where $H^2(0,1)$

denotes the standard Sobolev space on the unit interval. It can be shown that U has eigenvalues $\lambda_k = k^2\pi^2$, $k = 1, 2, \ldots$, and the corresponding eigenvectors $\phi_k(x) = \sqrt{2}\sin(k\pi x)$ constitute an orthonormal basis in $L^2(0,1)$ [19].

The damping operator M is defined by $Mv = \epsilon\langle g, v\rangle_{L^2(0,1)} g$, where $\epsilon > 0$ and

$$g = \sum_{k=1}^{\infty} \gamma_k \phi_k, \tag{8.2}$$

with γ_k satisfying $0 < |\gamma_k| \leq \frac{m}{\sqrt{\lambda_k}}$ (for example, $\gamma_k = \frac{1}{k}$) for some $m > 0$. Next we define the Hilbert space $Z = L^2(0,1) \times L^2(0,1)$ with the natural inner product $\langle \cdot, \cdot \rangle$ and introduce

$$z(t) = \begin{pmatrix} U^{\frac{1}{2}} v(x,t) \\ \frac{\partial v(x,t)}{\partial t} \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 0 & U^{\frac{1}{2}} \\ -U^{\frac{1}{2}} & M \end{pmatrix}. \tag{8.3}$$

Then (8.1) can be rewritten as $\dot{z}(t) = Az(t)$, $z(0) \in Z$, and it can be shown [19] that A generates a strongly (but not exponentially) stable C_0-semigroup $T_A(t)$ in Z. Furthermore, A is a Riesz spectral operator [6] with eigenvalues

$$\nu_k = ik\pi - \frac{\epsilon}{2}\gamma_{|k|}^2 + \mathcal{O}\left(\frac{\epsilon^2}{k^2\pi^2}\right), \quad k \neq 0, \tag{8.4}$$

all of which have negative real parts. The corresponding eigenvectors $(\psi_k)_{k\neq 0}$ form a Riesz basis in Z (the biorthogonal sequence is denoted by $(\psi_k^*)_{k\neq 0}$).

Consider, then, application of distributed control and observation to the system (8.1) in the following sense. For a control operator $B \in \mathcal{L}(\mathbb{C}, Z)$ and an observation operator $C \in \mathcal{L}(Z, \mathbb{C})$ the SISO plant is given as

$$\dot{z}(t) = Az(t) + Bu(t), \quad z(0) \in Z, \quad t \geq 0, \tag{8.5a}$$

$$y(t) = Cz(t), \quad t \geq 0. \tag{8.5b}$$

Here $Bu(t) = bu(t)$ for some $b = \sum\langle b, \psi_k^*\rangle\psi_k \in Z$ and $Cz(t) = \langle z(t), c\rangle$ for some $c = \sum\langle c, \psi_k\rangle\psi_k^* \in Z$. We study the FRP for p-periodic reference signals in the Sobolev space $\mathcal{H} = H(f_n, \omega_n)$ [16].

DEFINITION 8.2. *Let $I \subset \mathbb{Z}$, let $p > 0$, and let $\omega_n = \frac{2\pi n}{p}$ for every $n \in I$. Let $(f_n)_{n\in I} \subset \mathbb{R}$ such that $f_n \geq 1$ for each $n \in I$ and $(f_n^{-1})_{n\in I} \in \ell^2$. The Sobolev space $H(f_n, \omega_n)$ is defined as $\left\{ u : \mathbb{R} \to \mathbb{C} \;\middle|\; u(t) = \sum_{n\in I} a_n e^{i\omega_n t}$ for each $t \in \mathbb{R}$, $\sum_{n\in I} |f_n|^2|a_n|^2 < \infty$ and $(a_n)_{n\in I} \subset \mathbb{C} \right\}$.*

It has been shown in [16] that $\mathcal{H}$ is continuously embedded in $BUC(\mathbb{R}, \mathbb{C})$. Moreover, $H(f_n, \omega_n)$ is a Hilbert space with respect to the inner product $\langle u, v\rangle_f = \sum_{n\in I} a_n\overline{b_n}|f_n|^2$. Here $u(t) = \sum_{n\in I} a_n e^{i\omega_n t}$ and $v(t) = \sum_{n\in I} b_n e^{i\omega_n t}$ for every $t \in \mathbb{R}$, and $\overline{b_n}$ denotes the complex conjugate of b_n.

The transfer function of the system (8.5) is given for $\Re(s) \geq 0$ as $H(s) = CR(s, A)B$. We assume that $H(i\omega_n) \neq 0$ for each $n \in I$, i.e., that the transfer function of the system (8.5) does not have a zero at the frequencies of the reference signals.

Remark 8.3. One possible situation in which $H(i\omega) \neq 0$ for each $\omega \in \mathbb{R}$ is the following: The control and observation elements b and c are such that

$$\langle b, \psi_k^*\rangle > 0 \quad \text{and} \quad \langle \psi_k, c\rangle > 0 \quad \forall k \neq 0. \tag{8.6}$$

In fact, in this case we have

$$\begin{aligned} H(i\omega) &= CR(i\omega, A)B \\ &= \sum_n \frac{\langle b, \psi_n^* \rangle \langle \psi_n, c \rangle}{i\omega - \nu_n} = \sum_n \frac{d_n(\alpha_n - i\beta_n)}{\alpha_n^2 + \beta_n^2} \\ &= \sum_n \frac{d_n \alpha_n}{\alpha_n^2 + \beta_n^2} - i \sum_n \frac{d_n \beta_n}{\alpha_n^2 + \beta_n^2}, \end{aligned} \tag{8.7}$$

where $\omega \in \mathbb{R}$, $d_n = \langle b, \psi_n^* \rangle \langle \psi_n, c \rangle > 0$, $\alpha_n = \frac{\epsilon}{2}\gamma_{|n|}^2 - \delta_{n,\epsilon} > 0$ for some $\delta_{n,\epsilon}$, and $\beta_n = \omega - n\pi$. Clearly $\Re(CR(i\omega, A)B) > 0$ so that $H(i\omega) \neq 0$ for every $\omega \in \mathbb{R}$.

This allows us to solve the regulator equations (3.1) as in [15, 16]. If $(\theta_n)_{n\in I}$ is the canonical orthonormal basis of exponentials for $\mathcal{H} = H(f_n, \omega_n)$, then it is readily verified that a suitable pair (Π, Γ) of operators is

$$\Gamma y = \sum_{n\in I} y_n \frac{Q\theta_n}{H(i\omega_n)} \quad \forall y \in \mathcal{H}, \tag{8.8}$$

$$\Pi y = \sum_{n\in I} y_n R(i\omega_n, A)B\Gamma\theta_n \quad \forall y \in \mathcal{H}, \tag{8.9}$$

where $\omega_n = \frac{2\pi n}{p}$, $n \in I$, and y_n is the nth Fourier coefficient of $y \in \mathcal{H}$ in the basis $(\theta_n)_{n\in I}$. For a suitable weighting sequence $(f_n)_{n\in I}$ in the space $H(f_n, \omega_n)$ these operators Π and Γ are bounded [16]. Hence by Theorem 3.1 the feedforward control law which achieves the asymptotic tracking of $y \in \mathcal{H}$ for such $(f_n)_{n\in I}$ is given by

$$\begin{aligned} u(t) = \Gamma w(t) = \Gamma T_S(t)|_{\mathcal{H}} y = \Gamma \sum_{n\in I} y_n T_S(t)|_{\mathcal{H}} \theta_n = \sum_{n\in I} \frac{y_n}{H(i\omega_n)} Q T_S(t)|_{\mathcal{H}} \theta_n \\ = \sum_{n\in I} \frac{y_n}{H(i\omega_n)} \theta_n(t). \end{aligned} \tag{8.10}$$

In conclusion, in order to asymptotically track (a sufficiently smooth) p-periodic function y, we apply a p-periodic feedforward control whose nth Fourier coefficient is that of y multiplied by the reciprocal of the transfer function of the plant evaluated at $i\omega_n$. The use of an error feedback controller in this case may be difficult, because the closed loop system is difficult to stabilize.

Example 8.4. Let $a > 0$, $r \neq 0$, $\tau_1 > \tau_2 > 0$. Consider the following scalar delay differential equation with control and observation:

$$\dot{x}(t) = -ax(t) - b\big[x(t - \tau_1) + x(t - \tau_2)\big] + u(t), \tag{8.11a}$$

$$y(t) = rx(t), \quad t \geq 0. \tag{8.11b}$$

Our goal is to build, using Corollary 4.2, a dynamic controller (2.4) which solves the EFRP for this plant and p-periodic reference signals in the standard Sobolev spaces $\mathcal{H} = H^\alpha_{per}(0, p)$ for $\alpha > \frac{1}{2}$ [17].

Taking initial conditions for $x(\cdot)$ into account, the pair (8.11) can be formulated as a plant of the form (2.1) in which $D = 0$ and $\mathcal{U}_{dist} = 0$ [6]. Moreover, it can be shown (see, e.g., [6, Lemma 4.3.9]) that the transfer function $H(s) = CR(s, A)B$ of this plant is given by

$$H(s) = \frac{r}{s + a + b(e^{-s\tau_1} + e^{-s\tau_2})} \tag{8.12}$$

for those $s \in \mathbb{C}$ at which the denominator is not equal to zero.

The semigroup generated by A is exponentially stable if and only if $s + a + b(e^{-s\tau_1} + e^{-s\tau_2}) \neq 0$ for all $s \in \{ z \in \mathbb{C} \mid \Re(z) \geq 0 \}$ [6, Theorem 5.1.7]. Ruan and Wei [20] give a complete characterization (in terms of a, b, τ_1, and τ_2) of those instances in which all roots of equation $s + a + b(e^{-s\tau_1} + e^{-s\tau_2}) = 0$ have negative real parts. In their characterization, the parameter b lies on an interval (b_0^-, b_0^+). We assume that the semigroup generated by A is exponentially stable. By the above discussion, then $i\omega_n = i\frac{2\pi n}{p} \in \rho(A)$ and $H(i\omega_n) \neq 0$ for every $n \in \mathbb{Z}$.

It is evident that for every $\alpha > \frac{3}{2}$, $\sum_{n=-\infty}^{\infty} |H(i\omega_n)^{-1}|^2 (1 + \omega_n^2)^{-\alpha} < \infty$. Hence for $\alpha > \frac{3}{2}$ we can solve the regulator equations (3.1) for bounded operators Π and Γ as in [16]. We obtain

$$\Gamma y = \sum y_n \frac{Q\theta_n}{H(i\omega_n)} \quad \forall y \in \mathcal{H}, \tag{8.13}$$

$$\Pi y = \sum y_n R(i\omega_n, A) B\Gamma\theta_n \quad \forall y \in \mathcal{H}, \tag{8.14}$$

where $(\theta_n)_{n\in\mathbb{Z}}$ is the natural orthonormal basis for $\mathcal{H}$ and y_n is the nth Fourier coefficient of y with respect to this basis.

It is easy to see using Theorem 14 of [2] that $S|_{\mathcal{H}} - Q^*Q$ generates a strongly stable C_0-semigroup. Hence we may use Corollary 4.2 to deduce that an error feedback controller (2.4) with

$$F = \begin{pmatrix} A & B\Gamma \\ Q^*C & S|_{\mathcal{H}} - Q^*Q \end{pmatrix}, \quad J = \begin{pmatrix} 0 & \Gamma \end{pmatrix}, \quad \text{and} \quad G = \begin{pmatrix} 0 \\ -Q^* \end{pmatrix} \tag{8.15}$$

solves the EFRP.

Example 8.5. Let $H(s) = CR(s, A)B$ be a transfer function of an exponentially stable disturbance-free finite-dimensional SISO system (2.1) with $D = 0$. Let γ be a compact subset of $\mathbb{R}$. Consider reference signals in the space $\mathcal{H} = \{ f \in BUC(\mathbb{R}, \mathbb{C}) \mid sp(f) \subset \gamma \}$, where $sp(f)$ denotes the Carleman spectrum of f. Then $S|_{\mathcal{H}} \in \mathcal{L}(\mathcal{H})$ and $\sigma(S|_{\mathcal{H}}) = i\gamma$. Let δ be a smooth contour surrounding $\sigma(S|_{\mathcal{H}})$ such that it does not enclose any eigenvalues of A. Assume that $\inf_{i\omega\in\sigma(S|_{\mathcal{H}})} |H(i\omega)| > 0$. Then the linear operator $V \in \mathcal{L}(\mathcal{H})$ defined as

$$Vf = \frac{1}{2\pi i} \oint_\delta H(s) R(s, S|_{\mathcal{H}}) f ds \quad \forall f \in \mathcal{H} \tag{8.16}$$

is boundedly invertible (cf. Theorem 8.2 in [22]).

It is a straightforward calculation to show that the bounded operators Π and Γ defined as

$$\Gamma f = QV^{-1} \quad \forall f \in \mathcal{H}, \tag{8.17}$$

$$\Pi f = \frac{1}{2\pi i} \oint_\delta R(\lambda, A) B\Gamma R(s, S|_{\mathcal{H}}) f ds \quad \forall f \in \mathcal{H} \tag{8.18}$$

solve the regulator equations (3.1). Hence it is possible to track all reference signals in $\mathcal{H}$ using a feedforward control law $u(t) = \Gamma w(t)$. Such reference signals need not be periodic, but they are infinitely smooth because $S|_{\mathcal{H}} \in \mathcal{L}(\mathcal{H})$.

We remark that although the plant is exponentially stable and the regulator equations (3.1) can be solved, the corresponding EFRP may be difficult to solve. This is because $\sigma(S|_{\mathcal{H}})$, although compact, may be an uncountable set in $i\mathbb{R}$. The authors are not aware of any stabilization results in Banach spaces allowing for an uncountable spectrum in $\{ s \in \mathbb{C} \mid \Re(s) \geq 0 \}$; hence closed loop stability may be difficult to obtain.

9. Conclusions. In this article we have developed a state space output regulation theory for linear infinite-dimensional systems and bounded uniformly continuous exogenous reference/disturbance signals. We have solved both feedforward and error feedback output regulation problems; we have provided complete characterizations for the solvability of these problems in terms of certain regulator equations; and we have provided several examples to illustrate the theory. In this article we have also discussed strong stabilizability of the closed loop error feedback control system, which is an important problem in our case due to the following reasons:

- The exogenous signals are, in general, generated by an infinite-dimensional exosystem.
- The system operator of such an infinite-dimensional exosystem is often impossible to stabilize exponentially.
- The closed loop error feedback control system must (under certain assumptions) contain a copy of this exosystem operator in order that error feedback output regulation is possible.

The third item above is the well-known internal model principle, and we have generalized it in this article for infinite-dimensional plants and infinite-dimensional exosystems.

Interesting directions for future research include a detailed study of robustness issues and an extension of the results of this paper for unbounded operators B and C in the plant. In this way the output regulation theory of the present paper could be extended for many important boundary control and point observation problems not covered by the theory of this article.

Acknowledgments. The authors thank the reviewers for their many useful comments for improving the original manuscript.

REFERENCES

[1] W. Arendt, C. Batty, M. Hieber, and F. Neubrander, *Vector-Valued Laplace Transforms and Cauchy Problems*, Birkhäuser Verlag, Berlin, 2001.

[2] C. Batty and Q. Vu, *Stability of individual elements under one-parameter semigroups*, Trans. Amer. Math. Soc., 322 (1990), pp. 805–818.

[3] C. Byrnes, I. Laukó, D. Gilliam, and V. Shubov, *Output regulation for linear distributed parameter systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 2236–2252.

[4] C. Byrnes, D. Gilliam, V. Shubov, and J. Hood, *An example of output regulation for a distributed parameter system with infinite dimensional exosystem*, in Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2002), University of Notre Dame, Notre Dame, IN, 2000. http://www.nd.edu/~mtns/papers/22618_2.pdf.

[5] C. Byrnes, D. Gilliam, V. Shubov, and J. Hood, *Examples of output regulation for distributed parameter systems with infinite dimensional exosystem*, in Proceedings of the 40th IEEE Conference on Decision and Control, 2001, pp. 547–548.

[6] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[7] R. Curtain and H. Zwart, *Riccati equations and normalized coprime factorizations for strongly stabilizable infinite-dimensional systems*, Systems Control Lett., 28 (1996), pp. 11–22.

[8] E. Davison, *The robust control of a servomechanism problem for linear time-invariant multivariable systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 25–34.

[9] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.

[10] B. A. Francis, *The linear multivariable regulator problem*, SIAM J. Control Optim., 15 (1977), pp. 486–505.

[11] B. Francis and W. Wonham, *The internal model principle of control theory*, Automatica, 12 (1976), pp. 457–465.

[12] S. Hara, Y. Yamamoto, T. Omata, and M. Nakano, *Repetitive control system: A new type servo system for periodic exogenous signals*, IEEE Trans. Automat. Control, 33 (1988), pp. 659–668.

[13] Y. Hino, T. Naito, N. Van Minh, and J. Shin, *Almost Periodic Solutions of Differential Equations in Banach Spaces*, Taylor-Francis, New York, 2002.

[14] E. Immonen, *Some properties of infinite-dimensional systems capable of asymptotically tracking bounded uniformly continuous signals*, Math. Control Signals Systems, 18 (2006), pp. 323–344.

[15] E. Immonen and S. Pohjolainen, *Output regulation of periodic signals for DPS: An infinite-dimensional signal generator*, IEEE Trans. Automat. Control, 50 (2005), pp. 1799–1804.

[16] E. Immonen and S. Pohjolainen, *What periodic signals can an exponentially stabilizable linear control feedforward system asymptotically track?*, SIAM J. Control Optim., 44 (2006), pp. 2253–2268.

[17] R. Iorio and V. Iorio, *Fourier Analysis and Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 2001.

[18] Z.-H. Luo, B.-Z. Guo, and O. Morgul, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, London, 1999.

[19] J. Oostveen, *Strongly Stabilizable Infinite-Dimensional Systems*, Ph.D. Thesis, University of Groningen, The Netherlands, 1999.

[20] S. Ruan and J. Wei, *On the zeros of transcendental functions with applications to stability of delay differential equations with two delays*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 10 (2003), pp. 863–874.

[21] J. Schumacher, *Finite-dimensional regulators for a class of infinite-dimensional systems*, Systems Control Lett., 1 (1983), pp. 7–12.

[22] A. Taylor and D. Lay, *Introduction to Functional Analysis*, John Wiley & Sons, New York, 1980.

[23] Q. P. Vu, *The operator equation $AX - XB = C$ with unbounded operators A and B and related abstract Cauchy problems*, Math. Z., 208 (1991), pp. 567–588.

[24] Q. P. Vu and E. Schüler, *The operator equation $AX-XB = C$, admissibility, and asymptotic behavior of differential equations*, J. Differential Equations, 145 (1998), pp. 394–419.

[25] G. Weiss and M. Häfele, *Repetitive control of MIMO systems using H^∞ design*, Automatica, 35 (1999), pp. 1185–1199.

[26] W. Wonham, *Linear Multivariable Control. A Geometric Approach*, Springer-Verlag, New York, 1979.

[27] Y. Yamamoto, *Learning control and related problems in infinite-dimensional systems*, in Essays on Control: Perspectives in the Theory and Its Applications, H. Trentelman and J. Willems, eds., Birkhäuser, Boston, 1993, pp. 191–222.

[28] Y. Yamamoto and S. Hara, *Relationships between internal and external stability for infinite-dimensional systems with applications to a servo problem*, IEEE Trans. Automat. Control, 33 (1988), pp. 1044–1052.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1736–1757

ROBUSTNESS IN THE GRAPH TOPOLOGY OF A COMMON ADAPTIVE CONTROLLER*

MARK FRENCH†, ACHIM ILCHMANN‡, AND EUGENE P. RYAN§

Abstract. For any m-input, m-output, finite-dimensional, linear, minimum-phase plant P with first Markov parameter having spectrum in the open right half complex plane, it is well known that the adaptive output feedback control C, given by $u = -ky$, $\dot{k} = \|y\|^2$, yields a closed-loop system $[P, C]$ for which the state converges to zero, the signal k converges to a finite limit, and all other signals are of class L^2. It is first shown that these properties continue to hold in the presence of L^2-input and L^2-output disturbances. Working within the conceptual framework of the nonlinear gap metric approach to robust stability, and by establishing gain function stability of an appropriate closed-loop operator, it is proved that these properties also persist when the plant P is replaced with a stabilizable and detectable linear plant P_1 within a sufficiently small neighborhood of P in the graph topology, provided that the plant initial data and the L^2 magnitude of the disturbances are sufficiently small. Example 9 of Georgiou and Smith [*IEEE Trans. Automat. Control*, 42 (1997), pp. 1200–1221] is revisited. Unstable behavior for large initial conditions and/or large L^2 disturbances is shown, demonstrating that the bounds obtained from the L^2 theory are qualitatively tight: this contrasts with the L^∞-robustness analysis of Georgiou and Smith, which is insufficiently tight, to predict the stable behavior for small initial conditions and zero disturbances.

Key words. adaptive control, gap metric, robust stability

AMS subject classifications. 93D21, 93D09, 93C40, 93D25

DOI. 10.1137/0506336371

1. Introduction.

In an important paper in 1997, the well-established concept of the gap metric for linear systems (see, e.g., [19, 5, 6, 18, 2]) was extended to a nonlinear setting by Georgiou and Smith [7]. Further developments can be found in [8, 16, 1, 4]. The central property analyzed in the nonlinear gap framework is that of robust stability, i.e., the property that, if $\mathcal{W}$ is some requisite class (for example, L^∞ or L^2) to which the signals of a nominal closed-loop plant/controller configuration belong, then the closed-loop signals remain in $\mathcal{W}$ if the nominal plant is replaced with another plant which is sufficiently close in the gap sense. Gain function stability (gf-stability) a concept made precise in subsection 2.3, of the closed-loop operator mapping external disturbances to the input and output of the nominal plant provides a sufficient condition for robust stability (however, in contrast with the results in the linear setting, gf-stability is not a necessary condition for robust stability in the nonlinear setting). The nonlinear gap framework has been used to investigate the robustness (or lack of robustness) of certain classical adaptive controllers and variants thereof.

In an L^∞ setting, Example 9 of [7] (see also [9]) considers the controller (ubiqui-

*Received by the editors July 18, 2005; accepted for publication (in revised form) June 2, 2006; published electronically December 1, 2006. This research was supported by Deutsche Forschungsgemeinschaft grant IL 25/3-1.

http://www.siam.org/journals/sicon/45-5/63637.html

†School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK (mcf@ecs.soton.ac.uk).

‡Institut für Mathematik, Technische Universität Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (ilchmann@mathematik.tu-ilmenau.de).

§Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (epr@maths.bath.ac.uk).

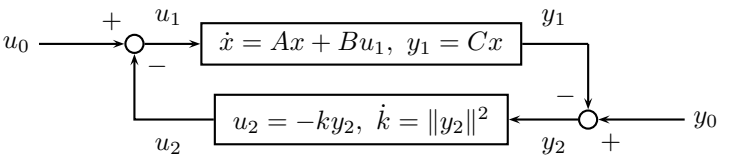


FIG. 1.1. *The adaptive closed-loop system.*

tous in the adaptive control literature)

$$u = -ky, \quad \dot{k} = y^2 \tag{1.1}$$

applied to the scalar linear plant $\dot{y} = ay + u$, for some $a \in \mathbb{R}$, and shows that the closed-loop operator mapping external disturbances onto the input and output of the nominal plant is not gain function stable (gf-stable). While the lack of gf-stability does not preclude robust stability, numerical and other informal evidence was presented which suggested that, with nonzero initial conditions, the closed-loop system is not robustly stable, even in the absence of disturbances. One consequence of the results of the present paper is to clarify the latter suggestion: we prove that, in the absence of disturbances, the closed-loop system is—with sufficiently small initial data—robustly stable but fails to be robustly stable for large initial data.

In [4], the nonlinear gap framework is used in an L^2 setting to establish robust stability properties of the controller $u = -k^{\frac{1}{4}}y$, $\dot{k} = y^2$ when applied to (a) single-input, single-output, linear, minimum-phase, relative-degree-one, nominal plants with positive high-frequency gain, and (b) a class of perturbed plants, where the gap metric distance between the nominal and perturbed plants is constrained by a function of the norms of the external disturbances. The present paper shows that the analysis developed in [4] can also be applied to the more familiar adaptive controller (1.1) (and its multivariable counterpart). This is considered to be important, as such controllers form the basis for many adaptive designs; see, e.g., [13, 14]. For example, adaptive backstepping designs [14, p. 100] reduce to (1.1) when applied to first-order linear plants. From an applied point of view, many application studies of such controllers exist in the literature (see, for example, [14] and the bibliography therein).

The analysis of the controller (1.1) differs significantly from that of the controllers in [4], requiring a different technique to establish the required gf-stability (Propositions 3.3, 3.5). In particular, in an L^2 setting, we establish a robust stability result for nominal m-input, m-output, finite-dimensional, stabilizable and detectable linear plants (A, B, C) which are minimum phase and are such that the first Markov parameter CB has its spectrum in the open right half complex plane (we denote the class of such plants by $\mathcal{M}$). With reference to Figure 1.1, in the absence of external disturbances (that is, with $u_0 = 0 = y_0$), it is well known (see, for example, [11, Theorem 4.2.1]) that, for every plant in $\mathcal{M}$ and all initial plant/controller data $(x^0, k^0) \in \mathbb{R}^n \times \mathbb{R}_+$ ($\mathbb{R}_+ := [0, \infty)$), the closed loop is such that (i) $u_1, y_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, (ii) $x(t) \to 0$ as $t \to \infty$, and (iii) $k(t) \to k^\infty \in \mathbb{R}_+$ as $t \to \infty$. First, we show that properties (i)–(iii) persist under external disturbances $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. Second, we consider the question of robust stability of the closed loop with respect to both external L^2 disturbances and perturbations of the plant (A, B, C), i.e., to what extent do the above properties (i)–(iii) persist if $(A, B, C) \in \mathcal{M}$ is perturbed to another m-input, m-output, linear, finite-dimensional, stabilizable and detectable plant $(A_p, B_p, C_p) \notin \mathcal{M}$? An appropriate conceptual framework in which to pose and answer such questions is provided by the gap metric. We show that properties (i)–(iii) persist if (A, B, C)

and (A_p, B_p, C_p) are sufficiently close in the gap metric. The associated bounds on the robust stability margin have a semiglobal nature insofar as they depend on the "size" of the external disturbances and initial data. In the case of zero initial conditions, the linear gap metric δ_0 measures the size of the smallest stable coprime factor perturbation between plants. Thus, the stability results of the present paper can be interpreted within the framework of linear robust control, where coprime factor perturbations form the widely accepted model for unstructured uncertainties.

We emphasize that the principal contribution of [4] and the present paper is to obtain robust stability results for adaptive controllers for this general description of unmodeled dynamics. This contrasts with other results in adaptive control (see, for example, [13, Chapters 8 and 9], [10]), which are far more restricted in the class of unmodeled dynamics considered and which typically consider robust "modifications" to the underlying adaptive law, which can introduce conservativeness (see [4] for a full discussion).

For purposes of illustration, one expression for the linear gap δ_0 is given in the frequency domain as follows. Let (A_1, B_1, C_1) and (A_2, B_2, C_2) be single-input, single-output stabilizable linear systems with respective transfer functions $\hat{P}_1$ and $\hat{P}_2$. Then $\hat{P}_1$ and $\hat{P}_2$ admit normalized right coprime factorizations over RH^∞, the class of rational functions that are analytic and bounded on the open half plane $\mathbb{C}_+ := \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) > 0\}$. In particular, there exist $N_i, D_i \in RH^\infty$ such that

$$\hat{P}_i = N_i D_i^{-1}, \qquad N_i^* N_i + D_i^* D_i = 1, \qquad i = 1, 2. \tag{1.2}$$

For $(i, j) = (1, 2), (2, 1)$, define the directed gap

$$\vec{\delta}_0(\hat{P}_i, \hat{P}_j) := \inf \left\{ \|(\Delta_N, \Delta_D)\|_{H^\infty} \,\middle|\, \Delta_N, \Delta_D \in RH^\infty,\ \hat{P}_j = (N_i + \Delta_N)(D_i + \Delta_D)^{-1} \right\}$$

(with the convention $\inf \emptyset := +\infty$). The linear gap between $\hat{P}_1$ and $\hat{P}_2$ is given by

$$\delta_0(\hat{P}_1, \hat{P}_2) \equiv \delta_0(\hat{P}_2, \hat{P}_1) = \max \left\{ \vec{\delta}_0(\hat{P}_1, \hat{P}_2), \vec{\delta}_0(\hat{P}_2, \hat{P}_1) \right\}. \tag{1.3}$$

We remark that the gap between the following plants $\hat{P}_1$ and $\hat{P}_2$ tends to zero as $\varepsilon \to 0$:

	$\hat{P}_1(s)$	$\hat{P}_2(s)$		Reference
(i)	$\dfrac{1}{s-\theta}$	$\dfrac{\lvert\lambda\rvert^2}{(s-\lambda)(s-\bar{\lambda})(s-\theta)}$,	$\mathrm{Re}(\lambda) \le -\varepsilon^{-1}$	[4]
(ii)	$\dfrac{1}{s-\theta}$	$\dfrac{N(M-s)}{(N+s)(M+s)(s-\theta)}$,	$N, M \ge \varepsilon^{-1}$	section 3.4, Example 3.10
(iii)	$\dfrac{1}{s-\theta}$	$\dfrac{(M-s)}{(M+s)(s-\theta)}$,	$M \ge \varepsilon^{-1}$	section 4.5; see also [7]

Example (i) is the classical Rohrs' example [15] which first drew the attention of the adaptive control community to the robustness issue. As observed in [4], example (ii) is of particular interest since $\hat{P}_2$ exhibits none of the classical assumptions of adaptive control: in particular, the sign of the high-frequency gain and the relative degree of $\hat{P}_2$ differ from those of the nominal plant $\hat{P}_1$ and, moreover, $\hat{P}_2$ is not minimum phase. Example (ii) is considered in more detail in section 3.4. Example (iii) is comprised of an all-pass factor in series with the nominal plant and is considered extensively in section 4. Example (iii) also coincides with Example 9 in [7], to which our general L^2 theory applies to conclude robust stability provided the initial data and L^2 disturbance

norms are sufficiently small. In section 4, we additionally prove the lack of robustness when the initial data or the L^2 disturbances are large. Moreover, we clarify some of the informal arguments in the L^∞ setting of [7].

The paper is structured as follows. In section 2 we introduce the notation and background theory: notions of well posedness and gf-stability of plant/controller interconnections are made precise, and the nonlinear gap is defined; this material is based on [7, section II] and [4, section 2]. Section 3 establishes the main robust stability result for the adaptive controller (1.1) (Theorem (3.8)) via the general framework of [4]. By considering example (iii) above in detail in section 4, we show that the bounds obtained in Theorem 3.8 are qualitatively tight in the sense that examples (i)–(iii) show that the robustness margin necessarily depends on the size of the disturbances and initial conditions.

2. Background concepts and terminology.

2.1. Preliminaries. While our goal is to establish stability of various configurations of plant and controller, the nonlinear nature of the controller is such that finite-time blow up of solutions of the closed-loop system cannot be ruled out a priori. To accommodate the potential for such behavior in the analysis, we introduce the following notions. Let $\mathcal{X}$ be a nonempty set and, for $0 < \omega \le \infty$, let $\mathcal{S}_\omega$ denote the set of locally integrable maps $[0,\omega) \to \mathcal{X}$. For simplicity, we write $\mathcal{S} := \mathcal{S}_\infty$. For $0 < \tau < \omega \le \infty$, $T_\tau : \mathcal{S}_\omega \to \mathcal{S}$ denotes the operator given by

$$T_\tau v \;:=\; \begin{cases} v(t), & t \in [0,\tau), \\ 0, & \text{otherwise.} \end{cases}$$

With $\mathcal{V} \subset \mathcal{S}$ we associate spaces as follows: $\mathcal{V}_e = \{v \in \mathcal{S} \big|\ T_\tau v \in \mathcal{V} \text{ for all } \tau > 0\}$, *the extended space*; $\mathcal{V}_\omega = \{v \in \mathcal{S}_\omega \big|\ T_\tau v \in \mathcal{V} \text{ for all } \tau \in (0,\omega)\}$, $0 < \omega \le \infty$; $\mathcal{V}_a = \cup_{\omega \in (0,\infty]} \mathcal{V}_\omega$, *the ambient space*. If $v, w \in \mathcal{V}_a$ with $v|_I = w|_I$ on $I = \mathrm{dom}(v) \cap \mathrm{dom}(w)$, then we write $v = w$. Note that $\mathcal{V} \subset \mathcal{V}_e \subset \mathcal{V}_a$ are strict inclusions and $\mathcal{V}_\infty = \mathcal{V}_e$. For $(f,g) \in \mathcal{V}_a \times \mathcal{V}_a$, the domains of f and g may be different; we adopt the convention $\mathrm{dom}(f,g) := \mathrm{dom}(f) \cap \mathrm{dom}(g)$. We define $\mathcal{V} \subset \mathcal{S}$ to be a *signal space* if and only if it is a vector space. In our applications, frequently $\mathcal{V}$ will be a normed signal space, such as $L^p(\mathbb{R}_+, \mathbb{R}^m)$ for $1 \le p \le \infty$, in which case $\mathcal{V}_e = L^p_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, $\mathcal{V}_\omega = L^p_{\mathrm{loc}}([0,\omega), \mathbb{R}^m)$ for $\omega \in (0,\infty]$, and $\mathcal{V}_a = \cup_{0<\omega\le\infty} L^p_{\mathrm{loc}}([0,\omega), \mathbb{R}^m)$. It is important to note that $\mathcal{V}_\omega \ne L^p([0,\omega), \mathbb{R}^m)$. Throughout the paper we consider only those normed signal spaces $\mathcal{V}$ which have the property that $\sup_{\tau \ge 0} \|T_\tau x\| < \infty$ implies $x \in \mathcal{V}$. We observe that $L^p(\mathbb{R}_+, \mathbb{R}^m)$ for $1 \le p \le \infty$ is such a normed signal space. We will often write $\|x\|_\tau = \|T_\tau x\|$. For a normed signal space $\mathcal{U}$ and a Euclidean space $\mathbb{R}^n$, we will also consider subsets of spaces of the form $\mathcal{V} = \mathbb{R}^n \times \mathcal{U}$ which, on identifying each $\theta \in \mathbb{R}^n$ with the constant signal $t \mapsto \theta$, can be thought of as a normed signal space with norm given by $\|(\theta, x)\| = \sqrt{|\theta|^2 + \|x\|^2_{\mathcal{U}}}$.

2.2. Well-posedness. A mapping $Q\colon \mathcal{X}_1 \to \mathcal{X}_2$ between signal spaces is said to be *causal* if and only if for all $\tau > 0$, $x, y \in \mathcal{X}_1$, $T_\tau x = T_\tau y$ implies $T_\tau Q x = T_\tau Q y$. Let $\mathcal{U}$ and $\mathcal{Y}$ be normed signal spaces and let $P\colon \mathcal{U}_a \to \mathcal{Y}_a$ and $C\colon \mathcal{Y}_a \to \mathcal{U}_a$ be causal mappings representing a plant and controller, respectively. Our central concern is the system of equations

$$[P,C] \;:\quad y_1 = P u_1, \quad u_2 = C y_2, \quad u_0 = u_1 + u_2, \quad y_0 = y_1 + y_2 \tag{2.1}$$

corresponding to the closed-loop feedback configuration as depicted in Figure 2.1. By a solution of (2.1) we mean the following. For $w_0 = (u_0, y_0) \in \mathcal{W} := \mathcal{U} \times \mathcal{Y}$, a pair

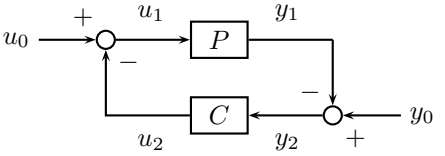


FIG. 2.1. *The closed-loop system* $[P, C]$.

$(w_1, w_2) = \big((u_1, y_1), (u_2, y_2)\big) \in \mathcal{W}_a \times \mathcal{W}_a$, $\mathcal{W}_a := \mathcal{U}_a \times \mathcal{Y}_a$, is a *solution* of (2.1) if and only if (2.1) holds on $\operatorname{dom}(w_1, w_2)$. The (possibly empty) set of all solutions is denoted by $\mathcal{X}_{w_0} := \big\{(w_1, w_2) \in \mathcal{W}_a \times \mathcal{W}_a |\ (w_1, w_2) \text{ solves } (2.1)\big\}$. The closed-loop system $[P, C]$, given by (2.1), is said to have (a) the *existence property* if and only if $\mathcal{X}_{w_0} \neq \emptyset$; (b) the *uniqueness property* if and only if for each $w_0 \in \mathcal{W}$,

$$(\hat{w}_1, \hat{w}_2), (\tilde{w}_1, \tilde{w}_2) \in \mathcal{X}_{w_0} \Rightarrow (\hat{w}_1, \hat{w}_2) = (\tilde{w}_1, \tilde{w}_2) \ \text{ on } \ \operatorname{dom}(\hat{w}_1, \hat{w}_2) \cap \operatorname{dom}(\tilde{w}_1, \tilde{w}_2).$$

Assume that $[P, C]$ has the existence and uniqueness properties. For each $w_0 \in \mathcal{W}$, define ω_{w_0}, $0 < \omega_{w_0} \leq \infty$, by the property $[0, \omega_{w_0}) := \cup_{(\hat{w}_1, \hat{w}_2) \in \mathcal{X}_{w_0}} \operatorname{dom}(\hat{w}_1, \hat{w}_2)$, and define $(w_1, w_2) \in \mathcal{W}_a \times \mathcal{W}_a$, with $\operatorname{dom}(w_1, w_2) = [0, \omega_{w_0})$, by the property $(w_1, w_2)|_{[0,t)} \in \mathcal{X}_{w_0}$ for all $t \in [0, \omega_{w_0})$. This construction induces an operator $H_{P,C} : \mathcal{W} \to \mathcal{W}_a \times \mathcal{W}_a$, $w_0 \mapsto (w_1, w_2)$. The closed-loop system $[P, C]$, given by (2.1), is said to be

- *locally well posed* if and only if it has the existence and uniqueness properties and the operator $H_{P,C} \colon \mathcal{W} \to \mathcal{W}_a \times \mathcal{W}_a$, $w_0 \mapsto (w_1, w_2)$, is causal;
- *globally well posed* if and only if it is locally well posed and $\operatorname{im} H_{P,C} \subset \mathcal{W}_e \times \mathcal{W}_e$;
- $\mathcal{W}$*-stable* if and only if it is locally well posed and $\operatorname{im} H_{P,C} \subset \mathcal{W} \times \mathcal{W}$;
- *regularly well posed* if and only if it is locally well posed and

$$\forall\, w_0 \in \mathcal{W} \quad \big[\ \omega_{w_0} < \infty \quad \Longrightarrow \quad T_{\omega_{w_0}} H_{P,C}(w_0) \notin \mathcal{W} \times \mathcal{W}\big]\,. \tag{2.2}$$

If $[P, C]$ is globally well posed, then for each $w_0 \in \mathcal{W}$ the solution $H_{P,C}(w_0)$ exists on the half line $\mathbb{R}_+$. Regular well posedness means that if the closed-loop system has a finite escape time $\omega > 0$ for some disturbance $(u_0, y_0) \in \mathcal{W}$, then at least one of the signals (u_1, y_1) or (u_2, y_2) is not a restriction to $[0, \omega)$ of a function in $\mathcal{W}$. If $[P, C]$ is regularly well posed and satisfies

$$\forall\, w_0 \in \mathcal{W} \quad \big[\ \omega_{w_0} < \infty \quad \Longrightarrow \quad T_{\omega_{w_0}} H_{P,C}(w_0) \in \mathcal{W} \times \mathcal{W}\big]\,,$$

then there does not exist a solution of $[P, C]$ with a finite escape time, and therefore $[P, C]$ is globally well posed. However, global well posedness does not guarantee that each solution belongs to $\mathcal{W} \times \mathcal{W}$; the latter is ensured by $\mathcal{W}$-stability of $[P, C]$. Note also that neither regular nor global well posedness implies the other. Our main concern will be the situation wherein the closed-loop system $[P, C]$ is generated by a system of (nonlinear) differential equations. In this context, a globally well-posed system is a system with the property of existence and uniqueness of solutions and for which finite-time blow up does not occur: all (forward) solutions have a maximal interval of existence $[0, \infty)$. Regular well posedness usually follows from standard existence theory for differential equations when $\mathcal{W} = L^\infty \times L^\infty$. However, when $\mathcal{W} \neq L^\infty \times L^\infty$ (in this paper we are primarily interested in $\mathcal{W} = L^2 \times L^2$), stronger properties of the underlying differential equations are required. As will be shown, all closed-loop systems considered in this paper are regularly well posed.

2.3. Graphs and gf-stability. In our investigation of robustness of stability properties of a closed-loop system, the concept of graphs and gf-stability will play a central role. Corresponding to a plant operator P (respectively, the controller operator C) is a subset of $\mathcal{W}$, called the *graph* of the plant $\mathcal{G}_P$ (respectively, the controller $\mathcal{G}_C$), defined as

$$\mathcal{G}_P = \left\{ \begin{pmatrix} u \\ Pu \end{pmatrix} \;\middle|\; u \in \mathcal{U},\ Pu \in \mathcal{Y} \right\} \subset \mathcal{W}, \quad \mathcal{G}_C = \left\{ \begin{pmatrix} Cy \\ y \end{pmatrix} \;\middle|\; Cy \in \mathcal{U},\ y \in \mathcal{Y} \right\} \subset \mathcal{W}.$$

Note that in general, $\mathcal{G}_P, \mathcal{G}_C \neq \mathcal{W}$. A causal operator $F\colon \mathcal{X} \to \mathcal{V}_a$, where $\mathcal{X}, \mathcal{V}$ are subsets of normed signal spaces, is said to be gf-stable if and only if $\operatorname{im} F \subset \mathcal{V}$ and if the following nonlinear so-called *gain function* is well defined:

$$(2.3) \quad g[F]\colon (r_0, \infty) \to \mathbb{R}_+,$$
$$r \mapsto g[F](r) = \sup \left\{ \|T_\tau F x\| \,\middle|\, x \in \mathcal{X}\ ,\ \|T_\tau x\| \in (r_0, r],\ \tau > 0 \right\},$$

where $r_0 := \inf_{x \in \mathcal{X}} \|x\| < \infty$. Observe that $\|Fx\|_\tau \leq g[F](\|x\|_\tau)$. A closed-loop system $[P, C]$ is said to be *gf-stable* if and only if it is globally well posed and $H_{P,C}\colon \mathcal{W} \to \mathcal{W}_e \times \mathcal{W}_e$ is gf-stable. Note the following facts: (i) global well posedness of $[P, C]$ implies that $\operatorname{im} H_{P,C} \subset \mathcal{W}_e \times \mathcal{W}_e$; (ii) gf-stability of $[P, C]$ implies $\mathcal{W}$-stability of $[P, C]$; (iii) if $[P, C]$ is $\mathcal{W}$-stable, then $H_{P,C} : \mathcal{W} \to \mathcal{G}_P \times \mathcal{G}_C$ is a bijective operator with inverse $H_{P,C}^{-1} : (w_1, w_2) \mapsto w_1 + w_2$. To see (iii), note that $\operatorname{im} H_{P,C} \subset \mathcal{W} \times \mathcal{W}$ implies that $\operatorname{im} H_{P,C} \subset \mathcal{G}_P \times \mathcal{G}_C$, and since for any $w_1 \in \mathcal{G}_P \subset \mathcal{W}$, $w_2 \in \mathcal{G}_C \subset \mathcal{W}$ we have $w_1 + w_2 \in \mathcal{W}$, it follows that $\operatorname{im} H_{P,C} \supset \mathcal{G}_P \times \mathcal{G}_C$. Therefore, we can think of a gf-stable $H_{P,C}$ as a surjective operator $H_{P,C}\colon \mathcal{W} \to \mathcal{G}_P \times \mathcal{G}_C$. The inverse of $H_{P,C} : \mathcal{W} \to \mathcal{G}_P \times \mathcal{G}_C$ is obviously $H_{P,C}^{-1} : (w_1, w_2) \mapsto w_1 + w_2$. Finally, with a closed-loop system $[P, C]$ we associate the following two parallel projection operators: $\Pi_{P//C} : \mathcal{W} \to \mathcal{W}_a$, $w_0 \mapsto w_1$ and $\Pi_{C//P} : \mathcal{W} \to \mathcal{W}_a$, $w_0 \mapsto w_2$. Clearly, $H_{P,C} = \left(\Pi_{P//C}\,, \Pi_{C//P}\right)$ and $\Pi_{P//C} + \Pi_{C//P} = I$. Therefore, gf-stability of one of the operators $\Pi_{P//C}$ and $\Pi_{C//P}$ implies the gf-stability of the other, and so gf-stability of either operator implies gf-stability of the closed-loop system $[P, C]$.

2.4. The nonlinear gap. The essence of the paper is a study of robust stability in a specific adaptive control context. Robust stability is the property that the stability properties of a globally well-posed closed-loop system $[P, C]$ persist under "sufficiently small" perturbations of the plant. In other words robust stability is the property that $[P_1, C]$ inherits the stability properties of $[P, C]$ when the plant P is replaced with any plant P_1 sufficiently "close" to P. In the context of this paper, plants P_1 and P_2 are deemed to be close if and only if the nonlinear gap [7] between P_1 and P_2 is small, where the nonlinear gap is defined as follows. Let $\Gamma := \left\{ P : \mathcal{U}_a \to \mathcal{Y}_a \,\middle|\, P \text{ is causal} \right\}$ and, for $P_1, P_2 \in \Gamma$, define the (possibly empty) set $\mathcal{O}_{P_1,P_2} := \left\{ \Phi\colon \mathcal{G}_{P_1} \to \mathcal{G}_{P_2} \,\middle|\, \Phi \text{ is causal, bijective, and } \Phi(0) = 0 \right\}$. Write

$$\vec{\delta}(P_1, P_2) := \inf_{\Phi \in \mathcal{O}_{P_1,P_2}} \sup_{x \in \mathcal{G}_{P_1} \setminus \{0\},\ \tau > 0} \left(\frac{\|(\Phi - I)|_{\mathcal{G}_{P_1}} x\|_\tau}{\|x\|_\tau} \right),$$

with the convention that $\vec{\delta}(P_1, P_2) := \infty$ if $\mathcal{O}_{P_1,P_2} = \emptyset$. The *nonlinear gap* δ is

$$(2.4) \quad \delta\colon \Gamma \times \Gamma \to [0, \infty], \qquad (P_1, P_2) \mapsto \delta(P_1, P_2) := \max\{\vec{\delta}(P_1, P_2), \vec{\delta}(P_2, P_1)\}.$$

The nonlinear gap provides a generalization of the standard definition of the linear gap δ_0 (previously discussed briefly in the introduction). To explain this, some notation

is needed. For $q, m \in \mathbb{N}$, let $R_{q,m}$ denote the set of proper, rational, $(q \times m)$-matrix-valued functions and let $H^\infty_{q,m}$ denote the set of analytic and bounded $\mathbb{C}^{q\times m}$-valued functions on the open right half plane $\mathbb{C}_+ := \{\lambda \in \mathbb{C} |\ \mathrm{Re}(\lambda) > 0\}$. By $RH^\infty_{q,m}$, we denote the class of functions in $R_{q,m}$ that are analytic in $\mathbb{C}_+$. It is known (see, for example, [17, pp. 74–75; 261–262]) that any $P \in R_{q,m}$ has a normalized right coprime factorization, that is, $P = ND^{-1}$, where $N \in RH^\infty_{q,m}$, $D \in RH^\infty_{m,m}$, D has an inverse in $R_{m,m}$, and $N^*N + D^*D = I_m$, where $N^*(s) := \overline{N(-\bar{s})}^T$. Let $\mathcal{U} = L^2(\mathbb{R}_+, \mathbb{R}^m)$, $\mathcal{Y} = L^2(\mathbb{R}_+, \mathbb{R}^q)$ for some $m, q \in \mathbb{N}$ and associate with $\hat{P} \in R_{q,m}$ the linear operator $P\colon \mathcal{U}_e \to \mathcal{Y}_e,\ u \mapsto y := \mathbf{L}^{-1}(\hat{P}) \star u$, where $\mathbf{L}$ denotes the Laplace transform and $\star$ denotes convolution. We refer to P as a linear plant with associated transfer function $\hat{P} \in R_{q,m}$. For $i = 1, 2$, let $P_i\colon \mathcal{U}_e \to \mathcal{Y}_e$ be linear plants with associated strictly proper rational transfer functions $\hat{P}_i = N_i D_i^{-1}$, where $(N_i, D_i) \in RH^\infty_{q,m} \times RH^\infty_{m,m}$ are right coprime factors (analogous to (1.2)), and let $\Pi_i : \mathcal{U}_e \times \mathcal{Y}_e \to \mathcal{G}_{P_i}$ denote the orthogonal projection onto the closed subspace $\mathcal{G}_{P_i}$. The linear gap between these plants is defined as in (1.3), with

$$\vec{\delta}_0(\hat{P}_i, \hat{P}_j) := \inf\big\{\|(\Delta_N, \Delta_D)\|_{H^\infty}\big|\ \ \Delta_N \in RH^\infty_{q,m},\ \Delta_D \in RH^\infty_{m,m},\ \hat{P}_j = (N_i + \Delta_N)(D_i + \Delta_D)^{-1}\big\}.$$

In [7, Proposition 5] it is shown that if $\max\{\|(\Pi_2 - \Pi_1)\,\Pi_1\|, \|(\Pi_1 - \Pi_2)\,\Pi_2\|\} < 1$, then $\vec{\delta}(P_1, P_2) = \|(\Pi_2 - \Pi_1)\,\Pi_1\|$, and in [5, Lemma 2] it is shown that $\|(\Pi_2 - \Pi_1)\,\Pi_1\| = \vec{\delta}_0(\hat{P}_1, \hat{P}_2)$ if $\vec{\delta}_0(\hat{P}_1, \hat{P}_2) < 1$, the conjunction of which yields

$$\vec{\delta}(P_1, P_2) \ = \ \vec{\delta}_0(\hat{P}_1, \hat{P}_2) \qquad \text{if } \ \delta_0(\hat{P}_1, \hat{P}_2) < 1. \tag{2.5}$$

The topology induced on $R_{q,m}$ by the gap δ_0 is called the graph topology [17, p. 235]; note that the graph topology on Γ induces the graph topology on $R_{q,m}$ via the subset topology and the Laplace transform $\mathbf{L}$. Robust stability with respect to gap perturbations has been extensively studied in the linear setting; see, e.g., [18] and the references therein. In a nonlinear setting, sufficient conditions for robust gain stability were given in [7]. The adaptive controllers considered in this paper and in [4] do not meet any of the sufficient conditions of [7] and hence require a further development of the robust stability theory in section 5 of [4], as given in Theorem 3.11 of the present paper.

3. System classes and the adaptive controller. We are interested in the control of linear m-input, m-output stabilizable n-dimensional state space realizations of transfer functions in $R_{m,m}$, i.e., systems of the form

$$\dot{x}(t) = A\,x(t) + B\,u_1(t), \qquad x(0) \in \mathbb{R}^n, \qquad y_1(t) = Cx(t). \tag{3.1}$$

Henceforth, we fix (arbitrarily) the number $m \geq 1$ of inputs/outputs but allow for a variation in the state space dimension n. We denote this system class by

$$\mathcal{P}_n \ = \ \{(A, B, C) \in E_n \ \mid \ n \geq m,\ (A, B, C) \ \text{ is stabilizable and detectable}\}, \tag{3.2}$$

where $E_n := \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{m\times n}$, and write $\mathcal{P} := \cup_{n\geq m}\mathcal{P}_n$. We define the subclass of minimum-phase systems with "high-frequency gain" CB having spectrum in $\mathbb{C}_+$,

$$\widetilde{\mathcal{M}}_n \ = \ \left\{(A, B, C) \in \mathcal{P}_n \ \middle|\ \ \sigma(CB) \subset \mathbb{C}_+, \ \ \det\begin{bmatrix} sI_n - A & B \\ C & 0 \end{bmatrix} \neq 0\ \forall s \in \overline{\mathbb{C}}_+\right\}. \tag{3.3}$$

Observe that for each $(A, B, C) \in \widetilde{\mathcal{M}}_n$, there exists an element of its similarity orbit $\{(TAT^{-1}, TB, CT^{-1}) |\ T \in \mathbb{R}^{n\times n} \text{ invertible}\}$ such that

$$TAT^{-1} = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, \quad TB = \begin{bmatrix} 0 \\ B_2 \end{bmatrix} = \begin{bmatrix} 0 \\ CB \end{bmatrix}, \quad CT^{-1} = \begin{bmatrix} 0 & I \end{bmatrix},$$

where $\sigma(B_2) \subset \mathbb{C}_+$ and, by the minimum-phase property, A_1 has a spectrum in the open left half complex plane $\mathbb{C}_-$. Therefore, we introduce

$$\mathcal{M}_n := \left\{ (A, B, C) \in \mathcal{P}_n \;\middle|\; \begin{array}{l} A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix},\ B = \begin{bmatrix} 0 \\ B_2 \end{bmatrix},\ C = [0 \ \ I], \\ B_2, A_4 \in \mathbb{R}^{m\times m},\ \sigma(A_1) \subset \mathbb{C}_-,\ \sigma(B_2) \subset \mathbb{C}_+ \end{array} \right\} \tag{3.4}$$

and write $\mathcal{M} := \cup_{n\geq m} \mathcal{M}_n$. For a system of class $\mathcal{M}_n$, (3.1) may be expressed in the equivalent form

$$\left.\begin{array}{ll} \dot{z}(t) = A_1 z(t) + A_2 y_1(t), & z(0) = z^0 \in \mathbb{R}^{n-m} \\ \dot{y_1}(t) = A_3 z(t) + A_4 y_1(t) + B_2 u_1(t), & y_1(0) = y_1^0 \in \mathbb{R}^m \end{array}\right\}, \quad (z^0, y_1^0) = x^0. \tag{3.5}$$

We will have occasion to identify $\mathcal{P}_n$ with a subspace of a Euclidean space $\mathbb{R}^{n^2+2mn}$ by identifying a plant $\theta = (A, B, C)$ with a vector θ consisting of the elements of the plant matrices, ordered lexicographically. With normed signal spaces $\mathcal{U}$ and $\mathcal{Y}$ and $(\theta, x^0) \in \mathcal{P}_n \times \mathbb{R}^n$, we associate the causal plant operator

$$\widetilde{P}(\theta, x^0) :\ \mathcal{U}_a \to \mathcal{Y}_a, \quad u_1 \mapsto \widetilde{P}(\theta, x^0)(u_1) := y_1\,, \tag{3.6}$$

where, for $u_1 \in \mathcal{U}_a$ with $\mathrm{dom}(u_1) = [0, \omega)$, we have $y_1 = Cx$, x being the unique solution of (3.1) on $[0, \omega)$. Note that $\widetilde{P}$ is a map from $\bigcup_{n\geq m}(\mathcal{P}_n \times \mathbb{R}^n)$ to the space of maps $\mathcal{U}_a \to \mathcal{Y}_a$. Furthermore, for $(\theta, x^0) = (A, B, C, x^0) \in \mathcal{P}_n \times \mathbb{R}^n$, $\widetilde{P}(\theta, x^0)$ corresponds to a stabilizable and detectable realization of $C(sI_n - A)^{-1}B \in R_{m,m}$. Our objective is to study, in the context of systems of form (3.1), the adaptive strategy

$$u_2(t) = -k(t)y_2(t), \quad \dot{k}(t) = \|y_2(t)\|^2, \quad k(0) = k^0 \in \mathbb{R}_+\,, \tag{3.7}$$

with the associated control operator, parameterized by k^0, denoted by

$$\widetilde{C}(k^0) : \mathcal{Y}_a \to \mathcal{U}_a, \qquad y_2 \mapsto \widetilde{C}(k^0)(y_2) := u_2. \tag{3.8}$$

Note that $\widetilde{C}$ is a map from $\mathbb{R}_+$ to the space of causal maps $\mathcal{Y}_a \to \mathcal{U}_a$. In particular and with reference to Figure 2.1, we will study properties of the closed-loop system $[\widetilde{P}(\theta, x^0), \widetilde{C}(k^0)]$, generated by the application of the controller (3.8) to system (3.1), in the presence of disturbances $(u_0, y_0) \in \mathcal{W}$ satisfying the interconnection equations

$$u_0 \ = \ u_1 + u_2, \qquad y_0 \ = \ y_1 + y_2\,. \tag{3.9}$$

Results will be given for systems (3.5) of class $\mathcal{M}_n$ (such systems will later play the role of the nominal plant) and for the more general class of systems (3.1), (3.2) (such systems will later play the role of the perturbed plant).

3.1. Properties of the interconnection of the adaptive controller with the general linear plant. In this section we investigate the interconnection of the adaptive controller (3.7) (with associated operator $\widetilde{C}(k^0)$) and any plant in the form (3.1) (with associated operator $\widetilde{P}(\theta, x^0)$), where $(\theta, x^0, k^0) \in \mathcal{P}_n \times \mathbb{R}^n \times \mathbb{R}_+$.

PROPOSITION 3.1. *Let $n \geq m$, $(\theta, x^0, k^0) \in \mathcal{P}_n \times \mathbb{R}^n \times \mathbb{R}_+$, and $u_0, y_0 \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$. Then the closed-loop initial-value problem given by* (3.1), (3.7), (3.9) *has the following properties:*

(i) *There exists a unique maximal solution $(x, k) : [0, \omega) \to \mathbb{R}^n \times \mathbb{R}_+$, $0 < \omega \leq \infty$;*
(ii) *if $k \in L^\infty([0, \omega), \mathbb{R}_+)$, then $\omega = \infty$;*
(iii) *if $y_2 \in L^2([0, \omega), \mathbb{R}^m)$, then $\omega = \infty$.*

Proof. (i) follows from the theory of ordinary differential equations.

(ii) If $k \in L^\infty([0, \omega), \mathbb{R}_+)$, then consider the following subsystem of the initial-value problem (3.1), (3.7): $\dot{x}(t) = Ax(t) + B\big[u_0(t) + k(t)y_2(t)\big]$. Integration, together with elementary estimates, yields the existence of constants $c_0, c_1 > 0$ such that

$$\|x(t)\| \leq c_0 \left(e^{c_1 t} + \int_0^t e^{c_1(t-s)} \big(\|u_0(s)\| + \|y_2(s)\|\big) \mathrm{d}s \right) \quad \forall\, t \in [0, \omega). \tag{3.10}$$

Suppose $\omega < \infty$. Since $y_2 \in L^2([0, \omega), \mathbb{R}^m)$ (which is equivalent to $k \in L^\infty([0, \omega), \mathbb{R}_+)$) and since $u_0 \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, it follows from the convolution in (3.10) that the right-hand side of (3.10) is bounded on $[0, \omega)$, which contradicts maximality of the solution x. Therefore, $\omega = \infty$.

(iii) By assumption, $y_2 \in L^2([0, \omega), \mathbb{R}^m)$, and so $t \mapsto k(t) = k^0 + \|y_2\|^2_{L^2([0,t), \mathbb{R}^m)}$ is bounded on $[0, \omega)$, and thus assertion (iii) follows from (ii). ∎

COROLLARY 3.2. *Let $\mathcal{U} = \mathcal{Y} = L^2(\mathbb{R}_+, \mathbb{R}^m)$, $n \geq m$, $(\theta, x^0, k^0) \in \mathcal{P}_n \times \mathbb{R}^n \times \mathbb{R}_+$. Then the closed-loop initial-value problem $[\widetilde{P}(\theta, x^0), \widetilde{C}(k^0)]$ given by* (2.1), (3.6), *and* (3.8) *is regularly well posed.*

Proof. Let $\mathcal{W} = L^2(\mathbb{R}_+, \mathbb{R}^m) \times L^2(\mathbb{R}_+, \mathbb{R}^m)$. The closed-loop $[\widetilde{P}(\theta, x^0), \widetilde{C}(k^0)]$ is locally well posed by Proposition 3.1(i). To prove that $[\widetilde{P}(\theta, x^0), \widetilde{C}(k^0)]$ is regularly well posed, it suffices to show that (2.2) holds. Let $w_0 = (u_0, y_0) \in \mathcal{W}$. Consider $(w_1, w_2) = H_{\widetilde{P}(\theta, x^0), \widetilde{C}(k^0)}(w_0)$, where $\mathrm{dom}\,(w_1, w_2) = [0, \omega)$ and is maximal. Suppose $T_\omega(w_1, w_2) \in \mathcal{W} \times \mathcal{W}$. Then we have $y_1 \in L^2([0, \omega), \mathbb{R}^m)$, which, in view of Proposition 3.1(iii) yields $\omega = \infty$. Hence as $w_0 \in \mathcal{W}$ is arbitrary, it follows that the closed-loop system is regularly well posed. ∎

3.2. Properties of the interconnection of the adaptive controller with the nominal plant. In this section we consider the closed-loop behavior of the nominal plant and controller interconnection given by (3.5), (3.7), and perturbations u_0, y_0 satisfying (3.9). The L^2 bounds obtained in [4] do not directly generalize to the controller (1.1), even in the single-input, single-output (SISO) case. Hence the results in this subsection are obtained via an alternative (and tighter) series of bounds than those considered in the results of [4].

PROPOSITION 3.3. *Let $n \geq m$, $(A, B, C) \in \mathcal{M}_n$, $(x^0, k^0) \in \mathbb{R}^n \times \mathbb{R}_+$, and $u_0, y_0 \in L^2([0, \infty), \mathbb{R}^m)$. Then the closed-loop initial-value problem* (3.5), (3.7), (3.9) *has the following properties:*

(i) *There exists a unique solution $(z, y_1, k) : [0, \infty) \to \mathbb{R}^n \times \mathbb{R}_+$;*
(ii) *the limit $\lim_{t\to\infty} k(t)$ exists and is finite;*
(iii) *$u_1, y_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, $z \in L^2(\mathbb{R}_+, \mathbb{R}^{n-m})$;*
(iv) *$\lim_{t\to\infty}(z(t), y_1(t)) = 0$.*

Proof. The closed-loop equations (3.5), (3.7), (3.9) may be expressed as

$$\left\{\begin{array}{l} \dot{z}(t) = A_1 z(t) + A_2 y_1(t)\,, \\ \dot{y_1}(t) = A_3 z(t) + A_4 y_1(t) + B_2\left(u_0(t) + k(t)\,[y_0(t) - y_1(t)]\right), \\ \dot{k}(t) = \|y_0(t) - y_1(t)\|^2, \\ u_1(t) = u_0(t) + k(t)\,[y_0(t) - y_1(t)]\,, \\ (z(0), y_1(0), k(0)) = (z^0, y_1^0, k^0) \in \mathbb{R}^{n-m} \times \mathbb{R}^m \times \mathbb{R}_+\,, \end{array}\right. \tag{3.11}$$

where $x^0 = (z^0, y_1^0)$. By Proposition 3.1 there exists a unique maximal solution $(z, y_1, k) : [0, \omega) \to \mathbb{R}^n \times \mathbb{R}_+$ of the initial-value problem (3.11) for some $\omega \in (0, \infty]$. To prove the proposition, we claim that it suffices to show that $y_1 \in L^2([0,\omega), \mathbb{R}^m)$. To argue this claim, assume that $y_1 \in L^2([0,\omega), \mathbb{R}^m)$ and first note that, by Proposition 3.1, $\omega = \infty$, and so assertion (i) holds. Assertion (ii) follows from the third of equations (3.11). Since $u_0, y_0, y_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ and k is bounded, we have $u_0 + k[y_0 - y_1] = u_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ and, since $\sigma(A_1) \subset \mathbb{C}_-$, it follows from the first of equations (3.11) that $z \in L^2(\mathbb{R}_+, \mathbb{R}^{n-m})$ and $z(t) \to 0$ as $t \to \infty$, whence we get assertion (iii). Finally, writing the second of equations (3.11) in the form

$$\dot{y}_1(t) = -y_1(t) + f(t), \quad f : t \mapsto (I + A_4) y_1(t) + A_3 z(t) + B_2\big(u_0(t) + k(t)[y_0(t) - y_1(t)]\big),$$

and noting that $f \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, we conclude $y_1(t) \to 0$ as $t \to \infty$, and so (iv) holds. We proceed to show that $y_1 \in L^2([0,\omega), \mathbb{R}^m)$. First, we assemble some useful inequalities. Recalling that $\sigma(A_1) \subset \mathbb{C}_-$, we have

$$M_p := \left(\int_{\mathbb{R}_+} \|e^{A_1 t}\|^p dt\right)^{1/p} < \infty \;\; \forall\, p \in [1, \infty), \quad M_\infty := \sup_{t \in \mathbb{R}_+} \|e^{A_1 t}\| < \infty,$$

and, by the first of equations (3.11) (together with elementary estimates),

$$\|z(t)\|^2 \;\le\; 2\left[M_\infty^2 \|z^0\|^2 + M_2^2 \|A_2\|^2 \|y_1\|^2_{L^2([0,t),\mathbb{R}^m)}\right] \qquad \forall\, t \in [0, \omega). \tag{3.12}$$

Introduce the convolution operator $L : L^2(\mathbb{R}_+, \mathbb{R}^{n-m}) \to L^2(\mathbb{R}_+, \mathbb{R}^{n-m})$, given by

$$(Lv)(t) := \int_0^t e^{A_1(t-\tau)} v(\tau) d\tau\,.$$

Then $\|Lv\|_{L^2(I,\mathbb{R}^{n-m})} \le M_1 \|v\|_{L^2(I,\mathbb{R}^{n-m})}$ for every bounded interval $I \subset \mathbb{R}_+$ and all $v \in L^2_e(\mathbb{R}_+, \mathbb{R}^{n-m})$, which, with the first of equations (3.11), yields

$$\|z\|^2_{L^2([s,t],\mathbb{R}^{n-m})} \le 2M_1^2\left[\|z(s)\|^2 + \|A_2\|^2 \|y_1\|^2_{L^2([s,t],\mathbb{R}^m)}\right]$$

for all s, t with $0 \le s \le t < \omega$. Writing

$$c_1 := \frac{1}{2} + M_1^2\left[1 + \|A_2\|^2\right], \tag{3.13}$$

we may infer

$$\begin{aligned} \int_s^t \|z(\tau)\|\|y_1(\tau)\| d\tau &\le \frac{1}{2}\left[\|z\|^2_{L^2([s,t],\mathbb{R}^{n-m})} + \|y_1\|^2_{L^2([s,t],\mathbb{R}^m)}\right] \\ &\le c_1\left[\|z(s)\|^2 + \|y_1\|^2_{L^2([s,t],\mathbb{R}^m)}\right] \;\text{ for all } s, t \text{ with } 0 \le s \le t < \omega. \end{aligned} \tag{3.14}$$

Also, observe that, for all s, t with $0 \le s \le t < \infty$,

$$k(t) = k(s) + \|y_0 - y_1\|^2_{L^2([s,t],\mathbb{R}^m)} \le k(s) + \|y_0\|^2_{L^2([s,t],\mathbb{R}^m)} + \|y_1\|^2_{L^2([s,t],\mathbb{R}^m)}. \tag{3.15}$$

Since $\sigma(B_2) \subset \mathbb{C}_+$, the Lyapunov equation $QB_2 + B_2^T Q - 2I = 0$ has a unique positive-definite symmetric solution Q. From the second of equations (3.11), noting that $\|QB_2\| = 1$ and invoking elementary estimates, we have

$$\begin{aligned}\langle y_1(t), Q\dot{y}_1(t)\rangle \le \|QA_3\|\|z(t)\|\|y_1(t)\| - \frac{1}{2}\big[k(t) - 2\|QA_4\| - 1\big]\|y_1(t)\|^2 \\ + \frac{1}{2}k(t)\|y_0(t)\|^2 + \frac{1}{2}\|u_0(t)\|^2 \ \forall\ t \in [0,\omega)\end{aligned}$$

which, on integration, using (3.14), (3.15) and invoking monotonicity of k, yields

$$\begin{aligned}0 \le \langle y_1(t), Qy_1(t)\rangle \le \langle y_1(s), Qy_1(s)\rangle + 2c_1\|QA_3\|\big(\|z(s)\|^2 + \|y_1\|^2_{L^2([s,t],\mathbb{R}^m)}\big) \\ + \big(k(s) + \|y_0\|^2_{L^2([s,t],\mathbb{R}^m)} + \|y_1\|^2_{L^2([s,t],\mathbb{R}^m)}\big)\|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + \|u_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} \\ - \big(k(s) - 2\|QA_4\| - 1\big)\|y_1\|^2_{L^2([s,t],\mathbb{R}^m)} \ \ \forall\ s, t \text{ with } 0 \le s \le t < \omega.\end{aligned}$$

Defining

$$c_2 := 2c_1\|QA_3\| + 2\|QA_4\| + 2 + \|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)}, \tag{3.16}$$

we have

$$\begin{aligned}0 \le \langle y_1(t), Qy_1(t)\rangle \le \langle y_1(s), Qy_1(s)\rangle + 2c_1\|QA_3\|\|z(s)\|^2 + \|u_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} \\ - \big(k(s) - c_2 + 1\big)\|y_1\|^2_{L^2([s,t],\mathbb{R}^m)} \ \forall\ s, t \text{ with } 0 \le s \le t < \omega.\end{aligned} \tag{3.17}$$

Next, observe that

$$\|y_1\|^2_{L^2([0,t),\mathbb{R}^m)} \le \|y_0\|^2_{L^2([0,t),\mathbb{R}^m)} + \|y_2\|^2_{L^2([0,t),\mathbb{R}^m)} = \|y_0\|^2_{L^2([0,t),\mathbb{R}^m)} + k^0 + k(t)$$

for all $t \in [0,\omega)$. We consider two possible cases.

Case 1. Assume $k(t) \le c_2$ for all $t \in [0,\omega)$. Then $\|y_1\|^2_{L^2([0,\omega),\mathbb{R}^m)} \le \|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + c_2$.

Case 2. Assume $k(\tau) = c_2$ for some $\tau \in [0,\omega)$. Then, by (3.17) with $s = \tau$, we have

$$\|y_1\|^2_{L^2([\tau,\omega),\mathbb{R}^m)} \le \langle y_1(\tau), Qy_1(\tau)\rangle + 2c_1\|QA_3\|\|z(\tau)\|^2 + \|u_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)}.$$

By monotonicity, $k(t) \le c_2$ for all $t \in [0,\tau]$, and so $\|y_1\|^2_{L^2([0,\tau],\mathbb{R}^m)} \le \|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + c_2$. By (3.17) with $s = 0$ and $t = \tau$, we now have

$$\begin{aligned}\langle y_1(\tau), Qy_1(\tau)\rangle \le c_3 := \langle y_1^0, Qy_1^0\rangle + 2c_1\|QA_3\|\|z^0\|^2 + \|u_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} \\ + (c_2 + 1)\big(\|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + c_2\big).\end{aligned} \tag{3.18}$$

By (3.12), we have

$$\|z(\tau)\|^2 \le c_4 := 2\big[M_\infty^2\|z^0\|^2 + M_2^2\|A_2\|^2\|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + c_2M_2^2\|A_2\|^2\big]. \tag{3.19}$$

We may now conclude that

$$\|y_1\|^2_{L^2([0,\tau],\mathbb{R}^m)} + \|y_1\|^2_{L^2([\tau,\omega),\mathbb{R}^m)} \leq c_5 := c_2 + c_3 + 2c_1c_4\|QA_3\| + \|(u_0,y_0)\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^{2m})}. \tag{3.20}$$

Therefore, in both Cases 1 and 2, we have $\|y_1\|^2_{L^2([0,\omega),\mathbb{R}^m)} \leq c_5$. ∎

Proposition 3.3 immediately implies the following.

COROLLARY 3.4. *Let $n \geq m$, $\mathcal{U} = \mathcal{Y} = L^2(\mathbb{R}_+,\mathbb{R}^m)$, $\theta = (A,B,C) \in \mathcal{M}_n$, and $(x^0,k^0) \in \mathbb{R}^n \times \mathbb{R}_+$. Then the closed loop $[\widetilde{P}(\theta,x^0),\widetilde{C}(k^0)]$ given by* (3.1) (*equivalently,* (3.5)), (3.6), (3.8), (3.9) *is globally well posed and $(\mathcal{U}\times\mathcal{Y})$-stable.*

PROPOSITION 3.5. *Let $n \geq m$ and define*

$$\mathcal{D} := \mathcal{M}_n \times \mathbb{R}^n \times \mathbb{R}_+ \times L^2(\mathbb{R}_+,\mathbb{R}^m) \times L^2(\mathbb{R}_+,\mathbb{R}^m). \tag{3.21}$$

There exists a continuous map $\nu : \mathcal{D} \to \mathbb{R}_+$ such that, for all

$$d = (A,B,C,x^0,k^0,u_0,y_0) \in \mathcal{D},$$

the closed-loop system (3.11) *is such that $\|(u_1,y_1)\|_{L^2(\mathbb{R}_+,\mathbb{R}^{2m})} \leq \nu(d)$.*

Proof. Observe that the parameters c_i, $i = 1,\dots,5$, defined in (3.13), (3.16), (3.18), (3.19), and (3.20), depend continuously on the data $d = (A,B,C,x^0,k^0,u_0,y_0)$. In particular, the map $\hat{\nu} : d \mapsto \sqrt{c_5}$ is continuous. Let $d \in \mathcal{D}$ be arbitrary. Then, by Proposition 3.3 (and recalling (3.20)), we have $\|y_1\|_{L^2(\mathbb{R}_+,\mathbb{R}^m)} \leq \hat{\nu}(d)$. Now,

$$k(t) = k^0 + \|y_0 - y_1\|^2_{L^2([0,t),\mathbb{R}^m)} \leq \nu^*(d) := k^0 + 2\|y_0\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + 2(\hat{\nu}(d))^2 \quad \forall\, t \in \mathbb{R}_+.$$

Therefore,

$$\begin{aligned}\|u_1\|_{L^2([0,t),\mathbb{R}^m)} &= \|u_0 + k(t)[y_0(t) - y_1(t)]\|_{L^2([0,t),\mathbb{R}^m)} \\ &\leq \tilde{\nu}(d) := \|u_0\|_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + \nu^*(d)\big(\|y_0\|_{L^2(\mathbb{R}_+,\mathbb{R}^m)} + \hat{\nu}(d)\big) \ \ \forall\, t \in [0,\omega).\end{aligned}$$

We may now infer that $\|(u_1,y_1)\|_{L^2(\mathbb{R}_+,\mathbb{R}^{2m})} \leq \nu(d) := \sqrt{(\hat{\nu}(d))^2 + (\tilde{\nu}(d))^2}$. ∎

Remark 3.6. It is worthwhile to note that $\nu(d) \to \infty$ as the data approach the boundary of $\mathcal{M}_n$: for example, if some eigenvalues of A_1 approach the imaginary axis, then $\|L\| \to \infty$ and so c_1, given by (3.13), grows unboundedly; if $\|B_2\| \to 0$, then $\|Q\| \to \infty$ and so c_2, given by (3.16), grows unboundedly. Specifically, there exists a *bounded* sequence (d_j) in $\mathcal{D}$ such that $\nu(d_j) \to \infty$ as $j \to \infty$. However, if $\Omega \subset \mathcal{M}_n$ is closed and (d_j) is a bounded sequence in $\Omega \times \mathbb{R}^n \times \mathbb{R}_+ \times L^2(\mathbb{R}_+,\mathbb{R}^m) \times L^2(\mathbb{R}_+,\mathbb{R}^m) \subset \mathcal{D}$, then $(\nu(d_j))$ is bounded.

3.3. Construction of a gain function. To establish gap margin results, we will need to construct augmented plant and controller operators, as in [4]. Reiterating earlier remarks, we may consider $\mathcal{M}_n$ to be a subset of the Euclidean space $E_n = \mathbb{R}^{n^2+2mn}$, with the standard Euclidean norm, by identifying a plant $\theta = (A,B,C) \in \mathcal{M}_n$ with a vector $\theta \in E_n$ consisting of the $n^2 + 2mn$ elements of the plant matrices ordered lexicographically. Note that $0 \notin \mathcal{M}_n$. Let $\mathcal{U} = \mathcal{Y} = L^2(\mathbb{R}_+,\mathbb{R}^m)$ and define $\widetilde{\mathcal{U}} := \mathbb{R}^{n^2+2mn} \times \mathcal{U}$ and $\widetilde{\mathcal{W}} := \widetilde{\mathcal{U}} \times \mathcal{Y}$, which can be considered as signal spaces by identifying $\theta \in \mathbb{R}^{n^2+2mn}$ with the constant function $t \mapsto \theta$ and endowing $\widetilde{\mathcal{U}}$ with the norm $\|(\theta,u)\|_{\widetilde{\mathcal{U}}} := \sqrt{\|\theta\|^2 + \|u\|^2_{L^2(\mathbb{R}_+,\mathbb{R}^m)}}$. For given $\widetilde{P}(\theta,0)$ as in (3.6), we define the (augmented) plant operator as

$$P : \widetilde{\mathcal{U}}_a \to \mathcal{Y}_a, \quad (\theta,u_1) = \widetilde{u}_1 \mapsto y_1 = P(\widetilde{u}_1) := \widetilde{P}(\theta,0)(\widetilde{u}_1). \tag{3.22}$$

Fix $k^0 \geq 0$ and define, for $\widetilde{C}(k^0)$ as in (3.8), the (augmented) controller operator as

$$C:\ \mathcal{Y}_a \to \widetilde{\mathcal{U}}_a, \quad y_2 \mapsto \widetilde{u}_2 = C(y_2) := (0, \widetilde{C}(k^0)(y_2)). \tag{3.23}$$

For each nonempty $\Omega \subset \mathcal{M}_n$, define $\mathcal{W}^\Omega := (\Omega \times \mathcal{U}) \times \mathcal{Y}$ and $H^\Omega_{P,C} := H_{P,C}|_{\mathcal{W}^\Omega}$. It easily follows from Corollary 3.4 that $H^\Omega_{P,C}\colon \mathcal{W}^\Omega \to \widetilde{\mathcal{W}} \times \widetilde{\mathcal{W}}$ is a causal operator for any $\Omega \subset \mathcal{M}_n$. We now establish gf-stability.

PROPOSITION 3.7. *Let $n \geq m$, $k^0 \geq 0$, and assume $\Omega \subset \mathcal{M}_n$ is closed. Then, for the closed-loop system $[P, C]$ given by* (2.1), (3.22), *and* (3.23), *the operator $H^\Omega_{P,C}$ is gf-stable.*

Proof. For $\nu : \mathcal{D} \to \mathbb{R}_+$ as in Proposition 3.5 we have, for all $(\theta, u_0, y_0) \in \mathcal{W}^\Omega$,

$$\begin{aligned}\|H^\Omega_{P,C}(\theta, u_0, y_0)\|_{\widetilde{\mathcal{W}}\times\widetilde{\mathcal{W}}} &\leq \|(\theta, u_0, y_0)\|_{\widetilde{\mathcal{W}}} + 2\|(\theta, u_1, y_1)\|_{\widetilde{\mathcal{W}}} \\ &\leq \|(u_0, y_0)\|_{\mathcal{W}} + 3\,\|\theta\| + 2\,\nu(\theta, 0, k^0, u_0, y_0)\,,\end{aligned}$$

and so, for $r_0 := \inf_{w\in\mathcal{W}^\Omega} \|w\|_{\widetilde{\mathcal{W}}}$ and $\alpha \in (r_0, \infty)$, closedness of Ω yields

$$\gamma(\alpha) := \sup\left\{ \|(u_0, y_0)\|_{\mathcal{W}} + 3\|\theta\| + 2\,\nu(\theta, 0, k^0, u_0, y_0) \,\middle|\, \begin{array}{l}(\theta, u_0, y_0) \in \mathcal{W}^\Omega, \\ \|(\theta, u_0, y_0)\|_{\widetilde{\mathcal{W}}} \leq \alpha\end{array} \right\} < \infty.$$

Therefore, a gain function for $H^\Omega_{P,C}$ exists, and the proof is complete. ∎

3.4. Robust stability. In Proposition 3.3 and Corollary 3.4 we have established that, for $k^0 \geq 0$, $(\theta, x^0) \in \mathcal{M}_n \times \mathbb{R}^n$ for some $n \geq m$, and $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, the closed-loop system $[\widetilde{P}(\theta, x^0), \widetilde{C}(k^0)]$ is globally well posed and has desirable stability properties. The purpose of this section is to determine conditions under which these properties are maintained when the plant $\widetilde{P}(\theta, x^0)$ is perturbed to a plant $\widetilde{P}(\theta_1, x^0_1)$, where $(\theta_1, x^0_1) \in \mathcal{P}_q \times \mathbb{R}^q$ for some $q \geq m$, in particular when $\theta_1 \notin \mathcal{M}_q$. The essence of the main result, Theorem 3.8, is that the stability properties persist if (a) the plants $\widetilde{P}(\theta_1, 0)$ and $\widetilde{P}(\theta, 0)$ are sufficiently close (in the gap sense) and (b) the initial data x^0_1 and disturbance $w_0 = (u_0, y_0)$ are sufficiently small.

THEOREM 3.8. *Let $m, n, q \in \mathbb{N}$ with $n \geq m$, $q \geq m$, $\mathcal{U} = \mathcal{Y} = L^2(\mathbb{R}_+, \mathbb{R}^m)$, $\mathcal{W} = \mathcal{U} \times \mathcal{Y}$, and $\theta \in \mathcal{M}_n$. For (ϑ, x^0) in $\mathcal{P}_q \times \mathbb{R}^q$ or $\mathcal{P}_n \times \mathbb{R}^n$ and $k^0 \geq 0$, consider $\widetilde{P}(\vartheta, x^0)\colon \mathcal{U}_a \to \mathcal{Y}_a$ and $\widetilde{C}(k^0)\colon \mathcal{Y}_a \to \mathcal{U}_a$ as defined by* (3.6) *and* (3.8), *respectively. There exist a continuous function $\eta\colon \mathbb{R}_+ \to (0, \infty)$ and a function $\lambda : \mathcal{P}_q \to (0, \infty)$ such that the following holds. For all $(\theta_1, x^0_1, w_0, r) \in \mathcal{P}_q \times \mathbb{R}^q \times \mathcal{W} \times (0, \infty)$,*

$$\left.\begin{array}{l}\vec{\delta}(\widetilde{P}(\theta, 0), \widetilde{P}(\theta_1, 0)) \leq\ \eta(r) \\ \lambda(\theta_1)\|x^0_1\| + \|w_0\|_{\mathcal{W}} \leq r\end{array}\right\} \quad \Longrightarrow \quad H_{\widetilde{P}(\theta_1, x^0_1), \widetilde{C}(k^0)}(w_0)\ \in\ \mathcal{W} \times \mathcal{W}. \tag{3.24}$$

Remark 3.9. In the setup of Theorem 3.8, if $H_{\widetilde{P}(\theta_1, x^0_1), \widetilde{C}(k^0)}(w_0) \in \mathcal{W} \times \mathcal{W}$ with $\theta_1 = (A, B, C) \in \mathcal{P}_q$, then the following hold:

(i) If $u_1, y_1 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, then detectability of (A, C) yields that the solution x of (3.1) belongs to $x \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. Since $x, \dot{x} \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, it follows that $x(t) \to 0$ as $t \to \infty$. Moreover, the monotone controller gain k, given by (3.7), converges to a finite limit.

(ii) If $u_0, y_0 \in (L^2 \cap L^\infty)(\mathbb{R}_+, \mathbb{R}^m)$, then $u_1, y_1 \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$. This follows from the fact that $x \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$ by (i), and so $y_1 \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$. Furthermore, $y_2 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, and so $k \in L^\infty(\mathbb{R}_+, \mathbb{R})$, which, by $u_2 = k\, y_2$, yields $u_2 \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$, whence $u_1 \in L^\infty(\mathbb{R}_+, \mathbb{R}^m)$.

(iii) If $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ and $\lim_{t\to\infty} u_0(t) = \lim_{t\to\infty} y_0(t) = 0$, then $\lim_{t\to\infty} u_1(t) = 0$. This is a consequence of (i), which gives $\lim_{t\to\infty} y_1(t) = 0$, and therefore $\lim_{t\to\infty} y_2(t) = 0$, which, together with boundedness of k, yields $\lim_{t\to\infty} u_2(t) = 0$ and $\lim_{t\to\infty} u_1(t) = 0$.

Example 3.10. As an illustrative example, we consider example (ii) from the table in the introduction, where P and P_1 are specified in the frequency domain by the associated transfer functions

$$\hat{P}_1(s) = \frac{1}{s-\theta}, \qquad \hat{P}_2(s) = \frac{N(M-s)}{(N+s)(M+s)(s-\theta)}, \quad N, M \geq \varepsilon^{-1}\,,\ \varepsilon > 0.$$

Note that $\hat{P}_1$ has a realization $\widetilde{P}(\theta, 0) \in \mathcal{M}_1$ and $\hat{P}_2$ has a realization $\widetilde{P}(\theta_1, 0) \in \mathcal{P}_3 \setminus \mathcal{M}_3$. We claim that $\vec{\delta}_0(\hat{P}_1, \hat{P}_2) \to 0$ as $\varepsilon \to 0+$. To prove this assertion, note that

$$A(s) = \frac{s-\theta}{s+\sqrt{\theta^2+1}}, \qquad B(s) = \frac{1}{s+\sqrt{\theta^2+1}}$$

satisfy $\hat{P}_1(s) = B(s)A(s)^{-1}$, $A, B \in H^\infty$, and $A^*(s)A(s) + B^*(s)B(s) = I$. Therefore, A, B form a normalized right coprime factorization of $\hat{P}_1$. Since

$$\hat{P}_2(s) = \frac{B(s) + \Delta_B(s)}{A(s)}, \quad \text{where} \qquad \Delta_B(s) = \left(\frac{N(M-s)}{(s+N)(s+M)} - 1\right) B(s),$$

and $\Delta_B \in H^\infty$, by (2.5) it suffices to show that

$$\left\| \begin{pmatrix} 0 \\ \Delta_B \end{pmatrix} \right\|_{H^\infty} = \left\| \left(\frac{N(M-s)}{(s+N)(s+M)} - 1 \right) B(s) \right\|_{H^\infty} \to 0 \quad \text{as } \varepsilon \to 0+, \tag{3.25}$$

and this follows from a routine calculation. Thus the claim is proved. To apply Theorem 3.8 to conclude robust stability, it would suffice to show that $\vec{\delta}(\hat{P}_1, \hat{P}_2) \to 0$ as $\varepsilon \to 0^+$. In view of the equivalence (2.5), the latter could be shown by establishing that the directed gap $\vec{\delta}_0(\hat{P}_2, \hat{P}_1)$ is less than 1 (recall that $\vec{\delta}_0(\hat{P}_1, \hat{P}_2) \to 0$ as $\varepsilon \to 0^+$). Alternatively, anticipating Lemma 4.5, we can adapt the proof of that lemma (to the case wherein A, B are defined as above and $A' = A$, $B' = B + \Delta_B$, $(V, U) := (A^*, B^*)$), and invoke (3.25), to conclude that

$$\vec{\delta}(\widetilde{P}(\theta, 0), \widetilde{P}(\theta_1, 0)) \leq \left\| \begin{pmatrix} 0 \\ \Delta_B \end{pmatrix} \right\| \cdot \|(A^*, B^*)\|_{H^\infty} = \left\| \begin{pmatrix} 0 \\ \Delta_B \end{pmatrix} \right\| \to 0 \quad \text{as } \varepsilon \to 0+.$$

Thus, under the conditions of Theorem 3.8, the controller $\widetilde{C}(k^0)\colon \mathcal{Y}_a \to \mathcal{U}_a$ defined by (3.6), (3.8) stabilizes any stabilizable and detectable realization of $\hat{P}_2$. As observed in the introduction, $\hat{P}_2$ is an example of a plant which violates all the classical assumptions of adaptive control.

To prove Theorem 3.8 we need to show how the gf-stability of the augmented closed loop $[P, C]$ as given in (3.22) and (3.23) yields the robustness property (3.24) for the unaugmented closed loop $[\widetilde{P}(\theta_1, x_1^0), \widetilde{C}(k^0)]$. This follows from the next result, which is a direct consequence of Theorems 5.2 and 5.3 in [4].

THEOREM 3.11. *Let $m, n, q \in \mathbb{N}$ with $n \geq m$, $q \geq m$, $\mathcal{U} = \mathcal{Y} = L^p(\mathbb{R}_+, \mathbb{R}^m)$, $1 \leq p \leq \infty$, and $\mathcal{W} = \mathcal{U} \times \mathcal{Y}$. Let $\widetilde{K}\colon \mathcal{Y}_a \to \mathcal{U}_a$ be causal, and consider $\widetilde{P}(\vartheta, 0)\colon \mathcal{U}_a \to \mathcal{Y}_a$ defined in* (3.6) *for (ϑ, x^0) in $\mathcal{P}_q \times \mathbb{R}_q$ or $\mathcal{P}_n \times \mathbb{R}_n$. Assume that $[\widetilde{P}(\vartheta, 0), \widetilde{K}]$ is regularly well posed for all $\vartheta \in \mathcal{P}_q$ and let $\Omega \subset \mathcal{M}_n$ be closed. Define*

$$\begin{aligned} P\colon \mathcal{P}_n \times \mathcal{U}_a \to \mathcal{Y}_a, \qquad (\vartheta, u_1) \mapsto P(\vartheta, u_1), &= \widetilde{P}(\vartheta, 0)(u_1), \\ C\colon \mathcal{Y}_a \to \mathcal{P}_n \times \mathcal{U}_a, \qquad y_2 \mapsto C(y_2) &= (0, \widetilde{K}(y_2)). \end{aligned}$$

If $H_{P,C}|_{\Omega\times\mathcal{W}}$ is gf-stable and $T_\tau\Pi_{P//C}$ is continuous for all $\tau>0$, then there exist a continuous function $\mu\colon \mathbb{R}_+\times\Omega\to(0,\infty)$ and a function $\lambda:\mathcal{P}_q\to(0,\infty)$ such that, for all $(\theta_1,\theta,x_1^0,w_0,r)\in\mathcal{P}_q\times\Omega\times\mathbb{R}^q\times\mathcal{W}\times(0,\infty)$,

$$\left.\begin{array}{l}\vec{\delta}(\widetilde{P}(\theta,0),\widetilde{P}(\theta_1,0))\le\mu\big(r\,,\,\theta\big)\\ \lambda(\theta_1)\|x_1^0\|+\|w_0\|_{\mathcal{W}}\le r\end{array}\right\}\quad\Longrightarrow\quad H_{\widetilde{P}(\theta_1,x_1^0),\widetilde{K}}(w_0)\;\in\;\mathcal{W}\times\mathcal{W}. \tag{3.26}$$

We are now in a position to prove Theorem 3.8.

Proof of Theorem 3.8. Let $\theta\in\mathcal{M}_n$ and define $\Omega=\{\theta\}$. Consider Theorem 3.11 with $p=2$ and $\widetilde{K}=\widetilde{C}(k^0)$, where $\widetilde{C}(k^0)$ is given by (3.8). Note that by Corollary 3.2, the closed loop $[\widetilde{P}(\vartheta,0),\widetilde{C}(k^0)]$ is regularly well posed for all $\vartheta\in\mathcal{P}_q$. For P and C as defined in Theorem 3.11, the operator $H^{\Omega}_{P,C}$ is gf-stable by Proposition 3.7. By, for example, the proof of Theorem 4.D in [20], $T_\tau\Pi_{\widetilde{P}(\theta,0)//\widetilde{C}(k^0)}$ is continuous for all $\tau>0$, and therefore $T_\tau\Pi_{P//C}|_{\Omega\times\mathcal{W}}$ is continuous for all $\tau>0$. Now all hypotheses of Theorem 3.11 are in place and so there exist a continuous function $\mu\colon\mathbb{R}_+\times\Omega\to(0,\infty)$ and a function $\lambda:\mathcal{P}_q\to(0,\infty)$ such that (3.26) holds. Statement (3.24) follows on setting $\eta(\cdot)=\mu(\cdot,\theta)$. ∎

4. Georgiou and Smith's example revisited. In this section we reconsider Example 9 in the paper by Georgiou and Smith [7] (see also [9]). This serves two purposes, which are to clarify some of the informal arguments therein in relation to robustness with respect to initial data, and to demonstrate that, in the L^2 setting of the present paper, the robustness bound in Theorem 3.8 is qualitatively tight in the sense that it is necessarily dependent on the data u_0, y_0, x_1^0. In those cases where we demonstrate a lack of stability, namely for large initial conditions, large L^2 disturbances, or L^∞ disturbances, the instability is caused by the adaptive gain k undergoing the well-known phenomena of "parameter drift" [13, p. 545], leading to k passing some critical value, thus inducing instability. The conditions given for stability, namely small initial conditions and small L^2 disturbances, prevent this instability mechanism by ensuring that the adaptive gain k never reaches criticality. In general, the issue of "drift-inducing" disturbances has led to various "robustifying" modifications to adaptive designs: one such modification is the introduction of a dead-zone in the second of equations (1.1), to become $\dot{k}=|y|\max\{0,|y|-\lambda\}$, with parameter $\lambda>0$. Such a modification has led to practical applications to, for example, exothermic chemical reactors [12]. The robustness in the graph topology of such modifications remains to be investigated.

4.1. The nominal and perturbed closed-loop systems. After appropriate rescaling and relabeling of variables, the first-order linear plant $\widetilde{P}(a,y_1^0)$, considered in [7, Example 9], can be parameterized by $a\in\mathbb{R}$ and $y_1^0\in\mathbb{R}$ and expressed as

$$\widetilde{P}\big(a,y_1^0\big)\colon\mathcal{U}_a\to\mathcal{Y}_a,\quad u_1\mapsto y_1,\quad\text{where }\dot{y}_1=ay_1+u_1,\;\;y_1(0)=y_1^0, \tag{4.1}$$

and so, for $u_1\in\mathcal{U}_a$, $y_1=\widetilde{P}(a,y_1^0)(u_1):\mathrm{dom}(u_1)\to\mathbb{R}$ is the unique maximal solution of the initial-value problem in (4.1). The controller, parameterized by $k^0\in\mathbb{R}_+$, is

$$\widetilde{C}(k^0)\colon\;\mathcal{Y}_a\to\mathcal{U}_a,\;y_2\mapsto u_2:=-k\,y_2,\quad\text{where }\dot{k}=y_2^2,\;\;k(0)=k^0, \tag{4.2}$$

and so, for $y_2\in\mathcal{Y}_a$, $-ky_2=u_2=\widetilde{C}(k^0)(y_2):\mathrm{dom}(y_2)\to\mathbb{R}$, where k is the unique maximal solution of the initial-value problem in (4.2). The closed-loop system $[\widetilde{P}(a,y_1^0),\widetilde{C}(k^0)]$ will be analyzed in the two settings of $\mathcal{U}=\mathcal{Y}=L^2(\mathbb{R}_+,\mathbb{R})$ and

$\mathcal{U} = \mathcal{Y} = L^\infty(\mathbb{R}_+, \mathbb{R})$. In view of Proposition 3.1(i), for all $a, y_1^0 \in \mathbb{R}$ and $k^0 \in \mathbb{R}_+$, the closed-loop system $[\widetilde{P}(a, y_1^0), \widetilde{C}(k^0)]$ is locally well posed in both settings; moreover, by Proposition 3.3, in the former L^2 setting, the closed-loop system is globally well posed, and the signals y_1 and k are bounded, with $y_1(t) \to 0$ and $k(t) \to k^\infty \in \mathbb{R}$ as $t \to \infty$. As in [7, Example 9], consider a perturbation of the plant $\widetilde{P}(a, y_1^0)$ consisting of the series connection of $1/(s-a)$ (i.e., the transfer function associated with $\widetilde{P}(a, y_1^0)$) and an all-pass factor $(M-s)/(M+s)$ with $M > 0$. As a realization of this series connection, consider

$$\dot{y}_1 = a\, y_1 + z - u_1\,, \quad \dot{z} = -M\,[z - 2\,u_1]\,, \quad (y_1(0), z(0)) = (y_1^0, z^0)\,.$$

This series connection is denoted by

$$\widetilde{P}_1\left(a, M, y_1^0, z^0\right) : \mathcal{U}_a \to \mathcal{Y}_a, \quad u_1 \mapsto y_1\,. \tag{4.3}$$

The closed-loop equations for $[\widetilde{P}_1\left(a, M, y_1^0, z^0\right), \widetilde{C}(k^0)]$ are given by

$$\left\{\begin{array}{ll} \dot{y}_1(t) = [a + k(t)]\, y_1(t) + z(t) - k(t) y_0(t) - u_0(t)\,, & y_1(0) = y_1^0, \\ \dot{z}(t) = -2Mk(t) y_1(t) - Mz(t) + 2Mk(t) y_0(t) + 2M u_0(t)\,, & z(0) = z^0, \\ \dot{k}(t) = (y_0(t) - y_1(t))^2\,, & k(0) = k^0, \\ u_1(t) = u_0(t) + k(t)\, y_0(t) - k(t)\, y_1(t). & \end{array}\right. \tag{4.4}$$

For fixed (but arbitrary) $a \in \mathbb{R}$ and $k^0 \in \mathbb{R}_+$, and applying Theorem 3.8, we may conclude the existence of a continuous function $\eta : \mathbb{R}_+ \to (0, \infty)$ and a function $\lambda_1 : \mathbb{R} \to (0, \infty)$ such that, if $\vec{\delta}(\widetilde{P}(a, 0), \widetilde{P}_1(a, M, 0, 0)) \leq \eta(r)$ for some $r > 0$, then, for all initial data $x_1^0 = (y_1^0, z^0)$ and all disturbances $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R})$ satisfying $\lambda_1(M)\|x_1^0\| + \|(u_0, y_0)\|_{L^2(\mathbb{R}_+, \mathbb{R}^2)} \leq r$, the closed-loop system is globally well posed and is such that $(y_1(t), z(t)) \to (0, 0)$ as $t \to \infty$, and the monotone gain k converges to a finite limit. At this point, we briefly digress to prove a technicality which will prove convenient in later analyses.

LEMMA 4.1. *Let $M > 0$, $a = 0$, $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R}) \cup L^\infty(\mathbb{R}_+, \mathbb{R})$, and $(y_1^0, z^0, k^0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$. Let $(y_1, z, k) : [0, \omega) \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$ be the unique maximal solution to the closed-loop initial-value problem* (4.4). *If there exists $T \in [0, \omega)$ such that $k(T) \geq 4M$, $(y_0(T) - y_1(T), z(T)) \neq (0, 0)$, and $(u_0(t), y_0(t)) = (0, 0)$ for all $t \in [T, \infty)$, then*

(i) *$u_1, y_1 \notin L^\infty([0, \omega), \mathbb{R})$ and*

(ii) *$k \notin L^\infty([0, \omega), \mathbb{R}_+)$.*

Proof. Writing $y_2 := y_0 - y_1$, then, by the hypotheses, we get

$$\left.\begin{array}{l} \dot{y}_2(t) = k(t) y_2(t) - z(t), \quad \dot{z}(t) = 2Mk(t) y_2(t) - Mz(t) \\ \dot{k}(t) = \big(y_2(t)\big)^2 \end{array}\right\} \quad \forall\ t \in [T, \omega). \tag{4.5}$$

Defining $\eta := y_2 - z/(2M)$, we have

$$\left.\begin{array}{l} \dot{y}_2(t) = \big[k(t) - 2M\big] y_2(t) + 2M\eta(t) \\ \dot{\eta}(t) = -My_2(t) + M\eta(t), \quad \dot{k}(t) = \big(y_2(t)\big)^2 \end{array}\right\} \quad \forall\ t \in [T, \omega). \tag{4.6}$$

Introduce $W : [T, \omega) \to \mathbb{R}_+$, $t \mapsto \frac{1}{2}\big[y_2^2 + 2\eta^2\big](t)$. By hypothesis, $(y_2(T), z(T)) \neq (0, 0)$ and so $W(T) > 0$. Moreover, since $k(T) \geq 4M$, we have

$$\dot{W}(t) = \big(k(t) - 2M\big)(y_2(t))^2 + 2M(\eta(t))^2 \geq 2MW(t) \quad \forall\ t \in [T, \omega)\,.$$

Therefore,

$$W(t) \geq e^{2M(t-T)}W(T) \quad \forall\, t \in [T,\omega), \qquad W(T) > 0. \tag{4.7}$$

Seeking a contradiction, suppose that y_2 is bounded on $[T,\omega)$. Then $\omega = \infty$ (to see this, simply note that, if ω is finite, then y_2 is square integrable and so k is bounded, which, together with Proposition 3.1(ii), yields a contradiction). Let $c_0 > 0$ be such that $\big(y_2(t)\big)^2 \leq c_0$ for all $t \in [T,\infty)$, and so $0 \leq k(t) \leq k(T) + c_0[t-T]$ for all $t \geq 0$. By (4.5), there exists $c_1 > 0$ such that $|z(t)| \leq c_1\big[1+t\big]$ for all $t \in [T,\infty)$. Hence, there exists $c_2 > 0$ such that $|\eta(t)| \leq c_2[1+t]$ for all $t \in [T,\infty)$. Therefore, it follows that $W(t) \leq \frac{1}{2}\big[c_0 + 2c_2^2(1+t)^2\big]$ for all $t \geq T$, which contradicts (4.7). Therefore, y_2 is unbounded on $[T,\omega)$ and so, since $y_1(t) = -y_2(t)$ for all $t \in [T,\omega)$, we have $y_1 \notin L^\infty([0,\omega),\mathbb{R})$. Finally, and again seeking a contradiction, suppose that k is bounded. Then, by Proposition 3.1(ii), $\omega = \infty$. By the third of equations (4.5), $y_2 \in L^2([T,\infty),\mathbb{R})$. By the second of equations (4.5), we may conclude that $z \in L^2([T,\infty),\mathbb{R})$ and $z(t) \to 0$ as $t \to \infty$. Rewriting the first of equations (4.5) in the form $\dot{y}_2(t) = -y_2(t) + \zeta(t)$, with $\zeta(t) := [1+k(t)]y_2(t) - z(t)$, and noting that $\zeta \in L^2(\mathbb{R}_+,\mathbb{R})$, it follows that $y_2(t) \to 0$ as $t \to \infty$. Therefore, $W(t) \to 0$ as $t \to \infty$, which contradicts (4.7). Therefore, k is unbounded, and so property (ii) holds. Since both k, y_1 are unbounded and k is monotone, and $u_1(t) = k(t)y_1(t)$ for all $t \geq T$, it follows that $u_1 \notin L^\infty([0,\omega),\mathbb{R})$. Therefore, property (i) holds. ∎

4.2. Nonrobustness with respect to large initial conditions.

PROPOSITION 4.2. *For $M, a, y_1^0, z^0, k^0 \in \mathbb{R}$ and $u_0, y_0 \in L^2(\mathbb{R}_+,\mathbb{R})$, consider the closed-loop system $[\widetilde{P}_1\left(M,a,y_1^0,z^0\right), \widetilde{C}(k^0)]$ defined by (2.1), (4.4), (4.2), (4.3) in the specific case wherein $M > 0$, $a = z^0 = k^0 = 0$, and $u_0 = y_0 = 0$.*

There exists $\chi > 0$ such that, if $(y_1^0)^2 > \chi$, then the unique maximal solution $(y_1, z, k) : [0,\omega) \to \mathbb{R}^2 \times \mathbb{R}_+$, $0 < \omega \leq \infty$, of the closed-loop system $[\widetilde{P}_1\left(M,0,y_1^0,0\right), \widetilde{C}(0)]$ has the following properties:

(i) *$u_1, y_1 \notin L^\infty([0,\omega),\mathbb{R})$;*

(ii) *$k \notin L^\infty([0,\omega),\mathbb{R}_+)$.*

Proof. Under the above hypothesis, the initial-value problem (4.4) is given by

$$\begin{cases} \dot{y}_1(t) = k(t)y_1(t) + z(t), \quad \dot{z}(t) = -2Mk(t)y_1(t) - Mz(t), \ \dot{k}(t) = (y_1(t))^2, \\ (y_1(0), z(0), k(0)) = (y_1^0, 0, 0). \end{cases} \tag{4.8}$$

Let $(y_1, z, k) : [0,\omega) \to \mathbb{R}^2 \times \mathbb{R}_+$ be the unique maximal solution of (4.8) with

$$\big(y_1^0\big)^2 > \chi := \big(32M^2\big(1+64M^2\big) + 4\big(1+8M^2\big)^2\big)/\big(1-e^{-4M}\big). \tag{4.9}$$

We will consider separately the two possible cases (a) $\omega < \infty$, and (b) $\omega = \infty$.

Case (a). Assume $\omega < \infty$. Then, by Proposition 3.1, the monotone function k is unbounded. This, in turn, implies that $y_1 \notin L^2([0,\omega),\mathbb{R})$, and so $y_1 \notin L^\infty([0,\omega),\mathbb{R})$. Therefore, properties (i) and (ii) hold.

Case (b). Now assume $\omega = \infty$. For later convenience, we observe that, by (4.8),

$$\|z\|_{L^2([0,t);\mathbb{R})} \leq 2\,\|ky_1\|_{L^2([0,t),\mathbb{R})} \quad \forall\, t \geq 0. \tag{4.10}$$

First, we will show that $k(1) > 4M$. For contradiction, suppose otherwise. Then, $\|y_1\|^2_{L^2([0,1],\mathbb{R})} = k(1) \leq 4M$ and, by monotonicity of k, $k(t) \in [0,4M]$ for all $t \in [0,1]$. From (4.10), it now follows that

$$\|y_1\|^2_{L^2([0,1],\mathbb{R})} + \|z\|^2_{L^2([0,1];\mathbb{R})} \leq 4M\big[1+64M^2\big]. \tag{4.11}$$

Define $V : [0,1] \to \mathbb{R}_+$, $t \mapsto \frac{1}{2}\big(y_1^2 + z^2\big)(t)$. Then,

$$\begin{aligned}\dot V(t) &= k(t)(y_1(t))^2 + \big(1 - 2Mk(t)\big)y_1(t)z(t) - M(z(t))^2\\ &\geq -(1+8M^2)|y_1(t)z(t)| - M(z(t))^2\\ &\geq -2M(z(t))^2 - (4M)^{-1}\big[(1+8M^2)y_1(t)\big]^2\\ &\geq -4MV(t) - (4M)^{-1}\big[(1+8M^2)y_1(t)\big]^2 \quad \forall\, t \in [0,1].\end{aligned}$$

Therefore,

$$\begin{aligned}\|V\|_{L^1([0,1],\mathbb{R}_+)} &\geq (4M)^{-1}\big(1 - e^{-4M}\big)\, V(0) - (4M)^{-2}(1+8M^2)^2\, \|y_1\|^2_{L^2([0,1],\mathbb{R})}\\ &\geq (8M)^{-1}\big(1 - e^{-4M}\big)\, \chi - (4M)^{-1}(1+8M^2)^2 > 2M\big(1+64M^2\big),\end{aligned}$$

which, in conjunction with (4.11), yields the contradiction

$$4M\big(1+64M^2\big) \geq \|y_1\|^2_{L^2([0,1],\mathbb{R})} + \|z\|^2_{L^2([0,1],\mathbb{R})} = 2\|V\|_{L^1([0,\tau],\mathbb{R}_+)} > 4M\big(1+64M^2\big).$$

Therefore, $k(1) > 4M$. Moreover, since $y_1^0 \neq 0$, we may infer from (4.8) that $(y_1(1), z(1)) \neq (0,0)$. The result follows by application of Lemma 4.1 (with $T = 1$). □

4.3. Nonrobustness with respect to large L^2 disturbances.

PROPOSITION 4.3. *For $M, a, y_1^0, z^0, k^0 \in \mathbb{R}$ and $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R})$, consider the closed-loop system $[\widetilde P_1\,(M, a, y_1^0, z^0)\,, \widetilde C(k^0)]$ defined by (2.1), (4.4), (4.2), (4.3) in the specific case wherein $M > 0$, $a = y_1^0 = z^0 = k^0 = 0$, and $y_0 = 0$.*

There exists $u_0 \in L^2(\mathbb{R}_+, \mathbb{R})$ such that the unique maximal solution $(y_1, z, k) : [0, \omega) \to \mathbb{R}^2 \times \mathbb{R}_+$ of the closed-loop system $[\widetilde P_1\,(M,0,0,0)\,, \widetilde C(0)]$ has the following properties:

(i) $y_1 \notin L^\infty([0,\omega), \mathbb{R})$;

(ii) $k \notin L^\infty([0,\omega), \mathbb{R}_+)$.

Proof. Let $a = y_1^0 = z^0 = k^0 = 0$, $M > 0$, and $y_0 = 0$. Fix $r \neq 0$, and denote, by $(\tilde y_1, \tilde z, \tilde k) : [0, \tilde\omega) \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$, the unique maximal solution of (4.4) with u_0 given by $u_0(t) = r$ for all $t \geq 0$, in which case we have

$$\left\{\begin{array}{l}\frac{d}{dt}\,[\tilde y_1(t) - r] = -[\tilde y_1(t) - r] + [1 + \tilde k(t)]\,\tilde y_1(t) + [\tilde z(t) - 2r],\\ \frac{d}{dt}\,[\tilde z(t) - 2r] = -M\,[\tilde z(t) - 2r] - 2M\tilde k(t)\,\tilde y_1(t),\\ \frac{d}{dt}\,\tilde k(t) = (\tilde y_1(t))^2, \qquad (\tilde y_1(0), \tilde z(0), \tilde k(0)) = (0,0,0).\end{array}\right. \tag{4.12}$$

For contradiction, suppose that the component $\tilde k$ is bounded. Then, by Proposition 3.1, $\tilde\omega = \infty$. Since $\tilde k \in L^\infty(\mathbb{R}_+)$, it follows that $\tilde y_1 \in L^2(\mathbb{R}_+)$ and so, by the second differential equation in (4.12), we may infer that $\tilde z(\cdot) - 2r \in L^2(\mathbb{R}_+)$. Noting that $[1 + \tilde k(\cdot)]\tilde y_1(\cdot) + [\tilde z(\cdot) - 2r] \in L^2(\mathbb{R}_+)$, it follows from the first equation in (4.12) that $\tilde y_1(t) \to r \neq 0$ as $t \to \infty$, which contradicts the fact that $\tilde y_1 \in L^2(\mathbb{R}_+)$. Therefore, the solution component $\tilde k$ is unbounded. Unboundedness of $\tilde k$, together with the third differential equation in (4.12), implies the existence of $T \in [0, \tilde\omega)$ such that $k(T) > 4M$ and $\tilde y_1(T) \neq 0$. Let $u_0 \in L^2(\mathbb{R}_+, \mathbb{R})$ be the piecewise constant function $u_0 := T_T r$ (viz. $u_0(t) = r$ on $[0,T)$ and $u_o(t) = 0$ on $[T, \infty)$), and denote by $(y_1, z, k) : [0,\omega) \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$ the unique maximal solution of

$$\begin{array}{ll}\dot y_1(t) = k(t)\, y_1(t) + z(t) - u_0(t), & y_1(0) = 0,\\ \dot z(t) = -2Mk(t)\, y_1(t) - Mz(t) + 2Mu_0(t), & z(0) = 0,\\ \dot k(t) = y_1(t)^2, & k(0) = 0.\end{array}$$

Clearly, $\tilde{\omega} \leq \omega$ and $(\tilde{y}_1(t), \tilde{z}(t), \tilde{k}(t)) = (y_1(t), z(t), k(t))$ for all $t \in [0, \tilde{\omega})$. Therefore, $y_1(T) = \tilde{y}_1(T) \neq 0$ and $k(T) = \tilde{K}(T) > 4M$. An application of Lemma 4.1 completes the proof. ∎

4.4. Nonrobustness with respect to L^∞ disturbances. The initial calculation in [7, Example 9] (see also [9]) shows that arbitrarily small L^∞ disturbances u_0, $y_0 \in L^\infty(\mathbb{R}_+, \mathbb{R})$ can cause $\|u_1\|_{L^\infty(\mathbb{R}_+,\mathbb{R})}$ to be arbitrarily large, and so $H_{P,C}$ is not gf-stable in an L^∞ sense, whence we get the claim that the (L^∞) "robustness margin... should be assigned the value zero." Note that the gf-stability of $H_{P,C}$ is only a sufficient condition for robust stability. In this context, we next show that any constant nonzero input disturbance (and zero output disturbance) leads to unbounded signals in the perturbed closed loop $[\widetilde{P}_1(a, M, y_1^0, z^0), \widetilde{C}(k^0)]$. This is not surprising in view of [7, Example 9], where it is shown that the unperturbed closed loop $[\widetilde{P}(a, y_1^0), \widetilde{C}(k^0)]$ is nonrobust with respect to L^∞ disturbances.

PROPOSITION 4.4. *For $M, a, y_1^0, z^0, k^0 \in \mathbb{R}$ and $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R})$, consider the closed-loop system $[\widetilde{P}_1\left(M, a, y_1^0, z^0\right), \widetilde{C}(k^0)]$ defined by (2.1), (4.4), (4.2), (4.3) in the specific case wherein $M > 0$, $a = y_1^0 = z^0 = k^0 = 0$, and $y_0 = 0$.*

For any $r > 0$, there exists $u_0 \in L^\infty(\mathbb{R}_+, \mathbb{R})$ with $\|u_0\|_{L^\infty(\mathbb{R}_+,\mathbb{R})} \leq r$, such that the unique maximal solution $(y_1, z, k) : [0, \omega) \to \mathbb{R}^2 \times \mathbb{R}_+$ of the closed-loop system $[\widetilde{P}_1\left(M, 0, 0, 0\right), \widetilde{C}(0)]$ has the following properties:

(i) *$u_1, y_1 \notin L^\infty([0, \omega), \mathbb{R})$;*

(ii) *$k \notin L^\infty([0, \omega), \mathbb{R}_+)$.*

Proof. This follows directly from the proof of Proposition 4.3, wherein it was shown that for each, $r > 0$, there exists $T \in (0, \omega)$ such that the disturbance $u_0 \in L^\infty(\mathbb{R}_+, \mathbb{R})$, given by $u_0 := T_T r$ and with norm $\|u_0\|_{L^\infty(\mathbb{R}_+,\mathbb{R})} = r$, is such that properties (i) and (ii) hold. ∎

Interestingly, our analysis in subsections 4.2, 4.3, and 4.4 has not established whether a finite escape time can occur within these closed loops. Simulation evidence presented in [7] makes this plausible, but the question remains open.

4.5. Robustness with respect to small L^2 disturbances and small initial conditions. Having noted in subsection 4.4 that the L^∞ robustness margin should be assigned the value zero, our next task is to show that, in the L^2 setting of the current paper, the situation is less pessimistic. In the L^∞ framework, one last remark is warranted. In [7, Example 9], and based on informal numerical evidence, there is a suggestion that—even with zero disturbances—the closed-loop system fails to be robustly stable if the initial conditions are nonzero. Proposition 4.2 confirms this in the case of *large* initial conditions. However, Proposition 4.6 below subsumes the following observation: With zero disturbances, the closed-loop system is robustly stable for *sufficiently small* initial conditions. As noted in [7, Example 8], in the L^∞ framework, $\vec{\delta}\big(\widetilde{P}(a, 0), \widetilde{P}_1(a, M, 0, 0)\big) \to 0$ as $M \to \infty$. We now show that this result also holds true in the L^2 framework.

LEMMA 4.5. *For $M, a, y_1^0, z^0 \in \mathbb{R}$ and $u_0, y_0 \in L^2(\mathbb{R}_+, \mathbb{R})$, consider $\widetilde{P}(a, y_1^0)$ and $\widetilde{P}_1\left(M, a, y_1^0, z^0\right)$ given by (4.1) and (4.3), respectively, in the specific case wherein $M > 0$ and $y_1^0 = z^0 = 0$. Then $\vec{\delta}\big(\widetilde{P}(a, 0), \widetilde{P}_1(a, M, 0, 0)\big) \to 0$ as $M \to \infty$.*

Proof. It will be convenient to utilize a frequency domain representation of linear operators. Firstly, let $c > 0$, and define the rational functions A, A', B, B' by $A(s) = A'(s) := (s-a)/(s+c)$, $B(s) := 1/(s+c)$, and $B'(s) := (M-s)/(s+c)(s+M)$. For $n, m \in \mathbb{N}$, let H^2 denote the set of all analytic functions $f : \mathbb{C}_+ \to \mathbb{C}^{n\times m}$ so that $\int_{\mathbb{R}} \|f(\alpha + i\beta\|^2 \mathrm{d}\beta$ is finite for all $\alpha > 0$. Since, by Paley and Wiener (see, for example,

[3, Thm. 11.9]), the extended Laplace transform $\mathbf{L}$ on $L^2(\mathbb{R}_+,\mathbb{R})$ yields an isometric isomorphism between $L^2(\mathbb{R}_+,\mathbb{R})$ and H^2, we may observe (see [17, p. 234]) that the graphs corresponding to $\widetilde{P}(a,0)$, $\widetilde{P}_1(M,a,0,0)$ can be written as:

$$\mathbf{L}\left(\mathcal{G}_{\widetilde{P}}\right) \;=\; \begin{pmatrix} A \\ B \end{pmatrix} H^2 \quad \text{and} \quad \mathbf{L}\left(\mathcal{G}_{\widetilde{P}_1}\right) \;=\; \begin{pmatrix} A' \\ B' \end{pmatrix} H^2.$$

Let $(V,U) := (1, c+a)$, and define the mappings $\Phi, \tilde{\Phi}$:

$$\tilde{\Phi}\colon \mathbf{L}\left(\mathcal{G}_{\widetilde{P}(a,0)}\right) \to \mathbf{L}\left(\mathcal{G}_{\widetilde{P}_1(M,a,0,0)}\right), \quad \begin{pmatrix} u \\ y \end{pmatrix} \mapsto \tilde{\Phi}\begin{pmatrix} u \\ y \end{pmatrix} := \begin{pmatrix} A' \\ B' \end{pmatrix}(V,U)\begin{pmatrix} u \\ y \end{pmatrix},$$
$$\Phi\colon \mathcal{G}_{\widetilde{P}(a,0)} \to \mathcal{G}_{\widetilde{P}_1(M,a,0,0)}, \qquad \Phi := \mathbf{L}^{-1}\tilde{\Phi}\mathbf{L}.$$

Since $(V,U)\binom{A}{B} = 1$, A', B' are coprime, and $\mathbf{L}$ is an isometric isomorphism, it follows that $\Phi \in \mathcal{O}_{\widetilde{P},\widetilde{P}_1}$. Additionally, since every element of $\mathbf{L}\left(\mathcal{G}_{\widetilde{P}}\right)$ is of the form $y = \binom{A}{B}x$, where $x \in H^2$, it follows that

$$(\tilde{\Phi} - I)|_{\mathbf{L}(\mathcal{G}_{\widetilde{P}})} y = \left(I - \begin{pmatrix} A' \\ B' \end{pmatrix}(V,U)\right)\begin{pmatrix} A \\ B \end{pmatrix} x = \begin{pmatrix} A - A' \\ B - B' \end{pmatrix} x = \begin{pmatrix} A - A' \\ B - B' \end{pmatrix}(V,U)y,$$

and since $(\Phi - I)|_{\mathcal{G}_{\widetilde{P}}} = \mathbf{L}^{-1}(\tilde{\Phi} - I)|_{\mathbf{L}(\mathcal{G}_{\widetilde{P}})}\mathbf{L}$, we have

$$\vec{\delta}(\widetilde{P},\widetilde{P}_1) \le \left\| \mathbf{L}^{-1}\left(\begin{pmatrix} A - A' \\ B - B' \end{pmatrix}(V,U)\right)\mathbf{L} \right\|_{L^2} = \left\| \begin{pmatrix} A - A' \\ B - B' \end{pmatrix}(V,U) \right\|_{H^\infty}.$$

Hence,

$$\vec{\delta}(\widetilde{P},\widetilde{P}_1) \le \left\| \begin{pmatrix} 0 \\ \frac{2s}{(s+c)(s+M)} \end{pmatrix}(1, c+a) \right\|_{H^\infty} \le 2\sqrt{1+(c+a)^2}\left\| \frac{s}{(s+c)(s+M)} \right\|_{H^\infty}.$$

A straightforward computation confirms that the right-hand side in the above goes to 0 as M tends to ∞. This completes the proof. ∎

The final result states that for any disturbance level, M can be chosen to ensure stability of the perturbed closed-loop system; furthermore, this stability is local with respect to initial conditions.

PROPOSITION 4.6. *Let $a \in \mathbb{R}$. For any $M > 0$, $y_1^0, z^0 \in \mathbb{R}$, $k^0 > 0$, consider the closed-loop system $[\widetilde{P}_1\left(M, a, y_1^0, z^0\right), \widetilde{C}(k^0)]$ as defined by* (2.1), (4.4), (4.3), *and* (4.2). *For any $r > 0$ there exists $\hat{M} > 0$, and for any $M \ge \hat{M}$ there exists $\varepsilon > 0$ such that, if $\|(y_1^0, z^0)\| \le \varepsilon$ and $\|(u_0, y_0)\|_{L^2(\mathbb{R}_+,\mathbb{R}^2)} \le r$, then the closed-loop system $[\widetilde{P}_1\left(M, a, y_1^0, z^0\right), \widetilde{C}(k^0)]$ has the following properties:*

(i) *There exists a unique solution $(y_1, z, k) : \mathbb{R}_+ \to \mathbb{R}^2 \times \mathbb{R}_+$;*
(ii) *$(u_1, y_1) \in L^2(\mathbb{R}_+, \mathbb{R}^2)$;*
(iii) *$\lim_{t\to\infty}(y_1(t), z(t)) = 0$;*
(iv) *$k \in L^\infty(\mathbb{R}_+, \mathbb{R}_+)$.*

Proof. Properties (i), (ii), and (iv) follow from Lemma 4.5 and Theorem 3.8. Therefore, $y_1, y_0, u_0, ky_0, ky_1 \in L^2(\mathbb{R}_+)$. Invoking the second differential equation of (4.4), we have $z \in L^2(\mathbb{R}_+)$ and $\lim_{t\to\infty} z(t) = 0$. By the first equation in (4.4) we have $\dot{y}_1 \in L^2(\mathbb{R}_+)$. It now follows that $\lim_{t\to\infty} y_1(t) = 0$, whence we get property (iii). ∎

5. Summary. In subsection 3.4, we developed a general result establishing a robust stability margin (whose size is dependent on the L^2 disturbance level and size of the initial condition) for the class of m-input, m-output, relative degree one, nonminimum phase plants, whose first Markov parameter lies in the open right half plane, when controlled by the "standard" adaptive output feedback controller (1.1). In section 4 we have given a qualitative analysis of a first-order system $\widetilde{P}(0, y_1^0)$ perturbed by an all-pass factor $\frac{M-s}{M+s}$ and controlled by a standard adaptive controller as considered by Georgiou and Smith [7, Ex. 9]. The results of section 4 are summarized in the following table:

Disturbances and initial data	Stability	Internal signals	Controller gain
For any small L^2 disturbance (u_0, y_0) and any small initial condition (y_1^0, z^0)	stable	$(u_1, y_1) \in L^2$	$k \in L^\infty$
There exists large L^2 disturbance $(u_0, 0)$	unstable	$(u_1, y_1) \notin L^\infty$	$k \notin L^\infty$
For any large initial condition $(y_1^0, 0)$	unstable	$(u_1, y_1) \notin L^\infty$	$k \notin L^\infty$
There exists an L^∞ disturbance $(u_0, 0)$ (of any nonzero size)	unstable	$(u_1, y_1) \notin L^\infty$	$k \notin L^\infty$

It is worth noting that the L^2 analysis in this paper provides a mechanism to prove the stability of the disturbance-free system in the presence of small initial conditions. The informal plausibility arguments presented in [7] for the lack of robustness of a closed-loop system in the presence of nonzero initial conditions do not predict the stable behavior of the closed-loop system when the initial conditions are small. This case study highlights the critical role played by the choice of signal space—alternative signal spaces may give different robust stability guarantees (in particular, in the disturbance-free case, we have seen that the L^∞ analysis does not give any indication of the robustness of the closed-loop system under gap perturbations with nonzero initial conditions; however, the L^2 analysis does establish robustness). This highlights the importance of an L^2 analysis for considering response to initial conditions. The second and third entries in the above table illustrate that the sufficient conditions for robust stability given by the gap analysis in section 3.4 cannot be improved qualitatively and emphasize the complementary role of initial conditions and disturbances. It is also important to note that zero L^∞ robustness margins are not inevitable. An example of an adaptive controller exhibiting a nonzero L^∞ margin is given in [4].

REFERENCES

[1] W. Bian and M. French, *Graph topologies, gap metrics, and robust stability for nonlinear systems*, SIAM J. Control Optim., 44 (2005), pp. 418–443.

[2] M. Cantoni and G. Vinnicombe, *Linear feedback systems and the graph topology*, IEEE Trans. Automat. Control, 47 (2002), pp. 710–719.

[3] P. L. Duren, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[4] M. French, *Adaptive control and robustness in the gap metric*, IEEE Trans. Automat. Control, 2006, to appear.

[5] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, Trans. Automat. Control, 35 (1990), pp. 673–686.

[6] T. T. Georgiou, *On the computation of the gap metric*, Systems Control Lett., 11 (1988), pp. 253–257.

[7] T. T. Georgiou and M. C. Smith, *Robustness analysis of nonlinear feedback systems: An input-output approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1200–1221.

[8] T. T. Georgiou and M. C. Smith, *Robustness of a relaxation oscillator*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 1005–1024.

[9] T. T. Georgiou and M. C. Smith, *Remarks on "Robustness Analysis of Nonlinear Feedback Systems"*, IEEE Trans. Automat. Control, 46 (2001), pp. 171–172.

[10] F. Ikhouane and M. Krstić, *Adaptive backstepping with parameter projection: Robustness and asymptotic performance*, Automatica, 34 (1998), pp. 429–435.

[11] A. Ilchmann, *Non-Identifier-Based High-Gain Adaptive Control*, Lecture Notes in Control Inform. Sci. 189, Springer-Verlag, London, 1993.

[12] A. Ilchmann, M. Thuto, and S. Townley, *Input constrained adaptive tracking with applications to exothermic chemical reaction models*, SIAM J. Control Optim., 43 (2004), pp. 154–173.

[13] P. A. Ioannou and J. Sun, *Robust Adaptive Control*, Prentice–Hall, New York, 1996.

[14] M. Krstić, I. Kanellakopoulos, and P. V. Kokotović, *Nonlinear and Adaptive Control Design*, Wiley, New York, 1995.

[15] C. Rohrs, L. S. Valavani, M. Athans, and G. Stein, *Robustness of continuous-time adaptive control algorithms in the presence of unmodeled dynamics*, IEEE Trans. Automat. Control, 30 (1985), pp. 881–889.

[16] M. R. James, M. C. Smith, and G. Vinnicombe, *Gap metrics, representations, and nonlinear robust stability*, SIAM J. Control Optim., 43 (2005), pp. 1535–1582.

[17] M. Vidyasagar, *Control System Synthesis*, MIT Press, Cambridge, MA, 1985.

[18] G. Vinnicombe, *Uncertainty and Feedback: $\mathcal{H}_\infty$ Loop-Shaping and the ν-gap Metric*, Imperial College Press, London, 2001.

[19] G. Zames and A. K. El-Sakkary, *Unstable systems and feedback: The gap metric*, in Proceedings of the 18th Annual Allerton Conference on Communication, Control, and Computing, University of Illinois, Urbana-Champaign, IL, 1980, pp. 380–385.

[20] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* I: *Fixed Point Theorems*, Springer, Berlin, 1986.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1758–1768

INSENSITIZING CONTROLS FOR THE 1-D WAVE EQUATION*

RENÉ DÁGER†

Abstract. We study the insensitizing controllability property of the one-dimensional wave equation observed in some open set in two cases: when the control acts in an interior region and when it acts on the boundary. In both cases, when the control time is sufficiently large the ϵ-insensitizing controllability holds. Moreover, for the boundary controlled equation the (exact) insensitizing controllability also holds.

Key words. wave equation, insensitizing control, controllability, unique continuation

AMS subject classifications. 93B05, 93B07, 35B37, 35B60, 35L05

DOI. 10.1137/060654372

1. Introduction.

Consider the controlled wave equations with partially known initial data,

$$\begin{cases} y_{tt} - \Delta y = \xi + f\mathbf{1}_\omega & \text{in } \mathbb{R}\times(0,1),\\ y = 0 & \text{in } \mathbb{R}\times\{0,1\},\\ y(0,x) = y^0 + \tau_0 z^0 & \text{in } (0,1),\\ y_t(0,x) = y^1 + \tau_1 z^1 & \text{in } (0,1), \end{cases} \tag{1}$$

where ω is an open subset of $[0,1]$, and

$$\begin{cases} y_{tt} - \Delta y = \xi & \text{in } \mathbb{R}\times(0,1),\\ y(t,0) = 0, \quad y(t,1) = g(t) & \text{in } \mathbb{R},\\ y(0,x) = y^0 + \tau_0 z^0 & \text{in } (0,1),\\ y_t(0,x) = y^1 + \tau_1 z^1 & \text{in } (0,1). \end{cases} \tag{2}$$

In both (1) and (2) y^0, y^1, and $\xi \in L^2(\mathbb{R}\times(0,1))$ are fixed functions. The last two equations in (1) or (2) indicate that the initial data are not exactly known. The terms $\tau_0 z^0, \tau_1 z^1$ with $\tau_0, \tau_1 \in \mathbb{R}$ and z^0, z^1 satisfying

$$\left\|z^0\right\|_a = \left\|z^1\right\|_b = 1 \tag{3}$$

(in appropriate norms $\|.\|_a$ and $\|.\|_b$ guaranteeing the well-posedness of the corresponding equations) are small unknown perturbations of the original initial data y^0, y^1. Finally, $f \in L^2(\mathbb{R}\times(0,1))$ and $g \in L^2(\mathbb{R})$ are control functions. The only difference between (1) and (2) is the location of the control: in (1) the control f acts in an interior region ω, while in (2) the control g acts through a part of the boundary.

Let Φ be the functional defined on the solutions of the controlled equations (1) or (2) by

$$\Phi(y) = \frac{1}{2}\int_0^T \int_{\mathcal{O}} y^2 dxdt,$$

*Received by the editors March 15, 2006; accepted for publication (in revised form) September 9, 2006; published electronically December 1, 2006.
http://www.siam.org/journals/sicon/45-5/65437.html

†Departamento de Fundamentos Matemáticos de la Tecnología Aeronáutica, Universidad Politécnica de Madrid, Pl. Cardenal Cisneros 3, 28040 Madrid, Spain (rene.dager@upm.es).

where $T > 0$ and $\mathcal{O}$ is an open subset of $[0,1]$. Note that Φ corresponds to observations of the solutions made in the interval $\mathcal{O}$ during a time interval of length T.

Given a control f (resp., g) we say that f (resp., g) insensitizes Φ if for every pair z^0, z^1 satisfying (3) the corresponding solution of (1) (resp., (2)) satisfies

$$\frac{d}{d\tau_0}\Phi(y)\bigg|_{\tau_0=\tau_1=0} = \frac{d}{d\tau_1}\Phi(y)\bigg|_{\tau_0=\tau_1=0} = 0. \tag{4}$$

The concept of insensitizing control was introduced by Lions in [9]. Later on, Bodart and Fabre in [3] proposed the weakened notion of ε-insensitizing control. Given $\varepsilon > 0$, the control f (resp., g) is said to ε-insensitize Φ if for every pair z^0, z^1 satisfying (3) the corresponding solution of (1) (resp., (2)) satisfies

$$\left|\frac{d}{d\tau_0}\Phi(y)\bigg|_{\tau_0=\tau_1=0}\right| \le \varepsilon, \quad \left|\frac{d}{d\tau_1}\Phi(y)\bigg|_{\tau_0=\tau_1=0}\right| \le \varepsilon. \tag{5}$$

In this paper we study the existence of insensitizing or ε-insensitizing controls for systems (1) and (2). Our main results are the following (see Theorems 7 and 13):

Let $\omega, \mathcal{O}$ be nonempty open subsets of $[0,1]$ and $T \ge 4$. Then

- *for any $\varepsilon > 0$ there exist controls that ε-insensitize Φ along the solutions of (1);*
- *there exist controls that insensitize Φ along the solutions of (2).*

To prove these results we use the technique proposed by Lions in [9]. For this concrete choice of Φ, finding insensitizing controls is equivalent to finding controls driving the solutions of an auxiliary cascade system to rest, i.e., to a usual controllability problem, which is reduced to proving observability properties for the adjoint control-free cascade system. In particular, the existence of ε-insensitizing controls follows from a unique continuation property for the adjoint system. The main difficulty in this latter problem arises due to the nonstandard coupling in the cascade system.

We have to note that insensitizing or ε-insensitizing controllability properties have been generally studied in considerably more complex situations, namely, nonlinear parabolic equations in $\mathbb{R}^n$ (see [1, 2, 3, 4, 6, 7, 8, 9, 10, 11]). In this context our problem looks fairly simple. However, all those papers assume the observation and control regions to have nonempty intersection ($\omega \cap \mathcal{O} \neq \varnothing$). In this sense, the wave equation provides the first example when ε-insensitizing controls exist for arbitrarily chosen observation and control regions. Up to now, the only known example of such situation was given in [10] for the one-dimensional (1-D) linear heat equation with interior control and a particular choice of ω and $\mathcal{O}$. Besides, as far as we know, this is the first attempt to consider insensitizing controllability problems for the wave equation.

On the other hand, the existence of insensitizing controls for all the initial data for (2) is an interesting fact, since up to now the insensitizing controls have been found as the limit of a sequence of ε-insensitizing controls for some particular choices of the initial data (see, e.g., [1, 2, 7, 8]).

2. Interior control. In this section we concentrate ourselves in system (1). This is precisely the choice of control made in all the preceding works related to the insensitizing controllability of parabolic equations.

We assume $y^0 \in H_0^1(0,1), y^1 \in L^2(0,1)$. Then system (1) has a unique solution of finite energy that satisfies, for every $T > 0$, $y \in C([0,T], H_0^1(0,1)) \cap$

$C^1([0,T],L^2(0,1))$. In this case, the norms appearing in (3) are $\|.\|_a = \|.\|_{H_0^1(0,1)}$ and $\|.\|_b = \|.\|_{L^2(0,1)}$. In what follows, to simplify notation we denote

$$\|.\|_1 := \|.\|_{H_0^1(0,1)}, \quad \|.\|_0 := \|.\|_{L^2(0,1)}, \quad \|.\|_{-1} := \|.\|_{H^{-1}(0,1)},$$
$$\|.\|_{1,0} := \left(\|.\|_1^2 + \|.\|_0^2\right)^{\frac{1}{2}}, \quad \|.\|_{0,-1} := \left(\|.\|_0^2 + \|.\|_{-1}^2\right)^{\frac{1}{2}}.$$

2.1. Reduction to a unique continuation problem. As stated above, the main technique for proving the existence of ε-insensitizing controls is to reduce the problem to proving a uniqueness result for the solutions of a cascade system. In this subsection we describe such a reduction. These facts are well known in the context of parabolic equations, and we use the same ideas here. Although this scheme is rather standard, we include the details here since, as far as we know, they have never been written for the wave equation.

The first step is to reduce the problem to an approximate controllability problem. Consider the following controlled cascade system of wave equations:

$$\begin{cases} \bar{y}_{tt} - \Delta \bar{y} = \xi + f\mathbf{1}_\omega & \text{in } \mathbb{R}\times(0,1), \\ \bar{y} = 0 & \text{in } \mathbb{R}\times\{0,1\}, \\ \bar{y}(0,x) = y^0, \quad \bar{y}_t(0,x) = y^1 & \text{in } (0,1), \end{cases} \tag{6}$$

$$\begin{cases} q_{tt} - \Delta q = \bar{y}\mathbf{1}_{\mathcal{O}} & \text{in } \mathbb{R}\times(0,1), \\ q = 0 & \text{in } \mathbb{R}\times\{0,1\}, \\ q(T,x) = 0, \quad q_t(T,x) = 0 & \text{in } (0,1). \end{cases} \tag{7}$$

Note that, since $\bar{y}\mathbf{1}_{\mathcal{O}} \in L^2(\mathbb{R}\times(0,1))$, system (7) has a unique solution $q \in C([0,T], H_0^1(0,1)) \cap C^1([0,T],L^2(0,1))$. In particular, $(q(0,.), q_t(0,.)) \in H_0^1(0,1)\times L^2(0,1)$.

The following proposition holds.

PROPOSITION 1. *Given $y^0 \in H_0^1(0,1), y^1 \in L^2(0,1)$, and $\xi \in L^2(\mathbb{R}\times(0,1))$, the control f ε-insensitizes Φ along the solutions of* (1) *if and only if the corresponding solution $(\bar{y},q)$ of the cascade system* (6)–(7) *satisfies*

$$\|q(0,x)\|_0 < \varepsilon, \|q_t(0,x)\|_{-1} < \varepsilon. \tag{8}$$

Proof. A simple calculation yields

$$\frac{d}{d\tau_0}\Phi(y)\Big|_{\tau_0=\tau_1=0} = \int_0^T \int_{\mathcal{O}} \bar{y}\theta dxdt, \tag{9}$$

where $\theta = \frac{dy}{d\tau_0}$ is the solution of the equation

$$\begin{cases} \theta_{tt} - \Delta\theta = 0 & \text{in } \mathbb{R}\times(0,1), \\ \theta = 0 & \text{in } \mathbb{R}\times\{0,1\}, \\ \theta(0,x) = z^0, \quad \theta_t(0,x) = 0 & \text{in } (0,1). \end{cases} \tag{10}$$

On the other hand, integrating by parts in account of the boundary and initial (final) conditions in (7) and (10) we get

$$\int_0^T \int_{\mathcal{O}} \bar{y}\theta dxdt = \int_0^T \int_0^1 \bar{y}\mathbf{1}_{\mathcal{O}}\theta dxdt = \int_0^T \int_0^1 (q_{tt} - q_{xx})\theta dxdt = -\int_0^1 q_t(0,x)z^0 dx,$$

and thus (9) becomes

$$\frac{d}{d\tau_0}\Phi(y)\Big|_{\tau_0=\tau_1=0} = -\int_0^1 q_t(0,x)z^0dx. \tag{11}$$

In the same way we can prove that

$$\frac{d}{d\tau_1}\Phi(y)\Big|_{\tau_0=\tau_1=0} = \int_0^1 q(0,x)z^1dx. \tag{12}$$

Finally, since z^0, z^1 are arbitrary functions satisfying

$$\left\|z^0\right\|_1 = \left\|z^1\right\|_0 = 1,$$

from (11) and (12) the assertion of the proposition holds. □

The next step is to characterize the approximate controllability property of Proposition 1 in terms of the adjoint system to (6)–(7):

$$\begin{cases} p_{tt} - \Delta p = 0 & \text{in } \mathbb{R}\times(0,1),\\ p = 0 & \text{in } \mathbb{R}\times\{0,1\},\\ p(0,x) = p^0, \quad p_t(0,x) = p^1 & \text{in } (0,1), \end{cases} \tag{13}$$

$$\begin{cases} z_{tt} - \Delta z = p\mathbf{1}_{\mathcal{O}} & \text{in } \mathbb{R}\times(0,1),\\ z = 0 & \text{in } \mathbb{R}\times\{0,1\},\\ z(T,x) = 0, \quad z_t(T,x) = 0 & \text{in } (0,1). \end{cases} \tag{14}$$

We assume $\left(p^0,p^1\right) \in H_0^1(0,1)\times L^2(0,1)$. Thus system (13)–(14) has a unique solution (p,z), which satisfies

$$p, z \in C([0,T], H_0^1(0,1)) \cap C^1([0,T], L^2(0,1)).$$

In particular, $z \in L^2([0,T]\times(0,1))$.

The following result is easily proved by multiplying (6) by z and integrating by parts.

PROPOSITION 2. *The following duality identity for the solutions of* (6)–(7), (13)–(14) *holds for every* $f \in L^2(\mathbb{R}\times(0,1))$, $\left(y^0,y^1\right), \left(p^0,p^1\right) \in H_0^1(0,1)\times L^2(0,1)$*:*

$$\begin{aligned}\int_{[0,T]\times\omega} fz dxdt + \int_{[0,T]\times(0,1)} \xi z dxdt = &\int_0^1 \left(y^0 z_t(0,x) - y^1 z(0,x)\right) dx\\ &+ \int_0^1 \left(p^0 q_t(0,x) - p^1 q(0,x)\right) dx.\end{aligned} \tag{15}$$

PROPOSITION 3. *The set* $\mathcal{B} = \left\{(q(0,.), q_t(0,.)) : f \in L^2(\mathbb{R}\times(0,1))\right\}$ *is dense in* $L^2(0,1)\times H^{-1}(0,1)$*, and then the approximate controllability property* (8) *holds if and only if the following unique continuation property is true for every* $\left(p^0,p^1\right) \in H_0^1(0,1)\times L^2(0,1)$*:*

$$z \equiv 0 \text{ in } [0,T]\times\omega \text{ implies } z \equiv p \equiv 0 \text{ in } [0,T]\times(0,1). \tag{16}$$

Proof. Assume that $\left(p^0,p^1\right) \in H_0^1(0,1)\times L^2(0,1)$ satisfies

$$\int_0^1 \left(p^0 q_t(0,x) - p^1 q(0,x)\right) dx = 0$$

for every $f \in L^2(\mathbb{R}\times(0,1))$ and $\left(y^0,y^1\right) \in H_0^1(0,1)\times L^2(0,1)$. Then, in view of (15),

$$\int_{[0,T]\times\omega} fz dxdt = -\int_{[0,T]\times(0,1)} \xi z dxdt + \int_0^1 \left(y^0 z_t(0,x) - y^1 z(0,x)\right) dx.$$

But the right-hand term of this equality does not depend on f, so choosing $f=0$ it follows that

$$\int_{[0,T]\times\omega} fz dxdt = 0 \quad \text{for every } f \in L^2(\mathbb{R}\times(0,1)),$$

and then, since $z \in L^2([0,T]\times\omega)$, it follows that $z=0$ in $[0,T]\times\omega$. Next, the unique continuation property implies $p\equiv 0$ in $[0,T]\times(0,1)$, and hence $p^0=p^1=0$. Therefore, from the Hahn–Banach theorem it follows that $\mathcal{B}$ is dense in $L^2(0,1)\times H^{-1}(0,1)$.

For the converse assertion, assume that there exists $\left(p^0,p^1\right) \neq (0,0)$ such that $z\equiv 0$ in $[0,T]\times\omega$, and pick $\left(y^0,y^1\right)$ such that

$$\int_{[0,T]\times(0,1)} \xi z dxdt = \int_0^1 \left(y^0 z_t(0,x) - y^1 z(0,x)\right) dx.$$

Then from (15) we obtain

$$\int_0^1 \left(p^0 q_t(0,x) - p^1 q(0,x)\right) dx = 0,$$

and this means that $\mathcal{B}$ is not dense in $L^2(0,1)\times H^{-1}(0,1)$. □

Gathering the results of Propositions 2 and 3 we obtain the following.

PROPOSITION 4. *If the solutions of* (13)–(14) *satisfy the unique continuation property* (16)*, then for every* $\varepsilon>0$, $\xi \in L^2(\mathbb{R}\times(0,1))$*, and* $\left(y^0,y^1\right) \in H_0^1(0,1)\times L^2(0,1)$ *there exists a control* $f^\varepsilon \in L^2(\mathbb{R}\times(0,1))$ *such that along the corresponding solution* y^ε *of system* (1) *the functional* Φ *verifies* (5)*.*

Remark 5. When $\omega\cap\mathcal{O}\neq\varnothing$ the unique continuation property (16) is reduced to a unique continuation property for the usual wave equation. Indeed, if $z\equiv 0$ in $[0,T]\times(\omega\cap\mathcal{O})$, so is p in view of (14). Therefore, if $[0,T]\times(\omega\cap\mathcal{O})$ is an appropriate unique continuation region for the wave equation (that is, if the solution vanishes in $[0,T]\times(\omega\cap\mathcal{O})$, then it vanishes in $[0,T]\times(0,1)$), then $p\equiv 0$ in $[0,T]\times(0,1)$ and thus $z\equiv 0$ in $[0,T]\times(0,1)$. The same occurs in the case of heat equations considered in the literature.

Note that for the wave equation in $\mathbb{R}^n$ the unique continuation property from a region $[0,T]\times U$ depends on both T and the geometry of U. For the 1-D wave equation considered here, unique continuation from a region $[t_1,t_2]\times U$ holds whenever $t_2-t_1\geq 2$ (twice the length of the interval $(0,1)$).

When $\omega\cap\mathcal{O}=\varnothing$ such arguments cannot be applied, and the problem becomes more difficult.

2.2. Unique continuation results. In this subsection we show that the unique continuation property (16) is true for sufficiently large T for arbitrarily chosen ω and $\mathcal{O}$, independently of whether their intersection is void or not. Indeed, the following holds.

PROPOSITION 6. *Let* $\omega,\mathcal{O}$ *be nonempty open subsets of* $[0,1]$ *and* $T\geq 4$*. Then the unique continuation property* (16) *holds for the solutions of system* (13)–(14)*.*

As a consequence of Propositions 4 and 6 we obtain the existence of ε-insensitizing controls.

THEOREM 7. *Let $\omega, \mathcal{O}$ be nonempty open subsets of $[0,1]$ and $T \geq 4$. Then, for every $\varepsilon > 0$, $\xi \in L^2(\mathbb{R} \times (0,1))$, and $\left(y^0, y^1\right) \in H_0^1(0,1) \times L^2(0,1)$, there exists a control $f^\varepsilon \in L^2(\mathbb{R} \times (0,1))$ that ε-insensitizes functional Φ along the solutions of* (1).

Proposition 6 is an immediate consequence of the following observability inequality.

PROPOSITION 8. *Let $\omega, \mathcal{O}$ be nonempty open subsets of $[0,1]$ and $T \geq 4$. Then there exists a positive constant C, such that for every solution of* (13)–(14),

$$C \left\|(p^0, p^1)\right\|_{1,0}^2 \int_0^T \int_\omega z^2 dt dx \geq \left\|(p^0, p^1)\right\|_{0,-1}^4 .$$

Proof. We use in an essential way the fact that the first equation in (13) is the 1-D linear homogeneous wave equation

$$\text{(17)} \qquad \begin{cases} p_{tt} - \Delta p = 0 & \text{in } \mathbb{R} \times (0,1), \\ p = 0 & \text{in } \mathbb{R} \times \{0,1\}, \\ p(0,x) = p^0, \quad p_t(0,x) = p^1 & \text{in } (0,1). \end{cases}$$

Let us recall some known facts that are used in the proof. The solutions of (17) are periodic in time with period equal to 2, which may be easily obtained, e.g., from the fact that the solutions are of the form

$$\text{(18)} \qquad p(t,x) = g(t+x) - g(t-x),$$

with g being a 2-periodic function.

Note that the derivatives p_x and p_t have the same periodicity property.

Moreover, from representation formula (18) one can prove the following well-known unique continuation property for the solutions of (17): *If $t^* - t_* \geq 2$, then any of the equalities $p = 0$ in $[t_*, t^*] \times \omega$ or $p_x(t,1) = 0$ in $[t_*, t^*]$ implies $p = 0$ in $[t_*, t^*] \times (0,1)$.* This fact may be quantified by means of observability inequalities: *There exist positive constants C_1, C_2 such that for every solution p of* (17)

$$\text{(19)} \qquad C_1 \int_{t_*}^{t^*} \int_\omega p^2 dx dt \geq \left\|(p^0, p^1)\right\|_{0,-1}^2 ,$$

$$\text{(20)} \qquad C_2 \int_{t_*}^{t^*} p_x^2(t,1) dt \geq \left\|(p^0, p^1)\right\|_{1,0}^2 .$$

In addition, the following conservation laws are true for the solutions of (17):

$$\text{(21)} \qquad \left\|(p(\mu,.), p_t(\mu,.))\right\|_{1,0} = \left\|(p^0, p^1)\right\|_{1,0} ,$$

$$\text{(22)} \qquad \left\|(p(\mu,.), p_t(\mu,.))\right\|_{0,-1} = \left\|(p^0, p^1)\right\|_{0,-1}$$

for every $\mu \in \mathbb{R}$

Now we proceed to the proof of the proposition. Let T_0 be the smallest positive number such that $T - T_0 \in 2\mathbb{Z}$. The inequality stated in the proposition is obtained from the following two inequalities:

$$\text{(23)} \qquad C_1 \int_0^T \int_\omega z^2 dt dx \geq \left\|(z(T_0), z_t(T_0))\right\|_{0,-1}^2 ,$$

$$\text{(24)} \qquad C_2 \left\|(z(T_0), z_t(T_0))\right\|_{0,-1} \left\|(p^0, p^1)\right\|_{1,0} \geq \left\|(p^0, p^1)\right\|_{0,-1}^2 ,$$

which hold for every solution (p,z) of (13)-(14) with positive constants C_1, C_2 independent of (p,z).

In order to prove inequality (23), we define the linear operator ℓ acting on a function $u(t)$ by

$$(\ell u)(t) = u(t-1) - u(t+1).$$

Note that ℓ is continuous from $L^2(\alpha,\beta)$ to $L^2(\alpha+1,\beta-1)$ for any $\alpha,\beta \in \mathbb{R}$ with $\beta \geq \alpha+2$.

Further, for the solution z of (14) we denote

$$\psi(t,x) = (\ell z)(t,x) = z(t+1,x) - z(t-1,x).$$

The continuity of ℓ implies

$$C\int_0^T \int_\omega z^2 dxdt \geq \int_1^{T-1} \int_\omega \psi^2 dxdt \tag{25}$$

for some positive constant C independent of z .

On the other hand, in view of the time periodicity of p, it follows that ψ is a solution of the equation

$$\begin{cases} \psi_{tt} - \Delta\psi = 0 & \text{in } \mathbb{R}\times(0,1), \\ \psi = 0 & \text{in } \mathbb{R}\times\{0,1\}, \end{cases}$$

and then, since $(T-1)-1 = T-2 \geq 2$, inequality (19) and (22) yield

$$C\int_1^{T-1} \int_\omega \psi^2 dxdt \geq \|(\psi(\mu,.),\psi_t(\mu,.))\|^2_{0,-1} \tag{26}$$

for every $\mu \in \mathbb{R}$.

Now denote $n = (T-T_0)/2$ and observe that

$$\begin{aligned} \|z(T)-z(T_0)\|^2 &= \left\| \sum_{k=0}^{n-1} (z(T-2k) - z(T-2k-2)) \right\|^2 \\ &\leq \left(\sum_{k=0}^{n-1} \|z(T-2k) - z(T-2k-2)\| \right)^2 \\ &\leq n \sum_{k=0}^{n-1} \|z(T-2k) - z(T-2k-2)\|^2 \\ &= n \sum_{k=0}^{n-1} \|\psi(T-2k-1)\|^2 . \end{aligned}$$

Thus, from inequality (26) we get

$$\|z(T)-z(T_0)\|_0^2 \leq n^2 C \int_1^{T-1} \int_\omega \psi^2 dxdt.$$

Exactly in the same way it follows that

$$\|z_t(T)-z_t(T_0)\|_{-1}^2 \leq n^2 C \int_1^{T-1} \int_\omega \psi^2 dxdt.$$

From the last two inequalities, the fact that $z(T,x) = z_t(T,x) = 0$, and (25) we obtain inequality (23).

To prove inequality (24) we multiply (14) by p and integrate over $[T_0, T] \times (0,1)$. After integration by parts it follows that

$$\int_{T_0}^{T} \int_{\mathcal{O}} p^2 dxdt = \int_0^1 (z_t(T_0)p(T_0) - z(T_0)p_t(T_0))\, dx. \tag{27}$$

Now observe that, on one hand,

$$\int_0^1 (z_t(T_0)p(T_0) - z(T_0)p_t(T_0))\, dx \le \|(z(T_0), z_t(T_0))\|_{0,-1} \|(p(T_0), p_t(T_0))\|_{1,0},$$

and, on the other hand, from inequality (19),

$$\int_{T_0}^{T} \int_{\mathcal{O}} p^2 dxdt \ge C \left\|(p^0, p^1)\right\|_{0,-1}^2.$$

Thus, from (27) inequality (24) is obtained. □

Remark 9. Although the observability inequality provided by Proposition 8 allows us to ensure the appropriate unique continuation property and then the approximate controllability of system (6)–(7), it is not good enough to guarantee the exact controllability of that system, since it does not allow us to estimate a norm of (p^0, p^1) in terms of the observed quantity $\int_0^T \int_\omega z^2 dtdx$.

3. Boundary control. In this section we consider the boundary controlled system (2). We follow the same scheme as in the previous section. The main difference is that in this case we are able to prove the existence of (exact) insensitizing controls.

3.1. Reduction to a controllability problem. The proofs of the assertions given below use essentially the same arguments as in section 2.

Consider the following controlled cascade system of wave equations:

$$\begin{cases} \bar{y}_{tt} - \Delta \bar{y} = \xi & \text{in } \mathbb{R} \times (0,1), \\ \bar{y}(t,0) = 0, \quad \bar{y}(t,1) = g(t) & \text{in } \mathbb{R}, \\ \bar{y}(0,x) = y^0, \quad \bar{y}_t(0,x) = y^1 & \text{in } (0,1), \end{cases} \tag{28}$$

$$\begin{cases} q_{tt} - \Delta q = \bar{y}\mathbf{1}_{\mathcal{O}} & \text{in } \mathbb{R} \times (0,1), \\ q = 0 & \text{in } \mathbb{R} \times \{0,1\}, \\ q(T,x) = 0, \quad q_t(T,x) = 0 & \text{in } (0,1). \end{cases} \tag{29}$$

Proposition 10. *Given $y^0 \in H_0^1(0,1), y^1 \in L^2(0,1)$, and $\xi \in L^2(\mathbb{R} \times (0,1))$, the control f insensitizes Φ along the solutions of* (2) *if and only if the corresponding solution $(\bar{y}, q)$ of the cascade system* (28)–(29) *satisfies*

$$q(0,x) = q_t(0,x) = 0. \tag{30}$$

The latter exact controllability property may be characterized with the help of the adjoint cascade system

$$\begin{cases} p_{tt} - \Delta p = 0 & \text{in } \mathbb{R} \times (0,1), \\ p = 0 & \text{in } \mathbb{R} \times \{0,1\}, \\ p(0,x) = p^0, \quad p_t(0,x) = p^1 & \text{in } (0,1), \end{cases} \tag{31}$$

$$(32)\qquad \begin{cases} z_{tt} - \Delta z = p\mathbf{1}_{\mathcal{O}} & \text{in } \mathbb{R}\times(0,1),\\ z = 0 & \text{in } \mathbb{R}\times\{0,1\},\\ z(T,x) = 0, \quad z_t(T,x) = 0 & \text{in } (0,1). \end{cases}$$

The following proposition holds.

PROPOSITION 11. *Given* $\left(y^0, y^1\right) \in H_0^1(0,1)\times L^2(0,1)$ *and* $g \in L^2(\mathbb{R})$, *the solution* $(\overline{y}, q)$ *of* (28)–(29) *satisfies* (30) *if and only if the following duality identity holds for every* $\left(p^0, p^1\right) \in L^2(0,1)\times H^{-1}(0,1)$*:*

$$(33)\qquad \int_0^T g(t) z_x(1,t)dt + \int_0^T\int_0^1 \xi z dxdt = \int_0^1 \left(y^0 z_t(0,x) - y^1 z(0,x)\right) dx.$$

3.2. Observability inequalities.

PROPOSITION 12. *Let* $\mathcal{O}$ *be a nonempty open subset of* $[0,1]$ *and let* $T \geq 4$. *Then there exists a positive constant* C, *such that for every solution of* (31)–(32),

$$C\int_0^T z_x^2(1,t)dt \geq \left\|(p^0,p^1)\right\|_{0,-1}^2.$$

Proof. The inequality given in the proposition is a consequence of the fact that there exist positive constants C_1, C_2, such that for every solution of (31)–(32),

$$C_1\int_0^T z_x^2(1,t)dt \geq \left\|(z(T_0), z_t(T_0))\right\|_{1,0}^2,$$
$$C_2\left\|(z(T_0), z_t(T_0))\right\|_{1,0} \geq \left\|(p^0,p^1)\right\|_{0,-1},$$

where T_0 is the smallest positive number such that $T - T_0 \in 2\mathbb{Z}$.

These inequalities are proved using the same arguments as in Proposition 8. ∎

3.3. Existence of insensitizing controls. Using Proposition 12 we can prove the following result.

THEOREM 13. *Let* $\mathcal{O}$ *be a nonempty open subset of* $[0,1]$ *and let* $T \geq 4$. *Then for every* $\xi \in L^2(\mathbb{R}\times(0,1))$ *and* $\left(y^0, y^1\right) \in H_0^1(0,1)\times L^2(0,1)$ *there exists a control* $g \in L^2(\mathbb{R}\times(0,1))$ *that insensitizes* Φ *along the solutions of* (2).

Proof. Consider the quadratic functional $J : L^2(0,1)\times H^{-1}(0,1) \to \mathbb{R}$ defined by

$$J(p^0,p^1) = \frac{1}{2}\int_0^T z_x^2(1,t)dt + \int_0^T\int_0^1 \xi z dxdt - \int_0^1 \left(y^0 z_t(0,x) - y^1 z(0,x)\right) dx,$$

where z is the corresponding solution of (32). Let $(\hat{p}^0, \hat{p}^1)$ be the minimizer of J provided by Proposition 14 (see below):

$$J(\hat{p}^0,\hat{p}^1) = \min\left\{J(p^0,p^1) : (p^0,p^1) \in L^2(0,1)\times H^{-1}(0,1)\right\}.$$

Then for every $(p^0,p^1) \in L^2(0,1)\times H^{-1}(0,1)$ the function $\phi(h) = J(\hat{p}^0 + hp^0, \hat{p}^1 + hp^1)$ has a minimum at $h = 0$ and thus

$$0 = \phi'(0) = \int_0^T \hat{z}_x(1,t) z_x(1,t)dt + \int_0^T\int_0^1 \xi\hat{z} dxdt - \int_0^1 \left(y^0\hat{z}_t(0,x) - y^1\hat{z}(0,x)\right) dx,$$

where $\hat{z}$ is the solution of (31)–(32) with initial data $(\hat{p}^0, \hat{p}^1)$. Consequently, if we choose $g(t) = \hat{z}_x(1,t)$, equality (33) is verified. Therefore, the control $g(t) = \hat{z}_x(1,t)$ insensitizes Φ. ∎

PROPOSITION 14. *The functional J has a minimizer $(\hat{p}^0, \hat{p}^1) \in L^2(0,1)\times H^{-1}(0,1)$.*

Proof. J in continuous and convex. So, to guarantee the existence of a minimizer it is sufficient to prove that J is coercive, i.e., $J \to \infty$ as $\left\|(p^0,p^1)\right\|_{0,-1} \to \infty$.

Let us note first that z verifies

$$\int_0^T \int_0^1 z^2 dxdt \le C \int_0^T \int_{\mathcal{O}} p^2 dxdt \le C \left\|(p^0,p^1)\right\|_{0,-1}^2,$$
$$\left\|(z(0), z_t(0))\right\|_{1,0} \le C \int_0^T \int_{\mathcal{O}} p^2 dxdt \le C \left\|(p^0,p^1)\right\|_{0,-1}^2.$$

Consequently,

$$\left|\int_0^T \int_0^1 \xi z dxdt\right| \le \left(\int_0^T \int_0^1 \xi^2 dxdt\right)^{\frac{1}{2}} \left(\int_0^T \int_0^1 z^2 dxdt\right)^{\frac{1}{2}} \le C_1 \left\|\xi\right\|_{L^2((0,T)\times(0,1))} \left\|(p^0,p^1)\right\|_{0,-1}, \tag{34}$$

$$\left|\int_0^1 \left(y^0 z_t(0) - y^1 z(0,x)\right) dx\right| \le \left\|(y^0,y^1)\right\|_{0,-1} \left\|(z(0), z_t(0))\right\|_{1,0} \le C_2 \left\|(y^0,y^1)\right\|_{0,-1} \left\|(p^0,p^1)\right\|_{0,-1}, \tag{35}$$

where C_1 and C_2 are positive constants.

From Proposition 12, (34), and (35) we get

$$\left|J(p^0,p^1)\right| \ge \left\|(p^0,p^1)\right\|_{0,-1} \left(C \left\|(p^0,p^1)\right\|_{0,-1} - C_1 \left\|\xi\right\|_{L^2((0,T)\times(0,1))} - C_2 \left\|(y^0,y^1)\right\|_{0,-1}\right)$$

and this implies that $J \to \infty$ as $\left\|(p^0,p^1)\right\|_{0,-1} \to \infty$. Hence the assertion of the proposition is proved. □

4. Final remarks. The condition that the observation time be large ($T \ge 4$), which does not appear in previous papers in the context of the heat equations, is natural for the wave equation due to the finite speed of propagation of its solutions. Indeed, it is possible to prove that, for every $T < 2$ one can choose an observation region $\mathcal{O}$ such that a certain nontrivial solution of (17) satisfies $p = 0$ in $[0,T]\times\mathcal{O}$, and then $z = 0$ in $[0,T]\times(0,1)$. This means that the unique continuation property is not verified. One may expect, however, the property to be true for $2 < T < 4$, and in fact we do not know whether the property is true in this case or not. Nevertheless, the condition $T \ge 4$ (twice the sum of the lengths of the strings) appears to be natural in the context of coupled wave equations. Several examples of this situation are given in [5] (see, for instance, section 4.9 in Chapter 4). Note, however, that for every particular choice of the observation interval $\mathcal{O}$ the minimal observation time to guarantee the unique continuation property is strictly smaller than 4.

Of course, it would be of interest to study the insensitizing properties of the controlled systems for other choice of the observation functionals, in particular,

$$\Phi(y) = \frac{1}{2} \int_0^T \int_{\mathcal{O}} \left(y_t^2 + y_x^2\right) dxdt,$$

which may be understood as the mean value of the energy of the portion $\mathcal{O}$ of the string in a time interval of length T. This problem turns out to be more difficult, at least at the level of reducing it to a controllability property.

It is clear that the method we have used to prove the unique continuation does not apply to other 1-D equations, for which the time periodicity property is not valid. However, it does not mean that the results are not valid for those equations. We hope that the existence of ε-insensitizing controls for nonintersecting observation and control regions is not restricted to the 1-D wave equation, and the example given in [10] suggests that this property is also true for the 1-D heat equation.

Acknowledgments. The author would like to express his gratitude to Enrique Zuazua for valuable discussions on this problem and also to the referees for their help in improving the presentation of the paper.

REFERENCES

[1] O. Bodart, M. González-Burgos, and R. Pérez-García, *Existence of insensitizing controls for a semilinear heat equation with a superlinear nonlinearity*, Comm. Partial Differential Equations, 29 (2004), pp. 1017–1050.

[2] O. Bodart, M. González-Burgos, and R. Pérez-García, *Insensitizing controls for a heat equation with a nonlinear term involving the state and the gradient*, Nonlinear Anal., 57 (2004), pp. 687–711.

[3] O. Bodart and C. Fabre, *Controls insensitizing the norm of the solution of a semilinear heat equation*, J. Math. Anal. Appl., 195 (1995), pp. 658–683.

[4] O. Bodart, M. González-Burgos, and R. Pérez-García, *A local result on insensitizing controls for a semilinear heat equation with nonlinear boundary Fourier conditions*, SIAM J. Control Optim., 43 (2004), pp. 955–969.

[5] R. Dáger and E. Zuazua, *Wave Propagation, Observation and Control in* 1*-d Flexible Multi-structures*, Math. Appl. 50, Springer-Verlag, Berlin, 2006.

[6] L. de Teresa, *Controls insensitizing the norm of the solution of a semilinear heat equation in unbounded domains*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 125–149.

[7] L. de Teresa, *Insensitizing controls for a semilinear heat equation*, Comm. Partial Differential Equations, 25 (2000), pp. 39–72.

[8] E. Fernández-Cara, G. C. Garcia, and A. Osses, *Controls insensitizing the observation of a quasi-geostrophic ocean model*, SIAM J. Control Optim., 43 (2005), pp. 1616–1639.

[9] J.-L. Lions, *Quelques notions dans l'analyse et le contrôle de systèmes à données incomplètes*, in Proceedings of the 11th Congress on Differential Equations and Applications/First Congress on Applied Mathematics (Málaga, 1989), Univ. Málaga, Málaga, Spain, 1990, pp. 43–54.

[10] S. Micu, J. H. Ortega, and L. de Teresa, *An example of ϵ-insensitizing controls for the heat equation with no intersecting observation and control regions*, Appl. Math. Lett., 17 (2004), pp. 927–932.

[11] S. Micu, J. H. Ortega, and L. de Teresa, *ε-insensitizing controls for pointwise observations of the heat equation*, Systems Control Lett., 51 (2004), pp. 407–415.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1769–1800

CONEWISE LINEAR SYSTEMS: NON-ZENONESS AND OBSERVABILITY*

M. KANAT CAMLIBEL†, JONG-SHI PANG‡, AND JINGLAI SHEN§

Abstract. Conewise linear systems are dynamical systems in which the state space is partitioned into a finite number of nonoverlapping polyhedral cones on each of which the dynamics of the system is described by a linear differential equation. This class of dynamical systems represents a large number of piecewise linear systems, most notably, linear complementarity systems with the P-property and their generalizations to affine variational systems, which have many applications in engineering systems and dynamic optimization. The challenges of dealing with this type of hybrid system are due to two major characteristics: mode switchings are triggered by state evolution, and states are constrained in each mode. In this paper, we first establish the absence of Zeno states in such a system. Based on this fundamental result, we then investigate and relate several state observability notions: short-time and T-time (or finite-time) local/global observability. For the short-time observability notions, constructive, finitely verifiable algebraic (both sufficient and necessary) conditions are derived. Due to their long-time mode-transitional behavior, which is very difficult to predict, only partial results are obtained for the T-time observable states. Nevertheless, we completely resolve the T-time local observability for the bimodal conewise linear system, for finite T, and provide numerical examples to illustrate the difficulty associated with the long-time observability.

Key words. piecewise linear systems, conewise linear systems, hybrid systems, Zeno behavior, observability

AMS subject classifications. 34C99, 34A40, 90C33, 93B07, 93C10

DOI. 10.1137/050645166

1. Introduction.

A conewise linear system (CLS) is a hybrid dynamical system consisting of a finite number of linear ordinary differential equations (ODEs) that are active on certain polyhedral cones which partition the whole Euclidean state space. Each of these cones is called a *mode* of the system; transitions between modes occur along a state trajectory. Many piecewise linear systems can be formulated as CLSs; among these, linear complementarity systems (LCSs) [15, 6] are perhaps the most prominent. Specifically, an LCS is defined by a linear time-invariant ODE containing an algebraic variable that is required to be a solution to a finite-dimensional linear complementarity problem (LCP). Collectively, these piecewise linear systems and their generalizations, such as the differential variational inequalities (DVIs) [25], have found a wide range of applications in nonsmooth mechanical systems, switched electrical

*Received by the editors November 14, 2005; accepted for publication (in revised form) May 21, 2006; published electronically December 11, 2006.
http://www.siam.org/journals/sicon/45-5/64516.html

†Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, and Department of Electronics and Communications Engineering, Dogus University, Istanbul, Turkey (k.camlibel@tue.nl). The work of this author is partially supported by the European Community through the Information Society Technologies thematic programme under the project SICONOS (IST-2001-37172) and by the Scientific and Technological Research Council of Turkey (TUBITAK) under grant 105E079.

‡Department of Mathematical Sciences and Department of Decision Science and Engineering Systems, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (pangj@rpi.edu). The research of this author was partially supported by a National Science Foundation Focused Research Group grant DMS 0353216.

§Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (shenj2@rpi.edu). The research of this author was partially supported by the National Science Foundation under grant DMS 0508986.

networks and control systems, and dynamic optimization in operations research and economics. See the two surveys [4, 31] and the recent papers [7, 8, 16, 17] as well as the references therein.

As with all switched dynamical systems, a critical issue associated with a CLS is whether infinitely many mode transitions exist in any finite time along a state trajectory, i.e., the *Zeno behavior* of the CLS. Such an issue was studied in a different setting for piecewise analytic systems in [5, 35] two decades ago. It has regained considerable attention and received extensive treatment in the hybrid system literature in the past few years, e.g., [19, 32, 37], due to its fundamental role in the study of numerical simulations and basic system and control properties of hybrid systems. Adding to the recent study of the Zeno issue [7] for complementarity systems, the paper [33] introduces several important notions of non-Zenoness and non-Zeno states of an LCS and establishes the "strong non-Zenoness" for an LCS with the P-property and the "weak non-Zenoness" for a broader class of LCSs. The paper [24] further extends the Zeno study to a nonlinear complementarity system (NCS) and to the DVI; it shows the strong non-Zenoness for an NCS satisfying the strong regularity condition and investigates certain system properties using these non-Zeno results.

Having its roots in the very early stages of modern control theory [20], observability is a fundamental concept in systems and control. Roughly speaking, observability refers to the ability of reconstructing the initial state from given output observations. This notion is well understood for linear systems [10]. However, characterization of observability of nonlinear systems (with control inputs) becomes a very hard problem. For instance, one has to analyze many different observability concepts of nonlinear systems, and only local sufficient conditions are available for small-time observability [23]; see the algebraic approach for analytic systems [2]. Moreover, checking these conditions can become a computationally untractable task [3].

The observability of hybrid systems has attracted growing attention in recent years. Mode and state observability of discrete-time switched linear systems are studied in [1], under the assumption that mode sequences are arbitrary; linear algebraic tests are provided, and the decidability is discussed. The paper [36] analyzes the observability of jump-linear systems and linear hybrid systems; necessary and sufficient conditions in terms of algebraic tests are given. Several observability notions are proposed for piecewise affine hybrid systems in [12]; sufficient conditions are obtained for the observability test and are used for observer design. Other related results include observability of Turing machines and its connection to hybrid systems [11]. For more discussions on observability analysis and observer design, see the references cited in the above-mentioned papers.

The present paper deals with the non-Zenoness and state observability of CLSs, assuming a linear system output. Compared with the observability results in the literature, there are two unique characteristics of the CLS that make the observability analysis challenging and different from the prior results: (i) mode switchings are triggered by state evolution instead of being arbitrarily chosen; (ii) the state is restricted to a cone in each mode. Due to the first property, the issues of well-posedness and Zenoness of system solutions become nontrivial. The second property implies that classical matrix rank conditions are insufficient to characterize observability properties. Moreover, a state trajectory is at best only once differentiable with respect to time and is not differentiable with respect to the initial condition. These properties necessitate the development of new tools to handle observability issues for this class of nonsmooth systems.

The organization of the paper is as follows. In section 2, we begin with the introduction of the main object of the study, CLSs, and discuss certain fundamental properties of the solutions. We then prove the non-Zenoness and piecewise analyticity of such solutions in section 3. Various kinds of observability notions are investigated, and the relation between them is discussed in section 4. It is shown that the linear dynamics, together with the conic state space partition, is instrumental in the derivation of constructive necessary and sufficient conditions for certain observability notions, particularly for short-time observability. Detailed investigation of T-time local observability of bimodal CLSs is given in section 5. For the long-time observability, we present examples to demonstrate several interesting properties that make this observability notion challenging to characterize, even for bimodal CLSs.

2. Conewise linear systems. Consider the ODE

$$\dot{x} = f(x), \tag{2.1}$$

where $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a piecewise affine (PA) function; i.e.,

(a) f is continuous, and
(b) a positive integer m and a family of affine functions $\{f_i\}_{i=1}^m$ exist such that $f(x) \in \{f_i(x)\}_{i=1}^m$ for all $x \in \mathbb{R}^n$.

We call systems of the form (2.1) *piecewise affine* systems. The representation (2.1) describes the system at hand in an implicit way via the component functions $\{f_i\}_{i=1}^m$. Alternatively, a geometric representation of (2.1) can be obtained by invoking well-known properties of PA functions (see, e.g., [14]). To elaborate on this, we recall that a finite collection of polyhedra in $\mathbb{R}^n$, denoted Ξ, is a *polyhedral subdivision* of $\mathbb{R}^n$ if

(a) the union of all polyhedra in Ξ is equal to $\mathbb{R}^n$,
(b) each polyhedron in Ξ is of dimension n, and
(c) the intersection of any two polyhedra in Ξ is either empty or a common proper face of both polyhedra.

For every PA function f, one can find a polyhedral subdivision of $\mathbb{R}^n$ and a finite family of affine functions $\{g_i\}$ such that f coincides with one of the functions $\{g_i\}$ on each polyhedron in Ξ [14, Proposition 4.2.1]. Let such a polyhedral subdivision be given by $\Xi = \{\mathcal{X}_i\}_{i=1}^m$, where each polyhedron $\mathcal{X}_i$, called a *piece* of the system, is described by a finite system of linear inequalities:

$$\mathcal{X}_i = \{x \mid C_i x + d_i \geqslant 0\} \tag{2.2}$$

for a certain matrix $C_i \in \mathbb{R}^{m_i \times n}$ and vector $d_i \in \mathbb{R}^{m_i}$; also write $g_i(x) = A_i x + b_i$ for some matrix A_i and vector b_i. With these definitions, we can write the system (2.1) in the equivalent form

$$\dot{x} = A_i x + b_i \quad \text{if } x \in \mathcal{X}_i. \tag{2.3}$$

In this case, continuity of the function f is equivalent to the following implication:

$$x \in \mathcal{X}_i \cap \mathcal{X}_j \Rightarrow A_i x + b_i = A_j x + b_j. \tag{2.4}$$

Since a PA function must be globally Lipschitz continuous (see, e.g., [14]), it follows from well-known ODE theory that the PA system (2.1) must admit a unique solution, which is denoted by $x(t, \xi)$, that is continuously differentiable (i.e., C^1) in time for any initial state $x(0) = \xi$. Moreover, it was recently proved in [26] that for each t, $x(t, \cdot)$

is a "semismooth" function on $\mathbb{R}^n$, meaning that it is "B(ouligand)-differentiable" (i.e., locally Lipschitz continuous and directionally differentiable) everywhere with the directional derivative along a prescribed direction given as the unique solution of a "first-order variational equation."

Throughout this paper, we focus on a particular type of PA system obtained by taking $b_i = d_i = 0$ for all i in (2.2) and (2.3). In this case, the system takes the form

$$\dot{x} = A_i x \quad \text{if } x \in \mathcal{X}_i \equiv \{x \mid C_i x \geqslant 0\}. \tag{2.5}$$

The continuity requirement of the right-hand side of (2.5) reduces to

$$x \in \mathcal{X}_i \cap \mathcal{X}_j \Rightarrow A_i x = A_j x; \tag{2.6}$$

i.e., $\mathcal{X}_i \cap \mathcal{X}_j \subseteq \ker(A_i - A_j)$, where ker denotes the kernel of a matrix. Since the pieces $\mathcal{X}_i$ are cones in this case, we call the system (2.5) a *conewise linear system* (CLS). Without loss of generality, we assume throughout that each matrix C_i contains no zero rows. Under this assumption, and by the fact that $\mathcal{X}_i$ is full dimensional, it follows that for each index $\ell = 1, \ldots, m_i$, there exists a vector $\widehat{x}^\ell \in \mathcal{X}_i$ such that $(C_i \widehat{x}^\ell)_\ell > 0$. Therefore, we must have

$$\varnothing \neq \operatorname{int} \mathcal{X}_i = \{x \mid C_i x > 0\}, \tag{2.7}$$

where int denotes the interior of a set. By property (c) of a polyhedral subdivision, it follows that $\mathcal{X}_j \cap \operatorname{int} \mathcal{X}_i = \varnothing$ for all $i \neq j$.

Associated with the "forward-time" system (2.5) is a *backward-time* system that allows us to obtain reverse-time results easily from a forward-time analysis. Specifically, for any given $T > 0$, define $x^{\mathrm{r}}(t) \equiv x(T - t)$. We have $x^{\mathrm{r}}(0) = x(T)$ and

$$\dot{x}^{\mathrm{r}} = \widetilde{A}_i \, x^{\mathrm{r}} \quad \text{if } x^{\mathrm{r}} \in \mathcal{X}_i, \tag{2.8}$$

where $\widetilde{A}_i \equiv -A_i$. Obviously, the latter system remains a CLS. The reverse-time system can be used to derive backward-time results pertaining to the forward-time trajectory. For instance suppose that $x(t_0, \xi) = x(t_0, \eta) = z^0$ for some $t_0 > 0$ and some ξ and η in $\mathbb{R}^n$. By considering the reverse-time trajectory starting at time t_0 and going backwards in time until the initial time $t = 0$ and by using the uniqueness of the solution to the reverse-time system given an initial condition, it follows that $\xi = x^{\mathrm{r}}(t_0, z^0) = \eta$. In words, this observation says that if two forward-time trajectories starting at two initial conditions ξ and η ever intersect at some common future time, then these two trajectories must in fact be identical at all times.

CLSs form a special class of linear hybrid systems (see, for instance, [22]). In fact, they can be cast as hybrid automata for which

(a) the vector fields in each location are linear,
(b) the invariant sets are solid polyhedral cones,
(c) the guard sets are the boundaries of these cones, and
(d) the reset maps are all identity.

In what follows, we look at two specific examples of CLSs.

Example 2.1. Bimodal CLSs are the simplest CLSs with only 2 pieces; i.e., $m = 2$ and $\Xi = \{\mathcal{X}_1, \mathcal{X}_2\}$. We claim that any such system can be described by the ODE

$$\dot{x} = Ax + b \max(0, c^T x) \tag{2.9}$$

for some $n \times n$ matrix A and n-vectors b and c. This is a nontrivial assertion; indeed, we need to show that given (2.5) and (2.6), we can identify the matrix A and the two

vectors b and c such that (2.9) is equivalent to the given bimodal CLS. The proof is as follows. Since $\Xi = \{\mathcal{X}_1, \mathcal{X}_2\}$ is a polyhedral division of $\mathbb{R}^n$, it follows that

$$\operatorname{int} \mathcal{X}_1 \cap \operatorname{int} \mathcal{X}_2 = \varnothing, \tag{2.10}$$

$$\mathcal{X}_1 \cup \mathcal{X}_2 = \mathbb{R}^n. \tag{2.11}$$

In view of (2.10), it follows from [34, Theorem 3.3.4] that there exists a hyperplane, say $\mathcal{H} = \{x \in \mathbb{R}^n \mid c^T x = 0\}$, such that $\mathcal{X}_1 \subseteq \mathcal{H}_+ \equiv \{x \in \mathbb{R}^n \mid c^T x \geqslant 0\}$ and $\mathcal{X}_2 \subseteq \mathcal{H}_- \equiv \{x \in \mathbb{R}^n \mid c^T x \leqslant 0\}$. We claim that $\mathcal{X}_1 = \mathcal{H}_+$. To see this, note that

$$(\operatorname{int} \mathcal{H}_+ \setminus \mathcal{X}_1) \cap \mathcal{X}_1 = \varnothing, \tag{2.12}$$

$$(\operatorname{int} \mathcal{H}_+ \setminus \mathcal{X}_1) \cap \mathcal{X}_2 \subseteq \operatorname{int} \mathcal{H}_+ \cap \mathcal{H}_- = \varnothing. \tag{2.13}$$

Then, (2.11) together with (2.12)–(2.13) implies that $\operatorname{int} \mathcal{H}_+ \setminus \mathcal{X}_1 = \varnothing$, i.e., $\operatorname{int} \mathcal{H}_+ \subseteq \mathcal{X}_1$. Since $\mathcal{X}_1$ is contained in $\mathcal{H}_+$ and it is a closed set, we get $\mathcal{X}_1 = \mathcal{H}_+$. In a similar fashion, one can show that $\mathcal{X}_2 = \mathcal{H}_-$. Then, we get $\mathcal{X}_1 \cap \mathcal{X}_2 = \mathcal{H}_+ \cap \mathcal{H}_- = \mathcal{H} = \ker(c^T)$. Hence, we can write (2.5)–(2.6) as

$$\dot{x} = \begin{cases} A_1 x & \text{if} \quad c^T x \geqslant 0, \\ A_2 x & \text{if} \quad c^T x \leqslant 0, \end{cases} \tag{2.14}$$

where A_1 and A_2 satisfy $c^T x = 0 \Rightarrow A_1 x = A_2 x$. Equivalently, $A_1 - A_2 = bc^T$ for some n-vector b. Thus, (2.14) becomes (2.9) with $A \equiv A_2$.

Example 2.2. A broad class of CLSs consists of the following *linear cone complementarity system* (LCCS):

$$\begin{aligned} \dot{x} &= Ax + Bz, \\ \mathcal{C} \ni z &\perp Cx + Dz \in \mathcal{C}^*, \end{aligned} \tag{2.15}$$

where $x \in \mathbb{R}^n$, $z \in \mathbb{R}^p$, $\mathcal{C}$ is a polyhedral cone and $\mathcal{C}^*$ is its dual (as in convex analysis [27]), and $a \perp b$ means $a^T b = 0$. A wealth of examples, from various areas of engineering as well as operations research, of these piecewise linear (hybrid) systems can be found in [31, 30, 16]. For references on the analysis of the general LCCS, we refer the reader to [8, 17, 7, 28, 29, 18]. A special case of interest emerges when $\mathcal{C} = \mathbb{R}^p_+$; the resulting LCCS is called simply a *linear complementarity system* (LCS). A fundamental subclass of the LCSs arises when all the principal minors of the matrix D are positive. Such matrices are called P-matrices in the literature of mathematical programming. The class of P-matrices is very broad (see [13]); in particular, it includes the class of positive definite (not necessarily symmetric) matrices. Most importantly, P-matrices play a fundamental role in the LCP, i.e., the problem of finding a p-vector z satisfying

$$0 \leqslant z \perp q + Dz \geqslant 0 \tag{2.16}$$

for a given p-vector q and a $p \times p$ matrix D. We denote the latter problem by $\mathrm{LCP}(q, D)$. It is well known that the $\mathrm{LCP}(q, D)$ admits a unique solution for all $q \in \mathbb{R}^p$ if and only if D is a P-matrix; see [13, Theorem 3.3.7]. Moreover, for each q there exists an index set $\alpha \subseteq \{1, 2, \ldots, p\}$ with complement $\bar{\alpha}$ such that

(a) $-(D_{\alpha\alpha})^{-1} q_\alpha \geqslant 0$ and $q_{\bar{\alpha}} - D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1} q_\alpha \geqslant 0$,

(b) the unique solution z of the LCP (q, D) is given by $z_\alpha = -(D_{\alpha\alpha})^{-1} q_\alpha$ and $z_{\bar{\alpha}} = 0$.

This shows that the solution mapping $q \mapsto z$ is a piecewise linear function on $\mathbb{R}^p$. Based on the above facts, we can rewrite the LCS (2.15) in the form of the CLS (2.5) as follows:

(2.17)
$$\dot{x} = (A - B_{\bullet\alpha}(D_{\alpha\alpha})^{-1}C_{\alpha\bullet})x \quad \text{if} \quad \begin{bmatrix} -D_{\alpha\alpha}^{-1} & 0 \\ -D_{\bar{\alpha}\alpha}(D_{\alpha\alpha})^{-1} & I_{\bar{\alpha}\bar{\alpha}} \end{bmatrix} \begin{bmatrix} C_{\alpha\bullet} \\ C_{\bar{\alpha}\bullet} \end{bmatrix} x \geqslant 0.$$

There are generalizations of the LCP results to the linear cone complementarity problem (LCCP), which can then be applied to the LCCS. In what follows, we discuss one such generalization that does not require the LCCP to have a unique solution. Let us denote by $\mathrm{SOL}(\mathcal{C}, q, D)$ the solution set of the LCCP

$$\mathcal{C} \ni z \perp q + Dz \in \mathcal{C}^*.$$

It has been observed [9, 25] that the LCCS (2.15) has a unique C^1 trajectory $x(t)$ for all initial conditions $x(0) = x^0 \in \mathbb{R}^n$ if and only if $B\mathrm{SOL}(\mathcal{C}, Cx, D)$ is a singleton for all $x \in \mathbb{R}^n$. If the latter singleton condition holds, then it follows that (2.15) is equivalent to

$$\dot{x} = Ax + B\mathrm{SOL}(\mathcal{C}, Cx, D), \tag{2.18}$$

where the right-hand side is a piecewise linear function of x on $\mathbb{R}^n$. Thus (2.18) is a CLS. A special case where $B\mathrm{SOL}(\mathcal{C}, Cx, D)$ is a singleton for all $x \in \mathbb{R}^n$ occurs when D is positive semidefinite (albeit not necessarily symmetric), $C\mathbb{R}^n \subseteq -D\mathcal{C} + \mathcal{C}^*$, and $(\mathcal{C} - \mathcal{C}) \cap \ker(D + D^T) \subseteq \ker B$. The first two conditions imply that $\mathrm{SOL}(\mathcal{C}, Cx, D)$ is a nonempty polyhedron for all $x \in \mathbb{R}^n$, and the last assumption ensures that $B\mathrm{SOL}(\mathcal{C}, Cx, D)$ is a singleton; see [14] for a proof of these facts.

Unlike the LCS with a P-matrix D, it is not straightforward to write down the pieces of the system (2.18); nevertheless, this can be achieved by introducing multipliers to the constraints defining the cone $\mathcal{C}$, which we write as

$$\mathcal{C} = \{ z \,|\, Gz \geqslant 0 \} \tag{2.19}$$

for some matrix $G \in \mathbb{R}^{r \times p}$. Letting $\lambda \in \mathbb{R}^r$ be the vector of multipliers corresponding to the constraint $Gz \geqslant 0$, the complementarity condition $\mathcal{C} \ni z \perp Cx + Dz \in \mathcal{C}^*$ is equivalent to

$$0 = Cx + Dz - G^T\lambda \quad \text{and} \quad 0 \leqslant \lambda \perp Gz \geqslant 0.$$

The pieces of the CLS (2.18) can be identified as follows. Define for each index subset α of $\{1, \ldots, r\}$, with complement $\bar{\alpha}$, the polyhedral cone $\widehat{\mathcal{X}}_\alpha \subset \mathbb{R}^n$ consisting of all vectors x for which there exist (z, λ_α) such that

$$\begin{aligned} 0 &= Cx + Dz - (G_{\alpha\bullet})^T\lambda_\alpha, \\ \lambda_\alpha \geqslant 0 &= G_{\alpha\bullet}z, \quad \text{and} \quad \lambda_{\bar{\alpha}} = 0 \leqslant G_{\bar{\alpha}\bullet}z. \end{aligned}$$

While there may be multiple pairs (z, λ_α) satisfying the above linear inequality system for a given x (which explains why $\widehat{\mathcal{X}}_\alpha$ is a polyhedral cone), the vector Bz is a constant among all such pairs, as long as $B\mathrm{SOL}(\mathcal{C}, Cx, D)$ is a singleton. Moreover, with some linear algebraic manipulations, it can be deduced that $Ax + B\mathrm{SOL}(\mathcal{C}, Cx, D) = \widehat{A}_\alpha x$ for some matrix $\widehat{A}_\alpha$ for all $x \in \widehat{\mathcal{X}}_\alpha$. Notice that the family $\{\widehat{\mathcal{X}}_\alpha\}$ for α ranging over all subsets of $\{1, \ldots, r\}$ may not form a polyhedral subdivision of $\mathbb{R}^r$ (for one thing, some of them could overlap); nevertheless, they are enough to show that the right-hand side of (2.18) is a piecewise linear function of x.

2.1. Some structural properties. In this subsection, we establish some basic structural properties of the CLS (2.5). First we review some well-known concepts. An ordered tuple $a \equiv (a_1, \ldots, a_k)$ of real numbers is said to be *lexicographically nonnegative* if either $a = 0$ or its first nonzero component is positive. In this case, we write $a \succcurlyeq 0$. If a is nonzero and lexicographically nonnegative, we say that a is lexicographically positive. In this case, we write $a \succ 0$. Sometimes, we also use the signs "$\preccurlyeq$" and "$\prec$" with the obvious meanings. The set of lexicographically nonnegative k-tuples forms a convex, albeit not closed, cone in $\mathbb{R}^k$. A finite collection of n-dimensional vectors $(y^1, y^2, \ldots, y^k)$ for some positive integer k is said to be lexicographically nonnegative (positive), denoted $(y^1, y^2, \ldots, y^k) \succcurlyeq (\succ) 0$, if for each $j = 1, \ldots, n$, the k-dimensional tuple $(y^1_j, \ldots, y^k_j)$ is lexicographically nonnegative (positive).

For a given $x^0 \in \mathbb{R}^n$, the solution trajectory $x(t, x^0)$ of the CLS (2.5) does not necessarily stay in the same piece as the initial state x^0 for all sufficiently small $t > 0$. To ensure the latter persistence property, we define the following sets: for all $i = 1, \ldots, m$,

$$\mathcal{Y}_i \equiv \left\{ x \in \mathbb{R}^n \,\middle|\, \left(C_i x, C_i A_i x, \ldots, C_i A_i^{n-1} x \right) \succcurlyeq 0 \right\}. \tag{2.20}$$

Obviously, $\mathcal{Y}_i$ is a convex, albeit not closed, cone in $\mathbb{R}^n$; it bears a close connection with the set $\mathcal{X}_i$ as described in the following result, whose proof is elementary and thus omitted. In the result, we let cl denote the closure of a set.

LEMMA 2.3. *Assume that every matrix C_i has no zero rows. The following statements hold for all $i = 1, \ldots, m$:* (a) $\mathcal{Y}_i \subseteq \mathcal{X}_i$; (b) $\operatorname{int} \mathcal{X}_i \subseteq \mathcal{Y}_i$; (c) $\operatorname{cl} \mathcal{Y}_i = \mathcal{X}_i$; *and* (d) $\mathcal{Y}_i - \mathcal{Y}_i = \mathcal{X}_i - \mathcal{X}_i = \mathbb{R}^n$.

The next lemma characterizes the elements of the set $\mathcal{Y}_i$; it shows in particular that the solution trajectory $x(t, x^0)$ stays in one piece, say $\mathcal{X}_i$, for all $t > 0$ sufficiently small if and only if $x^0 \in \mathcal{Y}_i$.

LEMMA 2.4. *The three statements below are equivalent for any vector $x^0 \in \mathbb{R}^n$:*

(a) $x^0 \in \mathcal{Y}_i$;

(b) *there exists a positive number ε such that $x(t, x^0) \in \mathcal{X}_i$ for all $t \in [0, \varepsilon]$;*

(c) *for some (equivalently, any) positive n-vector $c = (c_1, \ldots, c_n)$, there exists a number $\mu_0 > 0$ such that $\sum_{k=0}^{n-1} c_{k+1}\, \mu^k A_i^k x^0 \in \mathcal{X}_i$ for all $\mu \in [0, \mu_0]$.*

Proof. The equivalence of (a) and (b) is easy. We prove only the equivalence of (a) and (c). Observe that (c) is equivalent to $\sum_{k=0}^{n-1} c_{k+1}\, \mu^k (C_i A_i^k x^0)_\ell \geqslant 0$ for each index $\ell = 1, \ldots, m_i$ and all $\mu \in [0, \mu_0]$. Therefore, the positivity of $c_1, \ldots, c_n$ and μ implies $((C_i x^0)_\ell, \ldots, (C_i A_i^{n-1} x^0)_\ell) \succcurlyeq 0$ for all indices $\ell = 1, \ldots, m_i$. Hence $x^0 \in \mathcal{Y}_i$. Conversely, suppose (c) holds, but (a) does not. Then there exist an index $\ell \in \{1, \ldots, m_i\}$ and an integer $0 \leq k \leq n-1$ such that $(C_i A_i^k x^0)_\ell < 0$ and $(C_i A_i^j x^0)_\ell = 0$ for all $j = 0, \ldots, k-1$. Therefore, for sufficiently small $\mu > 0$, $\sum_{k=0}^{n-1} c_{k+1} \mu^k (C_i A_i^k x^0)_\ell < 0$. This contradicts (c). ∎

In general, a given initial state x^0 may be contained in multiple cones $\mathcal{X}_i$ and $\mathcal{Y}_i$. This motivates the definition of the following index sets. Given $\xi \in \mathbb{R}^n$, define

$$\mathcal{I}(\xi) \equiv \{i \mid \xi \in \mathcal{X}_i\} \quad \text{and} \quad \mathcal{J}(\xi) \equiv \{i \mid \xi \in \mathcal{Y}_i\}. \tag{2.21}$$

Basic relations between these sets are summarized in the following lemma.

LEMMA 2.5. *The following statements hold for any $\xi \in \mathbb{R}^n$:*

(a) *a neighborhood $\mathcal{N}$ of ξ exists such that $\mathcal{N} \subseteq \bigcup_{i \in \mathcal{I}(\xi)} \mathcal{X}_i$;*

(b) $\mathcal{J}(\xi) \subseteq \mathcal{I}(\xi)$;

(c) $A_i \xi = A_j \xi$ *if* $i, j \in \mathcal{I}(\xi)$;

(d) $A_i^k \xi = A_j^k \xi$ *for all positive integers k if* $i, j \in \mathcal{J}(\xi)$.

Proof. The proof is easy and omitted. ∎

With respect to the reverse-time system (2.8) where $\widetilde{A}_i = -A_i$, both sets

$$\mathcal{Y}_i^{\mathrm{r}} \equiv \left\{ x \in \mathbb{R}^n \,|\, (C_i x, C_i \widetilde{A}_i x, \ldots, C_i \widetilde{A}_i^{n-1} x) \succcurlyeq 0 \right\} \text{ and } \mathcal{J}^{\mathrm{r}}(\xi) \equiv \{ i \,|\, \xi \in \mathcal{Y}_i^{\mathrm{r}} \}$$

are not necessarily equal to the respective sets $\mathcal{Y}_i$ and $\mathcal{J}(\xi)$ that are defined with respect to the original forward-time system (2.5). Nevertheless, the forward-time trajectory and the reverse-time trajectory are equal in any interval. In particular, if $i \in \mathcal{J}^{\mathrm{r}}(x(t_*, \xi))$, where $t_* > 0$, then there exists $\varepsilon > 0$ such that $C_i x(t, \xi) \geqslant 0$ for all $t \in [t_* - \varepsilon, t_*]$.

3. Non-Zeno property of CLSs. In the hybrid systems literature, the occurrence of an infinite number of mode transitions within a finite time interval is called the *Zeno behavior* with reference to the ancient Greek philosopher Zeno's paradoxes.[1] Our goal in this section is to show that the CLS does not exhibit such behavior. At the end of the section, we will compare the main result specialized to the LCS with those obtained in [33].

DEFINITION 3.1. *The PA system* (2.1) *is said to satisfy the* (forward and backward) non-Zeno property *if for any* $x^0 \in \mathbb{R}^n$ *and any* $t' \geq 0$*, there exist* $\varepsilon_\pm > 0$ *and* $\mathcal{X}_{i^\pm} \in \Xi$ *such that* $x(t, x^0) \in \mathcal{X}_{i^+}$ *for all* $t \in [t', t' + \varepsilon_+]$ *(forward-time non-Zeno) and, for* $t' > 0$*,* $x(t, x^0) \in \mathcal{X}_{i^-}$ *for all* $t \in [t' - \varepsilon_-, t']$ *(backward-time non-Zeno).*

(Note: the backward-time non-Zeno property is not defined at the initial time $t' = 0$; since the trajectory $x(t, x^0)$ is in principle defined only for $t \geqslant 0$, in the backward-time non-Zeno property at the time $t' > 0$, the scalar $\varepsilon_- > 0$ is taken to be less than t'.)

The following result shows that the backward non-Zeno property of the forward-time CLS (2.5) is equivalent to the forward non-Zeno property of the reverse-time CLS (2.8). It allows us to focus our attention on the forward non-Zeno property subsequently.

PROPOSITION 3.2. *The system* (2.5) *has the backward non-Zeno property if and only if the system* (2.8) *has the forward non-Zeno property.*

Proof. Suppose that (2.5) has the backward non-Zeno property. Let $x^0 \in \mathbb{R}^n$ and $t' \geqslant 0$ be given. Consider the reverse-time trajectory $x^{\mathrm{r}}(t, x^0)$ beginning at x^0 and terminating at a state $\xi^0 \equiv x^{\mathrm{r}}(T, x^0)$ at time $T > t'$. We then have $x^{\mathrm{r}}(t', x^0) = x(T - t', \xi^0)$. It follows by the backward non-Zeno property of the forward CLS (2.5) that an $\varepsilon > 0$ and an $\mathcal{X}_{i^-}$ exist such that $x(t, x^0) \in \mathcal{X}_{i^-}$ for all $t \in [T - t' - \varepsilon, T - t']$. Hence $x^{\mathrm{r}}(\widetilde{t}, x^0) = x(T - \widetilde{t}, \xi^0) \in \mathcal{X}_{i^-}$ for all $\widetilde{t} \in [t', t' + \varepsilon]$. Therefore, the system (2.8) has the forward non-Zeno property. The converse can be proved similarly. ∎

The next lemma is the first step in showing that the CLSs do not have the Zeno behavior.

LEMMA 3.3. *The following three statements are equivalent.*

(a) *The CLS* (2.5) *satisfies the forward non-Zeno property.*

(b) $\cup_{i=1}^m \mathcal{Y}_i = \mathbb{R}^n$.

(c) *For every* $\xi \in \mathbb{R}^n$*,* $\mathcal{J}(\xi) \neq \varnothing$.

Proof. Suppose (a) holds. Since for an arbitrary $x^0 \in \mathbb{R}^n$, there exist an $\varepsilon > 0$ and a piece $\mathcal{X}_i$ such that $x(t, x^0) \in \mathcal{X}_i$ for all $t \in [0, \varepsilon]$, we have $x^0 \in \mathcal{Y}_i$ by Lemma 2.4. Thus $\mathbb{R}^n \subseteq \cup_{i=1}^m \mathcal{Y}_i$, which clearly yields (b). For the converse, let $x^0 \in \mathbb{R}^n$ and $t' \geqslant 0$ be given. By (b), the vector $\xi \equiv x(t', x^0)$ belongs to some $\mathcal{Y}_{i^+}$. Hence Lemma 2.4 implies the existence of $\varepsilon_+ > 0$ such that $x(t, \xi) = x(t' + t, x^0) \in \mathcal{X}_{i^+}$ for all $t \in [0, \varepsilon_+]$. Thus (a) and (b) are equivalent. The equivalence of (b) and (c) is clear. ∎

[1] The most well known of these four paradoxes dealing with counterintuitive aspects of space and time is Achilles and the turtle.

We need another technical lemma in order to prove the main result of this section.

LEMMA 3.4. *Let $\xi \in \mathbb{R}^n$ be arbitrary. For any polynomial $p : \mathbb{R} \to \mathbb{R}^n$ with $p(0) = \xi$, there exists $i \in \mathcal{I}(\xi)$ such that $p(\mu) \in \mathcal{X}_i$ for all sufficiently small $\mu > 0$.*

Proof. Let $p(\mu)$ be a polynomial satisfying $p(0) = \xi$. For each index $i \in \mathcal{I}(\xi)$, there are only three possible cases:

(i) $p(\mu) \in \mathcal{X}_i$ for all sufficiently small $\mu > 0$;
(ii) $p(\mu) \notin \mathcal{X}_i$ for all sufficiently small $\mu > 0$;
(iii) there exists an infinite sequence of positive scalars $\{\mu_k\}$ all distinct and converging to zero as $k \uparrow \infty$ such that, for all $k = 1, 2, \ldots$, $p(\mu_{2k-1}) \in \mathcal{X}_i$ and $p(\mu_{2k}) \notin \mathcal{X}_i$.

If the claim of the lemma does not hold, then for each index $i \in \mathcal{I}(\xi)$ either (ii) or (iii) must hold. We claim that (iii) must hold at least for one $i \in \mathcal{I}(\xi)$ in this case. To show this, it is enough to prove that (ii) cannot hold for all $i \in \mathcal{I}(\xi)$. Suppose, on the contrary, that (ii) holds for all $i \in \mathcal{I}(\xi)$. Then, one gets $p(\mu) \notin \cup_{i \in \mathcal{I}(\xi)} \mathcal{X}_i$ for all sufficiently small $\mu > 0$. This, however, contradicts part (a) of Lemma 2.5. Therefore, there exists $i \in \mathcal{I}(\xi)$ such that (iii) holds. Without loss of generality, we may assume that the sequence $\{\mu_k\}$ is strictly decreasing. For each k, since $p(\mu_{2k}) \notin \mathcal{X}_i$, there exists an index ℓ_k such that $(C_i p(\mu_{2k}))_{\ell_k} < 0$. Since there are only finitely many such indices ℓ_k, there exists an index ℓ_0 such that $(C_i p(\mu_{2k}))_{\ell_0} < 0$ for infinitely many k's. Without loss of generality, we may assume that $(C_i p(\mu_{2k}))_{\ell_0} < 0$ for all k. Since $(C_i p(\mu_{2k-1}))_{\ell_0} \geqslant 0$, it follows that for all k, there exists $\bar{\mu}_k \in [\mu_{2k-1}, \mu_{2k})$ such that $(C_i p(\bar{\mu}_k))_{\ell_0} = 0$. Since the $\bar{\mu}_k$'s are all distinct (because the sequence $\{\mu_k\}$ is strictly decreasing) and $(C_i p(\mu))_{\ell_0}$ is a polynomial in μ with finitely many roots, we have $(C_i p(\mu))_{\ell_0} \equiv 0$ for all μ. This is a contradiction. □

With the help of the last two lemmas, we can now formally state and prove the absence of Zeno behavior in the CLS.

THEOREM 3.5. *The CLS* (2.5) *has the non-Zeno property; i.e., $\cup_{i=1}^m \mathcal{Y}_i = \mathbb{R}^n = \cup_{i=1}^m \mathcal{Y}_i^{\mathrm{r}}$.*

Proof. In view of Proposition 3.2, it suffices to show that any CLS has the forward non-Zeno property. In turn, by Lemma 3.3, it is enough to show that $\cup_{i=1}^m \mathcal{Y}_i = \mathbb{R}^n$. Take any $\xi \in \mathbb{R}^n$. Set $\eta_0 = \xi$ and $\mathcal{I}_0 = \mathcal{I}(\xi)$. The continuity implies that $A_i \eta_0 = A_j \eta_0$ for all $i, j \in \mathcal{I}_0$. Define

$$\mathcal{I}_1 \equiv \{i \in \mathcal{I}_0 \mid \eta_0 + \mu\eta_1 \in \mathcal{X}_i \ \forall \text{ sufficiently small } \mu > 0\},$$

where $\eta_1 \equiv A_i \eta_0$ for any $i \in \mathcal{I}_0$. Lemma 3.4 guarantees that $\mathcal{I}_1 \neq \varnothing$. Note that

$$A_i(\eta_0 + \mu\eta_1) = A_j(\eta_0 + \mu\eta_1) \tag{3.1}$$

for all $i, j \in \mathcal{I}_1$ and for all sufficiently small $\mu > 0$. Since $\mathcal{I}_1 \subseteq \mathcal{I}_0$, (3.1) implies that $A_i \eta_1 = A_j \eta_1$ for all $i, j \in \mathcal{I}_1$. Define

$$\mathcal{I}_2 \equiv \{i \in \mathcal{I}_0 \mid \eta_0 + \mu\eta_1 + \mu^2\eta_2 \in \mathcal{X}_i \ \forall \text{ sufficiently small } \mu > 0\},$$

where $\eta_2 \equiv A_i \eta_1$ for any $i \in \mathcal{I}_1$. Again, Lemma 3.4 guarantees that this set is nonempty. We claim that $\mathcal{I}_2 \subseteq \mathcal{I}_1$. To see this, let $i \in \mathcal{I}_2$. We need to show that $C_i(\eta_0 + \mu\eta_1) \geqslant 0$ for all $\mu > 0$ sufficiently small. Since $i \in \mathcal{I}_2$, we must have $C_i \eta_0 \geqslant 0$. If ℓ is an index such that $(C_i \eta_0)_\ell = 0$, then we must have $(C_i \eta_1)_\ell \geqslant 0$. Hence the claim holds. Therefore, for all $i, j \in \mathcal{I}_2$, we have $A_i(\eta_0 + \mu\eta_1 + \mu^2\eta_2) = A_j(\eta_0 + \mu\eta_1 + \mu^2\eta_2)$ for all $\mu > 0$ sufficiently small. Since $\mathcal{I}_2 \subseteq \mathcal{I}_1$, we deduce $A_i \eta_2 = A_j \eta_2$. Next, define

$$\mathcal{I}_3 \equiv \{i \in \mathcal{I}_0 \mid \eta_0 + \mu\eta_1 + \mu^2\eta_2 + \mu^3\eta_3 \in \mathcal{X}_i \ \forall \text{ sufficiently small } \mu > 0\},$$

where $\eta_3 \equiv A_i\eta_2$ for any $i \in \mathcal{I}_2$. In a similar fashion, we can show that $\mathcal{I}_3 \subseteq \mathcal{I}_2$ and $A_i\eta_3 = A_j\eta_3$ for all $i, j \in \mathcal{I}_3$. Continuing this process, we can eventually define $\mathcal{I}_{n-1}$, which is nonempty and is contained in $\mathcal{I}_{n-2} \subseteq \cdots \subseteq \mathcal{I}_1 \subseteq \mathcal{I}_0$. We claim that $i \in \mathcal{I}_{n-1}$ implies that $\xi \in \mathcal{Y}_i$. To see this, note that

$$\eta_0 + \mu\eta_1 + \cdots + \mu^{n-1}\eta_{n-1} \;=\; \xi + \mu A_i\xi + \cdots + \mu^{n-1}A_i^{n-1}\xi$$

for any $i \in \mathcal{I}_{n-1}$ because of the nested inclusions of the index sets $\mathcal{I}_j$ for $j = 0, 1, \ldots, n-1$. Hence, $i \in \mathcal{I}_{n-1}$ implies that $\sum_{k=0}^{n-1} \mu^k A_i^k \xi \in \mathcal{X}_i$ for all $\mu > 0$ sufficiently small; thus condition (c) of Lemma 2.4 is satisfied with c being the vector of all ones. Consequently, $\xi \in \mathcal{Y}_i$. □

The non-Zeno property is closely related to the boundedness of the number of "mode transitions" defined as follows.

DEFINITION 3.6. *Let $x(t, x^0)$ be a solution trajectory of the CLS* (2.5) *over a time interval $[0, T]$, $T > 0$, and let $t_* \in (0, T)$. We say that t_* is* not *a switching time if there exist $i \in \{1, \ldots, m\}$ and $\varepsilon > 0$ such that $x(t, x^0) \in \mathcal{X}_i$ for all $t \in [t_* - \varepsilon, t_* + \varepsilon]$; otherwise, we say that t_* is a* switching time*, and that the CLS has a mode transition or* mode switching *at t_*.*

With this definition, we easily obtain the following result from the non-Zeno property of the CLS. The proof is by a compactness argument and resembles that of Proposition 8 in [33].

THEOREM 3.7. *Let $x(t, x^0)$ be a solution trajectory of the CLS* (2.5) *on an open time interval containing $[0, T]$. Then there is a finite number of switching times in $[0, T]$. Hence, any such trajectory $x(\bullet, x^0)$ is a continuous, piecewise analytic function on $[0, T]$.*

Proof. See the cited proposition for a proof of the assertion about switching times. To prove the piecewise analyticity assertion, it suffices to note that if $x(t, x^0) \in \mathcal{X}_j$ for some j and all t in a subinterval $[t_{i-1}, t_i]$, where t_{i-1} and t_i are any two consecutive switching times, then $x(t, x^0) = e^{A_j(t-t_{i-1})}x(t_{i-1}, x^0)$ for all t in this subinterval. Hence $x(t, x^0)$ is an analytic function for $t \in (t_{i-1}, t_i)$. Since there are finitely many such subintervals, the piecewise analyticity of $x(\bullet, x^0)$ follows. □

For the bimodal system (2.9), we can say more; see Proposition 5.3. For now, we specialize Theorem 3.7 to the CLS (2.18), obtaining the following corollary.

COROLLARY 3.8. *Let $\mathcal{C}$ be given by* (2.19)*. Assume that $B\mathrm{SOL}(\mathcal{C}, Cx, D)$ is a singleton for all $x \in \mathbb{R}^n$. For every $x^0 \in \mathbb{R}^n$ and $T > 0$, there exist a pair of functions $(z, \lambda) : [0, T] \mapsto \mathcal{C} \times \mathbb{R}^r_+$ with $Bz(t)$ being continuous, a partition*

$$0 \;=\; t_0 \;<\; t_1 \;<\; \cdots \;<\; t_{N-1} \;<\; t_N \;=\; T \tag{3.2}$$

of the interval $[0, T]$, and index subsets $\alpha_j \subseteq \{1, \ldots, r\}$ each with complement $\bar{\alpha}_j$ for $j = 1, \ldots, N$, such that on each subinterval $[t_j, t_{j+1}]$, the triple $(x(t, x^0), z(t), \lambda(t))$ satisfies the linear differential algebraic equation

$$\begin{aligned} \dot{x}(t, x^0) &= Ax(t, x^0) + Bz(t), \\ 0 \quad &= Cx(t, x^0) + Dz(t) - G^T\lambda(t), \\ 0 \quad &= G_{\alpha_j \bullet} z(t) \quad \text{and} \quad \lambda_{\bar{\alpha}_j}(t) \;=\; 0. \end{aligned}$$

It is interesting to compare the above corollary with the non-Zenoness results in [33], which address only the LCS:

$$\dot{x} \;=\; Ax + Bz \quad \text{and} \quad 0 \;\leqslant\; z \;\perp\; Cx + Dz \;\geqslant\; 0.$$

There are some obvious similarities and subtle differences that are worth noting. The most obvious similarity is that all the results assert the finite number of switch times of some kind. The major difference lies in the treatment of the algebraic variable z. In Corollary 3.8, which is based on a hybrid system approach, z is treated implicitly; whereas in the treatment of [33], which originates from the P-matrix case and focuses on the fundamental triple of index sets

$$\begin{aligned} \alpha(t) &\equiv \{\, j \,|\, z_j(t) \,>\, 0 \,=\, (\, Cx(t,x^0) + Dz(t)\,)_j \,\}, \\ \beta(t) &\equiv \{\, j \,|\, z_j(t) \,=\, 0 \,=\, (\, Cx(t,x^0) + Dz(t)\,)_j \,\}, \\ \gamma(t) &\equiv \{\, j \,|\, z_j(t) \,=\, 0 \,<\, (\, Cx(t,x^0) + Dz(t)\,)_j \,\}, \end{aligned} \tag{3.3}$$

the switch times are defined with regard to a given trajectory $z(t)$. As a result of this difference in the points of view, the former corollary asserts the *existence* of a trajectory $z(t)$ satisfying the specified switching property; in contrast, the results in [33] start with a fixed but arbitrary trajectory $z(t)$ and establish the finite number of switch times for the pair $(x(t,x^0), z(t))$. The latter treatment has a price associated with it, namely, a restriction placed on the triple (B, C, D); such a restriction is not needed here. In the special case where D is a P-matrix (thus the trajectory $z(t)$ is unique), Theorem 9 in [33] is stronger than Corollary 3.8 here (for the LCS) in that the former asserts the constancy of the triple of index sets $(\alpha(t), \beta(t), \gamma(t))$ on the subintervals, whereas the latter pays no attention to the degenerate index set $\beta(t)$. As it is well known from finite-dimensional complementarity theory [14], the elements of $\beta(t)$ are most critical when one is interested in the sensitivity analysis of the system subject to parameter perturbations. The detailed exploration of this issue in a dynamic setting is beyond the scope of the present paper.

Not surprisingly, we can also establish the constancy of index sets for the CLS (2.5) similar to that for the P-matrix case of the LCS. We first establish the following proposition that pertains to an individual state.

PROPOSITION 3.9. *Let $x(t,x^0)$ be a solution trajectory of the CLS* (2.5) *over a time interval $[0,T]$. The following two statements hold.*

(a) *For every $t_* \in [0,T)$, there exists $\varepsilon_+ > 0$ such that $\mathcal{J}(x(t,x^0)) = \mathcal{J}(x(t_*,x^0))$ for all $t \in [t_*, t_* + \varepsilon_+]$.*

(b) *For every $t_* \in (0,T]$, there exists $\varepsilon_- > 0$ such that $\mathcal{J}(x(t,x^0)) = \mathcal{J}^{\,\mathrm{r}}(x(t_*,x^0))$ for all $t \in [t_* - \varepsilon_-, t_*)$.*

Proof. We prove only statement (a); the proof of (b) is similar. Write $\xi \equiv x(t_*,x^0)$. For each $i \in \mathcal{J}(\xi)$, Lemma 2.4 implies that there exist $\varepsilon_i > 0$ and $\mathcal{X}_i \in \Xi$ such that $x(t,x^0) \in \mathcal{X}_i$ for all $t \in [t_*, t_* + \varepsilon_i]$. Hence $x(t,x^0) \in \mathcal{Y}_i$ or, equivalently, $i \in \mathcal{J}(x(t,x^0))$ for all $t \in [t_*, t_* + \varepsilon_i/2]$. By letting $\varepsilon_+ \equiv \min_{1 \leqslant i \leqslant |\mathcal{J}(\xi)|} \varepsilon_i/2$, where $|\mathcal{J}(\xi)|$ denotes the cardinality of the set $\mathcal{J}(\xi)$, it follows that $\mathcal{J}(\xi) \subseteq \mathcal{J}(x(t,x^0))$ for all $t \in [t_*, t_* + \varepsilon_+]$. Conversely, consider an index $j \notin \mathcal{J}(\xi)$. If $(C_j\xi, C_jA_i\xi, \ldots, C_jA_i^{n-1}\xi) \succcurlyeq 0$ for some $i \in \mathcal{J}(\xi)$, then $C_jx(t,x^0) = C_je^{A_i(t-t_*)}\xi \geqslant 0$ for all $t \geqslant t_*$ sufficiently near t_*. Hence $j \in \mathcal{J}(\xi)$, which is a contradiction. Thus, $(C_j\xi, C_jA_i\xi, \ldots, C_jA_i^{n-1}\xi) \not\succcurlyeq 0$ for all $i \in \mathcal{J}(\xi)$. Hence, an index ℓ_i exists such that $(C_j\xi, C_jA_i\xi, \ldots, C_jA_i^{n-1}\xi)_{\ell_i} \prec 0$. By the non-Zenoness property, $\xi \in \mathcal{Y}_{i_0}$ for some $i_0 \in \mathcal{J}(\xi)$, which implies that $x(t,x^0) = e^{A_{i_0}(t-t_*)}\xi$ for all $t > t_*$ sufficiently near t_*. Hence $C_jx(t,x^0) = C_je^{A_{i_0}(t-t_*)}\xi$ for all such t. Since the tuple $(C_j\xi, C_jA_{i_0}\xi, \ldots, C_jA_{i_0}^{n-1}\xi)_{\ell_{i_0}}$ is nonzero and its first nonzero component is negative, it follows that $(C_jx(t,x^0))_{\ell_{i_0}} < 0$ for all $t > t_*$ sufficiently near t_*. Hence $j \notin \mathcal{J}(x(t,x^0))$ for all such t. Consequently, we must have $\mathcal{J}(\xi) \supseteq \mathcal{J}(x(t,x^0))$, and thus $\mathcal{J}(\xi) = \mathcal{J}(x(t,x^0))$, for all $t \in [t_*, t_* + \varepsilon_+]$, provided that $\varepsilon_+ > 0$ is further restricted if necessary. ∎

Extending the above proposition to a compact interval and using the reverse-time trajectory (2.8), we have the following result.

COROLLARY 3.10. *Let $x(t, x^0)$ be a solution trajectory of the CLS* (2.5) *on an open time interval containing $[0, T]$. There exists a partition* (3.2) *of the interval $[0, T]$ such that for every $i = 0, 1, \ldots, N-1$, $\mathcal{J}(x(t, x^0))$ is a constant for all $t \in [t_i, t_{i+1})$.*

Proof. By Proposition 3.9, we deduce that for every $t \in [0, T]$, there exists $\varepsilon_t > 0$ such that $\mathcal{J}(x(t', x^0)) = \mathcal{J}(x(t, x^0))$ for every $t' \in [t, t + \varepsilon_t]$ and $\mathcal{J}(x(\widetilde{t}, x^0)) = \mathcal{J}^{\,\mathrm{r}}(x(t, x^0))$ for every $\widetilde{t} \in [t - \varepsilon_t, t)$. We can now employ the same covering argument as in [33, Proposition 8] to complete the proof of the corollary. ∎

Switching times can also be expressed in terms of forward-time and backward-time index sets shown as follows.

PROPOSITION 3.11. *Let $x(t, x^0)$ be a solution trajectory of the CLS* (2.5). *Then a time $t_* > 0$ is a switching time if and only if $\mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0)) = \varnothing$.*

Proof. "Sufficiency." Suppose $\mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0)) = \varnothing$ but t_* is not a switching time. Then by Definition 3.6, there exist $i \in \{1, \ldots, m\}$ and $\varepsilon > 0$ such that $x(t, x^0) \in \mathcal{X}_i$ for all $t \in [t_* - \varepsilon, t_* + \varepsilon]$. This implies that $x(t_*, x^0) \in \mathcal{Y}_i \cap \mathcal{Y}_i^{\mathrm{r}}$ by taking derivatives of the forward-time trajectory at t_* and of the reverse-time trajectory at t_*, respectively. Thus $i \in \mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$. This results in a contradiction.

"Necessity." Suppose t_* is a switching time but $\mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0)) \neq \varnothing$. Let $i \in \mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$. Then $x(t_*, x^0) \in \mathcal{Y}_i \cap \mathcal{Y}_i^{\mathrm{r}}$. By Lemma 2.4 and the reverse-time argument, we deduce the existence of $\varepsilon > 0$ such that $x(t, x^0) \in \mathcal{X}_i$ for all $t \in [t_* - \varepsilon, t_* + \varepsilon]$. This contradicts the assumption that t_* is a switching time. ∎

One interesting observation about the CLS (2.5) is that a state trajectory may have boundary crossing, i.e., crossing a boundary of one cone and entering another cone, at a *nonswitching time* in the sense of Definition 3.6. We illustrate this observation by the following example.

Example 3.12. Consider a 3-dimensional CLS with the polyhedral subdivision:

$$\mathcal{X}_1 = \mathbb{R}_- \times \mathbb{R}^2, \quad \mathcal{X}_2 = \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+, \quad \mathcal{X}_3 = \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_-,$$

where $\mathbb{R}_\pm$ denote the nonnegative and nonpositive rays on the real line, respectively, and

$$A_1 = A_2 = A_3 = A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

It is easy to show that for any $x^0 \in \mathbb{R}^3$, $x(t, x^0) = e^{At} x^0 = \left(x_1^0\, e^t, x_2^0, x_3^0 + x_2^0 t \right)$ for all t. Now consider $x^0 = \left(0, 1, -1\right)$. Thus $x(t, x^0) = \left(0, 1, t - 1\right)$. Hence, $x(t, x^0) \in \mathcal{X}_1$ for all t and $x(t, x^0) \in \mathcal{X}_2$ for all $t \geqslant 1$, but $x(t, x^0) \notin \mathcal{X}_2$ for all $t < 1$. Consequently, $t_* = 1$ is *not* a switching time, but $x(t, x^0)$ crosses the boundary of $\mathcal{X}_2$ and $\mathcal{X}_3$ at $t_* = 1$.

We further illustrate this property via the index sets. Recall from Proposition 3.9 that for any given t_*, the index set $\mathcal{J}(x(t, x^0))$ remains constant for all t sufficiently close to t_*, both in the forward-time direction and in the backward-time direction. Note that the two constant index sets $\mathcal{J}(x(t_*, x^0))$ and $\mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$ may not be equal in general. In fact, expressing in terms of these index sets for this example, we have $\mathcal{J}^{\,\mathrm{r}}(x(1, x^0)) = \{1, 3\}$ and $\mathcal{J}(x(1, x^0)) = \{1, 2\}$. Notice that $\mathcal{J}^{\,\mathrm{r}}(x(1, x^0)) \cap \mathcal{J}(x(1, x^0)) = \{\,1\,\}$, but $\mathcal{J}^{\,\mathrm{r}}(x(1, x^0)) \neq \mathcal{J}(x(1, x^0))$.

The following proposition, however, shows that the LCS with the P-property, which is a special class of CLSs discussed previously, does not have the problem shown above and therefore exhibits relatively "simpler" switching behavior than general CLSs.

PROPOSITION 3.13. *Consider the LCS* (2.17) *with the P-property. If, for any* $t_* > 0$, $\mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$ *is nonempty, then* $\mathcal{J}(x(t_*, x^0)) = \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$. *In other words, if* t_* *is not a switching time, then* $\mathcal{J}(x(t_*, x^0)) = \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$.

Proof. It is shown in [33] that the LCS satisfying the P-property possesses the strong non-Zenoness at each state; i.e., for any t_*, there exist $\varepsilon_t > 0$ and two triples of index sets, $(\alpha_+, \beta_+, \gamma_+)$ and $(\alpha_-, \beta_-, \gamma_-)$, such that

$$\big(\,\alpha(t),\,\beta(t),\gamma(t)\,\big) = (\alpha_-, \beta_-, \gamma_-) \quad \forall\, t \in [t_* - \varepsilon_t, t_*),$$

$$\big(\,\alpha(t),\,\beta(t),\gamma(t)\,\big) = (\alpha_+, \beta_+, \gamma_+) \quad \forall\, t \in (t_*, t_* + \varepsilon_t],$$

where the index triple (α, β, γ) is defined in (3.3) for the associated LCP. For notational convenience, we denote each complementary cone in (2.17) by $\mathcal{X}_\delta = \{x \in \mathbb{R}^n \mid C_\delta\, x \geqslant 0\}$, where

$$C_\delta \equiv \begin{bmatrix} -D_{\delta\delta}^{-1} & 0 \\ -D_{\bar\delta\delta}(D_{\delta\delta})^{-1} & I_{\bar\delta\bar\delta} \end{bmatrix} \begin{bmatrix} C_{\delta\bullet} \\ C_{\bar\delta\bullet} \end{bmatrix}$$

and δ is a subset of $\{1, \ldots, m\}$. By the uniqueness of the solution of the LCS, it is clear that for all $t \in [t_* - \varepsilon_t, t_*)$, $x(t, x^0)$ is only in the cones $\mathcal{X}_\delta$'s with $\delta = \alpha_- \cup \beta_-^1$ and $\bar\delta = \gamma_- \cup \big(\beta_- \setminus \beta_-^1\big)$, where β_-^1 is a subset of β_-. Hence,

$$\mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0)) \;=\; \{\, \delta \;=\; \alpha_- \cup \beta_-^1 \;\mid\; \beta_-^1 \;\subseteq\; \beta_- \,\}. \tag{3.4}$$

Similarly, we have

$$\mathcal{J}(x(t_*, x^0)) \;=\; \{\, \delta \;=\; \alpha_+ \cup \beta_+^1 \;\mid\; \beta_+^1 \;\subseteq\; \beta_+ \,\}. \tag{3.5}$$

Let $\delta_i \in \mathcal{J}(x(t_*, x^0)) \cap \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$ with $\mathcal{X}_{\delta_i} = \{x \in \mathbb{R}^n \mid C_{\delta_i}\, x \geqslant 0\}$ and $\dot x = A_{\delta_i} x$ being the corresponding cone and dynamics, respectively. By the time-continuity of the state trajectory, it is easy to verify that there exists an $\varepsilon' > 0$ such that $x(t, x^0) \in \mathcal{X}_{\delta_i}$ for all $t \in [\,t_* - \varepsilon',\, t_* + \varepsilon'\,]$. Letting $\xi^* = x(t_*, x^0)$, we have

$$\big(\,(C_{\delta_i}\xi^*)_{\alpha_-},\, (C_{\delta_i}(-A_{\delta_i})\xi^*)_{\alpha_-}, \ldots, (C_{\delta_i}(-A_{\delta_i})^{n-1}\xi^*)_{\alpha_-}\,\big) \succ 0,$$

$$\big(\,(C_{\delta_i}\xi^*)_{\beta_-},\, (C_{\delta_i}(-A_{\delta_i})\xi^*)_{\beta_-}, \ldots, (C_{\delta_i}(-A_{\delta_i})^{n-1}\xi^*)_{\beta_-}\,\big) = 0,$$

$$\big(\,(C_{\delta_i}\xi^*)_{\gamma_-},\, (C_{\delta_i}(-A_{\delta_i})\xi^*)_{\gamma_-}, \ldots, (C_{\delta_i}(-A_{\delta_i})^{n-1}\xi^*)_{\gamma_-}\,\big) \succ 0,$$

which further implies that

$$\big(\,(C_{\delta_i}\xi^*)_{\alpha_-},\, (C_{\delta_i}(A_{\delta_i})\xi^*)_{\alpha_-}, \ldots, (C_{\delta_i}(A_{\delta_i})^{n-1}\xi^*)_{\alpha_-}\,\big) \neq 0,$$
$$\big(\,(C_{\delta_i}\xi^*)_{\beta_-},\, (C_{\delta_i}(A_{\delta_i})\xi^*)_{\beta_-}, \ldots, (C_{\delta_i}(A_{\delta_i})^{n-1}\xi^*)_{\beta_-}\,\big) = 0,$$
$$\big(\,(C_{\delta_i}\xi^*)_{\gamma_-},\, (C_{\delta_i}(A_{\delta_i})\xi^*)_{\gamma_-}, \ldots, (C_{\delta_i}(A_{\delta_i})^{n-1}\xi^*)_{\gamma_-}\,\big) \neq 0.$$

The second equality shows that $\beta_+ \supseteq \beta_-$, and the first and third inequalities show that $\beta_+ \subseteq \beta_-$. Hence $\beta_+ = \beta_-$. Moreover, we deduce from $\delta_i \in \mathcal{J}(x(t_*, x^0))$ that

$$\big(\,(C_{\delta_i}\xi^*)_{\alpha_-},\, (C_{\delta_i}(A_{\delta_i})\xi^*)_{\alpha_-}, \ldots, (C_{\delta_i}(A_{\delta_i})^{n-1}\xi^*)_{\alpha_-}\,\big) \succ 0,$$

$$\big(\,(C_{\delta_i}\xi^*)_{\gamma_-},\, (C_{\delta_i}(A_{\delta_i})\xi^*)_{\gamma_-}, \ldots, (C_{\delta_i}(A_{\delta_i})^{n-1}\xi^*)_{\gamma_-}\,\big) \succ 0.$$

Thus, by the uniqueness of the solution pair $\big(x(t), z(t)\big)$ at each t, we deduce that there is an $\varepsilon_+ > 0$ such that $z_{\alpha_-}(t) > 0$ and $(Cx(t) + Dz(t))_{\gamma_-} > 0$ for all $t \in (t_*, t_* + \varepsilon_+]$. This suggests $\alpha_- \subseteq \alpha_+$ and $\gamma_- \subseteq \gamma_+$. Since $z_{\beta_-}(t) = z_{\beta_+}(t) = \big(Cx(t)+Dz(t)\big)_{\beta_-} = \big(Cx(t)+Dz(t)\big)_{\beta_+} = 0$ for all $t \in [t_*, t_*+\varepsilon_+]$, we must have $\alpha_+ \equiv \alpha_-$ and $\gamma_+ \equiv \gamma_-$. By (3.4) and (3.5), we conclude that $\mathcal{J}(x(t_*, x^0)) = \mathcal{J}^{\,\mathrm{r}}(x(t_*, x^0))$. The second statement thus easily follows from Proposition 3.11. □

4. Observability of CLSs. In this section, we treat another fundamental property of the CLS, namely, observability with respect to a linear output. In the recent paper [24], we have treated this property rather extensively for the LCS (2.15); the treatment herein extends the previous analysis in several major ways. One, we deal with a general conewise linear system; two, Theorem 4.5 when specialized to the LCS closes a gap that was unresolved in [24]; three, we also treat other observability notions in detail. To be fair to [24], the approach used there is based on a general result for a nonlinear ODE with a nondifferentiable right-hand side and is applicable to nonlinear systems such as the nonlinear complementarity system; in contrast, the approach used below takes full advantage of the (piecewise) linear structure of the CLS. Most importantly, the notion of lexicographic ordering that has played a fundamental role in [24] remains the key to the present extended treatment.

Throughout the rest of this paper, let $H \in \mathbb{R}^{r\times n}$ be a given matrix that induces the linear output $Hx(t,\xi)$ associated with the solution trajectory $x(t,\xi)$ of (2.5). With respect to this matrix H, we formally introduce the observability concepts (see Definitions 4.2 and 4.3) to be analyzed subsequently, all of which are based on the following indistinguishability definition, which is classical in systems theory.

DEFINITION 4.1. *We say that a pair of states* $(\xi, \eta) \in \mathbb{R}^{n+n}$ *is*

- short-time indistinguishable *if* $\varepsilon > 0$ *exists such that* $Hx(t,\xi) = Hx(t,\eta)$ *for all* $t \in [0, \varepsilon]$*;*
- T-time indistinguishable *for a given* $T > 0$ *if* $Hx(t,\xi) = Hx(t,\eta)$ *for all* $t \in [0, T]$*;*
- long-time indistinguishable *if* $Hx(t,\xi) = Hx(t,\eta)$ *for all* $t \geqslant 0$.

Clearly, long-time indistinguishability $\Rightarrow$ T-time indistinguishability $\Rightarrow$ short-time indistinguishability for any pair of states.

DEFINITION 4.2. *We say that a state* $\xi \in \mathbb{R}^n$ *is*

- short-time locally observable *if there exists a neighborhood* $\mathcal{N}$ *of* ξ *such that no pair* (ξ, η) *with* $\eta \in \mathcal{N} \setminus \{\xi\}$ *is short-time indistinguishable;*
- short-time globally observable *if there exists no state* $\eta \neq \xi$ *such that the pair* (ξ, η) *is short-time indistinguishable;*
- T-time locally observable *for a given* $T > 0$ *if there exists a neighborhood* $\mathcal{N}$ *of* ξ *such that no pair* (ξ, η) *with* $\eta \in \mathcal{N} \setminus \{\xi\}$ *is* T*-time indistinguishable;*
- T-time globally observable *for a given* $T > 0$ *if there exists no state* $\eta \neq \xi$ *such that the pair* (ξ, η) *is* T*-time indistinguishable;*
- long-time locally observable *if there exists a neighborhood* $\mathcal{N}$ *of* ξ *such that no pair* (ξ, η) *with* $\eta \in \mathcal{N} \setminus \{\xi\}$ *is long-time indistinguishable;*
- long-time globally observable *if there exists no state* $\eta \neq \xi$ *such that the pair* (ξ, η) *is long-time indistinguishable.*

Clearly, the following implications hold for any state $\xi \in \mathbb{R}^n$:

$$\begin{array}{ccc}
\text{short-time global observability} & \Rightarrow & \text{short-time local observability} \\
\Downarrow & & \Downarrow \\
T\text{-time global observability} & \Rightarrow & T\text{-time local observability} \\
\Downarrow & & \Downarrow \\
\text{long-time global observability} & \Rightarrow & \text{long-time local observability.}
\end{array}$$

The above definitions pertain to individual states. At the system level, we have the following concepts. For simplicity, we define only the short-time version of the concepts.

DEFINITION 4.3. *The CLS* (2.5) *is said to be*

- short-time locally observable *if all states are short-time locally observable.*
- short-time globally observable *if all states are short-time globally observable.*

Recall that for a given pair of matrices $M \in \mathbb{R}^{n\times n}$ and $N \in \mathbb{R}^{m\times n}$, the unobservable space of (N, M), denoted $\overline{O}(N, M)$, is the set of vectors $\xi \in \mathbb{R}^n$ such that $NM^j\xi = 0$ for all $j = 0, 1, 2, \ldots$. By the well-known Cayley–Hamilton theorem in linear algebra, it follows that $\xi \in \overline{O}(N, M)$ if and only if $NM^j\xi = 0$ for all $j = 0, 1, 2, \ldots, n-1$. Elements of the space $\overline{O}(N, M)$ are said to be *unobservable* with respect to the pair (N, M). If $\overline{O}(N, M)$ consists only of the zero vector, then (N, M) is called an *observable pair*. We remark that a vector $\xi \in \overline{O}(H, A)$ if and only if for every $t_* \geqslant 0$, a scalar $\varepsilon_* > 0$ exists such that $He^{A(t-t_*)}x(t_*, \xi) = 0$ for all $t \in [t_*, t_* + \varepsilon_*]$. More generally, for any two vectors u and v, any scalar $\varepsilon > 0$, and two square matrices A and B,

$$\left\{He^{A(t-t_*)}u = He^{B(t-t_*)}v \ \forall t \in [t_*, t_* + \varepsilon]\right\} \Leftrightarrow \left[HA^k u = HB^k v \ \forall k = 0, \ldots, n-1\right].$$

This follows easily by differentiating the expression involving the exponential functions and then substituting $t = t_*$. This equivalence allows one to check the left-hand condition, which involves a continuum of times t, by a finite set of linear equations.

4.1. Short-time observability. We begin our investigation of various observability properties with the discussion of state short-time observability.

THEOREM 4.4. *A state ξ is short-time globally observable for the CLS* (2.5) *if and only if there exists no triple (η, i, j) satisfying $\eta \neq \xi$, $i \in \mathcal{J}(\eta)$, $j \in \mathcal{J}(\xi)$, and*

$$\begin{pmatrix} \eta \\ \xi \end{pmatrix} \in \overline{O}\left(\begin{bmatrix} H & -H \end{bmatrix}, \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix} \right). \tag{4.1}$$

Proof. Suppose one such triple (η, i, j) exists. From Lemma 2.4, we know that there exists a positive number ε such that $x(t, \eta)$ and $x(t, \xi)$ lie in the cones $\mathcal{X}_i$ and $\mathcal{X}_j$ for all $t \in [0, \varepsilon]$, respectively. Therefore, $x(t, \eta) = \exp(A_i t)\eta$ and $x(t, \xi) = \exp(A_j t)\xi$ on the same interval. This, together with (4.1), implies that $Hx(t, \xi) = Hx(t, \eta)$ for all $t \in [0, \varepsilon]$. Hence, the pair (ξ, η) is short-time indistinguishable. Consequently, ξ is not short-time globally observable. Conversely, suppose that ξ is short-time globally unobservable. There must exist a state $\eta \neq \xi$ such that the pair (ξ, η) is short-time indistinguishable, i.e., $Hx(t, \xi) = Hx(t, \eta)$ on an interval $[0, \varepsilon']$ for some $\varepsilon' > 0$. Let $\eta \in \mathcal{Y}_i$ and $\xi \in \mathcal{Y}_j$ for some $i, j \in \{1, \ldots, m\}$. By Lemma 2.4, there exists a positive number ε such that $H \exp(A_i t)\eta = H \exp(A_j t)\xi$ on the interval $[0, \varepsilon]$. Taking derivatives and evaluating at $t = 0$ show that the membership (4.1) holds. This leads to a contradiction. ∎

Toward the characterization of short-time local observability, we define, for each $i = 1, \ldots, m$ and each subset $\mathcal{L}$ of $\{1, \ldots, m_i\}$,

$$\mathcal{Y}_{i,\mathcal{L}} \equiv \left\{ x \,\middle|\, \left((C_i x)_\ell, (C_i A_i x)_\ell, \ldots, (C_i A_i^{n-1} x)_\ell\right) \succcurlyeq 0, \ \forall \ell \in \mathcal{L} \right\} \supseteq \mathcal{Y}_i.$$

The equality $\mathcal{Y}_{i,\mathcal{L}} = \mathcal{Y}_i$ holds when $\mathcal{L} = \{1, \ldots, m_i\}$; by convention, we let $\mathcal{Y}_{i,\varnothing} = \mathbb{R}^n$. To obtain a local version of Theorem 4.4, we need to define several index sets

associated with a given state $\xi \in \mathbb{R}^n$. The first one is

$$\mathcal{K}(\xi) \equiv \left\{ i \in \mathcal{I}(\xi) \,\middle|\, \begin{pmatrix} \xi \\ \xi \end{pmatrix} \in \overline{O}\left(\begin{bmatrix} H & -H \end{bmatrix}, \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix} \right) \text{ for some } j \in \mathcal{J}(\xi) \right\}$$
$$= \left\{ i \in \mathcal{I}(\xi) \,\middle|\, \exists\, j \in \mathcal{J}(\xi) \text{ such that } HA_i^k\xi = HA_j^k\xi \;\forall k \geqslant 0 \right\}.$$

By part (d) of Lemma 2.5 and the above definition, the following inclusions are clear:

$$\mathcal{J}(\xi) \subseteq \mathcal{K}(\xi) \subseteq \mathcal{I}(\xi) \quad \forall \xi \in \mathbb{R}^n.$$

As it turns out (see Theorem 4.5), the pieces $\mathcal{X}_i$ for $i \notin \mathcal{K}(\xi)$ play no role in the short-time local observability of ξ. Indeed, the set $\mathcal{K}(\xi)$ is the key to a complete characterization of the short-time local observability of ξ; this set was not discovered in [24] for the LCS.

For each $i \in \mathcal{I}(\xi)$, define the index set $\mathcal{I}_{i0}(\xi) \equiv \{\ell \mid (C_i\xi)_\ell = 0\} \subseteq \{1, \ldots, m_i\}$. Note that if $\mathcal{I}_{i0}(\xi) = \varnothing$ (a case which we call *nondegenerate*), we must have $\xi \in \operatorname{int} \mathcal{X}_i$ by (2.7), which implies $\mathcal{I}(\xi) = \mathcal{J}(\xi) = \mathcal{K}(\xi) = \{i\}$. If $\mathcal{J}(\xi)$ is a proper subset of $\mathcal{K}(\xi)$, we define, for each $i \in \mathcal{K}(\xi) \setminus \mathcal{J}(\xi)$,

$$\vartheta_i(\xi) \equiv \{\, \ell \;\big|\; \big((C_i\xi)_\ell,\ (C_iA_i\xi)_\ell, \ldots, (C_iA_i^{n-1}\xi)_\ell\big) \prec 0 \,\},$$

which must be nonempty. Since $\mathcal{K}(\xi) \subseteq \mathcal{I}(\xi)$, we have $C_i\xi \geqslant 0$ for all $i \in \mathcal{K}(\xi) \setminus \mathcal{J}(\xi)$, which implies $(C_i\xi)_{\vartheta_i(\xi)} = 0$; thus $\vartheta_i(\xi) \subseteq \mathcal{I}_{i0}(\xi)$. For each $\ell \in \vartheta_i(\xi)$, we let μ_ℓ^i be the first nonnegative integer k such that $(C_iA_i^k\xi)_\ell < 0$. We must have $1 \leq \mu_\ell^i \leq n-1$ for all $\ell \in \vartheta_i(\xi)$. Let $\overline{\vartheta}_i(\xi)$ be the complement of $\vartheta_i(\xi)$ in $\{1, \ldots, m_i\}$. Clearly,

$$\xi \in \mathcal{Y}_{i,\overline{\vartheta}_i(\xi)} = \left\{ x \,\middle|\, \big((C_ix)_\ell,\ (C_iA_ix)_\ell, \ldots, (C_iA_i^{n-1}x)_\ell\big) \succcurlyeq 0 \quad \forall\, \ell \in \overline{\vartheta}_i(\xi) \right\}.$$

Finally, we define

$$\widetilde{\mathcal{Y}}_{i,\vartheta_i(\xi)} \equiv \left\{ x \,\middle|\, \big((C_ix)_\ell,\ (C_iA_ix)_\ell, \ldots, (C_iA_i^{\mu_\ell^i - 1}x)_\ell\big) \succ 0 \quad \forall\, \ell \in \vartheta_i(\xi) \right\}$$

for each $i \in \mathcal{K}(\xi) \setminus \mathcal{J}(\xi)$. Note that $0 \notin \widetilde{\mathcal{Y}}_{i,\vartheta_i(\xi)}$. Moreover, it is easy to see that the following implication holds:

$$\eta \in \widetilde{\mathcal{Y}}_{i,\vartheta_i(\xi)} \;\Leftrightarrow\; \xi + \tau\eta \in \mathcal{Y}_{i,\vartheta_i(\xi)} \quad \forall\, \tau > 0 \text{ sufficiently small.} \tag{4.2}$$

With the above preparation, we are ready to establish a necessary and sufficient condition for a given state of the CLS (2.5) to be short-time locally observable.

THEOREM 4.5. *A state ξ is short-time locally observable for the CLS* (2.5) *if and only if*

$$\begin{aligned} \overline{O}(H, A_i) \cap (\, \mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)} - \xi \,) &= \{\, 0 \,\} \quad \forall\, i \in \mathcal{J}(\xi), \\ \overline{O}(H, A_i) \cap \widetilde{\mathcal{Y}}_{i,\vartheta_i(\xi)} \cap (\, \mathcal{Y}_{i,\overline{\vartheta}_i(\xi)} - \xi \,) &= \varnothing \quad \forall\, i \in \mathcal{K}(\xi) \setminus \mathcal{J}(\xi). \end{aligned} \tag{4.3}$$

Proof. "Sufficiency." Suppose that the state ξ is not short-time locally observable. Since $\cup_{i=1}^m \mathcal{Y}_i = \mathbb{R}^n$, by Lemma 2.5(a), there exist an index $i \in \mathcal{I}(\xi)$ and a sequence $\{\xi^\nu\}$ converging to ξ such that $\xi \neq \xi^\nu \in \mathcal{Y}_i$ and the pair (ξ, ξ^ν) is short-time indistinguishable for all ν. We claim that, for all ν sufficiently large, the nonzero vector $\eta^\nu \equiv \xi^\nu - \xi$ violates one of the two conditions in (4.3). Let $j \in \mathcal{J}(\xi)$ such that $\xi \in \mathcal{Y}_j$. By the proof of Theorem 4.4, we deduce that, for all ν,

$$\begin{pmatrix} \xi^\nu \\ \xi \end{pmatrix} \in \overline{O}\left(\begin{bmatrix} H & -H \end{bmatrix}, \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix} \right).$$

By taking the limit $\nu \to \infty$, we get

$$\begin{pmatrix} \xi \\ \xi \end{pmatrix} \in \overline{O}\left(\begin{bmatrix} H & -H \end{bmatrix}, \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix} \right).$$

Thus, i must belong to $\mathcal{K}(\xi)$. This implies that, for all nonnegative integers k, $HA_i^k\xi^\nu = HA_j^k\xi = HA_i^k\xi$; thus η^ν belongs to $\overline{O}(H, A_i)$. Since $\xi \neq \xi^\nu \in \mathcal{Y}_i \subseteq \mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)}$, we see that the first condition in (4.3) is violated if $i \in \mathcal{J}(\xi)$. Now suppose i belongs to $\mathcal{K}(\xi)\setminus\mathcal{J}(\xi)$. To see that this contradicts the second condition in (4.3), it remains to verify that $\eta^\nu \in \widetilde{\mathcal{Y}}_{i,\vartheta_i(\xi)}$ and $\xi^\nu \in \mathcal{Y}_{i,\overline{\vartheta}_i(\xi)}$. The latter membership is obvious because $\xi^\nu \in \mathcal{Y}_i$ for all ν. To prove the former membership, suppose that an index $\bar{\ell} \in \vartheta_i(\xi)$ exists satisfying $((C_i\eta^\nu)_{\bar{\ell}}, (C_iA_i\eta^\nu)_{\bar{\ell}}, \ldots, (C_iA_i^{\mu^i_{\bar{\ell}}-1}\eta^\nu)_{\bar{\ell}}) \preceq 0$, where $\mu^i_{\bar{\ell}}$ is the first nonnegative integer k such that $(C_iA_i^k\xi)_{\bar{\ell}} < 0$ defined before. Since $((C_i\xi)_{\bar{\ell}}, (C_iA_i\xi)_{\bar{\ell}}, \ldots, (C_iA_i^{\mu^i_{\bar{\ell}}-1}\xi)_{\bar{\ell}}) = 0$, we deduce that $((C_i\xi^\nu)_{\bar{\ell}}, (C_iA_i\xi^\nu)_{\bar{\ell}}, \ldots, (C_iA_i^{\mu^i_{\bar{\ell}}-1}\xi^\nu)_{\bar{\ell}}) \preceq 0$. Hence $(C_iA_i^k\xi^\nu)_{\bar{\ell}} = 0$ for all $k = 0, 1, \ldots, \mu^i_{\bar{\ell}} - 1$. But since $(C_iA_i^{\mu^i_{\bar{\ell}}}\xi)_{\bar{\ell}} < 0$, which implies $(C_iA_i^{\mu^i_{\bar{\ell}}}\xi^\nu)_{\bar{\ell}} < 0$ for all ν sufficiently large, it follows that $((C_i\xi^\nu)_{\bar{\ell}}, (C_iA_i\xi^\nu)_{\bar{\ell}}, \ldots, (C_iA_i^{\mu^i_{\bar{\ell}}}\xi^\nu)_{\bar{\ell}}) \prec 0$, which contradicts $\xi^\nu \in \mathcal{Y}_i$.

"Necessity." We show in what follows that the violation of either one of the two conditions in (4.3) leads to a contradiction to short-time local observability of ξ. Suppose that there exist an index $i \in \mathcal{J}(\xi)$ and a nonzero vector $\eta \in \overline{O}(H, A_i) \cap (\mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)} - \xi)$. We have $\eta + \xi \in \mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)}$. Since $\xi \in \mathcal{Y}_i \subseteq \mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)}$ and $\mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)}$ is convex, it follows that $\xi + \tau\eta \in \mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)}$ for all $\tau \in [0, 1]$. Thus, for each such τ, there exists $\varepsilon_\tau > 0$ such that $(C_ie^{A_it}[\xi + \tau\eta])_{\mathcal{I}_{i0}(\xi)} \geq 0$ for all $t \in [0, \varepsilon_\tau]$. Since $(C_i\xi)_j > 0$ for all $j \in \{1, \ldots, m_i\}\setminus\mathcal{I}_{i0}(\xi)$, it follows that $(C_ie^{A_it}[\xi + \tau\eta])_j \geq 0$ for all such j and all $(t, \tau) > 0$ sufficiently small. Consequently, for every $\tau > 0$ sufficiently small, there exists $\varepsilon_\tau > 0$ such that $x(t, \xi+\tau\eta) = e^{A_it}[\xi+\tau\eta]$ for all $t \in [0, \varepsilon_\tau]$. Since $\eta \in \overline{O}(H, A_i)$, we deduce $Hx(t, \xi + \tau\eta) = Hx(t, \xi)$ for all such pairs (τ, t). Hence the pair $(\xi + \tau\eta, \xi)$ is short-time indistinguishable, contradicting the short-time local observability of ξ.

Next, suppose that a nonzero vector $\eta \in \overline{O}(H, A_i) \cap \widetilde{\mathcal{Y}}_{i,\vartheta_i(\xi)} \cap (\mathcal{Y}_{i,\overline{\vartheta}_i(\xi)} - \xi)$ exists for some $i \in \mathcal{K}(\xi) \setminus \mathcal{J}(\xi)$. By (4.2), we deduce that $\xi + \tau\eta \in \mathcal{Y}_{i,\vartheta_i(\xi)}$ for all $\tau > 0$ sufficiently small. Moreover, since $\xi + \eta \in \mathcal{Y}_{i,\overline{\vartheta}_i(\xi)}$ and $\xi \in \mathcal{Y}_{i,\overline{\vartheta}_i(\xi)}$, it follows that $\xi + \tau\eta \in \mathcal{Y}_{i,\overline{\vartheta}_i(\xi)}$ for all $\tau \in [0, 1]$. Consequently, $\xi + \tau\eta \in \mathcal{Y}_i$ for all $\tau > 0$ sufficiently small. We can now apply the same argument as before to deduce a contradiction to the short-time local observability of ξ. □

The next result is an immediate corollary of Theorem 4.5 that pertains to a nondegenerate state $\xi \in \operatorname{int}\mathcal{X}_i$ for some i. No proof is needed.

COROLLARY 4.6. *A nondegenerate state $\xi \in \operatorname{int}\mathcal{X}_i$ is short-time locally observable for the CLS* (2.5) *if and only if $\overline{O}(H, A_i) = \{0\}$.*

We apply Theorem 4.5 to the bimodal CLS (2.9) with the two pieces $\mathcal{X}_1 = \{x | c^Tx \geqslant 0\}$ and $\mathcal{X}_2 = \{x | c^Tx \leqslant 0\}$, and the two matrices $A_1 = A + bc^T$ and $A_2 = A$. Let $\xi \in \mathbb{R}^n$ be arbitrary. The cases where $c^T\xi > 0$ and $c^T\xi < 0$ are covered by Corollary 4.6. We focus on the case where $c^T\xi = 0$. In this case, we have $\mathcal{I}(\xi) = \{1, 2\}$, $\mathcal{I}_{10}(\xi) = \{1\}$, $\mathcal{I}_{20}(\xi) = \{2\}$,

$$\mathcal{Y}_{1,\mathcal{I}_{10}(\xi)} = \left\{ x \,|\, (\, c^Tx, c^T(A + bc^T)x, \ldots, c^T(A + bc^T)^{n-1}x \,) \succcurlyeq 0 \right\} = \mathcal{Y}_1$$

and

$$\mathcal{Y}_{2,\mathcal{I}_{20}(\xi)} = \{x | (c^Tx, c^TAx, \ldots, c^TA^{n-1}x) \preceq 0\} = \mathcal{Y}_2.$$

The tuple $Y(\xi) \equiv (c^T\xi, c^TA\xi, \ldots, c^TA^{n-1}\xi)$ plays a central role in the following corollary of Theorem 4.5.

COROLLARY 4.7. *The following statements hold for the bimodal CLS* (2.9).

(a) *If* $Y(\xi) \succ 0$, *then* ξ *is short-time locally observable if and only if* $\overline{O}(H, A + bc^T) = \{0\}$.

(b) *If* $Y(\xi) \prec 0$, *then* ξ *is short-time locally observable if and only if* $\overline{O}(H, A) = \{0\}$.

(c) *If* $Y(\xi) = 0$, *then* ξ *is short-time locally observable if and only if* $\overline{O}(H, A + bc^T) = \{0\}$ *and* $\overline{O}(H, A) = \{0\}$.

Proof. It suffices to consider the case where $c^T\xi = 0$. We prove (a) only as the proofs of (b) and (c) are similar. Suppose $Y(\xi) \succ 0$, which implies $\mathcal{J}(\xi) = \{1\}$ and $\xi \notin \overline{O}(c^T, A)$. It can be verified by the definition of $\mathcal{Y}_{1,\mathcal{I}_{10}(\xi)}$ that, for any $\eta \in \mathbb{R}^n$, either $\xi + \tau\eta \in \mathcal{Y}_{1,\mathcal{I}_{10}(\xi)}$ for all $\tau > 0$ sufficiently small or $\xi - \tau\eta \in \mathcal{Y}_{1,\mathcal{I}_{10}(\xi)}$ for all $\tau > 0$ sufficiently small. Hence, for any linear subspace S of $\mathbb{R}^n$, $S \cap (\mathcal{Y}_{1,\mathcal{I}_{10}(\xi)} - \xi) = \{0\}$ if and only if $S = \{0\}$. From this observation, it follows that the first condition in (4.3) is equivalent to

$$\overline{O}(H, A + bc^T) \cap (\mathcal{Y}_{1,\mathcal{I}_{10}(\xi)} - \xi) = \{0\} \Leftrightarrow \overline{O}(H, A + bc^T) = \{0\}. \tag{4.4}$$

There are two subcases to consider: (i) $b \notin \overline{O}(H, A)$ and (ii) $b \in \overline{O}(H, A)$. In subcase (i), it is easy to show, using $\xi \notin \overline{O}(c^T, A)$, that there must exist a positive integer k such that $H(A + bc^T)^k\xi \neq HA^k\xi$. This means that $\mathcal{K}(\xi) = \mathcal{J}(\xi) = \{1\}$. In subcase (ii), we have $H(A + bc^T)^k v = HA^k v$ for all nonnegative integers k and all $v \in \mathbb{R}^n$; thus $\overline{O}(H, A + bc^T) = \overline{O}(H, A)$, which further implies $\mathcal{K}(\xi) = \{1, 2\}$ and $\overline{\vartheta}_2(\xi) = \varnothing$. In both subcases, assertion (a) follows readily from Theorem 4.5 using (4.4), and $\overline{O}(H, A) \cap \widetilde{\mathcal{Y}}_{2,\vartheta_2(\xi)} = \overline{O}(H, A) \setminus \{0\}$. ∎

Corollary 4.7 recovers Proposition 19 in [24], which was obtained by specializing a theory for nonsmooth systems that in turn was based on a differential approach. The purpose of including the above proof of Corollary 4.7 is to illustrate the application of Theorem 4.5 in the case of a bimodal CLS. The corollary also identifies the key vector $Y(\xi)$ that was not explicitly employed in [24]. It follows from this corollary that if both $\overline{O}(H, A) \neq \{0\}$ and $\overline{O}(H, A + bc^T) \neq \{0\}$, then the bimodal CLS (2.9) has no short-time locally observable state; see Theorem 4.9 below for a general result.

We can employ the state short-time local/global observability characterizations to deduce some corresponding system short-time local/global observability results. The first such result pertains to short-time global observability and requires no proof.

THEOREM 4.8. *The CLS* (2.5) *is short-time globally observable if and only if*

$$\overline{O}\left(\begin{bmatrix} H & -H \end{bmatrix}, \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix} \right) \bigcap (\mathcal{Y}_i \times \mathcal{Y}_j) \subseteq \{(\xi, \xi) \,|\, \xi \in \mathbb{R}^n\} \tag{4.5}$$

for all i *and* j.

It turns out that the characterization of system short-time local observability is quite simple, involving only linear subspace conditions that are easily verifiable.

THEOREM 4.9. *The CLS* (2.5) *is short-time locally observable if and only if*

$$\overline{O}(H, A_i) = \{0\}, \quad i \in \{1, \ldots, m\}. \tag{4.6}$$

Proof. The sufficiency is clear. For the necessity, let ξ be an arbitrary interior point of the cone $\mathcal{X}_i$ for each $i = 1, \ldots, m$. For such a vector, we have $\mathcal{J}(\xi) = \mathcal{I}(\xi) = \mathcal{K}(\xi) = \{i\}$ and $\mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)} = \mathbb{R}^n$. The necessity of (4.6) now follows readily. ∎

4.2. Finite verification. The characterizations of short-time observation beg the question of whether the necessary and sufficient conditions in Theorems 4.4, 4.5, and 4.8 can be verified by a finite procedure (it is obvious for Theorem 4.9). Note that we are not concerned about the computational complexity of the procedure, knowing that any such procedure is very likely to be exponential in the case of the LCS. We begin with the first condition in (4.3).

For each $i \in \mathcal{J}(\xi)$ and each $\ell \in \mathcal{I}_{i0}(\xi)$ for a given $\xi \in \mathbb{R}^n$, let μ_ℓ^i be the observability degree of the pair $((C_i)_{\ell\bullet}, A_i)$ at ξ (i.e., μ_ℓ^i is the first positive integer k such that $(C_i)_{\ell\bullet}A_i^k\xi > 0$); we set $\mu_\ell^i = n$ if $(C_i)_{\ell\bullet}A_i^k\xi = 0$ for all k. We claim that for each $i \in \mathcal{J}(\xi)$,

$$\overline{O}(H, A_i) \cap (\mathcal{Y}_{i,\mathcal{I}_{i0}(\xi)} - \xi) = \{0\} \Leftrightarrow \overline{O}(H, A_i) \bigcap \left(\bigcap_{\ell \in \mathcal{I}_{i0}(\xi)} \mathcal{Y}^\ell_{i,\mathcal{I}_{i0}(\xi)} \right) = \{0\}, \tag{4.7}$$

where $\mathcal{Y}^\ell_{i,\mathcal{L}} \equiv \left\{ v | ((C_i)_{\ell\bullet}v, (C_i)_{\ell\bullet}A_i v, \ldots, (C_i)_{\ell\bullet}A_i^{\mu_\ell^i - 1}v) \succcurlyeq 0 \right\}$ for any $\mathcal{L} \subseteq \{1, \ldots, m_i\}$ and each $\ell \in \mathcal{L}$. The claim (4.7) is a direct consequence of the first statement of the following lemma; the second statement of the lemma is used in the subsequent development (see Proposition 5.10).

LEMMA 4.10. *Let $\mathcal{L} \subseteq \{1, \ldots, m_i\}$, $\xi \in \mathcal{Y}_{i,\mathcal{L}}$, and $\ell \in \mathcal{L}$. It holds that*

(a) *$v \in \cap_{\ell \in \mathcal{L}} \mathcal{Y}^\ell_{i,\mathcal{L}} \Leftrightarrow \xi + \tau v \in \mathcal{Y}_{i,\mathcal{L}}$ for all $\tau > 0$ sufficiently small;*

(b) *for any $v \in \cap_{\ell \in \mathcal{L}} \mathcal{Y}^\ell_{i,\mathcal{L}}$, there exist $\tau_0 > 0$ and $\varepsilon_0 > 0$ (possibly dependent on τ_0) such that $\left(C_i e^{A_i t}[\xi + \tau v]\right)_{\mathcal{L}} \geqslant 0$ for all $(t, \tau) \in [0, \varepsilon_0] \times (0, \tau_0]$.*

Proof. The first statement is obvious, following from an argument similar to (4.2). We next show the second statement. Consider the case where the vector v is nonzero and $\mu_\ell^i < n$ for all ℓ; the other cases can easily be shown in a similar fashion. For each $\ell \in \mathcal{L}$, $v \in \bigcap_{\ell \in \mathcal{L}} \mathcal{Y}^\ell_{i,\mathcal{L}}$ implies that there exists an integer $0 \leqslant k \leqslant \mu_i^\ell - 1$ such that $(C_i A_i^k v)_\ell > 0$ and $(C_i A_i^j v)_\ell = 0$ for all $j = 0, \ldots, k-1$. Hence,

$$\left(C_i e^{A_i t}[\xi + \tau v]\right)_\ell = \tau \sum_{s=k}^{\mu_\ell^i - 1} \frac{(C_i A_i^s v)_\ell}{s!}\, t^s \;+\; \sum_{j=\mu_\ell^i}^{\infty} \frac{(C_i A_i^j [\xi + \tau v])_\ell}{j!}\, t^j.$$

The first summation is positive for all $\tau > 0$ and $t > 0$ sufficiently small, and since $(C_i A_i^{\mu_\ell^i}\xi)_\ell > 0$, it follows that for some positive ε_ℓ and τ_ℓ, the second summation

$$\sum_{j=\mu_\ell^i}^{\infty} \frac{(C_i A_i^j [\xi + \tau v])_\ell}{j!}\, t^j = t^{\mu_\ell^i} \left\{ \frac{(C_i A_i^{\mu_\ell^i}\xi)_\ell}{j!} + O(t) + \tau \left[\frac{(C_i A_i^{\mu_\ell^i} v)_\ell}{j!} + O(t) \right] \right\} \geqslant 0$$

for all $(t, \tau) \in [0, \varepsilon_\ell] \times (0, \tau_\ell]$. Hence, $\left(C_i e^{A_i t}[\xi + \tau v]\right)_\ell \geqslant 0$ for all $(t, \tau) \in [0, \varepsilon_\ell] \times (0, \tau_\ell]$. Finally, letting $\varepsilon_0 = \min_{\ell \in \mathcal{L}} \varepsilon_\ell$ and $\tau_0 = \min_{\ell \in \mathcal{L}} \tau_\ell$, we obtain the desired result. ∎

As explained in [24] in the context of the "semi-unobservable cones," checking the right-hand condition in (4.7) can be accomplished by solving finitely many linear programs; hence so can the first condition in (4.3). Indeed, a vector v belongs to $\mathcal{Y}^\ell_{i,\mathcal{I}_{i0}(\xi)}$ if and only if either $(C_i)_{\ell\bullet}A_i^k v = 0$ for all $k \in \{0, 1, \ldots, \mu_\ell^i - 1\}$ or there exists $k' \in \{0, 1, \ldots, \mu_\ell^i - 1\}$ such that $(C_i)_{\ell\bullet}A_i^k v = 0$ for all $k \in \{0, 1, \ldots, k' - 1\}$ and $(C_i)_{\ell\bullet}A_i^{k'} v > 0$. Hence, one can easily formulate finitely many linear inequality systems to determine if the right-hand condition in (4.7) holds. A similar procedure can be applied to check the conditions (4.1), (4.5), and the second condition in (4.3).

4.3. *T*-time and long-time observability. It has been pointed out at the beginning of section 4 that if a state $\xi = x(0,\xi)$ is short-time locally observable, then ξ is T-time locally observable for all $T > 0$, including $T = \infty$. The following result extends this observation to the case where a certain state $x(t,\xi)$ along the nominal solution trajectory is short-time locally observable for some $t \in (0,T)$.

PROPOSITION 4.11. *Let $0 < T \leqslant \infty$ and $\xi \in \mathbb{R}^n$ be given. If for some t_0 in $[0,T)$, the state $x(t_0,\xi)$ is short-time locally observable, then ξ is T-time locally observable.*

Proof. The short-time local observability of the state $x^0 \equiv x(t_0,\xi)$ means that there exists a neighborhood $\widetilde{\mathcal{N}}$ of x^0 such that for all $\widetilde{x} \in \widetilde{\mathcal{N}}$,

$$[\, Hx(t,x^0) = Hx(t,\widetilde{x}) \;\forall\, t \geqslant 0 \text{ sufficiently small}\,] \;\Rightarrow\; x^0 = \widetilde{x}.$$

Since the CLS is an ODE with a globally Lipschitz continuous right-hand side, there is a constant $L > 0$ such that $\|x(t,\xi) - x(t,\xi')\| \leqslant e^{Lt}\|\xi - \xi'\|$ for all $t > 0$. Hence, a neighborhood $\mathcal{N}_0$ of ξ exists such that $x(t_0,\xi') \in \widetilde{\mathcal{N}}$ for all $\xi' \in \mathcal{N}_0$. Let $\xi' \in \mathcal{N}_0$ be such that $Hx(t,\xi) = Hx(t,\xi')$ for all $t \in [0,T]$. Hence, we have, for all $\tau \in [0, T - t_0]$,

$$Hx(\tau,x^0) = Hx(\tau,x(t_0,\xi)) = Hx(t_0+\tau,\xi) = Hx(t_0+\tau,\xi') = Hx(\tau,x(t_0,\xi')).$$

Therefore, it follows that $x^0 = x(t_0,\xi) = x(t_0,\xi')$. Hence by considering the reverse-time system (2.8) starting at x^0 and noting that both ξ and ξ' are states on this reverse-time trajectory at time t_0, we easily obtain $\xi = \xi'$. ∎

Proposition 4.11 can be used to show, via the example below, that T-time local observability of a state does not imply short-time local observability of the given state.

Example 4.12. Consider the LCS(A,B,C,D) with the P-property where

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix},\; B = \begin{bmatrix} 0 & b_{12} \\ 0 & 0 \end{bmatrix},\; C = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},\; D = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix},\; H = \begin{bmatrix} h_1 & 0 \end{bmatrix}$$

with $\lambda_2 > \lambda_1 > 0$, and both b_{12} and h_1 nonzero. As shown in (2.17), the LCS is in the form of the CLS with four pieces:

$$\mathcal{X}_1 = \{x \in \mathbb{R}^2 \mid x_1 \geq 0,\, x_1 + x_2 \geq 0\}, \qquad \mathcal{X}_2 = \{x \in \mathbb{R}^2 \mid x_1 \leq 0,\, x_2 \geq 0\},$$

$$\mathcal{X}_3 = \{x \in \mathbb{R}^2 \mid 2x_1 + x_2 \leq 0,\, x_2 \leq 0\}, \qquad \mathcal{X}_4 = \{x \in \mathbb{R}^2 \mid 2x_1 + x_2 \geq 0,\, x_1 + x_2 \leq 0\},$$

which correspond to $\alpha = \varnothing$, $\alpha = \{1\}$, $\alpha = \{1,2\}$, $\alpha = \{2\}$ in (2.17), respectively, and their respective state matrices are

$$A_1 = A_2 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} \lambda_1 & -\frac{1}{2}b_{12} \\ 0 & \lambda_2 \end{bmatrix}, \quad A_4 = \begin{bmatrix} \lambda_1 & -b_{12} \\ 0 & \lambda_2 \end{bmatrix}.$$

Moreover, $\overline{O}(H,A_i) = \operatorname{span}\{(0,1)\}$ for $i = 1,2$ and $\overline{O}(H,A_i) = \{0\}$ for $i = 3,4$. Consider $\xi = (\xi_1,\xi_2)$ with $\xi_1 > 0 > \xi_2$ and $\xi_1 + \xi_2 > 0$. Since $\xi \in \operatorname{int}\mathcal{X}_1$ and $\overline{O}(H,A_1) \neq \{0\}$, Corollary 4.6 implies that ξ is not short-time locally observable. However, it can be seen that $t_1 > 0$ exists such that $x(t,\xi) \in \operatorname{int}\mathcal{X}_4$ for all $t > t_1$ sufficiently close to t_1. Hence, the condition $\overline{O}(H,A_4) = \{0\}$ implies the short-time local observability at $\widehat{x} = x(t_*,\xi)$ for some $t_* > t_1$. Consequently, ξ is T-time locally observable for any $T \in (t_1,\infty]$ by Proposition 4.11.

In light of Proposition 4.11, the challenge in establishing the T-time local observability of a given state $\xi \in \mathbb{R}^n$ occurs when none of the states $x(t,\xi)$ for $t \in [0,T)$ is short-time locally observable. In general, this is a rather difficult case to analyze fully,

due to the mode switchings along the nominal state trajectory $x(t,\xi)$ and perturbed state trajectories in the interval $[0,T]$. Our approach to dealing with this challenge is to invoke a result in [24] that pertains to an ODE with a B-differentiable right-hand side, which includes the CLS (2.5) as a special case. In what follows, after presenting a slight improvement of this result, we identify a class of initial states for which necessary and sufficient conditions for T-time local observability can be derived; see Proposition 4.13.

As mentioned in section 2, the solution map $\xi \mapsto x(t,\xi)$ is B-differentiable for all fixed $t \geqslant 0$; in particular, the directional derivative

$$x'_\xi(t,\xi;\eta) \equiv \lim_{\tau \downarrow 0} \frac{x(t,\xi+\tau\eta) - x(t,\xi)}{\tau}$$

of $x(t,\cdot)$ at $\xi \in \mathbb{R}^n$ along any direction $\eta \in \mathbb{R}^n$ exists and satisfies a certain first-order time-dependent variational ODE. In terms of such a derivative, define the set

$$\mathcal{Z}_T^\xi \equiv \left\{ v \in \mathbb{R}^n \,|\, Hx'_\xi(t,\xi;v) = 0 \;\forall\, t \in [0,T] \right\},$$

which is a closed, albeit not necessarily convex, cone. It was proved in [24, Theorem 10] that $\mathcal{Z}_T^\xi = \{0\}$ is a sufficient condition for the T-time local observability of ξ.

We next derive an improvement of the above result via the introduction of the set Ω_T^ξ consisting of all T-time indistinguishable states from ξ. Thus $\eta \in \Omega_T^\xi$ if and only if $Hx(t,\xi) = Hx(t,\eta)$ for all $t \in [0,T]$. Clearly, $\xi \in \Omega_T^\xi$. As such we can speak of the *feasible cone*, denoted $\mathcal{F}(\Omega_T^\xi,\xi)$, and the *tangent cone*, denoted $\mathcal{T}(\Omega_T^\xi,\xi)$, of Ω_T^ξ at ξ. Specifically, v is an element of the former cone if a $\bar\tau > 0$ exists such that $Hx(t,\xi+\tau v) = Hx(t,\xi)$ for all $(t,\tau) \in [0,T] \times [0,\bar\tau]$; w is an element of the latter cone if a sequence of vectors $\{\eta^k\} \subset \Omega_T^\xi$ converging to ξ and a sequence of positive scalars $\{\tau_k\}$ converging to zero exist such that $w = \lim_{k\to\infty} \frac{\eta^k - \xi}{\tau_k}$. We have

$$\mathcal{F}(\Omega_T^\xi,\xi) \subseteq \mathcal{T}(\Omega_T^\xi,\xi) \subseteq \mathcal{Z}_T^\xi. \tag{4.8}$$

Indeed, the first inclusion holds with Ω_T^ξ replaced by any set containing ξ; the second inclusion holds by the approximation

$$Hx(t,\eta) = Hx(t,\xi) + Hx'_\xi(t,\xi;\eta-\xi) + o_t(\|\,\eta-\xi\,\|), \tag{4.9}$$

where the error function $o_t(\tau)$ satisfies $\lim_{\tau\downarrow 0} \frac{o_t(\tau)}{\tau} = 0$. In turn, (4.9) is the consequence of the B-differentiability of the solution map $x(t,\cdot)$. We have the following result.

PROPOSITION 4.13. *Let $0 < T \leqslant \infty$ and $\xi \in \mathbb{R}^n$ be given. The following implications hold for the solution trajectory $x(t,\xi)$ of the CLS* (2.5)*:*

$$\mathcal{T}(\Omega_T^\xi,\xi) = \{0\} \;\Rightarrow\; \xi \text{ is } T\text{-time locally observable} \;\Rightarrow\; \mathcal{F}(\Omega_T^\xi,\xi) = \{0\}. \tag{4.10}$$

Hence, the following two statements are equivalent.

(a) $\mathcal{F}(\Omega_T^\xi,\xi) = \mathcal{T}(\Omega_T^\xi,\xi)$ *and* ξ *is T-time locally observable.*

(b) $\mathcal{F}(\Omega_T^\xi,\xi) = \mathcal{T}(\Omega_T^\xi,\xi) = \{0\}$.

Proof. Based on the same proof as in [24, Theorem 10], the first implication in (4.10) follows from (4.9). The equivalence of the two statements (a) and (b) is obvious. ∎

Admittedly, the condition $\mathcal{F}(\Omega_T^\xi, \xi) = \mathcal{T}(\Omega_T^\xi, \xi)$, while simple, is practically not easy to verify, due to the difficulty of complete characterization of the set Ω_T^ξ, except in special cases; see Theorem 5.6 and Corollary 5.7. Thus, rather than investigating this condition in its full generality, we devote section 5 to a detailed study of the bimodal CLS (2.9).

Before ending the discussion on the general CLS, we state the following sufficient condition for a state to be long-time locally observable.

COROLLARY 4.14. *Suppose that there exists* $T \in (0, \infty)$ *such that* $\mathcal{T}(\Omega_T^\xi, \xi) = \{0\}$*; then* ξ *is long-time locally observable for the CLS* (2.5).

Proof. This is obvious because $\Omega_\infty^\xi \subseteq \Omega_T^\xi$ for any finite $T > 0$. Thus the assumption implies $\mathcal{T}(\Omega_\infty^\xi, \xi) = \{0\}$ which holds because $\mathcal{T}(\Omega_\infty^\xi, \xi) = \{0\} \subseteq \mathcal{T}(\Omega_T^\xi, \xi) = \{0\}$. The long-time local observability of ξ now follows from Proposition 4.13. ∎

5. Bimodal CLSs. Currently, the results for T-time and long-time observability of a general CLS are limited to those in subsection 4.3. Further results, in particular, complete characterizations and finite verifications, appear difficult. Nevertheless, much more can be obtained for the bimodal CLS (2.9), whose detailed analysis is the subject of this section that is divided into several subsections. As we will see, even this simplified case is not easy to analyze, and some unsolved issues remain.

5.1. *T*-time local observability. For the analysis to be of interest, we make the blanket assumption throughout this section that $b \neq 0$ and do not repeat the assumption. We begin by giving two necessary conditions for the bimodal system (2.9) to have a T-time locally observable state for any $T \in [0, \infty]$. The first condition (a) is a minor variant of Proposition 13 in [24] specialized to the bimodal CLS. A proof of this part can be found in the reference.

PROPOSITION 5.1. *For the bimodal system* (2.9) *to have a* T*-time locally observable state, for any* $T \in [0, \infty]$*, it is necessary that* (a) $\overline{O}(H, A) \cap \overline{O}(c^T, A) = \{0\}$ *and* (b) $b \notin \overline{O}(H, A)$.

Proof. We prove only (b). Assume that $0 \neq b \in \overline{O}(H, A)$. By induction, it can be shown that $HA^k = H(A+bc^T)^k$ for all nonnegative k. (Indeed, assume that this holds for some i; we have $HA^{i+1} = H(A + bc^T)^i A = H(A + bc^T)^{i+1} - HA^i bc^T = H(A + bc^T)^{i+1}$, completing the inductive step.) Hence, $\overline{O}(H, A) = \overline{O}(H, A + bc^T)$. Thus $He^{At}\xi = He^{(A+bc^T)t}\xi$ for all $t \geqslant 0$ and all $\xi \in \mathbb{R}^n$, which implies $Hx(t, \xi) = He^{At}\xi$ for all $t \geqslant 0$ and all $\xi \in \mathbb{R}^n$. Since $0 \neq b \in \overline{O}(H, A) = \overline{O}(H, A + bc^T)$, we can similarly deduce $He^{At}\xi = He^{At}(\xi + \tau b) = He^{(A+bc^T)t}(\xi + \tau b)$ for all $t \geqslant 0$, all $\xi \in \mathbb{R}^n$, and all $\tau \in \mathbb{R}$. Consequently, $Hx(t, \xi) = Hx(t, \xi + \tau b)$ for all such triples (t, ξ, τ). Since $b \neq 0$, no state ξ can be T-time globally/locally observable. ∎

It is useful to record a corollary of the above proposition, which becomes yet another necessary condition for (2.9) to have a T-time locally observable state.

PROPOSITION 5.2. *Under conditions* (a) *and* (b) *of Proposition* 5.1, *it holds that* $\overline{O}(H, A) \cap \overline{O}(H, A + bc^T) = \{0\}$. *Conversely, if the latter holds, then* $b \notin \overline{O}(H, A)$.

Proof. Assume by way of contradiction that $0 \neq v \in \overline{O}(H, A) \cap \overline{O}(H, A + bc^T)$. Then $v \notin \overline{O}(c^T, A)$ by (a). Hence, a first nonnegative integer ℓ exists such that $c^T A^\ell v \neq 0$. Similarly, since $b \notin \overline{O}(H, A)$, we have the first nonnegative integer k satisfying $HA^k b \neq 0$. By expanding $(A + bc^T)^{k+\ell+1}$, it can easily be verified that $H(A + bc^T)^{k+\ell+1} v = (HA^k b)(c^T A^\ell v) \neq 0$ by the choice of k and ℓ. This contradicts $v \in \overline{O}(H, A + bc^T)$. The second assertion of the proposition can be easily argued as follows. If $b \in \overline{O}(H, A)$, then by the proof of Proposition 5.1, we have $0 \neq b \in \overline{O}(H, A) = \overline{O}(H, A) \cap \overline{O}(H, A + bc^T)$. ∎

We next establish two basic properties of a solution trajectory to the bimodal system (2.9). Stated in Proposition 5.3, the first property pertains to an individual trajectory $x(t,\xi)$; the second property pertains to perturbed trajectories $x(t,\eta)$ where the initial condition η is sufficiently close to ξ.

PROPOSITION 5.3. *If $x(t,\xi)$ is a solution of the bimodal CLS (2.9) such that $c^T x(t,\xi)$ is not identically zero on the interval $[0,T]$ with $T \in (0,\infty)$, then a partition (3.2) of the interval $[0,T]$ exists such that $c^T x(t,\xi)$ does not have a zero, and thus is persistently positive or negative, in each of the open subintervals (t_{i-1},t_i) for all $i = 1,\ldots,N$. Moreover, for every positive $\varepsilon < \frac{1}{2}\min_{1\leqslant i\leqslant N}(t_i - t_{i-1})$, there exists a neighborhood $\mathcal{N}$ of ξ such that for any $\eta \in \mathcal{N}$, $\min_{1\leqslant i\leqslant N}\min_{t\in[t_{i-1}+\varepsilon,t_i-\varepsilon]}\big(c^T x(t,\xi)\big)\big(c^T x(t,\eta)\big) > 0$.*

Proof. By the partition in Theorem 3.7, we deduce the existence of finitely many time instants t_i for $i = 1,\ldots,N$ with $t_0 = 0$ and $t_N = T$ such that $c^T x(t,\xi) = c^T e^{A_i(t-t_{i-1})}x(t_{i-1},\xi)$ for all $t \in [t_{i-1},t_i]$, where A_i is either A or $A + bc^T$. Since the right-hand function is analytic on the real line, it has finitely many zeros in the compact subinterval $[t_{i-1},t_i]$, unless the function is identically zero. In the latter case, $c^T x(t,\xi)$ is identically zero on $[t_{i-1},t_i]$, which implies that $x(t_{i-1},\xi) \in \overline{O}(c^T,A_i) = \overline{O}(c^T,A)$. Proceeding forward and backward in time, we can establish that $\xi \in \overline{O}(c^T,A)$ so that $c^T x(t,\xi)$ is identically zero on the entire interval $[0,T]$, contradicting the assumption. Thus, by refining the partition (3.2) of the interval $[0,T]$ if necessary, we readily obtained the desired conclusions of the proposition. ∎

Noticing that the right-hand side of the bimodal ODE (2.9) is the sum of a linear function and the max function whose directional derivative is trivial to write down, we can invoke the results in [26] to obtain the directional derivative $x'_\xi(t,\xi;\eta)$ of the solution function $x(t,\cdot)$ at a vector $\xi \in \mathbb{R}^n$ along a direction $\eta \in \mathbb{R}^n$. To describe this derivative succinctly, consider the function $g^\xi : [0,T]\times\mathbb{R}^n \to \mathbb{R}^n$ defined by

$$g^\xi(t,y) \equiv \begin{cases} Ay & \text{if } c^T x(t,\xi) < 0, \\ Ay + b\max(0,c^T y) & \text{if } c^T x(t,\xi) = 0, \\ (A + bc^T)y & \text{if } c^T x(t,\xi) > 0 \end{cases}$$

and the time-dependent ODE

$$\dot{y}(t) = g^\xi(t,y), \quad y(0) = \eta. \tag{5.1}$$

Note that the function $g^\xi(t,y)$ is only piecewise continuous in t. In fact, by Proposition 5.3, if $c^T x(t,\xi)$ is not identically zero on the interval $[0,T]$, where $T \in [0,\infty)$, there exists a partition (3.2) of the interval $[0,T]$ such that $c^T x(t,\xi)$ has no zero in (t_{i-1},t_i) and $c^T x(t,\xi) = 0$ for all $t \in \{t_1,\ldots,t_{N-1}\}$. The case where $c^T x(t,\xi)$ is identically equal to zero can be made to be part of this treatment by taking $N = 0$ (the vacuous partition). Note that the case where $c^T x(t,\xi)$ does not change sign, but can have (isolated) zeros, in $[0,T]$ is clearly permitted. We call the subinterval (t_{i-1},t_i) *positive* (*negative*) if $c^T x(t,\xi)$ is positive (negative, respectively) throughout (t_{i-1},t_i). Note that this terminology refers to the given trajectory $x(t,\xi)$. In terms of the partition (3.2), we can write

$$g^\xi(t,y) \equiv \begin{cases} Ay & \text{if } t \text{ is in a negative subinterval } (t_{i-1},t_i), \\ Ay + b\max(0,c^T y) & \text{if } t \in \{t_1,\ldots,t_{N-1}\}, \\ (A + bc^T)y & \text{if } t \text{ is in a positive subinterval } (t_{i-1},t_i), \end{cases}$$

which shows that $g^\xi(t,y)$ is in general discontinuous at the times t_i, $i = 1,\ldots,N-1$. By Theorem 3.2 in [21], the time-varying ODE (5.1) has a unique solution, which we

denote $y^\xi(t;\eta)$, for every initial η that is continuous on $[0,T]$; moreover, by Theorem 7 in [26], this unique solution is equal to $x'_\xi(t,\xi;\eta)$. Therefore, we obtain

$$x'_\xi(t,\xi;\eta) = \begin{cases} e^{A(t-t_i)}x'_\xi(t_{i-1},\xi;\eta) & \text{if } t \text{ is in a negative subinterval } (t_{i-1},t_i), \\ e^{(A+bc^T)(t-t_i)}x'_\xi(t_{i-1},\xi;\eta) & \text{if } t \text{ is in a positive subinterval } (t_{i-1},t_i). \end{cases}$$

From this expression, we deduce that $x(t,\xi)+x'_\xi(t,\xi;\eta-\xi) = y^\xi(t;\eta)$ for all $t \in [0,T]$. This equality remains valid if $c^T x(t,\xi)$ is identically equal to zero on the entire interval $[0,T]$. Note that $y^\xi(t;\eta) \neq x(t,\eta)$ in general. As proved in Proposition 5.5 below, these two functions will coincide if the states ξ and η are T-time mode consistent with respect to the bimodal CLS (2.9) as defined below.

DEFINITION 5.4. *Two states ξ and η are said to be* mode consistent *on an interval $\mathcal{I}$ with respect to the bimodal CLS* (2.9) *if $(c^T x(t,\xi))(c^T x(t,\eta)) \geqslant 0$ for all times $t \in \mathcal{I}$. If $\mathcal{I} = [0,T]$, we say that ξ and η are* T-time mode consistent.

According to Proposition 5.3, η is mode consistent with ξ on the subintervals $[t_{i-1}+\varepsilon, t_i-\varepsilon]$ for all $i = 1,\ldots,N$, provided that η is sufficiently near ξ. The difficulty with analyzing the perturbed trajectory $x(t,\eta)$ lies in the subintervals $(t_i-\varepsilon, t_i+\varepsilon)$ which are ε-neighborhoods of the critical times t_i. Due to its importance, we introduce the notation $\mathcal{M}_T^\xi$ to denote the set of all states that are T-time mode consistent with ξ. The next result asserts that $\mathcal{M}_T^\xi \cap \Omega_T^\xi = \mathcal{M}_T^\xi \cap (\xi + \mathcal{Z}_T^\xi)$.

PROPOSITION 5.5. *Two T-time mode consistent states ξ and η of the bimodal CLS* (2.9) *are T-time indistinguishable if and only if $Hx'_\xi(t,\xi;\eta-\xi) = 0$ for all $t \in [0,T]$.*

Proof. Under the mode consistency assumption, we have $x(t,\eta) = x(t,\xi) + x'_\xi(t,\xi;\eta-\xi)$ for all $t \in [0,T]$. The desired equivalence follows readily. ∎

Based on Propositions 5.1 and 5.5, the following result pertains to a special class of initial states ξ such that the trajectory $x(t,\xi)$ remains on the boundary of the two pieces at all times.

THEOREM 5.6. *Suppose $c^T x(t,\xi) = 0$ for all $t \in [0,T]$ with $0 < T \leqslant \infty$. The following six statements are equivalent.*

(a) *ξ is T-time globally observable for the bimodal CLS* (2.9).
(b) $\mathcal{Z}_T^\xi = \{0\}$.
(c) *$Hx(t,\eta) = 0$ for all $t \in [0,T]$ implies $\eta = 0$; i.e., the zero state is T-time globally (or locally) observable.*
(d) $\mathcal{F}(\Omega_T^\xi,\xi) = \mathcal{T}(\Omega_T^\xi,\xi) = \{0\}$.
(e) *ξ is T-time locally observable.*
(f) $\overline{O}(H,A) \cap \overline{O}(H,A+bc^T) = \{0\}$,

$$\begin{aligned} \overline{O}(H,A) \cap \left\{ v \mid c^T e^{At} v \leqslant 0 \ \forall\ t \in [0,T] \right\} &= \{0\}, \\ \overline{O}(H,A+bc^T) \cap \left\{ v \mid c^T e^{(A+bc^T)t} v \geqslant 0 \ \forall\ t \in [0,T] \right\} &= \{0\}. \end{aligned} \tag{5.2}$$

Proof. Under the assumption, all states are T-time mode consistent with ξ. Therefore, the equivalence of (a) and (b) follows from Proposition 5.5. Moreover, $x'_\xi(t,\xi;\eta) = x(t,\eta)$ for all $t \in [0,T]$. Therefore, (b) and (c) are equivalent. Statement (b) clearly implies (d) by (4.8); (d) implies (e) by Proposition 4.13. Finally, we show that (e) implies (b). Suppose (e) holds, but $\mathcal{Z}_T^\xi$ has a nonzero vector v. Then $Hx'_\xi(t,\xi;v) = 0$ for all $t \in [0,T]$. Let $\eta \equiv \xi + \tau v$, where $\tau > 0$ is chosen so that η falls in the neighborhood of ξ where its T-time local observability holds. Since

$Hx'_\xi(t,\xi;\eta-\xi)=0$ for all $t\in[0,T]$, Proposition 5.5 implies that η is T-time indistinguishable from ξ. By the T-time local observability of ξ, we deduce that $\eta-\xi$, and thus v is equal to zero, which is a contradiction. Hence (b) holds. The first five statements of the corollary are thus equivalent.

Suppose that any one of the five equivalent statements (a)–(e) holds. By Propositions 5.1 and 5.2, $\overline{O}(H,A)\cap\overline{O}(H,A+bc^T)=\{0\}$. Moreover, if either condition in (5.2) does not hold, then we have a nonzero vector v belonging to the left-hand set of one of these two expressions. It follows that $Hx(t,v)=0$ for all $t\in[0,T]$, which contradicts (c). Conversely, suppose that (f) holds. It suffices to show that any state η such that $Hx(t,\eta)=0$ for all $t\in[0,T]$ must satisfy $c^Tx(t,\eta)\geqslant 0$ for all $t\in[0,T]$ or $c^Tx(t,\eta)\leqslant 0$ for all $t\in[0,T]$; i.e., $c^Tx(t,\eta)$ remains in one piece over $[0,T]$. Suppose the claim does not hold. Then there is a switching time $t_*\in(0,T)$. This means that $\delta>0$ exists such that $c^Tx(t,\eta)$ is of one nonzero sign in $(t_*-\delta,t_*)$ and of a different nonzero sign in $(t_*,t_*+\delta)$. This implies in particular that $x(t,\eta)\neq 0$ for all $t\geqslant 0$. Without loss of generality, we may assume that $c^Tx(t,\eta)>0$ for all $t\in(t_*-\delta,t_*)$ and $c^Tx(t,\eta)<0$ for all $t\in(t_*,t_*+\delta)$. Hence, the indistinguishability condition $Hx(t,\eta)=0$ for all $t\in[0,T]$ yields $x(t_*,\eta)\in\overline{O}(H,A+bc^T)\cap\overline{O}(H,A)$, which is a contradiction because $x(t_*,\eta)\neq 0$. □

It is interesting to note that conditions (c) and (f) are independent of the state ξ. Clearly, elements of the set $\{v\mid c^Te^{At}v\leqslant 0\ \forall\, t\in[0,T]\}$ are vectors v in the half-plane $\mathcal{X}_2=\{v\mid c^Tv\leqslant 0\}$ such that a trajectory, when initiated at v, remains in the same half-plane. Thus the first condition in (5.2) stipulates that the zero vector is the only such vector that also lies in the unobservable space of the pair (H,A), which is the mode to which the trajectory in question belongs. A similar interpretation applies to the second condition. A local version of Theorem 5.6 is as follows.

COROLLARY 5.7. *Suppose that all states sufficiently near ξ are T-time mode consistent with ξ. The following statements are equivalent.*

(a) *ξ is T-time locally observable for the bimodal CLS* (2.9).

(b) $\mathcal{Z}_T^\xi=\{0\}$.

(c) $\mathcal{F}(\Omega_T^\xi,\xi)=\mathcal{T}(\Omega_T^\xi,\xi)=\{0\}$.

Proof. (a) ⇒ (b). This basically follows the same proof as in the previous proof. Suppose $Hx'_\xi(t,\xi;v)=0$ for all $t\in[0,T]$. Let $\eta\equiv\xi+\tau v$, where $\tau>0$ is chosen so that η is T-time mode consistent with ξ and that η falls in the neighborhood of ξ where its T-time local observability holds. Proposition 5.5 implies that η is T-time indistinguishable from ξ. By the T-time local observability of ξ, we then have $v=0$. The implications (b) ⇒ (c) ⇒ (a) require no proof. □

The condition that $c^Tx(t,\xi)=0$ for all $t\in[0,T]$ with $0<T\leqslant\infty$ in Theorem 5.6 is equivalent to $\xi\in\overline{O}(c^T,A)$, which is further equivalent to $x(t,\xi)\in\overline{O}(c^T,A)$ for all $t\in[0,T]$. The T-time (local or global) observability of such a state is completely resolved by the corollary. Note that for a state $\xi\in\overline{O}(c^T,A)$, the set $\mathcal{M}_T^\xi=\mathbb{R}^n$. The next result treats the case where $\xi\notin\overline{O}(c^T,A)$; the key condition in this result is the generalization of condition (c) in Theorem 5.6 to condition (b) in the theorem below.

THEOREM 5.8. *Suppose that $\xi\notin\overline{O}(c^T,A)$. The two conditions,*

(a) $\overline{O}(H,A)\cap\overline{O}(H,A+bc^T)=\{0\}$,

(b) *a neighborhood $\mathcal{N}$ of ξ exists such that $\mathcal{M}_T^\xi\cap(\,\xi+\mathcal{Z}_T^\xi\,)\cap\mathcal{N}=\{\xi\}$,*

are necessary for ξ to be T-time locally observable for the bimodal CLS (2.9) *for any $T\in(0,\infty]$, and sufficient for any $T\in(0,\infty)$.*

We first establish a lemma that is the key to the proof of the sufficiency part of Theorem 5.8.

LEMMA 5.9. *If $\xi \notin \overline{O}(c^T, A)$, then, under condition* (a) *in Theorem* 5.8, *for every finite $T > 0$, there exists a neighborhood $\mathcal{N}_0$ of ξ such that $\Omega_T^\xi \cap \mathcal{N}_0 \subseteq \mathcal{M}_T^\xi$.*

Proof. Since $\xi \notin \overline{O}(c^T, A)$, it follows that $x(t, \xi) \notin \overline{O}(c^T, A)$ for all $t \geqslant 0$. Furthermore, by Proposition 5.2, $b \notin \overline{O}(H, A)$. Consequently, for each $t \geqslant 0$, there exists an integer $\ell_t \geqslant 0$ such that $HA^{\ell_t}x(t, \xi) \neq H(A + bc^T)^{\ell_t}x(t, \xi)$. We claim that a neighborhood $\widetilde{\mathcal{N}}$ of ξ exists such that for every $t \in [0, T]$ and every $\eta \in \widetilde{\mathcal{N}}$, an integer $\ell \in \{0, 1, \ldots, n-1\}$ exists such that

$$HA^\ell x(t, \xi) \neq H(A + bc^T)^\ell x(t, \eta) \quad \text{and} \quad HA^\ell x(t, \eta) \neq H(A + bc^T)^\ell x(t, \xi).$$

Indeed, if no such neighborhood exists, then there exist a sequence of vectors $\{\eta^k\}$ converging to ξ and a sequence of times $\{t_k\} \subset [0, T]$ such that for each $\ell \in \{0, 1, \ldots, n-1\}$, either $HA^\ell x(t_k, \xi) = H(A+bc^T)^\ell x(t_k, \eta^k)$ or $HA^\ell x(t_k, \eta^k) = H(A+bc^T)^\ell x(t_k, \xi)$. The sequence $\{t_k\}$ accumulates to some time t_* in $[0, T]$; for any such time, we must have $HA^\ell x(t_*, \xi) = H(A + bc^T)^\ell x(t_*, \xi)$ for all nonnegative integers ℓ. This contradiction establishes the claim.

Let $t_i, i = 1, \ldots, N$ be finitely many time instants in $[0, T]$ described in Proposition 5.3 such that $c^T x(t_i, \xi) = 0$. Let ε be a positive scalar and $\widehat{\mathcal{N}}$ be a neighborhood of ξ such that every $\eta \in \widehat{\mathcal{N}}$ is mode consistent with ξ on all the subintervals $[t_{i-1} + \varepsilon, t_i - \varepsilon]$ for all $i = 1, \ldots, N$; the existence of ε and $\widehat{\mathcal{N}}$ is due to Proposition 5.3. Let $\mathcal{N}_0 \equiv \widetilde{\mathcal{N}} \cap \widehat{\mathcal{N}}$, where $\widetilde{\mathcal{N}}$ is established above. By way of contradiction, suppose that there exists a sequence of vectors $\{\eta^k\} \in \mathcal{N}_0$ converging to ξ such that each η^k is T-time indistinguishable from but not T-time mode consistent with ξ. Hence, there exists $\widetilde{t}_k \in (t_{i_k} - \varepsilon, t_{i_k} + \varepsilon)$ for some $i_k \in \{1, \ldots, N-1\}$ or $\widetilde{t}_k \in [0, \varepsilon)$ or $\widetilde{t}_k \in (T - \varepsilon, T]$ such that $(c^T x(\widetilde{t}_k, \xi))(c^T x(\widetilde{t}_k, \eta^k)) < 0$ for every k. We may assume without loss of generality that $\widetilde{t}_k \in [0, T)$ is such that for some $\delta_k > 0$, $c^T x(t, \xi) > 0$ and $c^T x(t, \eta^k) < 0$ for all $t \in (\widetilde{t}_k, \widetilde{t}_k + \delta_k)$ (if $\widetilde{t}_k = T$, we consider the interval $(\widetilde{t}_k - \delta_k, \widetilde{t}_k)$ and use the reverse-time argument). Hence, we have $x(t, \xi) = e^{(A+bc^T)(t-\widetilde{t}_k)}x(\widetilde{t}_k, \xi)$ and $x(t, \eta) = e^{A(t-\widetilde{t}_k)}x(\widetilde{t}_k, \eta^k)$ for all such t. Thus $He^{(A+bc^T)(t-\widetilde{t}_k)}x(\widetilde{t}_k, \xi) = He^{A(t-\widetilde{t}_k)}x(\widetilde{t}_k, \eta^k)$ for all $t \in (\widetilde{t}_k, \widetilde{t}_k + \delta_k)$, which yields $H(A + bc^T)^\ell x(\widetilde{t}_k, \xi) = HA^\ell x(\widetilde{t}_k, \eta^k)$ for all nonnegative integers ℓ. But this contradicts the claim established above because $\eta^k \in \widetilde{\mathcal{N}}$. ∎

Proof of Theorem 5.8. Suppose that ξ is T-time locally observable for $T \in [0, \infty]$. By Propositions 5.1 and 5.2, condition (a) is necessary. To prove (b), choose a neighborhood $\mathcal{N}$ of ξ so that $\Omega_T^\xi \cap \mathcal{N} = \{\xi\}$. Since $\mathcal{M}_T^\xi \cap (\xi + \mathcal{Z}_T^\xi) = \Omega_T^\xi$ by Proposition 5.5, we have

$$\mathcal{M}_T^\xi \cap (\xi + \mathcal{Z}_T^\xi) \cap \mathcal{N} = \mathcal{M}_T^\xi \cap \Omega_T^\xi \cap \mathcal{N} = \{\xi\}.$$

Conversely, let $T < \infty$. Assume that (a) and (b) hold and let $\mathcal{N}$ be the neighborhood of ξ described in (b). By Lemma 5.9, a neighborhood $\mathcal{N}_0$ of ξ exists such that $\Omega_T^\xi \cap \mathcal{N}_0 \subseteq \mathcal{M}_T^\xi$. So, by (b),

$$\Omega_T^\xi \cap \mathcal{N} \cap \mathcal{N}_0 \subseteq \Omega_T^\xi \cap \mathcal{M}_T^\xi \cap \mathcal{N} \cap \mathcal{N}_0 = (\xi + \mathcal{Z}_T^\xi) \cap \mathcal{M}_T^\xi \cap \mathcal{N} \cap \mathcal{N}_0 = \{\xi\}.$$

Hence ξ is T-time observable within the neighborhood $\mathcal{N} \cap \mathcal{N}_0$. ∎

It is important to note that the sufficient part of Theorem 5.8 requires T to be finite. We will discuss more about the case where $T = \infty$ in subsection 5.3.

5.2. More on mode consistency. From condition (b) in Theorem 5.8, the important role of mode consistency in T-time local observability is amply evident. In what follows, we discuss this condition in greater detail. Specifically, our goal is to generalize the two conditions in (5.2) to the case where $\xi \notin \overline{O}(c^T, A)$. As it turns out, such a generalization is not trivial because we need to deal with various mode transitions, which necessitate the introduction of the "mode-transition matrices" $\Phi_\xi^{\mathrm{n}}(t)$ associated with the nominal trajectory $x(t,\xi)$.

Recall that Proposition 5.3 implies the existence of a partition of the finite interval $[0,T]$, where $T \in (0,\infty)$, such that $c^T x(t,\xi)$ is persistently positive or negative in (t_{i-1}, t_i) for all $i = 1,\dots,N$. In other words, $c^T x(t,\xi) = 0$ only at t_i's. For each i, define

$$(\mathbf{c}^i, S_i) \equiv \begin{cases} (-c^T, A) & \text{if } (t_i, t_{i+1}) \text{ is a negative subinterval,} \\ (c^T, A + bc^T) & \text{if } (t_i, t_{i+1}) \text{ is a positive subinterval,} \end{cases} \qquad i = 0,\dots,N-1,$$

and $\mathcal{Y}(S_i) \equiv \left\{ x \,\middle|\, \left(\mathbf{c}^i x,\ \mathbf{c}^i S_i x, \dots, \mathbf{c}^i S_i^{n-1} x \right) \succcurlyeq 0 \right\}$. The nominal state trajectory on $[0,T]$ can be written as $x(t,\xi) = \Phi_\xi^{\mathrm{n}}(t)\xi$, where

$$\Phi_\xi^{\mathrm{n}}(t) \equiv \begin{cases} \text{the identity matrix} & \text{for } t = 0, \\ e^{S_k(t-t_k)} \displaystyle\prod_{i=1}^{k} e^{S_{i-1}(t_i - t_{i-1})} & \text{for } t \in [t_k, t_{k+1}],\ k = 0,\dots,N-1, \end{cases}$$

is a *mode-transition matrix*. It is clear that $\Phi_\xi^{\mathrm{n}}(t)$ is invertible for all t.

PROPOSITION 5.10. *Let $T \in (0,\infty)$. If $\xi \notin \overline{O}(H,A)$, then Theorem* 5.8(b) *is equivalent to the following: with $\widehat{x}_i = \Phi_\xi^{\mathrm{n}}(t_i)\xi$ for $i = 0,\dots,N$, and $S_{-1} = S_N \equiv 0$,*

$$\bigcap_{i=0}^{N} \left\{ \left[\Phi_\xi^{\mathrm{n}}(t_i)\right]^{-1} \left[\overline{O}(H,S_i) \cap \left(\mathcal{Y}(S_i) - \widehat{x}_i\right) \cap \overline{O}(H,-S_{i-1}) \cap \left(\mathcal{Y}(-S_{i-1}) - \widehat{x}_i\right) \right] \right\} = \{0\}.$$

Proof. Suppose that Theorem 5.8(b) does not hold. Then for any neighborhood $\mathcal{N}$ of ξ, there is $\eta \in \mathcal{N}$ other than ξ such that $\eta \in \mathcal{M}_T^\xi \cap (\xi + \mathcal{Z}_T^\xi)$. Hence, by Proposition 5.5, it follows that ξ and η are T-time mode consistent states and $Hx'_\xi(t,\xi;\eta-\xi) = 0$ for all $t \in [0,T]$. This implies that $x(t,\eta) = \Phi_\xi^{\mathrm{n}}(t)\eta$ and $Hx'_\xi(t,\xi;\eta-\xi) = H\Phi_\xi^{\mathrm{n}}(t)[\eta-\xi] = 0$ on $[0,T]$. Let $v = \eta - \xi$. The mode consistency condition further implies that for all $i = 1,\dots,N$,

$$\Phi_\xi^{\mathrm{n}}(t_i)\eta = \Phi_\xi^{\mathrm{n}}(t_i)[v+\xi] = \Phi_\xi^{\mathrm{n}}(t_i)v + \widehat{x}_i \in \mathcal{Y}(S_i).$$

Similarly, $\Phi_\xi^{\mathrm{n}}(t_i)\eta = \Phi_\xi^{\mathrm{n}}(t_i)v + \widehat{x}_i \in \mathcal{Y}(-S_{i-1})$. Moreover, for any $t \in (t_i, t_{i+1})$ and $i = 0,\dots,N-1$, $H\Phi_\xi^{\mathrm{n}}(t)v = He^{S_i(t-t_i)}\Phi_\xi^{\mathrm{n}}(t_i)v$. Hence, $H\Phi_\xi^{\mathrm{n}}(t)v = 0$ for all $t \in [0,T]$ implies that $\Phi_\xi^{\mathrm{n}}(t_i)v \in \overline{O}(H,S_i)$. Using the reverse-time argument, one can also deduce $\Phi_\xi^{\mathrm{n}}(t_i)v \in \overline{O}(H,-S_{i-1})$. Consequently, we have

$$0 \neq v \in \bigcap_{i=1}^{N} \left\{ \left[\Phi_\xi^{\mathrm{n}}(t_i)\right]^{-1} \left[\overline{O}(H,S_i) \cap \left(\mathcal{Y}(S_i) - \widehat{x}_i\right) \cap \overline{O}(H,-S_{i-1}) \cap \left(\mathcal{Y}(-S_{i-1}) - \widehat{x}_i\right) \right] \right\},$$

which is a contradiction.

Conversely, suppose Theorem 5.8(b) holds, but there is a nonzero vector v belonging to the above intersection. We therefore have $\Phi_\xi^{\mathrm{n}}(t_i)[\xi+v] \in \mathcal{Y}(S_i) \cap \mathcal{Y}(-S_{i-1})$ for

all $i = 0, \ldots, N-1$. Since $\Phi^{\rm n}_\xi(t_i)\xi \in \mathcal{Y}(S_i) \cap \mathcal{Y}(-S_{i-1})$, it is clear that $\Phi^{\rm n}_\xi(t_i)[\xi + \kappa v] \in \mathcal{Y}(S_i) \cap \mathcal{Y}(-S_{i-1})$ for all $0 \leqslant \kappa \leqslant 1$, following the same argument in the proof of Theorem 4.5. For each i, we can apply the second statement of Lemma 4.10 to obtain two positive pairs $(\varepsilon_{i+}, \tau_{i+})$ and $(\varepsilon_{i-}, \tau_{i-})$ such that $\mathbf{c}^i e^{S_i(t-t_i)}\Phi^{\rm n}_\xi(t_i)[\xi + \kappa v] \geqslant 0$ for all $(t, \kappa) \in [t_i, t_i + \varepsilon_{i+}] \times (0, \tau_{i+}]$ and $\mathbf{c}^i e^{S_{i-1}(t_i-t)}\Phi^{\rm n}_\xi(t_i)[\xi + \kappa v] \geqslant 0$ for all $(t, \kappa) \in [t_i - \varepsilon_{i-}, t_i] \times (0, \tau_{i-}]$. Define $\varepsilon' = \min_{1 \leqslant i \leqslant N} \varepsilon_{i\pm}$ and $\tau' = \min_{1 \leqslant i \leqslant N} \tau_{i\pm}$. As a result, for all $t \in \cup_{1 \leqslant i \leqslant N}[t_i - \varepsilon', t_i + \varepsilon']$ and all $0 \leqslant \kappa \leqslant \tau'$,

$$\big(c^T x(t, \xi)\big)\big(c^T x(t, \xi + \kappa v)\big) \geqslant 0. \tag{5.3}$$

By Proposition 5.3, a neighborhood $\mathcal{N}$ of ξ and a positive number $\tau_0 \leqslant \tau'$ exist such that $\xi + \kappa v \in \mathcal{N}$ for all $\kappa \in [0, \tau_0]$ and that $\big(c^T x(t, \xi)\big)\big(c^T x(t, \xi + \kappa v)\big) > 0$ for all $t \in \cup_{1 \leqslant i \leqslant N}[t_{i-1} + \varepsilon', t_i - \varepsilon']$. Since (5.3) also holds for $0 \leqslant \kappa \leqslant \tau_0$, we have $\big(c^T x(t, \xi)\big)\big(c^T x(t, \xi + \kappa\, v)\big) \geqslant 0$ for all $t \in [0, T]$ and $0 \leqslant \kappa \leqslant \tau_0$. Hence $(\xi, \xi + \kappa\, v)$ are two T-time mode consistence states for all $0 \leqslant \kappa \leqslant \tau_0$. Thus, for all $0 \leqslant \kappa \leqslant \tau_0$, $Hx(t, \xi + \kappa v) = H\Phi^{\rm n}_\xi(t)[\xi + \kappa v] = H\Phi^{\rm n}_\xi(t)\xi + \kappa H\Phi^{\rm n}_\xi(t)v$. Notice that for any $t \in [t_i, t_{i+1}]$, we have $H\Phi^{\rm n}_\xi(t)v = He^{S_i(t-t_i)}\Phi^{\rm n}_\xi(t_i)v$. Since $\Phi^{\rm n}_\xi(t_i)v \in \overline{O}(H, S_i) \cap \overline{O}(H, -S_i)$, $He^{S_i(t-t_i)}\Phi^{\rm n}_\xi(t_i)v \equiv 0$ for all $t \in [t_i, t_{i+1}]$ and all $i = 0, \ldots, N-1$. Consequently, $H\Phi^{\rm n}_\xi(t)v \equiv 0$ on $[0, T]$. This suggests that $(\xi, \xi + \kappa\, v)$ are two T-time indistinguishable states for all $0 \leqslant \kappa \leqslant \tau_0$. By Proposition 5.5, $v \in \mathcal{Z}^\xi_T$ and thus $(\xi + \kappa v) \in \mathcal{M}^\xi_T \cap (\xi + \mathcal{Z}^\xi_T) \cap \mathcal{N}$ for all $0 \leqslant \kappa \leqslant \tau_0$. As a result, for any small neighborhood of $\mathcal{V} \subseteq \mathcal{N}$, $(\xi + \kappa v) \in \mathcal{M}^\xi_T \cap (\xi + \mathcal{Z}^\xi_T) \cap \mathcal{V}$ for some $0 < \kappa \leqslant \tau_0$. This contradicts Theorem 5.8(b). ∎

5.3. Long-time observability. The finiteness of T is needed to ensure the applicability of Lemma 5.9 in the proof of the sufficiency part of Theorem 5.8. In what follows, we establish the long-time (i.e., $T = \infty$) local observability of ξ for the bimodal CLS (2.9) under various conditions, one of which postulates that the trajectory $x(t, \xi)$ has a switching time $t_* > 0$; this is a time for which $\delta > 0$ exists such that $c^T x(t, \xi)$ is of one nonzero sign in $(t_* - \delta, t_*)$ and of a different nonzero sign in $(t_*, t_* + \delta)$. Note that any trajectory $x(t, \xi)$ such that $c^T x(t, \xi)$ changes sign at least once must have a switching time, provided that $\xi \notin \overline{O}(c^T, A)$.

PROPOSITION 5.11. *Any of the conditions* (a), (b), *and* (c) *below is sufficient for* ξ *to be long-time locally observable for the bimodal system* (2.9)*:*

(a) $\overline{O}(H, A) = \{0\}$ *and there exists* $t_* \geqslant 0$ *such that* $c^T x(t_*, \xi) < 0$*;*
(b) $\overline{O}(H, A + bc^T) = \{0\}$ *and there exists* $t_* \geqslant 0$ *such that* $c^T x(t_*, \xi) > 0$*;*
(c) $\overline{O}(H, A) \cap \overline{O}(H, A + bc^T) = \{0\}$ *and the trajectory* $x(t, \xi)$ *has a switching time.*

Proof. Under condition (a) or (b), the state $x(t_*, \xi)$ is short-time, and thus long-time, locally observable, by Corollary 4.7 and Proposition 4.11. To prove the same under condition (c), we assume without loss of generality that for some $\delta > 0$, $c^T x(t, \xi)$ is positive on $[t_* - \delta, t_*)$ and negative on $(t_*, t_* + \delta]$. Localizing the proof of Lemma 5.9 to the compact interval $\mathcal{T}_* \equiv [t_* - \delta, t_* + \delta]$ and using the time invariance of the bimodal system, we deduce the existence of a neighborhood $\mathcal{N}$ of ξ such that any state $\eta \in \mathcal{N}$ that is indistinguishable from ξ in the interval $\mathcal{T}_*$ must be mode consistent with ξ on the same interval. Suppose that ξ is not long-time locally observable. There exists a sequence of vectors $\{\eta^k\} \subset \Omega^\xi_\infty$ converging to ξ such that $\eta^k \neq \xi$ for all k. It follows that for k sufficiently large, η^k must be mode consistent with ξ on the interval $\mathcal{T}_*$.

Hence we have

$$He^{(A+bc^T)(t-t_*+\delta)}x(t_*-\delta,\xi) = He^{(A+bc^T)(t-t_*+\delta)}x(t_*-\delta,\eta^k) \quad \forall\,[\,t_*-\delta,t_*\,),$$
$$He^{A(t-t_*)}x(t_*,\xi) = He^{A(t-t_*)}x(t_*,\eta^k) \qquad \forall\,(\,t_*,t_*+\delta\,].$$

It follows that $x(t_*,\xi)-x(t_*,\eta^k) \in \overline{O}(H,A+bc^T)\cap\overline{O}(H,A)$. Hence $x(t_*,\xi)=x(t_*,\eta^k)$, which yields $\xi=\eta^k$, a contradiction! □

We close this section by presenting numerical examples to illustrate several counter-intuitive properties of the long-time observability of an initial state ξ as well as the states along the trajectory $x(t,\xi)$. These examples also demonstrate the difficulties of designing constructive tests for long-time observability. The first example shows that in order for ξ to be long-time locally observable, it is possible that no state on the trajectory $x(t,\xi)$ is short-time observable. This example strengthens Example 4.12 which concerns mainly properties of the initial state.

Example 5.12. Consider a bimodal system in $\mathbb{R}^3$ with

$$A=\begin{bmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{bmatrix},\ b=\begin{pmatrix}b_1\\0\\0\end{pmatrix},\ c=\begin{pmatrix}c_1\\c_2\\c_3\end{pmatrix},\ \text{and } H=\begin{bmatrix}h_1 & h_2 & 0\end{bmatrix},$$

where $\lambda_3>\lambda_2>\lambda_1>0$, $b_1\neq 0$, $c_1c_2c_3\neq 0$, $h_1h_2\neq 0$,

$$\lambda_2\,\lambda_3\,(\,\lambda_3-\lambda_2\,)+\lambda_1\,\lambda_2\,(\,\lambda_2-\lambda_1\,)+\lambda_3\,\lambda_1\,(\,\lambda_1-\lambda_3\,)\neq 0, \tag{5.4}$$

and $h_2(\lambda_1-\lambda_2+b_1c_1)=h_1b_1c_2$. Straightforward computations show that

$$\overline{O}(H,A)=\text{span}\,\{(0,0,1)\} \quad \text{and} \quad \overline{O}(H,A+bc^T)=\text{span}\{(h_2,-h_1,0)\}.$$

Notice that $\overline{O}(H,A)\cap\overline{O}(H,A+bc^T)=\{0\}$. Since both of the above two subspaces contain nonzero states, it follows from Corollary 4.7 that the bimodal CLS has no short-time local observable states. However, the condition (5.4) ensures that $\overline{O}(c^T,A)=\{0\}$. Hence for any $\xi\neq 0$, the trajectory $c^Tx(t,\xi)$ can have only isolated zeros. Pick an initial state ξ such that $c^T\xi<0$ but $c_3\xi_3>0$. Then $c^Tx(t,\xi)<0$ for all $t\geqslant 0$ sufficiently small, but $c^Tx(t,\xi)$ must eventually become positive. Thus the trajectory $x(t,\xi)$ must have a switching time. By Proposition 5.11, such an initial state must be long-time locally observable.

The next example shows that the long-time local observability of a state does not imply its long-time global observability.

Example 5.13. Consider a bimodal system in $\mathbb{R}^2$ with

$$A=\begin{bmatrix}\lambda & 0\\ 0 & \lambda-b_2c_2\end{bmatrix},\ b=\begin{pmatrix}0\\b_2\end{pmatrix},\ c=\begin{pmatrix}c_1\\c_2\end{pmatrix},\ \text{and } H=\begin{bmatrix}1 & 1\end{bmatrix},$$

where $\lambda>0$, $b_2\neq 0$, $c_1c_2<0$, and $b_2c_2>0$. It is clear that $\overline{O}(H,A)=\{0\}$. Choose $\xi=(\xi_1,0)$ with $c_1\xi_1<0$. Hence, $c^Te^{At}\xi=e^{\lambda t}c_1\xi_1\leqslant c_1\xi_1<0$ for all $t\geq 0$, implying that $x(t,\xi)=e^{At}\xi=(e^{\lambda t}\xi_1,0)$ for all $t\geqslant 0$. Hence such a ξ is short-time, and thus long-time locally observable. Yet, ξ is not long-time globally observable because by considering $\eta=(0,\xi_1)$, we have $c^T\eta=c_2\xi_1>0$. Since $c^Te^{(A+bc^T)t}\eta=e^{\lambda t}c_2\xi_1>0$, implying that $x(t,\eta)=(0,e^{\lambda t}\xi_1)$ for all $t\geqslant 0$. Thus, ξ and η are long-time indistinguishable. Hence ξ is not long-time globally observable.

Our last example shows that for a state to be long-time locally observable, it is not necessary for this state and any future state along the nominal trajectory to be T-time locally observable for any finite $T \geqslant 0$. In other words, even if an initial state (and any future state along the nominal trajectory) is not T-time locally observable for any finite $T \geqslant 0$, it is still possible that the CLS is long-time locally observable at the initial state.

Example 5.14. Consider a bimodal system in $\mathbb{R}^3$ with

$$A = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \alpha & \omega \\ 0 & -\omega & \alpha \end{bmatrix},\ b = \begin{pmatrix} b_1 \\ 0 \\ 0 \end{pmatrix},\ c = \begin{pmatrix} c_1 \\ c_2 \\ 0 \end{pmatrix},\ \text{and } H = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix},$$

where $\lambda < 0$, $\alpha > 0$, $\omega > 0$, $b_1 \neq 0$, and c_1 and c_2 are both nonzero. Let $\xi = (\xi_1, 0, 0)$ with $c_1\xi_1 < 0$. It follows that $x(t, \xi) = (e^{\lambda t}\xi_1, 0, 0)$ and $c^T x(t, \xi) = e^{\lambda t}c_1\xi_1 < 0$ for all $t \geqslant 0$. Let $v_I = (0, 1, 0)$ and $v_R = (0, 0, 1)$. It is easy to see that $\overline{O}(H, A) = \text{span}\{v_I, v_R\}$ and $\overline{O}(H, A + bc^T) = \{0\}$. Note that any $\eta \in \overline{O}(H, A)$ with $\|\eta\|_2 = 1$ can be written as $\eta(\phi) = v_I \sin\phi + v_R \cos\phi$ for some $\phi \in [0, 2\pi)$ and that

$$e^{At}\eta(\phi) = e^{\alpha t}\big[(-v_I \sin\phi + v_R \cos\phi)\sin(\omega t) + (v_R \sin\phi + v_I \cos\phi)\cos(\omega t)\big].$$

Moreover, any initial state that is indistinguishable over the time interval when the corresponding state trajectory is in the mode characterized by A must be of the form $\xi + \tau\eta(\phi)$ for some $\tau \in \mathbb{R}$ and $\phi \in [0, 2\pi)$. Hence, for any given $T \geqslant 0$, there is a $\tau_{_T} > 0$ such that for all $(t, \tau, \phi) \in [0, T] \times [-\tau_{_T}, \tau_{_T}] \times [0, 2\pi)$,

$$\begin{aligned} c^T x(t, \xi + \tau\eta(\phi)) &= c_1\xi_1 e^{\lambda t} + \tau c_2 e^{\alpha t}\big[-\sin\phi\sin(\omega t) + \cos\phi\cos(\omega t)\big] \\ &= c_1\xi_1 e^{\lambda t} + \tau c_2 e^{\alpha t}\cos(\omega t + \phi) < 0. \end{aligned}$$

Note that $Hx(t, \xi) = Hx(t, \xi + \tau\eta(\phi))$ for all such triples (t, τ, ϕ). Hence, ξ is not T-time locally observable. The same argument applies to any future state along the nominal trajectory. However, notice that no matter how small $|\tau| > 0$ is and what $\phi \in [0, 2\pi)$ is, $c^T x(t, \xi + \tau\eta(\phi))$ will become positive in some finite time. Hence, $x(t, \xi + \tau\eta(\phi))$ will switch to the mode characterized by $(A + bc^T)$ at some $t_* > 0$; i.e., $c^T x(t_*, \xi + \tau\eta(\phi)) = 0$ but $c^T x(t, \xi + \tau\eta(\phi)) > 0$ for all $t > t_*$ sufficiently close to t_*. Let $\widetilde{x} = x(t_*, \xi + \tau\eta(\phi))$. Note that $0 \neq \widetilde{x} \notin \overline{O}(c^T, A)$, which implies that $H(A+bc^T)^\ell\widetilde{x} \neq HA^\ell x(t_*, \xi)$ for some nonnegative integer ℓ. We have $x(t, \xi+\tau\eta(\phi)) = e^{(A+bc^T)(t-t_*)}\widetilde{x}$ for all $t > t_*$ sufficiently close to t_*. By the proof of Lemma 5.9, $Hx(t, \xi) \equiv Hx(t, \xi + \tau\eta(\phi))$ for all $t \geqslant 0$ only when $\tau = 0$. This establishes that ξ is long-time locally observable.

REFERENCES

[1] M. Babaali and M. Egerstedt, *Observability of switched linear systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2993, R. Alur and G. J. Pappas, eds., Springer, Berlin, 2004, pp. 48–63.

[2] Z. Bartosiewicz, *Local observability of nonlinear systems*, Systems Control Lett., 22 (1995), pp. 295–298.

[3] A. Bemporad, G. Ferrari-Trecate, and M. Morari, *Observability and controllability of piecewise affine and hybrid systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1864–1876.

[4] B. Brogliato, *Some perspectives on analysis and control of complementarity systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 918–935.

[5] P. Brunovsky, *Regular synthesis for the linear-quadratic optimal control problem with linear control constraints*, J. Differential Equations, 38 (1980), pp. 344–360.
[6] M. K. Camlibel, *Complementarity Methods in the Analysis of Piecewise Linear Dynamical Systems*, Ph.D. thesis, Center for Economic Research, Tilburg University, Tilburg, The Netherlands, 2001.
[7] M. K. Camlibel, W. P. M. H. Heemels, and J. M. Schumacher, *On linear passive complementarity systems*, Eur. J. Control, 8 (2002), pp. 220–237.
[8] M. K. Camlibel, W. P. M. H. Heemels, A. J. van der Schaft, and J. M. Schumacher, *Switched networks and complementarity*, IEEE Trans. Circuits Syst. I, 50 (2003), pp. 1036–1046.
[9] M. K. Camlibel, J.-S. Pang, and J. Shen, *Lyapunov stability of complementarity and extended systems*, SIAM J. Optim., 17 (2006), pp. 1056–1101.
[10] C. T. Chen, *Linear System Theory and Design*, Oxford University Press, Oxford, UK, 1984.
[11] P. Collins and J. H. van Schuppen, *Observability of hybrid systems and Turing machines*, in Proceedings of the 43th IEEE Conference on Decision and Control, Paradise Islands, Bahamas, 2004.
[12] P. Collins and J. H. van Schuppen, *Observability of piecewise affine hybrid systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2993, R. Alur and G. J. Pappas, eds., Springer, Berlin, 2004, pp. 250–264.
[13] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.
[14] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems* I, Springer, New York, 2003.
[15] W. P. M. H. Heemels, *Linear Complementarity Systems: A Study in Hybrid Dynamics*, Ph.D. thesis, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, 1999.
[16] W. P. M. H. Heemels and B. Brogliato, *The complementarity class of hybrid dynamical systems*, Eur. J. Control, 9 (2003), pp. 311–349.
[17] W. P. M. H. Heemels, M. K. Camlibel, and J. M. Schumacher, *On the dynamic analysis of piecewise-linear networks*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 49 (2002), pp. 315–327.
[18] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *Linear complementarity systems*, SIAM J. Appl. Math., 60 (2000), pp. 1234–1269.
[19] K. H. Johansson, M. Egersted, J. Lygeros, and S. Sastry, On the regularization of Zeno hybrid automata, Systems Control Lett., 38 (1999), pp. 141–150.
[20] R. E. Kalman, *On the general theory of control systems*, in Proceedings of the 1st World Congress of the International Federation of Automatic Control, 1960, pp. 481–493.
[21] H. K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 2002.
[22] G. Lafferriere, G. J. Pappas, and S. Yovine. *Reachability computation for linear hybrid systems*, in Proceedings of the 14th IFAC World Congress, Vol. E, Beijing, 1999, pp. 7–12.
[23] H. Nijmeijer and A. J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer, New York, 1990.
[24] J. S. Pang and J. Shen, *Strongly regular differential variational systems*, IEEE Trans. Automat. Control, to appear.
[25] J. S. Pang and D. Stewart, *Differential variational inequalities*, Math. Program., Ser. A, to appear.
[26] J. S. Pang and D. Stewart, *Solution dependence on initial conditions in differential variational inequalities*, Math. Program., Ser. B, to appear.
[27] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[28] A. J. van der Schaft and J. M. Schumacher, *The complementary-slackness class of hybrid systems*, Math. Control Signals Systems, 9 (1996), pp. 266–301.
[29] A. J. van der Schaft and J. M. Schumacher, *Complementarity modeling of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 483–490.
[30] A. J. van der Schaft and J. M. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Springer, London, 2000.
[31] J. M. Schumacher, *Complementarity systems in optimization*, Math. Program., Ser. B, 101 (2004), pp. 263–295.
[32] S. N. Simic, K. H. Johansson, J. Lygeros, and S. Sastry, *Towards a geometric theory of hybrid systems*, Dyn. Contin. Discrete Impuls. Syst. Ser. B. Appl. Algorithms, 12 (2005), pp. 649–687.
[33] J. Shen and J.-S. Pang, *Linear complementarity systems: Zeno states*, SIAM J. Control Optim., 44 (2005), pp. 1040–1066.

[34] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions* I, Springer, Berlin, 1970.

[35] H. J. Sussmann, *Bounds on the number of switchings for trajectories of piecewise analytic vector fields*, J. Differential Equations, 43 (1982), pp. 399–418.

[36] R. Vidal, A. Chiuso, S. Soatto, and S. Sastry, *Observability of linear hybrid systems*, in Hybrid Systems: Computation and Control, Lecture Notes in Comput. Sci. 2623, O. Maler and A. Pnueli, eds., Springer, Berlin, 2003, pp. 526–539.

[37] J. Zhang, K. H. Johansson, J. Lygeros, and S. Sastry, *Zeno hybrid systems*, Internat. J. Robust Nonlinear Control, 11 (2001), pp. 435–451.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1801–1821

SPECTRAL FACTORIZATION OF A CLASS OF MATRIX-VALUED SPECTRAL DENSITIES*

HENDRA I. NURDIN†

Abstract. Recently, a necessary and sufficient uniform log-integrability condition has been established for the canonical spectral factorization mapping to be sequentially continuous. However, this condition, along with several other equivalent conditions, is not straightforward to verify. In this paper, we first derive a new set of easily checkable sufficient conditions which guarantee uniform log-integrability. Based on the newly derived conditions, we establish the existence of certain convergent rational approximations for a class of matrix-valued spectral densities. We then propose a new spectral factorization algorithm and provide convergence results. Our approach does not require the spectral density to be coercive. Numerical examples are given to illustrate the effectiveness and convergence of the proposed algorithm. In particular, we compute approximate spectral factors of the noncoercive and nonrational Kolmogorov and von Karman power spectra which arise in the study of turbulence.

Key words. spectral factorization, nonrational spectral density, rational approximation, rational covariance extension, second order stochastic processes

AMS subject classifications. 30E10, 41A20, 65D15, 60G12, 93E03, 93E25

DOI. 10.1137/040621089

1. Background and motivation.

It is known that a discrete-time second order wide sense stationary stochastic process with a power spectral density, or simply a *spectral density* or *spectrum*, satisfying a certain Szegö or Paley–Wiener condition can be modeled as the output of a discrete-time causal linear time invariant system (i.e., a "shaping filter") driven by white noise [28]. If the spectral density is rational, then determining a shaping filter is possible by spectral factorization of the spectral density, and there are practical algorithms to do this. In the case when the spectral density is nonrational, obtaining a spectral factor is much more difficult, and explicit spectral factorization can be done only in special cases. Apart from deriving shaping filters, computation of spectral factors plays an important role in the theory of optimal and robust control. A survey of spectral factorization methods for both rational and nonrational spectral densities is given in [26].

Given a scalar spectral density W, a particular rational spectral density W_n, known as a *maximum entropy* spectrum, can be constructed whose first n terms of the covariance sequence match those of W. Matching of partial covariance sequences is a natural approach since every spectral density is associated with a unique covariance sequence, and both give a complete and equivalent characterization of some purely nondeterministic wide sense stationary stochastic process [24, 28]. H_n, the canonical

*Received by the editors December 19, 2004; accepted for publication (in revised form) June 2, 2006; published electronically December 11, 2006. This work is supported by National ICT Australia, Ltd. (NICTA). National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence Program. A preliminary version of this work appeared in *Proceedings of the Joint 44th IEEE Conference on Decision and Control* (*CDC*) *and European Control Conference* (*ECC*), Seville, Spain, 2005, pp. 5929–5934.

http://www.siam.org/journals/sicon/45-5/62108.html

†Department of Information Engineering, Research School of Information Sciences and Engineering (Building 115), Australian National University, Canberra ACT 0200, Australia, and National ICT Australia, Ltd. (NICTA), Locked Bag 8001, Canberra ACT 2601, Australia (Hendra.Nurdin@anu.edu.au).

spectral factor of W_n, can be constructed recursively via the Szegö–Levinson algorithm (see, e.g., [11, 24]), and it is a well known that H_n converges to H, the canonical spectral factor of W, in $\mathcal{H}^2$ (see section 2 for notation) as $n \to \infty$; therefore also W_n converges to W in $\mathcal{L}^1$. However, H_n obtained in this way is an all-pole transfer function. Since H_n has no zeros, it has long been observed that $H_1, H_2, \ldots$ converges slowly to H when the latter has zeros on or close to the unit circle; some examples can be found in, e.g., [14, sect. IV], [5, pp. 214–217], and [6, section 6]. Intuitively, this is due to the inability of W_n to reproduce valleys of W (i.e., points for which W has a small value) for small or medium n. Similar slow convergence is also true when the matrix version of the Szegö–Levinson algorithm (see [29, 10]) is applied to a matrix-valued spectrum W with transmission zeros on or close to the unit circle [29]. More formally, it was shown in [13, 14] that if W is scalar and rational, then its zeros close to or on the unit circle decrease the rate of decay of the Schur parameters of W: the convergence rate decreases as a zero approaches the unit circle (see also [20]). In particular, when W has roots on the unit circle, the rate of convergence is no longer geometric. Consequently, good approximations can be achieved only for large n. This limitation of the maximum entropy spectra motivated research into the development of rational spectral densities which (1) matches the first few terms of the covariance sequence of W, and (2) has zeros at specified locations on the complex plane. A new theory of rational covariance extension with a degree constraint was later developed, which makes possible the construction of such spectra. Indeed, it has been demonstrated in [5, 6] that a finite dimensional spectrum constructed via the new theory, by suitably placing zeros on the unit disc, is a better estimate than the corresponding maximum entropy estimate, in the sense that it is able to better capture features of the true spectrum.

This paper develops a new approach to spectral factorization of a spectral density W based on constructing a sequence of rational approximations $\{W_n\}$ with freely specified zeros and which match partial covariance sequences of W. The construction is achieved by taking advantage of a recent result on continuity of the spectral factorization mapping given in [3] and the theory of degree constrained rational covariance extensions [13, 14, 7, 5, 6, 22, 21]. We derive theoretical results on convergence of this scheme for continuous W, and conditions on the zeros for convergence to be achieved, under further mild assumptions on W. Convergence of covariance matching approximations with freely specified zeros has not been studied in the literature; convergence results have been established only for the case when W_n has no zeros, i.e., the maximum entropy method/Szegö–Levinson algorithm. In particular, our results weaken some conditions previously derived by Anderson [1], Caines and Baykal-Gürsoy [8], and Mari, Dahlèn, and Lindquist [20] for convergence of $\{H_n\}$ to H in $\mathcal{H}^2$ or $\mathcal{H}^\infty$. Then in section 7 of this paper, several numerical examples are given that show the advantage of the proposed approach over the popular maximum entropy method: lower degree approximations with lower approximation error (defined in a certain sense).

The organization of this paper is as follows. In section 2 we introduce the notation and recall some definitions and results from the literature. Following that, in section 3 we discuss a result on sequential continuity of the spectral factorization mapping. In section 4 we derive a new set of easily checkable and sufficient conditions for uniform log-integrability of a sequence of spectral densities. In sections 5 and 6 we develop the theoretical foundation of a new approach to spectral factorization and introduce a new spectral factorization algorithm for a class of matrix-valued spectral densities. We then present a number of numerical examples using the proposed spectral factorization algorithm in section 7. Finally, in section 8 we give the conclusions of this paper and discuss potential applications of the results as well as directions for future research.

2. Main notation and basic definitions. In this section we introduce the main notation used throughout the paper and also recall some definitions and relevant results from the literature.

- $\mathbb{R}$, $\mathbb{C}$, $\mathbb{D}$, and $\mathbb{T}$ denote the set of real numbers, complex numbers, the open unit disc $= \{z \in \mathbb{C} : |z| < 1\}$, and the unit circle (the boundary of $\mathbb{D}$), respectively.
- $\overline{A}$ denotes the completion of A.
- A^* and $\Re\{A\} = A + A^*$ denote the Hermitian transpose and Hermitian part of a complex matrix A, respectively.
- $\mathbb{N}$ denotes the set of natural numbers $1, 2, 3, \ldots$.
- $\log c$ denotes the natural logarithm of c.
- A *pseudopolynomial* is a matrix-valued function f of the form

$$f(z) = \sum_{i=-m}^{n} A_i z^i,$$

  where $0 \le m, n < \infty$, and $A_i \in \mathbb{C}^{l\times l}$ $(l \in \mathbb{N})$ for $i = -m, \ldots, n$.
- The $\|\cdot\|_p$ norm of a matrix $A \in \mathbb{C}^{m\times n}$ is defined as [3]

$$\|A\|_p = \begin{cases} \left(\operatorname{tr}\{(A^*A)^{p/2}\}\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \sup_{v\in\mathbb{C}^n, \|v\|\le 1} \|Av\| & \text{if } p = \infty. \end{cases}$$

- μ denotes the Lebesque measure on $\mathbb{T}$.
- $\mathcal{L}^p_{m\times n}$, $1 \le p \le \infty$, denotes the space of measurable functions mapping from $\mathbb{T}$ to $\mathbb{C}^{m\times n}$ with a finite $\|\cdot\|_p$ norm defined by

$$\|f\|_p = \begin{cases} \left(\frac{1}{2\pi}\int_{\mathbb{T}} \|f(z)\|_p^p d\mu\right)^{1/p} & \text{if } 1 \le p < \infty, \\ \operatorname{ess\,sup}_{z\in\mathbb{T}} \|f(z)\|_\infty & \text{if } p = \infty. \end{cases}$$

  If $n = 1$, we write $\mathcal{L}^p_{m\times n}$ simply as $\mathcal{L}^p_m$.
- $\mathcal{H}^p_{m\times n}$, $1 \le p \le \infty$, denotes the subspace of functions in $\mathcal{L}^p_{m\times n}$ having an analytic continuation from $\mathbb{T}$ to $\mathbb{D}$. If $n = 1$, we write $\mathcal{H}^p_{m\times n}$ simply as $\mathcal{H}^p_m$.
- H_* denotes the *para-Hermitian conjugate* of H: $H_*(z) = H(z^{*-1})^*$.

If H is a rational element of $\mathcal{H}^p_{n\times n}$ or $\mathcal{L}^p_{n\times n}$, then the degree of H, denoted by $\deg(H)$, is defined to be the McMillan degree of H. Let $\mathcal{P}_n$ denote the linear space of $\mathbb{C}^n$-valued trigonometric polynomials on $\mathbb{T}$. It is well known that this space is dense in $\mathcal{L}^p_n$ for all $p \in [1, \infty)$. In a similar fashion we define the linear space $\mathcal{P}^+_n$ to be the set of $\mathbb{C}^n$-valued polynomials on $\mathbb{C}$. We may view $\mathcal{P}^+_n$ as a linear subspace of $\mathcal{P}_n$. A function $\rho \in \mathcal{H}^2_{n\times n}$ is said to be *outer* if $\overline{\rho\mathcal{P}^+_n} = \mathcal{H}^2_n$, i.e., the set of products $\rho\mathcal{P}^+_n$ is dense in $\mathcal{H}^2_n$ [3]. In the special case when $n = 1$ (the scalar case) and ρ is a rational function, it is known that ρ is outer if and only if all its zeros and poles lie in $\mathbb{C}\backslash\mathbb{D}$.

A function W which maps from $\mathbb{T}$ to $\mathbb{C}^{n\times n}$ is a *spectral density* if (1) it is in $\mathcal{L}^1_{n\times n}$, and (2) there exists an outer function $H \in \mathcal{H}^2_{n\times n}$ such that $W(e^{i\theta}) = H(e^{i\theta})^* H(e^{i\theta})$. Note that the definition implies that $W^* = W$ and W is nonnegative definite almost everywhere (a.e.) $\mathbb{T}$. The function H is called a *spectral factor* of W. A spectral factor can be uniquely specified if a condition is imposed on its value at the origin. We call the *unique* spectral factor, which is positive definite at the origin, the *canonical spectral factor* (CSF). We say that W is *rational* if each element W_{ij} is of the form $W_{ij}(e^{i\theta}) = \frac{P_{ij}(e^{i\theta})}{Q_{ij}(e^{i\theta})}$ for some scalar pseudopolynomials P_{ij} and Q_{ij}, and *coercive* if

W has no zeros on $\mathbb{T}$. A precise characterization of spectral densities is given in the following classical result.

THEOREM 2.1 (see [27, 18, 28, 26]). *A nonnegative definite function $W \in \mathcal{L}^1_{n\times n}$ is a spectral density if and only if $\int_{\mathbb{T}} |\log\det W(z)|\,\mu(dz) < \infty$.*

For a function $f : \mathbb{T} \to \mathbb{R}$ we write $f > 0$ ($f \geq 0$) if f is positive (nonnegative) definite a.e. $\mathbb{T}$, and $f > g$ ($f \geq g$) will be taken to mean $f - g > 0$ ($f - g \geq 0$).

3. Sequential continuity of the spectral factorization mapping. Let W be a spectral density and let $\Phi(W)$ denote its unique CSF. Then the mapping $\Phi : W \mapsto \Phi(W)$ is called the *spectral factorization mapping*. It was recently shown in [3] that the mapping Φ is *sequentially continuous*; see the following theorem.

THEOREM 3.1. *Let W be a spectral density, and let $\{W_r\}_{r\in\mathbb{N}}$ be a sequence of spectral densities such that $W_r \to W$ in $\mathcal{L}^1_{n\times n}$ as $r \to \infty$. Then the following are equivalent:*

1. *The sequence $\{\log\det W_r\}_{r\in\mathbb{N}}$ is uniformly integrable.*
2. *$\Phi(W_r) \to \Phi(W)$ in $\mathcal{H}^2_{n\times n}$ as $r \to \infty$.*

Recall that a family of scalar measurable functions $\{X_\gamma \mid \gamma \in \Gamma\}$ parametrized by a nonempty set Γ on a measurable space $(\Omega, \mathcal{F})$ with measure M is said to be *uniformly integrable* if $\lim_{\alpha\to\infty} \sup_{\gamma\in\Gamma} \int_{\{\omega\in\Omega||X_\gamma(\omega)|>\alpha\}} |X_\gamma(\omega)|M(d\omega) = 0$.

Remark 3.2. We shall refer to the condition in point 1 of Theorem 3.1 as *uniform log-integrability*.

Several conditions which are equivalent to uniform log-integrability are given in [3, Prop. 4.2]. However, these conditions are general and do not indicate how to construct a uniformly log-integrable sequence $\{W_r\}_{r\in\mathbb{N}}$ which converges to W in $\mathcal{L}^1_{n\times n}$. For this reason, we shall shortly develop more explicit sufficient conditions.

4. A sufficient and verifiable set of conditions for uniform log-integrability. In this section we shall derive a new set of conditions on the sequence of convergent spectral densities *and* the limiting spectral density which ensures that the uniform log-integrability condition of Theorem 3.1 is satisfied. To this end, for $\alpha \geq 0$, let us define

$$\begin{aligned} A_r(\alpha) &= \{z \in \mathbb{T} \mid |\log\det W_r(z)| > \alpha\}, \\ A_{r+}(\alpha) &= \{z \in \mathbb{T} \mid \det W_r(z) > e^{\alpha}\}, \\ A_{r-}(\alpha) &= \left\{z \in \mathbb{T} \mid \det W_r(z) < e^{-\alpha}\right\}, \end{aligned}$$

and note that $A_{r+}(\alpha) \cap A_{r-}(\alpha) = \phi$ and $A_r(\alpha) = A_{r+}(\alpha) \cup A_{r-}(\alpha)$. The set $A_{r+}(\alpha)$ is the collection of points at which $\det W_r$ has "large" values and which may grow to ∞ as $\alpha \to \infty$, while $A_{r-}(\alpha)$ is the set of points where $\det W_r$ take on "small" values and can diminish to 0 as $\alpha\to\infty$. Then we have the following inequality:

$$\begin{aligned} \sup_{r\in\mathbb{N}} \int_{A_r(\alpha)} |\log\det W_r(z)|d\mu \leq\ & \sup_{r\in\mathbb{N}} \int_{A_{r+}(\alpha)} \log\det W_r(z)d\mu \\ & + \sup_{r\in\mathbb{N}} \int_{A_{r-}(\alpha)} -\log\det W_r(z)d\mu. \end{aligned} \tag{4.1}$$

The main idea here is to derive sufficient conditions for each of the two terms on the right-hand side of (4.1) such that they go to 0 as $\alpha \to \infty$. It turns out that finding conditions to guarantee the desired effect on the first term is relatively easy. As for the second term, the conditions are more complicated. To have that term go to 0 as $\alpha \to \infty$, the idea is to impose conditions which exclude the existence of a set

of positive Lebesque measure on which $\det W_r$ decays to zero as $r \to \infty$. Before going into the formal details, we note the following matrix inequality.

LEMMA 4.1. *For any nonnegative definite matrix* $A \in \mathbb{C}^{n\times n}$, $\log\det A \le \|A\|_1$.

Proof. Note that the result is trivial if A is singular, since in this case we have $\log\det A = -\infty$. Therefore we assume that A is *positive definite*. Let $\sigma_1, \sigma_2, \ldots, \sigma_n$ be the singular values of A, with $\sigma_1 \ge \sigma_2 \ge \cdots > 0$. Since A is positive definite, we have that $\det(A^*) = \det(A)$ and $\log\det A = \frac{1}{2}\log\det(AA^*) = \sum_{k=1}^n \log\sigma_k$. On the other hand, we also have that $\|A\|_1 = \operatorname{tr}((AA^*)^{\frac{1}{2}}) = \sum_{k=1}^n \sigma_k$, and the result follows since $\log\sigma_k \le \sigma_k$ for $k = 1, \ldots, n$. ∎

First, let us make the following assumptions.

A1. $\operatorname{ess\,sup}_{z\in\mathbb{T}} \|W_r(z)\|_1 < \infty$ for all $r \in \mathbb{N}$.

A2. The sequence $\{W_r\}_{r\in\mathbb{N}}$ converges in $\mathcal{L}^1_{n\times n}$ to W as $r \to \infty$.

Now we can show the following result.

LEMMA 4.2. *Under* assumptions A1–A2*,*

$$\lim_{\alpha\to\infty} \sup_{r\in\mathbb{N}} \int_{A_{r+}(\alpha)} \log\det W_r(z)\mu(dz) = 0.$$

Proof. By assumptions A1 and A2 we have (i) $\lim_{\alpha\to\infty}\sup_{r\in\mathbb{N}}\int_{A_{r+}(\alpha)} \|W(z) - W_r(z)\|_1\mu(dz) = 0$, and (ii) $\lim_{\alpha\to\infty}\sup_{r\in\mathbb{N}}\mu(A_{r+}(\alpha)) = 0$ (if the latter is not true, then we have $\sup_{r\in\mathbb{N}}\|W_r\|_1 = \infty$, which contradicts A1 and A2). It then follows from (ii) that (iii) $\lim_{\alpha\to\infty}\sup_{r\in\mathbb{N}}\int_{A_{r+}(\alpha)}\|W(z)\|_1\mu(dz) = 0$. Since $\|W_r(z)\|_1 \le \|W(z) - W_r(z)\|_1 + \|W(z)\|_1$, we get from (i) and (iii) that $\lim_{\alpha\to\infty}\sup_{r\in\mathbb{N}}\int_{A_{r+}(\alpha)} \|W_r(z)\|_1\mu(dz) = 0$. Lemma 4.1 then gives $\lim_{\alpha\to\infty}\sup_{r\in\mathbb{N}}\int_{A_{r+}(\alpha)}\log\det W_r(z)\mu(dz) = 0$, as desired. ∎

Let us impose three further assumptions on $\{W_r\}_{r\in\mathbb{N}}$.

A3. $W_r(e^{i\theta})$ is a piecewise continuous function of θ for each $r \in \mathbb{N}$.

A4. Let Z_a be the set defined by

$$Z_a = \Big\{ z_0 \in \mathbb{T} \mid \liminf_r f_r = 0 \ \forall \text{ neighborhoods } U \text{ of } z_0, f_r = \inf_{z\in U\cap\mathbb{T}} \det W_r(z) \Big\}.$$

Then the cardinality of Z_a is finite.

A5. Let Z_r be the set of all zeros of $\det W_r$ (i.e., all points $z_0 \in \mathbb{T}$ for which $\inf_{z\in U\cap\mathbb{T}}\det W_r(z) = 0$ for all neighborhoods U of z_0). Then there exists $M_1, M_2, \Delta_1, \Delta_2 > 0$ such that for any $r \in \mathbb{N}$ and any $\theta_{0,r} \in (-\pi, \pi]$ such that $e^{i\theta_{0,r}} \in Z_r \cup Z_a$,

$$\det W_r(e^{i\theta}) \ge M_1 |\theta - \theta_{0,r}|^{M_2} \ \forall \theta \in [\theta_{0,r} - \Delta_1, \theta_{0,r} + \Delta_2] \cap (-\pi, \pi]. \tag{4.2}$$

Remark 4.3. Assumption A5 implies that the cardinality of Z_r is uniformly bounded (away from ∞).

We have the following result.

LEMMA 4.4. *Under assumptions* A3–A5*,*

$$\lim_{\alpha\to\infty} \sup_{r\in\mathbb{N}} \int_{A_{r-}(\alpha)} -\log\det W_r(z)\mu(dz) = 0.$$

Proof. Let $\theta_{r,1},\dots,\theta_{r,n_r}$ be the angles (in $(-\pi,\pi]$) of elements of $Z_r \cup Z_a$. Then $n_r \le L$ for all r, where L is some finite positive integer. Define

$$\tilde{A}_{r-}(\alpha) = \left\{\theta \in (-\pi,\pi] \mid e^{i\theta} \in A_{r-}(\alpha)\right\} \text{ and}$$
$$\tilde{A}_{r-,k}(\alpha) = \tilde{A}_{r-}(\alpha) \bigcap \{\theta \in (-\pi,\pi] \mid -\Delta_1 \le \theta - \theta_{r,k} \le \Delta_2\}$$

for $i = 1,\dots,n_r$. Note that $\theta_{r,k} \in \tilde{A}_{r-}(\alpha)$ for $k = 1,\dots,n_r$ and that $\tilde{A}_{r-,k}$ can be empty for some k's. Clearly, assumptions A3–A5 imply that for some α_0 large enough and $\alpha > \max\{\alpha_0, -\log(M_1(\min\{\Delta_1,\Delta_2\})^{M_2})\}$, $\tilde{A}_{r-,k}(\alpha)$ are disjoint for $k = 1,\dots,n_r$, independently of r, and $\tilde{A}_{r-}(\alpha) = \bigcup_{k=1}^{n_r} \tilde{A}_{r-,k}(\alpha)$. Furthermore, without loss of generality we may take M_1 to have a value less than 1. Hence the following holds:

$$\begin{aligned}
\int_{A_{r-}(\alpha)} -\log\det W_r(z)d\mu &= \sum_{k=1}^{n_r} \int_{\tilde{A}_{r-,k}(\alpha)} -\log\det W_r(e^{i\theta})d\theta, \\
&\le \sum_{k=1}^{n_r} \int_{\tilde{A}_{r-,k}(\alpha)} -\log\left(M_1|\theta-\theta_{r,k}|^{M_2}\right) d\theta, \\
&\le -\mu(A_{r-}(\alpha))\log M_1 \\
&\quad + M_2 \sum_{k=1}^{n_r} \int_{\tilde{A}_{r-,k}(\alpha)} -\log|\theta-\theta_{r,k}|\, d\theta. \qquad (4.3)
\end{aligned}$$

Let $\alpha_1 = \max\{\alpha_0, -\log(M_1(\min\{\Delta_1,\Delta_2\})^{M_2})\}$. Assumptions A3–A5 also imply that for $\alpha > \alpha_1$ there exists a number $\epsilon(\alpha) > 0$, dependent on α, such that $\lim_{\alpha\to\infty} \epsilon(\alpha) = 0$ and $\tilde{A}_{r-,k}(\alpha) \subset \theta_{r,k} + B(\alpha) = \{\theta \in (-\pi,\pi] \mid \theta = \theta_{r,k} + \omega; \omega \in B(\alpha)\}$, where $B(\alpha)$ is a set independent of r defined by $B(\alpha) = \{\theta \in (-\pi,\pi] \mid -\epsilon(\alpha)\Delta_1 \le \theta \le \epsilon(\alpha)\Delta_2\}$. Therefore from (4.3) we have

$$\begin{aligned}
\int_{A_{r-}(\alpha)} -\log\det W_r(z)d\mu &\le -\sum_{k=1}^{n_r} \Lambda(\theta_{r,k} + B(\alpha))\log M_1 \\
&\quad + M_2 \sum_{k=1}^{n_r} \int_{\theta_{r,k}+B(\alpha)} -\log|\theta-\theta_{r,k}| d\theta, \\
&\le -L\Lambda(B(\alpha))\log M_1 + LM_2 \int_{B(\alpha)} -\log|\theta| d\theta,
\end{aligned}$$

where Λ denotes the Lebesque measure on $(-\pi,\pi]$. Since the right-hand side of the last inequality $\to 0$ as $\alpha \to \infty$ independently of r, we conclude that

$$\lim_{\alpha\to\infty} \sup_{r\in\mathbb{N}} \int_{A_{r-}(\alpha)} -\log\det W_r(z)\mu(dz) = 0,$$

which is the statement we had set out to prove. ∎

A direct consequence of Lemmas 4.2 and 4.4 is the following theorem.

THEOREM 4.5. *Under assumptions* A1–A5*,*

$$\lim_{\alpha\to\infty} \sup_{r\in\mathbb{N}} \int_{A_r(\alpha)} |\log\det W_r(z)|\mu(dz) = 0,$$

i.e., the sequence $\{\log\det W_r\}_{r\in\mathbb{N}}$ *is uniformly integrable.*

Proof. This follows directly from Lemmas 4.2 and 4.4 by taking the limit $\alpha \to \infty$ on both sides of inequality (4.1). ∎

The above theorem has the following important corollary.

COROLLARY 4.6. *Let $W \in \mathcal{L}^1_{n\times n}$ with $\|\det W - f\|_1 = 0$ for some spectral density $f \in \mathcal{L}^1$ having a finite number of zeros on $\mathbb{T}$ (a zero is as defined in assumption* A5*). If $\{W_r\}_{r\geq 1}$ is a sequence of piecewise continuous spectral densities such that $\lim_{r\to\infty} \operatorname{ess\,sup}_{z\in\mathbb{T}} \|W(z) - W_r(z)\|_1 = 0$, then $\lim_{r\to\infty} \|\Phi(W) - \Phi(W_r)\|_2 = 0$.*

The corollary is a simple but useful result and relaxes the requirement $W > 0$ in [1, Thm. 1] (or [3, Cor. 6.2] with $p = 2$). We shall prove later on that a sequence satisfying the conditions of the corollary can be explicitly constructed under some regularity conditions on W.

5. Construction of convergent rational spectral densities with converging CSFs. In this section we give the main ideas for the construction of a sequence of rational spectral densities with CSFs converging to the true CSF. Let $\{W_r\}_{r\in\mathbb{N}}$ be a sequence of rational spectral densities having no poles on $\mathbb{T}$. We define

$$c_k = \frac{1}{2\pi}\int_{\mathbb{T}} W(z) z^{-k} \mu(dz) \quad \text{and} \quad c_{k,r} = \frac{1}{2\pi}\int_{\mathbb{T}} W_r(z) z^{-k} \mu(dz), \quad k = 0, 1, \dots .$$

The sequences $\{c_k\}_{k\in\mathbb{N}}$ and $\{c_{k,r}\}_{k\in\mathbb{N}}$ are the unique covariance sequences associated with W and W_r, respectively. By the Riemann–Lebesque lemma, $c_k \to 0$ as $k \to \infty$. The covariance sequence $c_{k,r}$ has the form

$$c_{k,r} = C_r A_r^k B_r + \sum_{m=0}^{m_r} D_{m,r} \Delta(k - m), \tag{5.1}$$

where A_r, B_r, C_r, and $D_{0,r}, D_{1,r}, \dots, D_{m_r,r}$ are $n \times n$ matrices with A_r having eigenvalues in $\mathbb{D}$, (A_r, B_r, C_r) is a minimal realization, and

$$\Delta(m) = \begin{cases} 1 & \text{if } m = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The central idea of our construction is to require the sequence $\{W_r\}_{r\in\mathbb{N}}$ to satisfy

$$\deg \Phi(W_r) \leq n d_r, \tag{5.2a}$$

$$c_{k,r} = c_k \quad \text{for} \quad k = 0, 1, \dots, d_r, \tag{5.2b}$$

where $\{d_r\}_{r\in\mathbb{N}}$ is an increasing sequence of positive integers. That a sequence $\{W_r\}_{r\in\mathbb{N}}$ satisfying (5.2) exists and is computable is the content of the theory of a rational covariance extension with a degree constraint [14, 7, 6, 15, 22, 21]. Since $\|W - W_r\|_1 \geq \sup_{k\geq 0} \|c_k - c_{k,r}\|_1 \geq \sup_{0\leq k\leq d_r} \|c_k - c_{k,r}\|_1$, we see that the discrepancy between the first few terms of the covariance sequence of W and W_r yields a lower bound for the approximation error in $\mathcal{L}^1_{n\times n}$. Therefore, it makes sense to impose condition (5.2b). Moreover, since it is desirable to have W_r be as "simple" as possible, the constraint (5.2a) is also well motivated. Plugging the Fourier series expansion of W_r into the

definition of $\|W - W_r\|_1$, we obtain

$$\begin{aligned}\int_{\mathbb{T}} \|W(z) - W_r(z)\|_1 \mu(dz) &\le \int_{\mathbb{T}} \left\| W(z) - \sum_{k=0}^{d_r} \Re\{c_k z^k\} \right\|_1 \mu(dz) \\ &\quad + \left\| \int_{\mathbb{T}} \Re\Big\{ C_r A_r^{d_r+1} (I - A_r z)^{-1} B_r \right. \\ &\quad \left. + I_{\{d_r \le m_r - 1\}}(d_r) \sum_{m=d_r+1}^{m_r} D_{m,r} z^m \Big\} \mu(dz) \right\|_1 \\ &\le \int_{\mathbb{T}} \left\| W(z) - \sum_{k=0}^{d_r} \Re\{c_k z^k\} \right\|_1 \mu(dz) + \mathrm{R}(W_r, d_r), \end{aligned} \tag{5.3}$$

where $I_A(x)$ is the indicator function for the set A and

$$\mathrm{R}(W_r, d_r) = \left\| \Re\Big\{ C_r A_r^{d_r+1} \int_{\mathbb{T}} (I - A_r z)^{-1} \mu(dz) B_r \Big\} \right\|_1.$$

If W satisfies $\operatorname{ess\,sup}_{z\in\mathbb{T}} \|W(z)\|_1 < \infty$, the Fourier series of W converges to W in $\mathcal{L}^2_{n\times n}$, hence also in $\mathcal{L}^1_{n\times n}$. Therefore, the first term on the right-hand side of (5.3) goes to 0 as $r \to \infty$, and the following theorem is immediate.

THEOREM 5.1. *Suppose* $\operatorname{ess\,sup}_{z\in\mathbb{T}} \|W(z)\|_1 < \infty$ *and let* $\{W_r\}_{r\in\mathbb{N}}$ *be a sequence of rational spectral densities satisfying assumptions* A4–A5 *and the interpolation constraints of* (5.2). *If* $\lim_{r\to\infty} \mathrm{R}(W_r, d_r) = 0$, *then assumption* A2 *holds and* $\lim_{r\to\infty} \|\Phi(W) - \Phi(W_r)\|_2 = 0$.

It is reasonable to expect, at least intuitively, that there could be "many" sequences which satisfy the condition of Theorem 5.1 if the spectral density W is not too "irregular." Indeed, we see later in Corollaries 5.3 and 5.5 particular instances where this is true. Moreover, we will show that the approximating sequence $\{W_r\}$ can be constructed explicitly under some further, yet mild, assumptions on W.

5.1. The scalar case. We shall give a constructive proof of the following result.

THEOREM 5.2. *Let* W *be a continuous scalar spectral density and* $\|W - U/V\|_\infty = 0$, *where* U *and* V *are, respectively, continuous and Lipschitz spectral densities. If* $\{U_r\}_{r\ge1}$ *is a sequence of nonnegative definite pseudopolynomials converging uniformly to* U, *then the following statements hold:*

1. *If* $V > 0$, *there is a sequence* $\{V_r\}_{r\ge1}$ *of nonnegative pseudopolynomials such that* $W_r = U_r (V_r)^{-1}$ *satisfies* (5.2b) (*with* $n = 1$) *for all* r *sufficiently large.*

2. *If* $V \not> 0$, *but* (i) U (*hence also* V) *is zero only at a finite number of points on* $\mathbb{T}$, (ii) $W(e^{i\theta}) = 0$ *whenever* $V(e^{i\theta}) = 0$, (iii) $U_r(e^{i\theta}) = 0$ *only if* $U(e^{i\theta}) = 0$, *and* (*iv*) $\sup_r \|U_r/U\|_\infty < \infty$, *there is a sequence* $\{V_r\}_{r\ge1}$ *of nonnegative pseudopolynomials such that* $W_r = U_r(V_r)^{-1}$ *satisfies* $\frac{1}{2\pi}\int_{\mathbb{T}} W_r(z) z^{-k} \mu(dz) = c_k - s_{k,r}$, *where* $s_{k,r} = \sum_{l=1}^{k_r} K_l e^{ik\theta_l}$, $k_r \in \mathbb{N}$, *is at most* $d_r - 1$, $K_1, \dots, K_{k_r}$ *are some nonnegative constants, and* $\theta_1, \dots, \theta_{k_r} \in (-\pi, \pi]$ *with* $\theta_i \ne \theta_j$ *if* $i \ne j$.

In either case, $\{V_r\}_{r\ge1}$ *and* $\{W_r\}_{r\ge1}$ *converge in* $\mathcal{L}^\infty$ *to* V *and* W, *respectively. Moreover, if* $\deg U_r \le 2d_r$, *then* W_r *also satisfies* (5.2a).

Let l^2 denote the (standard) set of all square-summable infinite sequences. Let Γ_c denote the real linear space of all real-valued continuous functions on $\mathbb{T}$. Define the real linear space $\mathcal{F}_c$ as

$$\mathcal{F}_c = \left\{ (q_0, q_1, \dots) \in l^2 \,\middle|\, q_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ik\theta} W(e^{i\theta}) d\theta \ \forall k \ge 0, \text{ for some } W \in \Gamma_c \right\}.$$

Since a continuous function f is uniquely determined by its Fourier coefficients [19, Theorem 2.4] and since the negative Fourier coefficients are merely conjugates of the positive Fourier coefficients whenever f is a real-valued function, $\mathcal{F}_c$ is actually isomorphic to Γ_c. Therefore, we may uniquely identify any element of $\mathcal{F}_c$ with an element of Γ_c, and vice versa. By endowing Γ_c with a topology induced by the supremum norm and endowing $\mathcal{F}_c$ with a topology induced by the norm

$$\|q\| = \operatorname*{ess\,sup}_{\theta\in(-\pi,\pi]} \left| \sum_{k=0}^{\infty} \Re\{q_k e^{ik\theta}\} \right|$$

(since any element of Γ_c is continuous, the infinite sum converges pointwisely for almost all $z \in \mathbb{T}$ [19, Chapter 19]), we in fact have a homeomorphism from $\mathcal{F}_c$ to Γ_c. Moreover, $\mathcal{F}_c$ is then a closed set since Γ_c is. In what follows, we denote the homeomorphic map from $\mathcal{F}_c$ to Γ_c by Q. The map Q is linear: $Q(a_1 f_1 + a_2 f_2) = a_1 Q(f_1) + a_2 Q(f_2)$ for any $a_1, a_2 \in \mathbb{R}$ and any $f_1, f_2 \in \mathcal{F}_c$. We now define some relevant convex subsets of Γ_c and $\mathcal{F}_c$. Define Γ_c^+ to be the convex cone consisting of elements of Γ_c which are nonnegative on $\mathbb{T}$. We define the convex cone $\mathcal{F}_c^+$ analogously to $\mathcal{F}_c$ by replacing Γ_c with Γ_c^+. In a similar manner, we see that $\mathcal{F}_c^+$ is isomorphic to Γ_c^+. Endowing Γ_c^+ (resp., $\mathcal{F}_c^+$) with a topology derived from Γ_c (resp., $\mathcal{F}_c$), we also get that $\mathcal{F}_c^+$ is homeomorphic to Γ_c^+ under Q; i.e., if Q^+ is the restriction of Q to $\mathcal{F}_c^+$, then Q^+ is a homeomorphism from $\mathcal{F}_c^+$ to Γ_c^+. Define D_r to be the subset of $\mathcal{F}_c^+$ consisting of all $q = (q_0, q_1, \dots) \in \mathcal{F}_c^+$ such that $q_k = 0$ for all $k > r$ and the trigonometric polynomial $\sum_{k=0}^{r} \Re\{q_k e^{ik\theta}\} \geq 0$ for all $\theta \in (-\pi, \pi]$. Clearly, $D_s \supset D_r$ if $s > r$. Moreover, since the partial Fejér sums of any $f \in \Gamma_c^+$ are nonnegative pseudopolynomials and approximate f arbitrarily closely (for details on Fejér sums, see [19]), it is immediate that $\overline{\cup_{r\geq 0} D_r} = \mathcal{F}_c^+$.

Recall that a function f on $\mathbb{T}$ is *Lipschitz* if $\|f(e^{i\theta}) - f(e^{i\psi})\|_1 \leq K|\theta - \psi|$ for all $\theta, \psi \in (-\pi, \pi]$ for some positive constant K, and observe that a scalar spectral density $W \in \Gamma_c^+$ can be written as $W = U/V$ a.e., where V is any Lipschitz scalar spectral density and $U = WV$. Let W have the covariance sequence $\mathbf{c} = \{c_0, c_1, \dots\} \in l^2$. Define $\mathbf{c}_r = \operatorname{col}(c_0, c_1, \dots, c_r)$ to be the partial covariance sequence of $\mathbf{c}$ up to the rth term, and let $\mathbf{d}_r$ be an arbitrary element of D_r. The functional $\mathbb{J}^r : D_r \to \mathbb{R} \cup \{\infty\}$, parametrized by $\mathbf{c}_r$ and $\mathbf{d}_r$ is defined as

$$\mathbb{J}^r(q; \mathbf{c}_r, \mathbf{d}_r) = \Re\left\{ \sum_{k=0}^{r} \min\{k+1, 2\} c_k^* q_k - \langle Q_r(\mathbf{d}_r), \log Q_r(q) \rangle \right\}, \tag{5.4}$$

where $\langle f, g \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\theta}) g_*(e^{i\theta}) d\theta$ and Q_r is a map with domain D_r defined by

$$Q_r(q_0, q_1, \dots, q_r, 0, 0, \dots)(e^{i\theta}) = \sum_{k=0}^{r} \Re\{q_k e^{ik\theta}\}.$$

Notice that Q_r can be viewed as the restriction of Q to D_r. The functional $\mathbb{J}^r$ first appeared in [5, 6] and was studied further in [22, 21], so we shall not repeat the analysis of its properties here (actually, our formulation is slightly different. However, it causes no difficulty since the functionals studied in [5, 6, 22, 21] can be recovered by application of the linear invertible transformation $(q_0, q_1, \dots, q_r) \mapsto (q_0, \frac{1}{2}q_1, \dots, \frac{1}{2}q_r)$ and redefining D_r in an obvious way). $\mathbb{J}^r$ has the following properties:

P1. $\mathbb{J}^r$ is strictly convex on D_r and is continuous at all points except the origin.

P2. $\mathbb{J}^r$ has compact sublevel sets and a unique minimizer in D_r.

P3. $\mathbb{J}^r$ is infinitely differentiable along any line lying in the interior of D_r. Moreover, if $\mathbb{J}^r$ has a minimizer q_s which is stationary (i.e., the gradient is zero at q_s) then $\frac{Q(d_r)}{Q(q_s)}$ satisfies $\langle \frac{Q(d_r)}{Q(q_s)}, g_k\rangle = c_k$ for $k = 0, 1, \ldots, r$, where $g_k(z) = z^k$. Note that q_s is always a stationary point whenever $Q(\mathbf{d}_r)$ is positive definite [6, Thm. 4.10].

Let us now consider another functional $\mathbb{J} : \mathcal{F}_c^+ \to \mathbb{R} \cup \{\infty\}$ parametrized by the covariance sequence $\mathbf{c}$ and an infinite sequence $\mathbf{d} \in \mathcal{F}_c^+$ such that $Q(\mathbf{d}) = U$. It is defined as

$$\mathbb{J}(q; \mathbf{c}, \mathbf{d}) = \Re\{\mathrm{H}(q; \mathbf{c}) - \langle Q(\mathbf{d}), \log Q(q)\rangle\}, \tag{5.5}$$

where $\mathrm{H}(\cdot; \mathbf{c})$ is a linear function on $\mathcal{F}_c$, parametrized by $\mathbf{c}$, defined by

$$\mathrm{H}(q; \mathbf{c}) = \lim_{r\to\infty} \sum_{k=0}^{r} \min\{k+1, 2\} c_k^* q_k. \tag{5.6}$$

Since $q \in l^2$ whenever $q \in \mathcal{F}_c$ (recall the definition of $\mathcal{F}_c$) and $\mathbf{c} \in l^2$, it follows that $\sum_{k=0}^{\infty} \min\{k+1, 2\} c_k^* q_k < \infty$. Therefore, $\mathrm{H}(\cdot; \mathbf{c})$ is well defined for all $q \in \mathcal{F}_c$. Let us define the convex set $D_e(\mathbb{J}) = \{q \in \mathcal{F}_c^+ \mid \int_{\mathbb{T}} \log^- Q(q)(z)\mu(dz) < \infty\}$ (where $\log^- x = \max\{0, -\log x\}$); $D_e(\mathbb{J})$ is actually the *effective domain* (see, for example, [25]) of $\mathbb{J}$. Then clearly $\cup_{r\geq 1} D_r \subset D_e(\mathbb{J})$. Since $\|\cdot\|$ is also a norm on $D_e(\mathbb{J})$, we endow $D_e(\mathbb{J})$ with the topology induced by the $\|\cdot\|$ norm (this is precisely the relative topology of $D_e(\mathbb{J})$ as a subset of $\mathcal{F}_c^+$: open sets in $D_e(\mathbb{J})$ are sets of the form $D_e(\mathbb{J}) \cap O$ for any O which is an open set of $\mathcal{F}_c^+$). Continuing along the same line of argument as for $\mathbb{J}^r$, we may verify that $\mathbb{J}$ has property P4 (given below) and property P3 with $\mathbb{J}^r$ and D_r replaced by $\mathbb{J}$ and $D_e(\mathbb{J})$, respectively (an analogue of property P2 need not hold for $\mathbb{J}$ and will not be required in the following).

P4. $\mathbb{J}$ is strictly convex on $D_e(\mathbb{J})$ and continuous on the interior of $D_e(\mathbb{J})$.

In the remaining analysis, let us view $\mathbb{J}_\Psi$ as a convex functional that maps from the convex set $D_e(\mathbb{J})$ to $\mathbb{R}$. We now derive an expression for the directional derivatives of $\mathbb{J}$ following [22, 21]. Define

$$\mathcal{M}_{\mathbf{d}} = \Big\{ q \in D_e(\mathbb{J}) \mid \operatorname*{ess\,sup}_{\theta\in(-\pi,\pi]} Q(\mathbf{d})(e^{i\theta})(Q(q)(e^{i\theta}))^{-1} < \infty \Big\}.$$

Let $q' \in \mathcal{F}_c^+$, $0 < h < 1$, and suppose that $q \in \mathcal{M}_{\mathbf{d}}$. We observe that if $Q(q + h(q'-q))(z) = (1-h)Q(q)(z) + hQ(q')(z_0) = 0$ for some $z_0 \in \mathbb{T}$ and all $0 < h < 1$, then $Q(q)$ and $Q(q')$ must share a zero at the point z_0. On the other hand, if $Q(q+h(q'-q))(z) > 0$ for all $z \in \mathbb{T}$, then $Q(q)$ and $Q(q')$ cannot possibly have a zero in common on $\mathbb{T}$. As a result, by the mean-value theorem of calculus, we obtain

$$\begin{aligned} &Q(\mathbf{d})(z)\frac{\log Q(q+h(q'-q))(z) - \log Q(q)(z)}{h} \\ &\quad = Q(\mathbf{d})(z)\frac{\partial}{\partial v}\log Q(q+v(q'-q))(z)\Big|_{v=\eta(h,z)} \\ &\quad = Q(\mathbf{d})(z)\frac{Q(q') - Q(q)}{Q(q)(z) + \eta(h,z)Q(q'-q)(z)} \end{aligned} \tag{5.7}$$

for all $z \in \mathbb{T}$ such that $Q(q)(z) > 0$ (hence for almost all $z \in \mathbb{T}$ since $q \in \mathcal{M}_{\mathbf{d}}$), where $0 < \eta(h,z) \leq h$. Moreover, $q \in \mathcal{M}_{\mathbf{d}}$ implies that

$$\operatorname*{ess\,sup}_{(h,z)\in[0,\frac{1}{2}]\times\mathbb{T}} \frac{Q(\mathbf{d})(z)}{Q(q)(z) + \eta(h,z)Q(q'-q)(z)} < \infty. \tag{5.8}$$

Now, the directional derivative $\nabla_{q'-q}\mathbb{J}$ at q in the direction $q'-q$ is defined as

$$\nabla_{q'-q}\mathbb{J}(q;\mathbf{c},\mathbf{d}) = \lim_{h\downarrow 0}\frac{\mathbb{J}(q+h(q'-q);\mathbf{c},\mathbf{d})-\mathbb{J}(q;\mathbf{c},\mathbf{d})}{h}. \tag{5.9}$$

Plugging the definition of $\mathbb{J}$ into (5.9), and using (5.7) and (5.8) with the Lebesque dominated convergence theorem to bring h under the integral (this is essentially the same argument used in [22]), and finally evaluating the limit as $h \downarrow 0$, we obtain $\nabla_{q'-q}\mathbb{J}(q;\mathbf{c},\mathbf{d}) = H(q'-q) - \langle \frac{\Psi}{Q(q)}, Q(q') - Q(q)\rangle$. Let $\mathcal{F}_{c,\infty}^{+}$ denote the set of all elements a in $D_e(\mathbb{J})$ for which $Q(a)(e^{i\theta})$ is infinitely differentiable with respect to θ, and suppose that $Q(q)$ is Lipschitz and $Q(q')$ in $\mathcal{F}_{c,\infty}^{+}$. Then $\sum_{k=0}^{r} q_k e^{ik\theta}$ (resp., $\sum_{k=0}^{r} q_k' e^{ik\theta}$) converges uniformly to $Q(q)$ (resp., $Q(q')$) [9, Thm. 2, p. 142]. By plugging in the definition of H and using another application of the Lebesque dominated convergence theorem, we get

$$\nabla_{q'-q}\mathbb{J}(q;\mathbf{c},\mathbf{d}) = \lim_{r\to\infty}\sum_{k=0}^{r}\min\{k+1,2\}\Re\left\{\left(c_k^* - \int_{\mathbb{T}}\frac{Q(\mathbf{d})(z)}{Q(q)(z)}z^k\mu(dz)\right)(q_k'-q_k)\right\}$$

for all $q \in \mathcal{M}_{\mathbf{d}}$ and all $q' \in \mathcal{F}_{c,\infty}^{+}$. However, since $\mathcal{F}_{c,\infty}^{+}$ is dense in $\mathcal{F}_{c}^{+}$ (hence also in $D_e(\mathbb{J})$), the preceding expression for $\nabla_{q'-q}\mathbb{J}(q;\mathbf{c},\mathbf{d})$ is valid for all $q \in \mathcal{M}_{\mathbf{d}}$ and all $q' \in \mathcal{F}_{c}^{+}$. Now, setting $q_{\#} = Q^{-1}(V)$ (hence $Q(q_{\#})$ is Lipschitz and $q_{\#} \in \mathcal{M}_{\mathbf{d}}$), we obtain

$$\sum_{k=0}^{r}\min\{k+1,2\}\left(c_k - \int_{\mathbb{T}}\frac{Q(\mathbf{d})(z)}{Q(q_{\#})(z)}z^{-k}\mu(dz)\right)^{*} = 0\,\forall\, r.$$

It therefore follows that $\nabla_{q'-q_{\#}}\mathbb{J}(q_{\#};\mathbf{c},\mathbf{d}) = 0$ for all $q' \in \mathcal{F}_{c}^{+}$, so $q_{\#}$ is a stationary point of $\mathbb{J}$ and, by property P4, must also be the unique minimizer of $\mathbb{J}$.

Suppose now that $\{\mathbf{d}_r\}_{r\in\mathbb{N}}$ is such that $\lim_{r\to\infty}\|\mathbf{d}_r - \mathbf{d}\| = 0$ with $\mathbf{d} = Q^{-1}(U)$ as defined previously. Let $\mathbb{J}\big|_{D_r}\colon D_r \to \mathbb{R}\cup\{\infty\}$ denote the restriction of $\mathbb{J}$ to D_r defined by

$$\mathbb{J}\big|_{D_r}(q;\mathbf{c},\mathbf{d}) = \Re\left\{\sum_{k=0}^{r}\min\{k+1,2\}c_k^* q_k - \langle Q(\mathbf{d}), \log Q_r(q)\rangle\right\}.$$

Since $Q(d) = U$ is continuous on $\mathbb{T}$ and is zero only on a subset of $\mathbb{T}$ of μ-measure zero, the analysis of $\mathbb{J}^r$ in [5, 6, 22] readily carries over to $\mathbb{J}\,|_{D_r}$ to show that the latter also has properties P1, P2, and P3 (with $\mathbb{J}^r$ replaced by $\mathbb{J}\big|_{D_r}$). Moreover, defining

$$s_r = \arg\min_{q\in D_r}\mathbb{J}\big|_{D_r}(q;\mathbf{c},\mathbf{d}),$$

it is then clear that $\lim_{r\to\infty} s_r = q_{\#}$. Hence, for any $\epsilon > 0$ we will have $\|q_{\#} - s_r\| < \frac{\epsilon}{2}$ by taking large enough r. Then also $\|q - q_{\#}\| \le \|q - s_r\| + \|s_r - q_{\#}\| \le \epsilon$ whenever $\|q - s_r\| \le \frac{\epsilon}{2}$ (recall that $q, q_{\#}, s_r \in D_e(\mathbb{J})$). Therefore, for all r large enough, $\{q \in D_r \mid \|q - s_r\| \le \frac{\epsilon}{2}\} \subset \{q \in D_e(\mathbb{J}) \mid \|q - q_{\#}\| \le \epsilon\}$ and

$$\begin{aligned}\sup_{\{q\in D_r, \|q-s_r\|\le\frac{\epsilon}{2}\}}|\mathbb{J}\big|_{D_r} - \mathbb{J}^r| &= \sup_{\{q\in D_r \mid \|q-s_r\|\le\frac{\epsilon}{2}\}}|\langle Q(\mathbf{d}) - Q(\mathbf{d}_r), \log Q_r(q)\rangle| \\ &\le \sup_{\{q\in D_e(\mathbb{J})\mid\|q-q_{\#}\|\le\epsilon\}}|\langle Q(\mathbf{d}) - Q(\mathbf{d}_r), \log Q(q)\rangle| \\ &\le \|Q(\mathbf{d}) - Q(\mathbf{d}_r)\|D,\end{aligned}$$

where $D = \max_{\{q\in D_e(\mathbb{J})|\|q-q_\#\|\le\epsilon\}} |\int_{-\pi}^{\pi} \log Q(q)(e^{i\theta})d\theta|$. Since $\lim_{r\to\infty} \|Q(\mathbf{d})-Q(\mathbf{d}_r)\| = 0$, it then follows that

$$\lim_{r\to\infty} \max_{\{q\in D_r|\|q-s_r\|\le\frac{\epsilon}{2}\}} \left|\mathbb{J}\big|_{D_r}(q;\mathbf{c},\mathbf{d}) - \mathbb{J}^r(q;\mathbf{c}_r,\mathbf{d}_r)\right| = 0$$

for any $\epsilon > 0$. Due to properties P1, P2, and P3 of $\mathbb{J}\big|_{D_r}$ and $\mathbb{J}^r$, the preceding limit implies that $\lim_{r\to\infty}\|s_r - q_{\#,r}\| = 0$, where $q_{\#,r} = \arg\min_{q\in D_r} \mathbb{J}^r(q;\mathbf{c}_r,\mathbf{d}_r)$; for the details see a proof of [22, Theorem 13] (replace Ψ, Ψ_k, $\mathbb{J}_\Psi$, and $\mathbb{J}_{\Psi_k}$ with $Q(\mathbf{d})$, $Q(\mathbf{d}_k)$, $\mathbb{J}\big|_{D_k}$ and $\mathbb{J}^k$, respectively). Furthermore, since $\|q_\# - q_{\#,r}\| \le \|q_\# - s_r\| + \|s_r - q_{\#,r}\|$ and both terms on the right tend to 0 as $r\to\infty$, we get $\lim_{r\to\infty}\|q_\# - q_{\#,r}\| = 0$ and $\lim_{r\to\infty}\|Q(q_\#) - Q(q_{\#,r})\| = 0$. Consequently, if $V > 0$, then $q_\#$ is in the interior of $D_e(\mathbb{J})$ and the same is true for $q_{\#,r}$ for all r sufficiently large. Therefore, W_r satisfies (5.2) for all r sufficiently large, and $W_r = \frac{Q(\mathbf{d}_r)}{Q(q_{\#,r})} \to W = \frac{Q(\mathbf{d})}{Q(q_\#)}$ uniformly as $r\to\infty$.

If $V \not> 0$, but assumptions (i)–(iv) in Theorem 5.2 (2) are satisfied, convergence of $\{W_r\}$ to W in $\mathcal{L}^\infty$ can again be established. Let $V_r = Q(q_{\#,r})$. Since $V_r \to V$ uniformly, we have that for any $\delta > 0$ $\exists R(\delta)\in\mathbb{N}$ such that $-\delta < V_r(e^{i\theta}) - V(e^{i\theta}) < \delta$ for all θ and for all $r > R(\delta)$. However, since $V, V_r \ge 0$ and V, V_r are continuous for all r, and $\exists R'\in\mathbb{N}$ such that $V_r(e^{i\theta}) > 0$ for all $r > R'$ whenever $V(e^{i\theta}) > 0$ (due to assumptions (i) and (iii) and the observation that all zeros of V_r on $\mathbb{T}$ are also zeros of U_r [22, 21] and $V_r \to V$ uniformly), there is a *continuous* function $f_\delta : \mathbb{T}\to[0,1]$ such that (a) $V - \delta f_\delta \ge 0$, (b) $(V-\delta f_\delta)(e^{i\theta}) = 0$ if and only if $V(e^{i\theta}) = 0$, and (c) $-\delta f_\delta < V_r - V < \delta$ for all θ and for all $r > \max\{R(\delta), R'\}$. In particular, letting $S(\delta) = \max\{R(\delta), R'\}$ we may always take f_δ to be $f_\delta = \sup_{r>S(\delta)}(V - V_r)/\delta$ (and we shall do so in what follows) and satisfy all the requirements. Then $V/V_r < V/(V-\delta f_\delta) = 1/(1-\delta f_\delta/V)$ a.e. $\mathbb{T}$ for all $r > S(\delta)$. Moreover, $1/(1-g_\delta/V)\in\mathcal{L}^\infty$ for all δ sufficiently small, where $g_\delta = \delta f_\delta = \sup_{r>S(\delta)}(V - V_r)$. To see this, first observe that property (a) of f_δ implies $g_\delta/V \le 1$ a.e. $\mathbb{T}$. Then we observe that $g_\delta \downarrow 0$ (i.e., g converges monotonically to 0) uniformly on $\mathbb{T}$ as $\delta\downarrow 0$ and hence, since also $g_\delta \le V$ for all $\delta > 0$, for sufficiently small δ we will have $g_\delta < V$ for all $z\in\mathbb{T}$ except those for which $V(z) = 0$. Therefore, $\|g_\delta/V\|_\infty < 1$ and $1/(1-g_\delta/V) < \infty$ a.e. for sufficiently small δ, as claimed, and it follows that $\sup_{r>S(\delta)}\|V/V_r\|_\infty < \infty$. Now, let $N_{\delta'}(z) = \{y\in\mathbb{T} \mid |z-y| < \delta'\}$ for any $z\in\mathbb{T}$ and $\delta' > 0$. Then we note that V/V_r (resp., $V/V_r - 1$) converges uniformly to 1 (resp., 0) on $\mathbb{T}\setminus\cup_{k=1}^m N_{\delta'}(z_k)$, where $z_1,\ldots,z_m\in\mathbb{T}$ are all zeros of V, for δ' small enough such that $\cup_{k=1}^m N_{\delta'}(z_k)$ is a strict subset of $\mathbb{T}$. By uniform convergence of U_r to U and assumption (iv), an analogous remark is also true for U_r/U. Next, we make the observation that $|W_r - W| = |U_r/V_r - U/V| \le W|U_r/U - 1|V/V_r + W|V/V_r - 1|$. Then the properties of V/V_r and U_r/U just stated, along with assumptions (i)–(ii) and the continuity of W, imply that both $W|U_r/U-1|V/V_r$ and $W|V/V_r-1|$ converge uniformly to 0 as $r\to\infty$. To see this, let us consider the term $W|U_r/U-1|V/V_r$ and let R'' be large enough such that $M = \sup_{r>R''}\|U_r/U-1\|_\infty\|V/V_r\|_\infty < \infty$ (recall that $\sup_{r>S(\delta)}\|V/V_r\|_\infty < \infty$). Then, by assumption (ii) and the continuity of W, for any $\epsilon > 0$ we may choose $\delta' > 0$ small enough such that $\sup_{\{z\in\cup_{k=1}^m N_{\delta'}(z_k)\}} W(z) < \epsilon/M$ a.e. followed by choosing $r > R''$ which is large enough such that $\sup_{\{z\in\mathbb{T}\setminus\cup_{k=1}^m N_{\delta'}(z_k)\}}|U_r(z)/U(z)-1|V(z)/V_r(z) < \epsilon/\|W\|_\infty$. In other words, for any $\epsilon > 0$ $\exists R'''(\epsilon)$ such that $\|W|U_r/U-1|V/V_r\|_\infty < \epsilon$ for all $r > R'''(\epsilon)$. The same line of arguments may then be applied to $W|V/V_r-1|$. In conclusion, we again have $W_r \overset{\mathcal{L}^\infty}{\to} W$ as $r\to\infty$. The fact that $\frac{1}{2\pi}\int_{\mathbb{T}} W_r(z)z^{-k}\mu(dz) = c_k - s_{k,r}$ as stated in the theorem has been shown in the proof of [21, Thm. 8].

Finally, we note that the preceding analysis remains valid if $\mathbf{d}_r \in D_r$ is replaced with $\mathbf{d}_r \in \cup_{k>r} D_k$. To see this, let $\mathbf{d}_r$ be any pseudopolynomial, *not necessarily* of degree at most r, and define $\mathbb{J}^r$ as in (5.4). Then it may be verified that $\mathbb{J}^r$ again has properties P1, P2, and P3, and if q_s is a stationary point minimizer of $\mathbb{J}^r$, then $\frac{Q(d_r)}{Q(q_s)}$ once more satisfies $\langle \frac{Q(d_r)}{Q(q_s)}, g_k \rangle = c_k$ for $k = 0, 1, \ldots, r$. The rest of the analysis follows mutatis mutandis. This completes the proof of Theorem 5.2. □

An important consequence of the theorem combined with Corollary 4.6 and [1, Theorem 2] is the following.

COROLLARY 5.3. *Suppose $W \in \mathcal{L}^1$ is a continuous spectral density with a finite number of zeros on $\mathbb{T}$ and let $\{W_r\}_{r\geq 1}$ be a sequence as defined in Theorem* 5.2. *Then* $\lim_{r\to\infty} \|\Phi(W_r) - \Phi(W)\|_2 = 0$. *If, in addition, $W > 0$ and $\frac{d}{d\theta} W(e^{i\theta}) \in \mathcal{L}^2$, then also* $\lim_{r\to\infty} \|\Phi(W_r) - \Phi(W)\|_\infty = 0$.

Notice that the corollary gives a weaker condition for convergence in the $\|\cdot\|_\infty$ norm than analyticity (resp., rationality and boundedness) and positivity of W given in [8, Thm. 1] (resp., [20, Thm. 3.4]) for the Szegö–Levinson algorithm and does not restrict $\Phi(W_r)$ to have all its zeros at the origin. Note that we say W is analytic if it can be continued analytically from all points in $\mathbb{T}$.

5.2. The matrix case. For a matrix-valued spectral density W, the situation is slightly more complicated. If W is a matrix-valued Lipschitz spectral density, then we may write $W = (W^{-1})^{-1} = \det(W)\mathrm{adj}(W)^{-1}$, where $\mathrm{adj}(W)$ denotes the adjoint of W. Define $U = P \det W$ and $V = P\,\mathrm{adj}(W)$ for any arbitrary scalar spectral density P which is Lipschitz and positive definite. Then U and V are Lipschitz. The representation $W = UV^{-1}$ a.e. can be viewed as the matricial counterpart of the scalar fractional representation. If W is positive definite, then so is V and, in this case, by suitably redefining the sets Γ_c, $\mathcal{F}_c$, Γ_c^+, $\mathcal{F}_c^+$, D_r and the associated norms with their respective matricial counterparts, as well as suitably modifying the functionals $\mathbb{J}^r$ and $\mathbb{J}$ (see [4, eq. (V.5), p. 2180]), it is a relatively straightforward, but tedious, exercise to adapt the analysis developed in deriving Theorem 5.2 to the matrix case. Then we may show the following counterpart of Theorem 5.2.

THEOREM 5.4. *Let $W = UV^{-1} \in \mathcal{L}^1_{n\times n}$ be a matrix-valued positive definite Lipschitz spectral density, where $U = P \det W$ and $V = P\mathrm{adj}(W)$ for some positive definite Lipschitz scalar spectral density P. If $\{U_r\}_{r\geq 1}$ is a sequence of positive definite pseudopolynomials converging uniformly to U, then there exists a (unique) sequence $\{V_r\}_{r\geq 1}$ of positive definite pseudopolynomials such that*

1. *$W_r = U_r(V_r)^{-1}$ satisfies* (5.2b) *for all r. If, in addition,* $\deg U_r \leq 2nd_r$, *then* (5.2a) *is also satisfied.*
2. *$\{V_r\}_{r\geq 1}$ and $\{W_r\}_{r\geq 1}$ converge uniformly to V and W, respectively.*

We then get the following from Corollary 4.6 and [1, Thm. 2].

COROLLARY 5.5. *Let $\{W_r\}_{r\in\mathbb{N}}$ be a sequence as defined in Theorem* 5.4. *Then* $\lim_{r\to\infty} \|\Phi(W_r) - \Phi(W)\|_2 = 0$. *If, in addition, $W > 0$ and $\frac{d}{d\theta} W(e^{i\theta}) \in \mathcal{L}^2_{n\times n}$, then also* $\lim_{r\to\infty} \|\Phi(W_r) - \Phi(W)\|_\infty = 0$.

It is plausible that Theorem 5.4 and Corollary 5.5 can be extended to the case where U has zeros on $\mathbb{T}$. However, to do this, we must allow some spectral zeros (see [4]) of W_r to be on $\mathbb{T}$. This is currently an open problem.

6. A spectral factorization algorithm. We now introduce a new algorithm for spectral factorization of a special class $\mathcal{W}_n$ of spectral densities. $\mathcal{W}_1$ denotes the set of spectral densities $W \in \mathcal{L}^1$ which can be continued analytically from every point $z \in \mathbb{T}$, except from a finite number of points $w_k = e^{i\theta_k}$, $k = 1, \ldots, M$, for which

$W(w_k)=0$ and $\lim_{z\in\mathbb{T},z\to w_k}\frac{|z-w_k|^{m_k}}{W(z)}<\infty$ for some integer $m_k\geq 1$. For $n>1$, $\mathcal{W}_n$ denotes the set of spectral densities in $\mathcal{L}^1_{n\times n}$ which are positive definite and can be continued analytically from every point on $\mathbb{T}$. We state the algorithm below followed by a discussion of the steps involved and a convergence analysis.

6.1. The algorithm.

Given: A spectral density $W\in\mathcal{W}_n$, the desired accuracy $\epsilon>0$ and maximum number of iterations $r_{\max}$.

Initialize: Normalize W so that $c_0=I$. Let $e^{i\lambda_1},\dots,e^{i\lambda_L}\in\mathbb{T}$ be the local minima of $\det W$ satisfying $0\leq\det W(e^{i\lambda_l})\leq\alpha$ ($\alpha\in\mathbb{R}$, $\alpha\geq 0.2$. The rule of thumb is $\alpha=0.2$). Let $\mathcal{V}_1$ be all points in $\{e^{i\lambda_1},\dots,e^{i\lambda_L}\}$ from which $\det W$ does not have an analytic continuation and $\mathcal{V}_2=\{e^{i\lambda_1},\dots,e^{i\lambda_L}\}\backslash\mathcal{V}_1$. For $l=1,\dots,L$, define $m_l=\min\{k\in\mathbb{N}\mid\lim_{z\in\mathbb{T},z\to e^{i\lambda_l}}\frac{|z-e^{i\lambda_l}|^k}{W(z)}<\infty\}$ if $e^{i\lambda_l}\in\mathcal{V}_1$ and $m_l=\min\{k\in\mathbb{N}\mid D^k_\theta\det W(e^{i\lambda_l})\neq 0\}$ if $e^{i\lambda_l}\in\mathcal{V}_2$ (here $D^m_\theta\det W(e^{i\lambda_l})=\frac{d^m\det W(e^{i\theta})}{d\theta^m}\big|_{\theta=\lambda_l}$). Let $\eta_0(z)=\prod_{l=1}^L(z-v_l)^{\frac{m_l}{2}}$, where $v_l=\max\{0,r_l\}e^{i\lambda_l}$ $\big($with $r_l=1-\big(\frac{\det W(e^{i\lambda_l})}{D^{m_l}_\theta\det W(e^{i\lambda_l})}\big)^{\frac{1}{m_l}}\big)$ if $e^{i\lambda_l}\in\mathcal{V}_2$, and $v_l=e^{i\lambda_l}$ if $e^{i\lambda_l}\in\mathcal{V}_1$. Set $r=1$, $d_0=L$, and compute $c_0,c_1,\dots,c_L$ and the outer polynomial matrix $R_0=\phi(V_0)$, where $V_0=Q(\arg\min_{q\in D_{d_0}}\mathbb{J}^{d_0}(q))$ (see section 5).

Step 1. Select a point $z_r\in\mathbb{D}$. Then

1. if W is symmetric (i.e., $W(e^{-i\theta})=W(e^{i\theta})$) or $\theta_r\notin\{0,\pi\}$, set $d_r=d_{r-1}+2$ and $\eta_r=\eta_{r-1}(z-z_r)(z-z_r^*)$; otherwise
2. set $d_r=d_{r-1}+1$, and $\eta_r=\eta_{r-1}(z-z_r)$.

Step 2. Compute $c_{d_{r-1}+1},\dots,c_{d_r}$ and the outer matrix polynomial $R_r=\phi(V_r)$, where $V_r=Q(\arg\min_{q\in D_{d_r}}\mathbb{J}^{d_r}(q))$.

Step 3. Compute $e_r=\frac{1}{2}\|W-W_r\|_1+\frac{1}{2}\|\Phi(W_r)-\Phi(W_{r-1})\|_2$, where $W_r=\eta_{r*}\eta_r(R_rR_{r*})^{-1}$. If $e_r>\epsilon$ and $r\leq r_{\max}$, set $r=r+1$ and return to Step 1.

End: $z^{d_r}\eta_{r*}(R_r)^{-1}\sqrt{c_0}$ is the approximate CSF.

Computation of the polynomial matrix R_r, $r=0,1,2,\dots$, is given in [12, 4]. By the analysis of [22], the method of [12] is applicable to the case when η_r has zeros on $\mathbb{T}$ if the optimization problem defined in [12] has a stationary point minimizer (which is true whenever $V_r>0$); that the minimizer is in fact always stationary is shown in [23]. The main idea of the algorithm is to find a sequence $z_1,z_2,\dots\in\overline{\mathbb{D}}$ such that $W_r=U_rV_r^{-1}$ satisfies (5.2) and $W_r\to W$ in $\mathcal{L}^\infty_{n\times n}$, where $U_r(z)=\eta_0\eta_{0*}\Pi_{k=1}^{d_r-L}(z-z_k)(z-z_k)_*$. It works as follows. Suppose $e^{i\lambda_l}\in\mathcal{V}_2$; then $\det W$ has an analytic continuation to some open set containing $e^{i\lambda_l}$. Moreover, if $D^m_\theta\det W(e^{i\lambda})=0$ for $m=1,\dots,l$, then also $(\det W)^{(m)}(e^{i\theta})=(\det W)^{(m)}(z)\big|_{z=e^{i\lambda_l}}=0$ ($(\det W)^{(m)}$ denotes the mth derivative of the analytic continuation of $\det W$). Since $e^{i\lambda_l}$ is a local minimum, we have that $D^{m_l}_\theta\det W(e^{i\lambda_l})>0$. Let us take care of points $z\in\mathbb{T}$ for which $\det W(z)\approx 0$. We take these to be the points $e^{i\lambda_1},\dots,e^{i\lambda_L}$ as defined in the algorithm. For $e^{i\lambda_l}\in\mathcal{V}_2$, the Taylor series expansion of $\det W(z)$ about $e^{i\lambda_l}$ gives $\det W(z)\approx\det W(e^{i\lambda_l})+(-ie^{-i\theta})^{m_l}D^{m_l}_\theta\det W(e^{i\lambda_l})(z-e^{i\lambda_l})^{m_l}$ for z sufficiently close to $e^{i\lambda_l}$. To estimate a zero of $\det W(z)$ about $e^{i\lambda_l}$, we set $\det W(z)=0$ to get $|z-e^{i\lambda_l}|\approx\big(\frac{\det W(e^{i\lambda_l})}{D^{m_l}_\theta\det W(e^{i\lambda_l})}\big)^{\frac{1}{m_l}}$. Assuming $z_l=r_le^{i\lambda_l}$ with $0\leq r_l\leq 1$ for our zero estimate, we obtain $|1-r_l|=\big(\frac{\det W(e^{i\lambda_l})}{D^{m_l}_\theta\det W(e^{i\lambda_l})}\big)^{\frac{1}{m_l}}$. Thus, we choose $r_l=1-\big(\frac{\det W(e^{i\lambda_l})}{D^{m_l}_\theta\det W(e^{i\lambda_l})}\big)^{\frac{1}{m_l}}$ and set $v_l=\max\{0,r_l\}e^{i\lambda_l}$ (hence, automatically $v_l=e^{i\lambda_l}$ if

$\det W(e^{i\lambda_l}) = 0$). As elaborated in section 1, points $z \in \mathbb{T}$ for which $\det W(z) \approx 0$ slow down convergence significantly due to slow decay of the so-called Schur parameters [14]. The main idea in the algorithm is to reduce their influence by suitably placing a zero of η_0 in their vicinity as in [14, 5, 6], but here we allow the degree of the approximation W_r to increase as required.

Remark 6.1. $v_1, \dots, v_L$ in $\mathbb{D}$ actually serve as estimates of zeros of $\det W$ in some open annulus $\{z \in \mathbb{C} \mid 1 - \delta < |z| < 1\}$ $(0 < \delta < 1)$. As such, other schemes can be used to determine these points. The "rule of thumb" $\alpha = 0.2$ is based on the subjective view that it is "not too small" and "not too large." If convergence of the algorithm is slow, say, $e_r > 10^{-2}$ in the first few (5–10) iterations, one may try restarting the algorithm with α increased, or the next remark may be taken into consideration.

Remark 6.2. If $\det W$ has thin and sharp "spectral line"–like peaks, then the algorithm may perform poorly. This is because such a peak indicates the possible presence of a (noncancelling) pole and zero close to each other and to the unit circle, while the zero is not included in η_0 [14]. To remedy the situation, let $H \in \mathcal{H}^2$ be a scalar notch filter with narrow stop bands around frequencies corresponding to the peaks, $P = H_* H$, and apply the algorithm to $W' = WP$. Then $\Phi(W) \approx \frac{\Phi(W')}{H}$ in $\mathcal{H}^2$.

The following theorem gives a requirement on $z_1, z_2, \dots$ for convergence:

THEOREM 6.3. *Let η_r be as defined in the algorithm. Suppose that the polynomial $\rho_r(z) = z^{d_r} \frac{\eta_{r*}(z)}{\eta_{0*}(z)} \overset{\mathcal{L}^\infty}{\to} \rho$, where ρ is continuous and has no zeros on $\mathbb{T}$. Let $W \in \mathcal{W}_1$ (resp., $\mathcal{W}_n$, $n > 1$), $U_r(z) = \eta_{r*}(z)\eta_r(z)$, and let V_r be as defined in Theorem* 5.2 *(resp., Theorem* 5.4*). Then $z^{d_r}\eta_{r*}(\Phi(V_r))^{-1}$ converges to $\Phi(W)$ in $\mathcal{H}^2$ (resp., $\mathcal{H}^2_{n\times n}$) and also in $\mathcal{H}^\infty$ (resp., $\mathcal{H}^\infty_{n\times n}$) if $W > 0$ and, when $n = 1$, $\frac{d}{d\theta}W(e^{i\theta}) \in \mathcal{L}^2$.*

Proof. Assume that W has been normalized so that $c_0 = I$. Let $U = \eta_0 \rho_* \rho \eta_{0*}$ and define V by $V(e^{i\theta}) = \lim_{\lambda \to \theta} U_*(e^{i\lambda})U(e^{i\lambda})W(e^{i\lambda})^{-1}$. Then, by definition, $V \in \mathcal{L}^\infty_{n\times n}$ and $\|W - UV^{-1}\|_\infty = 0$. Note that $\eta_r = z^{d_r}\eta_0\rho_{r*}$ and let $U_r = \eta_{r*}\eta_r$. Then, since $\rho_r \overset{\mathcal{L}^\infty}{\to} \rho$, we have that $U_r \overset{\mathcal{L}^\infty}{\to} U$. By Theorem 5.2 and Corollary 5.3, or Theorem 5.4 and Corollary 5.5 (whichever pair is applicable), it follows that V_r and $W_r = U_r V_r^{-1}$ converge in $\mathcal{L}^\infty_{n\times n}$, respectively, to V and W, and $\Phi(W_r)$ converges to $\Phi(W)$ in $\mathcal{H}^2_{n\times n}$ and also in $\mathcal{H}^\infty_{n\times n}$ if $W > 0$ and, when $n = 1$, $\frac{d}{d\theta}W(e^{i\theta}) \in \mathcal{L}^2$ (by the definition of $\mathcal{W}_n$, $\frac{d}{d\theta}W(e^{i\theta}) \in \mathcal{L}^2_{n\times n}$ is automatically satisfied when $W \in \mathcal{W}_n$, $n > 1$). Since $\Phi(W_r) = \Phi(U_r)(\Phi(V_r))^{-1} = z^{d_r}\eta_{r*}(\Phi(V_r))^{-1}$, scaling back by multiplication of both $\Phi(W)$ and $\Phi(W_r)$ on the right with $\sqrt{c_0}$ gives the desired result. ∎

Remark 6.4. Clearly, if $z_r = 0$ for all $r > R$ $(R \in \mathbb{N})$, then ρ_r converges uniformly to the analytic function $\rho = \prod_{k=1}^{R}(1 - z_k^* z)$, and the algorithm converges.

6.2. General approximation strategy. The algorithm requires $W \in \mathcal{W}_n$ for some $n \in \mathbb{N}$. If this is not the case, but W is continuous and has a finite number of zeros, then the strategy would be to first construct an approximating analytic spectral density (which need not be rational) in $\mathcal{L}^\infty_{n\times n}$. Then we apply the spectral factorization algorithm to the approximation to obtain an approximate CSF of $\Phi(W)$. The fact that the analytic approximation does not have to be rational affords us flexibility in choosing a set of basis functions for the approximation.

6.3. Heuristic scheme for selection of spectral zeros. In Theorem 6.3 we gave an explicit condition on the spectral zeros $z_1, z_2, \dots$ for the spectral factorization algorithm to converge, and we mentioned a particular situation when this condition is automatically met. In the following we give an intuitive heuristic scheme for choosing

$z_1, z_2, \ldots$ for scalar W. The idea goes as follows. For each r (including $r = 0$) we have at Step 1 that $\int_{-\pi}^{\pi} \big(W_r(e^{i\theta}) - W(e^{i\theta})\big)\, d\theta = 0$. If $W_r - W$ is not identically zero (for which the algorithm then terminates), then it is easy to show, using the mean-value theorem of calculus, that $\exists\theta$ such that $W_r(e^{i\theta}) - W(e^{i\theta}) > 0$. Since a zero of W_r can decrease the magnitude of W_r in certain regions of $\mathbb{T}$, the main idea now is to try to reduce the excess (or overshoot) of W_r over W at a point θ_r for which the excess is relatively large (preferably the largest). If W is not symmetric or $\theta_r \in \{0, \pi\}$, then we place a zero at $z_r = Re^{i\theta_r}$ (with $0 < R < 1$ so that $z_r \in \mathbb{D}$) such that $\frac{W_{r-1}(z)|z-z_r|^2}{W(z)}\big|_{z=e^{i\theta_r}} = \frac{W_{r-1}(e^{i\theta_r})}{W(e^{i\theta_r})}(1-R)^2 = 1$. From the last equality we obtain the required value of R for Step 1. In the case when W is symmetric and $\theta_r \notin \{0, \pi\}$, we must place two zeros at z_r and z_r^* to ensure W_r is also symmetric. By a procedure similar to the symmetric case, we find that a quartic equation $|1-R|^2|1-Re^{-i2\theta_r}|^2 - \frac{W_{r-1}(e^{i\theta_r})}{W(e^{i\theta_r})} = 0$ must be solved for R and a real solution satisfying $0 < R < 1$ is chosen. It is easy to see, since $\frac{W_{r-1}(e^{i\theta_r})}{W(e^{i\theta_r})} < 1$, that the quartic solution always has such a solution. It is not theoretically guaranteed that spectral zeros chosen by the scheme satisfy the requirements of Theorem 6.3 for convergence. However, in accordance with Remark 6.1, we may always proceed with the heuristic for a *finite* number of steps before terminating the selection by setting $z_r = 0$ for the remaining iterations. Simulation results, to be given in section 7, indicate however that this heuristic seems to work reasonably well.

6.4. Reduction of computational time. The computationally intensive part in the proposed algorithm is Step 2 for computing R_r. This is because the homotopy continuation algorithms described in [12, 4] involve solving a finite sequence of convex optimization problems. However, it is important to note that the computation can be substantially reduced at higher iterations down to solving only *one* convex optimization problem. To see this, consider the case when the algorithm is convergent and $E_r = \|z^{d_r}\eta_{r*}R_r^{-1} - z^{d_{r-1}}(\eta_{r-1})_*R_{r-1}^{-1}\|_\infty \to 0$ as $r \to \infty$. Since R_r is invertible a.e. $\mathbb{T}$ and $\sup_{r\geq 1}\|R_r\|_\infty < \infty$, we also have $E_r' = \|z^{d_r}\eta_{r*}R_{r-1} - z^{d_{r-1}}(\eta_{r-1})_*R_r\|_\infty \to 0$ as $r \to \infty$ (by noting $E_r' \leq \|R_{r-1}\|_\infty E_r \|R_r\|_\infty$). Assuming for the moment that $z_r \in \mathbb{R}$ and recalling that $\eta_r(z) = (z - z_r)\eta_{r-1}(z)$ and $d_r = d_{r-1} + 1$, we get that $\|(1 - z_r^* z)R_{r-1} - R_r\|_\infty \to 0$. Therefore, $\|(1 - z_r^* z)R_{r-1} - R_r\|_\infty$ will be small for all r sufficiently large. In that case, we simply set the homotopy step-size parameter ρ (resp., λ) in [12] (resp., [4]) to 1 and use the coefficients of $(1 - z_r^* z)R_{r-1}$ as an initial point in the algorithm for solving the *single* convex optimization problem which gives the coefficients of R_r. If $z_r \notin \mathbb{R}$, then replace $(1 - z_r^* z)$ with $(1 - z_r z)(1 - z_r^* z)$. This reduction scheme can be executed when $e_r \leq \delta$ for some small $\delta > 0$.

If required, further reduction is possible. We note that the Hessian of the functional to be minimized has a Hankel-plus-Toeplitz structure which can be inverted (or solved if it is the coefficient matrix in a system of linear equations) with fast algorithms given in [16, 17]. More important, however, is that these algorithms have *parallel* (i.e., Schur-type) versions which can be implemented on parallel computers.

7. Numerical examples. In this section we apply the new spectral factorization algorithm and heuristic of the last section to compute approximate CSFs of some rational and nonrational noncoercive spectral densities. In each example, three different simulations are carried out as follows:

1. Simulation A: The spectral zero selection heuristic is applied at Step 1 until termination of the algorithm.

TABLE 7.1
Simulation results.

	Number of iterations	Final value of e_r	Degree of approximation	Running time (seconds)
Simulation A	10	4.12909×10^{-5}	10	8.53
Simulation B	10	4.14906×10^{-5}	10	7.593
Simulation C	151	1.98136×10^{-3}	151	73.063

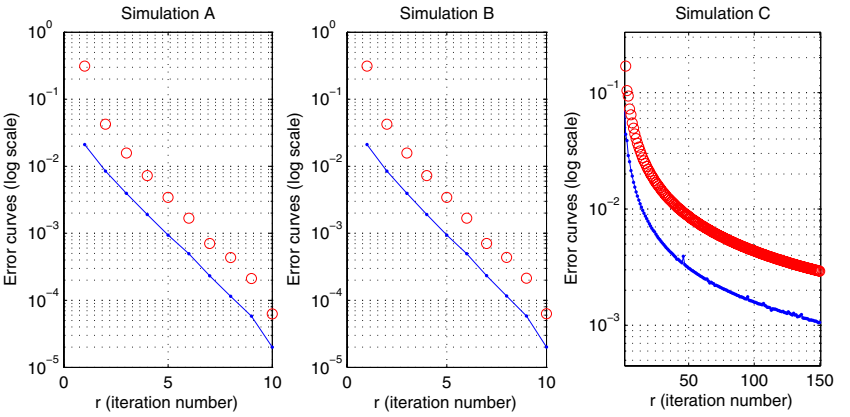


FIG. 7.1. *Plots of Simulation A (left), B (center), C (right):* $\|W - W_r\|_1$ *(dashed-dotted lines),* $\|\Phi(W_r) - \Phi(W_{r-1})\|_2$ *(circle).*

2. Simulation B: The spectral zero selection heuristic is applied at Step 1 for a finite, prespecified number of steps after which z_r is set to 0.

3. Simulation C: All of **Initialize** is skipped except the computation of c_0, z_r is set to 0 in Step 1 for all r, and R_r in Step 2 is computed recursively via the Szegö–Levinson algorithm. Step 3 is unaltered.

We set $\epsilon = 10^{-4}$ in all simulations and apply the computational reduction scheme of section 6.4 in Simulation A and B when $e_r \leq 10^{-2}$ is satisfied. The algorithm was implemented in MATLAB and executed on a computer with a Pentium 4 processor with a clock speed of 3.2 GHz and 1 GB of RAM.

Example 7.1. Consider the rational spectral density $W(e^{i\theta}) = \frac{2+\cos\theta-2\cos 2\theta}{24.1-18.9\cos\theta+2\cos 2\theta}$ which is noncoercive with a zero at $z = -1$. The exact CSF of W is known to be $\Phi(W)(z) = -\sqrt{10}\frac{(z-2)(z+1)}{(z-4)(z-5)}$. The results of Simulation A, B, and C are shown in Table 7.1 and Figure 7.1 (in simulation B, only five zeros are selected with the heuristic scheme). The exact error $\|\Phi(W_{10}) - \Phi(W)\|_2$ for Simulation A was 4.74526×10^{-5}.

Example 7.2. The *Kolmogorov* spectral density, which is the spectral density of a continuous time stochastic process arising in the study of turbulence, is defined along the imaginary axis as $W_K(i\omega;\sigma) = \frac{1}{\sqrt{1-\sigma(i\omega)^2}}$, where σ is a positive parameter. To use our approach, we first transform the spectral density from the imaginary axis to the unit circle via the (invertible) bilinear transformation $e^{i\theta} = \frac{1-i\omega}{1+i\omega}$. After applying the transformation we get a spectral density W_K^d on $\mathbb{T}$ given by $W_K^d(e^{i\theta};\sigma) = \sqrt{\frac{1+\cos\theta}{1+\cos\theta+\sigma(1-\cos\theta)}}$. Notice that W_K^d has a zero at $z = -1$ and cannot be continued analytically from that point. Setting $\sigma = 2$, we show results from Simulation A, B, and C are shown in Table 7.2 and Figure 7.2 (in Simulation B, only 10 zeros are selected with the heuristic scheme).

TABLE 7.2
Simulation results.

	Number of iterations	Final value of e_r	Degree of approximation	Running time (seconds)
Simulation A	22	9.32203×10^{-5}	43	239.269
Simulation B	29	9.31331×10^{-5}	39	294.741
Simulation C	151	2.11347×10^{-3}	151	44.86

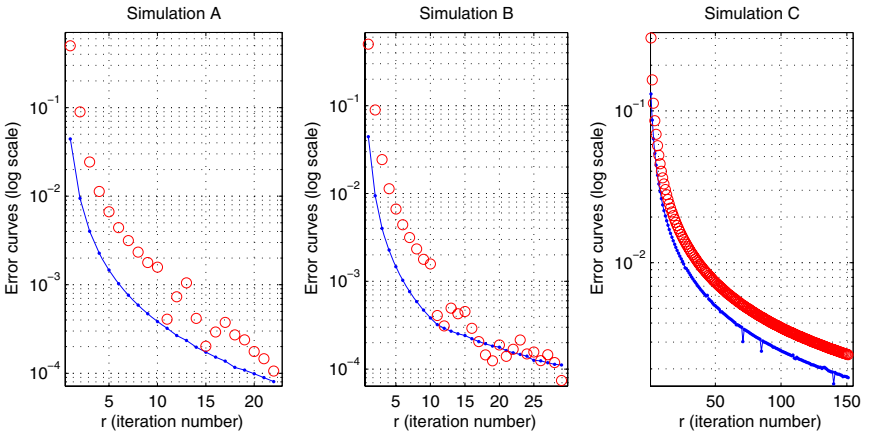


FIG. 7.2. *Plots of Simulation A (left), B (center), C (right):* $\|W - W_r\|_1$ *(dashed-dotted lines),* $\|\Phi(W_r) - \Phi(W_{r-1})\|_2$ *(circles).*

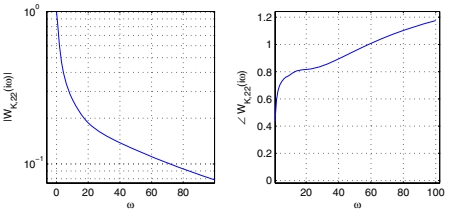


FIG. 7.3. *Frequency response of approximate CSF from Simulation A.*

Transforming the approximate CSF of Simulation A from the unit circle back to the imaginary axis gives us the frequency response shown in Figure 7.3.

Example 7.3. The *von Karman* spectral density [2, p. 73] is the spectral density of a continuous time stochastic process defined along the imaginary axis as

$$W_{vK}(i\omega;\sigma) = 2\sigma \frac{1 - \frac{8}{3}\sigma^2(1.339)^2(i\omega)^2}{(1 - \sigma^2(1.339)^2(i\omega)^2)^{\frac{11}{6}}},$$

where σ is a positive parameter. It is often used as a substitute for the Kolmogorov power spectral density of the previous example. After a transformation from the real line to the unit circle, we obtain a spectral density W^d_{vK} on $\mathbb{T}$ given by

$$W^d_{vK}(e^{i\theta};\sigma) = 2\sigma \frac{1 + \cos\theta + \frac{8}{3}\sigma^2(1.339)^2(1 - \cos\theta)}{(1 + \cos\theta + \sigma^2(1.339)^2(1 - \cos\theta))^{\frac{11}{6}}} (1 + \cos\theta)^{\frac{5}{6}}.$$

TABLE 7.3
Simulation results.

	Number of iterations	Final value of e_r	Degree of approximation	Running time (seconds)
Simulation A	30	6.54517×10^{-5}	49	926.656
Simulation B	47	8.97805×10^{-5}	47	1014.2
Simulation C	151	7.98211×10^{-3}	151	64.392

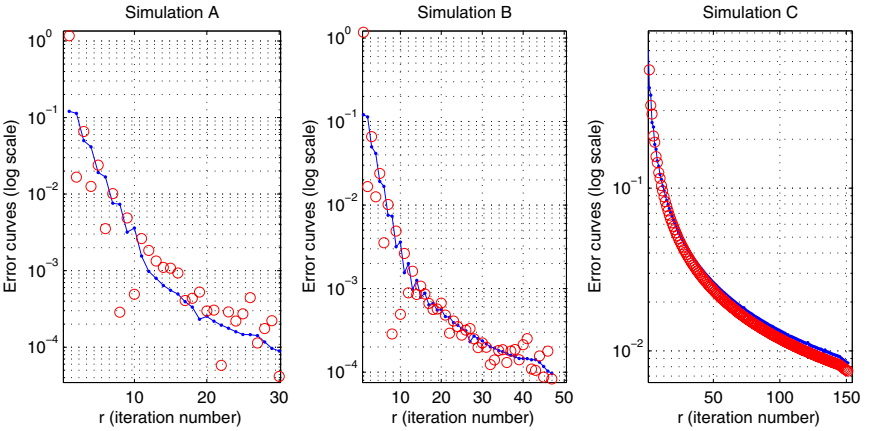


FIG. 7.4. *Plots of Simulation A (left), B (center), C (right):* $\|W - W_r\|_1$ *(dashed-dotted lines),* $\|\Phi(W_r) - \Phi(W_{r-1})\|_2$ *(circles).*

W_{vK}^d has a zero at $z = -1$ and cannot be continued analytically from that point. Setting $\sigma = 2$, the results of Simulation A, B, and C are shown in Table 7.3 and Figure 7.4 (in Simulation B, only 10 zeros are selected with the heuristic scheme).

Transforming the approximate CSF of Simulation A from the unit circle to the imaginary axis gives us the frequency response as shown in Figure 7.5.

All examples show that both Simulations A and B give better results than Simulation C (the Szegö–Levinson algorithm). Despite producing an approximation of substantially higher order, Simulation C gives a final error e_r of magnitude 10^2 higher, and it would seem many hundred more iterations are required to achieve $e_r < 10^{-4}$ as in Simulations A and B. In Example 7.1 for a simple second-order spectral factor, Simulation C also runs much longer. Simulation A runs faster than B but gives an approximation of slightly higher degree. The latter is not unexpected since Simulation B selects more real-valued zeros (i.e., at the origin). The simulations do suggest that

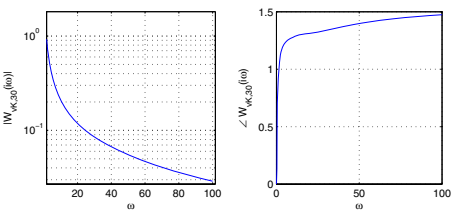


FIG. 7.5. *Frequency response of approximate CSF from Simulation A.*

the heuristic is practically useful and quite effective, regardless of whether a limited or indefinite number of spectral zeros is selected.

8. Conclusions and further research. This paper makes three primary contributions. First and foremost, we have derived a set of sufficient, easy to verify conditions for uniform log-integrability of a sequence of matrix-valued spectral densities. Second, we establish theoretical results on the existence of certain approximating rational sequences for a class of matrix-valued spectral densities. Finally, we propose a new spectral factorization algorithm for a more specific class of matrix-valued spectral densities based on degree constrained rational covariance extensions, and establish convergence results. Our approach does not require the spectral density to be coercive. There is freedom to choose a sequence of spectral zeros in the algorithm, and a heuristic has been proposed for choosing them. The performance of the new algorithm is demonstrated in a number of numerical examples, where it performed favorably compared to the popular Szegö–Levinson algorithm/maximum entropy method. In particular, the algorithm was successfully applied to the nonrational and noncoercive Kolmogorov and von Karman spectral densities. Possible topics for future research include development of fast algorithms for computing degree constrained covariance extensions, relaxations of the conditions of this paper, and development of better heuristics for selection of spectral zeros.

The results and algorithm of this paper may be useful in applications in which spectral factorization plays a prominent role, such as in computation of approximate solutions of algebraic Riccati equations (AREs) in optimal control of linear systems, or in which signals with nonrational power spectra is a central theme (e.g., control of aircraft subject to windgust, adaptive optics, and laser scintillation [2]). It may also prove to be useful in spectral estimation and system identification research.

Acknowledgments. The author thanks the reviewers for their feedback, and the associate and corresponding editors for their role in the peer-review process. He also thanks Prof. John B. Moore for helpful remarks on the first draft of this paper.

REFERENCES

[1] B. D. O. Anderson, *Continuity of the spectral factorization operation*, Mat. Apl. Comp., 4 (1985), pp. 139–156.
[2] A. V. Balakrishnan, *Introduction to Random Processes in Engineering*, John Wiley & Sons, New York, 1995.
[3] S. Barclay, *Continuity of the spectral factorization mapping*, J. London Math. Soc., 70 (2004), pp. 763–779.
[4] A. Blomqvist, A. Lindquist, and R. Nagamune, *Matrix-valued Nevanlinna–Pick interpolation with complexity constraint: An optimization approach*, IEEE Trans. Automat. Control, 48 (2003), pp. 2172–2190.
[5] C. I. Byrnes, S. V. Gusev, and A. Lindquist, *A convex optimization approach to the rational covariance extension problem*, SIAM J. Control Optim., 37 (1998), pp. 211–229.
[6] C. I. Byrnes, S. V. Gusev, and A. Lindquist, *From finite covariance windows to modeling filters: A convex optimization approach*, SIAM Rev., 43 (2001), pp. 645–675.
[7] C. I. Byrnes, A. Lindquist, S. V. Gusev, and A. S. Matveev, *A complete parametrization of all positive rational extensions of a covariance sequence*, IEEE Trans. Automat. Control, 40 (1995), pp. 1841–1857.
[8] P. Caines and M. Baykal-Gürsoy, *On the L^∞ consistency of L^2 estimators*, Systems Control Lett., 12 (1989), pp. 71–76.
[9] H. F. Davis, *Fourier Series and Orthogonal Functions*, Allyn and Bacon, Inc., Boston, 1963.
[10] P. Delsarte, Y. V. Genin, and Y G. Kamp, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits Systems, 25 (1978), pp. 149–160.

[11] P. Dewilde, A. C. Viera, and T. Kailath, *On a generalized Szegö-Levinson realization algorithm for optimal linear predictors based on a network synthesis approach*, IEEE Trans. Circuits Systems, 25 (1978), pp. 663–675.

[12] P. Enqvist, *A homotopy approach to rational covariance extension with degree constraint*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 1173–1201.

[13] T. T. Georgiou, *Partial Realization of Covariance Sequences*, Ph.D. thesis, Department of Electrical Engineering, Center for Mathematical Systems Theory, University of Florida, Gainesville, FL, 1983.

[14] T. T. Georgiou, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoust. Speech Signal Process., 42 (1987), pp. 1049–1070.

[15] T. T. Georgiou, *The interpolation problem with degree constraint*, IEEE Trans. Automat. Control, 44 (1999), pp. 631–635.

[16] G. Heining, *Inversion of Toeplitz-plus-Hankel matrices with arbitrary rank profile*, in Fast Algorithms for Structured Matrices: Theory and Applications Contemp. Math. 323, American Mathematical Society, Providence, RI, SIAM, Philadelphia, 2003, pp. 75–89.

[17] G. Heining and K. Rost, *New fast algorithms for Toeplitz-plus-Hankel matrices*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 842–857.

[18] H. Helson and D. Lowdenslager, *Prediction theory and Fourier series in several variables*, Acta Math., 99 (1958), pp. 165–202.

[19] T. W. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, UK, 1988.

[20] J. Mari, A. Dahlèn, and A. Lindquist, *A covariance extension approach to identification of time series*, Automatica J. IFAC, 36 (2000), pp. 379–398.

[21] H. I. Nurdin, *New results on the rational covariance extension problem with degree constraint*, Systems Control Lett., 55 (2006), pp. 530–537.

[22] H. I. Nurdin and A. Bagchi, *On the solutions of the rational covariance extension problem corresponding to pseudopolynomials having boundary zeros*, IEEE Trans. Automat. Control, 51 (2006), pp. 350–355.

[23] H. I. Nurdin and J. B. Moore, *Computation of degree constrained rational interpolants with non-strictly positive parametrizing functions via homotopy continuation*, in Proceedings of the 45th IEEE Conference on Decision and Control (San Diego, CA, Dec. 13–15, 2006), to appear.

[24] M. Pourahmadi, *Foundations of Time Series Analysis and Prediction Theory*, Wiley Ser. Probab. Statist. Appl. Probab. Statist., Wiley Interscience, New York, 2001.

[25] R. T. Rockafellar, *Conjugate Duality and Optimization*, CBMS Reg. Conf. Ser. Appl. Math. 16, SIAM, Philadelphia, 1974.

[26] A. H. Sayed and T. Kailath, *A survey of spectral factorization methods*, Numer. Linear Algebra Appl., 8 (2001), pp. 467–496.

[27] N. Wiener and P. Masani, *The prediction theory of multivariate stochastic processes*: I. *The regularity condition*, Acta Math., 98 (1957), pp. 111–150.

[28] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.

[29] D. C. Youla and N. N. Kazanjian, *Bauer-type factorization of positive matrices and the theory of matrix polynomials orthogonal on the unit circle*, IEEE Trans. Circuits Systems 25 (1978), pp. 57–69.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1822–1845

OPTIMAL CONTROL FOR A STATIONARY MHD SYSTEM IN VELOCITY-CURRENT FORMULATION*

ROLAND GRIESSE† AND KARL KUNISCH‡

Abstract. An optimal control problem for the equations governing the stationary problem of magnetohydrodynamics (MHD) is considered. Control mechanisms by external and injected currents and magnetic fields are treated. An optimal control problem is formulated. First order necessary and second order sufficient conditions are developed. An operator splitting scheme for the numerical solution of the MHD state equations is analyzed.

Key words. optimal control, magnetohydrodynamics, necessary optimality conditions, operator splitting

AMS subject classifications. 49J20, 49K20, 76D55, 35Q30, 35Q60

DOI. 10.1137/050624236

1. Introduction. Magnetohydrodynamics, or MHD, deals with the mutual interaction of electrically conducting fluids and magnetic fields. In particular the magnetic field interacts with the current in the fluid by exerting a Lorentz force. This feature renders it phenomenally attractive for exploitation, especially in metallurgical processes. The Lorentz force offers a unique possibility of generating a volume force in the fluid and hence to control its motion in a contactless fashion and without any mechanical interference. Therefore MHD technology is used routinely today by engineers, for instance, to stir molten metals during solidification, dampen their undesired convection-driven flow during casting, filter out impurities, and melt and even levitate metals.

With the present paper, we wish to contribute to the application of the powerful methods from mathematical optimization to compute tailored magnetic fields for MHD flow control. Although this work intends primarily to lay the mathematical foundations of MHD optimal control, we believe we have chosen a problem setup of practical relevance, allowing our results to be directly exploited in numerical methods and applications.

Before we turn to the problem description, let us put our work into perspective. Throughout we always refer to stationary incompressible MHD involving viscous fluids. Instationary problems will require an investigation in their own right, and compressible MHD mostly occurs in the realm of plasma physics, whereas we focus on the engineering aspects of MHD. While a remarkable amount of attention in the past decade was devoted to the analysis of optimal control of the Navier–Stokes equations (see, e.g., [9, 10, 11, 14] and the references therein), we are aware of only a few contributions so far concerning the optimal control of the MHD system [13, 16, 17, 18, 19, 24]. The majority of these papers treat the case of low magnetic Reynolds numbers or use either the velocity-potential or the velocity-magnetic field

*Received by the editors October 10, 2005; accepted for publication (in revised form) June 7, 2006; published electronically December 11, 2006.
http://www.siam.org/journals/sicon/45-5/62423.html
†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A–4040 Linz, Austria (roland.griesse@oeaw.ac.at).
‡Institute for Mathematics and Scientific Computing, Heinrichstraße 36, A–8010 Graz, Austria (karl.kunisch@uni-graz.at).

formulation, which both result in the necessity of using artificial boundary conditions. The MHD state equations alone have been investigated in a number of papers, including [22, 21, 12].

We organized the material as follows: In the remainder of this section, we briefly recall the stationary MHD state equations and their velocity-current formulation. In section 2 the variational form of the state equations following [22] are introduced. Our main results are given in section 3, where we consider an optimal control problem for the MHD system. We derive the first order necessary optimality system and establish second order sufficient conditions. In section 4, we analyze an operator splitting scheme for the solution of the MHD state equation which makes use of existing solvers for the Navier–Stokes equations and div-curl systems. We conclude with an outlook on follow-up work in section 5.

Essentially, the MHD system consists of the Navier–Stokes equation with Lorentz force, yielding the fluid velocity $\boldsymbol{u}$ and its pressure p, plus Maxwell's equations describing the interaction of the electric field $\boldsymbol{E}$ and the magnetic field $\boldsymbol{B}$.[1] In the stationary case, the complete MHD system is given by

$$\begin{aligned} \nabla \cdot \boldsymbol{J} &= 0 && \nabla \times \boldsymbol{E} = 0 \\ &\text{(charge conservation)} && \text{(Faraday's law)}, \end{aligned} \tag{1.1}$$

$$\begin{aligned} \nabla \cdot \boldsymbol{B} &= 0 && \nabla \times (\mu^{-1}\boldsymbol{B}) = \boldsymbol{J} \\ &\text{(No magnetic monopoles)} && \text{(Ampère's Law)}, \end{aligned} \tag{1.2}$$

$$\begin{aligned} \boldsymbol{J} &= \sigma(\boldsymbol{E} + \boldsymbol{u} \times \boldsymbol{B}) \\ &\text{(Ohm's law)}, \end{aligned} \tag{1.3}$$

together with the Navier–Stokes system with Lorentz force

$$\varrho(\boldsymbol{u} \cdot \nabla)\boldsymbol{u} - \eta\Delta\boldsymbol{u} + \nabla p = \boldsymbol{J} \times \boldsymbol{B}, \tag{1.4}$$

$$\nabla \cdot \boldsymbol{u} = 0. \tag{1.5}$$

We refer to [25, 4] for more details. Here μ denotes the magnetic permeability of the matter occupying a certain point in space, and ϱ, η, and σ denote the fluid's density, viscosity, and conductivity. All of these numbers are positive. We emphasize that we consider μ to be constant throughout space; hence we assume a nonmagnetic fluid and no magnetic material present in its relevant vicinity.

It is an outstanding feature in magnetohydrodynamics that, from the set of state variables $(\boldsymbol{u}, p, \boldsymbol{E}, \boldsymbol{B}, \boldsymbol{J})$, the electric and magnetic fields $\boldsymbol{E}$ and $\boldsymbol{B}$ extend to all of $\mathbb{R}^3$, whereas the velocity $\boldsymbol{u}$ and pressure p are confined to the bounded region $\Omega \subset \mathbb{R}^3$ occupied by the fluid. The current density $\boldsymbol{J}$ is defined within the fluid region and possibly also in external conductors.

Rather than treating the full set of variables $(\boldsymbol{u}, p, \boldsymbol{E}, \boldsymbol{B}, \boldsymbol{J})$, researchers often describe MHD systems by a properly chosen subset, which is frequently taken as the pair of primal variables $(\boldsymbol{u}, \boldsymbol{B})$. This entails that either $\boldsymbol{B}$ has to be considered on all of $\mathbb{R}^3$, or that artificial shielding boundary conditions have to be assigned on $\partial\Omega$ so that the coupled system can be considered on the fluid region Ω alone. Physically, shielding boundary conditions represent a fluid being surrounded on all sides by a perfectly conducting vessel. Such boundary conditions exclude the control action

[1] Strictly speaking, $\boldsymbol{B}$ should be called the magnetic induction, while $\boldsymbol{H} = \mu^{-1}\boldsymbol{B}$ is the magnetic field. It is, however, common usage in MHD literature to call $\boldsymbol{B}$ the magnetic field.

by means of distant magnetic fields. This is an especially attractive feature of MHD control. In what appear to be the practically more relevant cases, where the outside of the fluid region Ω is finitely conducting or nonconducting, hence permitting control by distant magnetic fields, the proper boundary condition for $\boldsymbol{B}$ is an interface condition requiring $\boldsymbol{B}$ to be continuous across $\partial\Omega$ in both its normal and tangential components, i.e.,

$$[\boldsymbol{B}]_{\partial\Omega} = 0, \tag{1.6}$$

where $[\cdot]_{\partial\Omega}$ denotes the jump of any quantity when going from the interior of Ω to its exterior. As a consequence, $\boldsymbol{B}$ has to be considered on all of $\mathbb{R}^3$.

These shortcomings of the $(\boldsymbol{u}, \boldsymbol{B})$ formulation are not present in the velocity-current formulation in the variables $(\boldsymbol{u}, \boldsymbol{J})$ of the state equation system (1.1)–(1.5) as introduced in [22]. In this formulation, the magnetic field $\boldsymbol{B}$ is eliminated by means of a solution operator $\widetilde{\mathcal{B}}(\boldsymbol{J})$ which solves the div-curl system (1.2) for divergence-free currents $\boldsymbol{J}$ and respects the interface condition (1.6). With the condition that it vanishes at infinity, the solution is unique. Moreover, the irrotational electric field $\boldsymbol{E}$ is replaced with its potential ϕ (unique only up to a constant). In our case of constant permeability μ, the operator $\widetilde{\mathcal{B}}(\boldsymbol{J})$ is given by the Biot–Savart law,

$$\widetilde{\mathcal{B}}(\boldsymbol{J})(\boldsymbol{x}) = -\frac{\mu}{4\pi}\int_{\mathbb{R}^3} \frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3} \times \boldsymbol{J}(\boldsymbol{y})\, d\boldsymbol{y}. \tag{1.7}$$

Inserting $\boldsymbol{B} = \widetilde{\mathcal{B}}(\boldsymbol{J})$ into (1.1)–(1.5) results in the velocity-current formulation of the stationary MHD system,

$$\varrho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} - \eta\Delta\boldsymbol{u} + \nabla p - \boldsymbol{J}\times\widetilde{\mathcal{B}}(\boldsymbol{J}) = 0, \qquad \nabla\cdot\boldsymbol{u} = 0, \tag{1.8}$$

$$\sigma^{-1}\boldsymbol{J} + \nabla\phi - \boldsymbol{u}\times\widetilde{\mathcal{B}}(\boldsymbol{J}) = 0, \qquad \nabla\cdot\boldsymbol{J} = 0 \tag{1.9}$$

for the unknowns $(\boldsymbol{u}, p, \boldsymbol{J}, \phi)$. Here $\boldsymbol{u}$ and p and the electric potential ϕ are confined to the region Ω occupied by our conducting fluid, while $\boldsymbol{J}$ may additionally extend to external conductors.

In general, the total magnetic field $\boldsymbol{B}$ is a superposition of the induced magnetic field $\widetilde{\mathcal{B}}(\boldsymbol{J})$ and other magnetic fields, for instance, those belonging to permanent magnets or fields generated by *given* electric currents; see (3.7). We note in passing that in the case of weakly conducting fluids, such as, for example, salt water, or more generally, in the case of low magnetic Reynolds numbers $R_m = \mu\sigma u l$ (where u and l are typical velocity and length scales), the magnetic field associated with the induced current is negligible in comparison with an imposed field [4]. Hence, $\widetilde{\mathcal{B}}(\boldsymbol{J})$ can be replaced with a *given* field $\boldsymbol{B}_0$. We refer to [19, 24, 17] concerning the optimal control of weakly conducting fluids and to [15] for control approaches concerning the instationary von Kármán flow for a weakly conducting fluid with given Lorentz force.

2. Function spaces and operators. In this section, we present the proper functional analytic setting for the stationary MHD problem following [22]. Throughout, let Ω denote a bounded domain in $\mathbb{R}^3$ with Lipschitzian boundary, and let $L^2(\Omega)$, $H^1(\Omega)$, and $H^1_0(\Omega)$ and, in general, $W^{1,p}(\Omega)$ and $W^{1,p}_0(\Omega)$ denote the usual Sobolev spaces [1] for $1 < p < \infty$. In addition, for $l = 1, 2$, let $V^l(\mathbb{R}^3)$ stand for the completion of $H^l(\mathbb{R}^3)$ with respect to the seminorm which measures only the lth order derivatives [21]. Furthermore, $H^{1/2}(\partial\Omega)$ is the trace space of $H^1(\Omega)$, endowed with the norm

$$\|\phi\|_{H^{1/2}(\partial\Omega)} = \inf \|\Phi\|_{H^1(\Omega)},$$

where the infimum extends over all Φ whose trace coincides with ϕ. The norm duals of $H_0^1(\Omega)$ and $H^{1/2}(\partial\Omega)$ are $H^{-1}(\Omega)$ and $H^{-1/2}(\partial\Omega)$, respectively. The norm dual of $W_0^{1,p}(\Omega)$ is $W^{-1,p'}(\Omega)$, where p' is the dual of p, i.e., $p' = p/(p-1)$. Boldface notation indicates the triple Cartesian product of a space with itself, e.g., $\boldsymbol{L}^2(\Omega) = [L^2(\Omega)]^3$, and the symbol $\boldsymbol{L}^2_{\mathrm{div}}(\Omega)$ denotes the subspace of divergence-free (solenoidal) functions in $\boldsymbol{L}^2(\Omega)$. Finally, we denote by $A^\star$ the adjoint of a bounded linear operator A and by X' the dual of a normed linear space X.

We will consider solutions

$$\begin{aligned} &\boldsymbol{u} \in \boldsymbol{H}^1(\Omega) \cap \boldsymbol{L}^2_{\mathrm{div}}(\Omega), && \boldsymbol{J} \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega), \\ &p \in L^2(\Omega)/\mathbb{R}, && \phi \in H^1(\Omega)/\mathbb{R}, \end{aligned}$$

which satisfy (1.8)–(1.9) in variational form. In order to obtain the variational formulation, we multiply (1.8)–(1.9) by smooth test functions $(\boldsymbol{v}, q, \boldsymbol{K}, \psi)$, where $\boldsymbol{v}$ has zero Dirichlet boundary values. Integration by parts yields

(2.1)
$$\begin{aligned} &\varrho \int_\Omega (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\cdot\boldsymbol{v} + \eta \int_\Omega \nabla\boldsymbol{u} \,:\, \nabla\boldsymbol{v} - \int_\Omega (\nabla\cdot\boldsymbol{v})\,\mathcal{P}(p) - \int_\Omega (\boldsymbol{J}\times\widetilde{\mathcal{B}}(\boldsymbol{J}))\cdot\boldsymbol{v} - \int_\Omega (\nabla\cdot\boldsymbol{u})\,q \\ &\qquad + \sigma^{-1}\int_\Omega \boldsymbol{J}\cdot\boldsymbol{K} + \int_\Omega \boldsymbol{K}\cdot(\nabla\phi) - \int_\Omega (\boldsymbol{u}\times\widetilde{\mathcal{B}}(\boldsymbol{J}))\cdot\boldsymbol{K} + \int_\Omega \boldsymbol{J}\cdot\nabla\psi = \int_{\partial\Omega} j\,\psi, \end{aligned}$$

where $j = \boldsymbol{J}\cdot\boldsymbol{n}$ denotes the given boundary values in the normal direction for the current $\boldsymbol{J}$. In (2.1),

$$\int_\Omega \nabla\boldsymbol{u} \,:\, \nabla\boldsymbol{v} = \sum_{i=1}^3 \int_\Omega \nabla\boldsymbol{u}_i\cdot\nabla\boldsymbol{v}_i,$$

and $\mathcal{P}(p)$ denotes the projection of p on the functions with zero mean,

$$\mathcal{P}(p) = p - \frac{1}{|\Omega|}\int_\Omega p.$$

Based on (2.1), we introduce the bilinear forms

$$\begin{aligned} &a_1(\boldsymbol{u},\boldsymbol{v}) = \eta\int_\Omega (\nabla\boldsymbol{u} \,:\, \nabla\boldsymbol{v}), && a_2(\boldsymbol{J},\boldsymbol{K}) = \sigma^{-1}\int_\Omega \boldsymbol{J}\cdot\boldsymbol{K}, \\ &d_1(\boldsymbol{u},p) = -\int_\Omega (\nabla\cdot\boldsymbol{u})\,\mathcal{P}(p), && d_2(\boldsymbol{J},\phi) = \int_\Omega \boldsymbol{J}\cdot(\nabla\phi) \end{aligned}$$

and trilinear forms

$$b(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) = \varrho\int_\Omega (\boldsymbol{u}\cdot\nabla)\boldsymbol{v}\cdot\boldsymbol{w}, \qquad c(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}) = \int_\Omega (\boldsymbol{u}\times\boldsymbol{v})\cdot\boldsymbol{w},$$

where $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w} \in \boldsymbol{H}^1(\Omega)$, $\boldsymbol{J},\boldsymbol{K} \in \boldsymbol{L}^2(\Omega)$, $p \in L^2(\Omega)/\mathbb{R}$, and $\phi \in H^1(\Omega)/\mathbb{R}$. Note that c is defined on any $\boldsymbol{L}^{p_1}(\Omega)\times\boldsymbol{L}^{p_2}(\Omega)\times\boldsymbol{L}^{p_3}(\Omega)$, where $1/p_1 + 1/p_2 + 1/p_3 = 1$.

Let us now turn to the forms introduced above. Besides the obvious continuity properties, they satisfy the following.

LEMMA 2.1 (LBB conditions [7]). *The constraint forms d_1 and d_2 satisfy the following Ladyzhenskaya–Babuska–Brezzi (LBB) conditions on $\boldsymbol{H}^1_0(\Omega) \times L^2(\Omega)/\mathbb{R}$, and on $\boldsymbol{L}^2(\Omega) \times H^1(\Omega)/\mathbb{R}$, respectively:*

$$\inf_{p \in L^2(\Omega)/\mathbb{R}} \sup_{\boldsymbol{u} \in \boldsymbol{H}^1_0(\Omega)} \frac{d_1(\boldsymbol{u}, p)}{\|\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)} \|p\|_{L^2(\Omega)/\mathbb{R}}} \geq \beta_1,$$
$$\inf_{\phi \in H^1(\Omega)/\mathbb{R}} \sup_{\boldsymbol{J} \in \boldsymbol{L}^2(\Omega)} \frac{d_2(\boldsymbol{J}, \phi)}{\|\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)} \|\phi\|_{H^1(\Omega)/\mathbb{R}}} \geq \beta_2$$

for some $\beta_1, \beta_2 > 0$.

Let us define the following spaces associated with the constraint forms d_1 and d_2:

$$V_1 = \{\boldsymbol{v} \in \boldsymbol{H}^1_0(\Omega) \,:\, d_1(\boldsymbol{v}, p) = 0 \text{ for all } p \in L^2(\Omega)/\mathbb{R}\},$$
$$V_1^0 = \{\Phi_1 \in \boldsymbol{H}^{-1}(\Omega) \,:\, \langle \Phi_1, \boldsymbol{v} \rangle = 0 \text{ for all } \boldsymbol{v} \in V_1\},$$
$$V_2 = \{\boldsymbol{K} \in \boldsymbol{L}^2(\Omega) \,:\, d_2(\boldsymbol{K}, \phi) = 0 \text{ for all } \phi \in H^1(\Omega)/\mathbb{R}\},$$
$$V_2^0 = \{\Phi_2 \in \boldsymbol{L}^2(\Omega)' \,:\, \langle \Phi_2, \boldsymbol{K} \rangle = 0 \text{ for all } \boldsymbol{K} \in V_2\}.$$

Note that [26]

$$V_1 = \{\boldsymbol{v} \in \boldsymbol{H}^1_0(\Omega) \,:\, \nabla \cdot \boldsymbol{v} = 0 \text{ on } \Omega\},$$
$$V_2 = \{\boldsymbol{K} \in \boldsymbol{L}^2(\Omega) \,:\, \nabla \cdot \boldsymbol{K} = 0 \text{ on } \Omega \text{ and } \boldsymbol{K} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\}.$$

LEMMA 2.2 (properties of constraint forms). *If $\Phi_1 \in V_1^0$ and $\Phi_2 \in V_2^0$, then the equations*

$$d_1(\boldsymbol{v}, p) = \langle \Phi_1, \boldsymbol{v} \rangle \text{ for all } \boldsymbol{v} \in \boldsymbol{H}^1_0(\Omega),$$
$$d_2(\boldsymbol{K}, \phi) = \langle \Phi_2, \boldsymbol{K} \rangle \text{ for all } \boldsymbol{K} \in \boldsymbol{L}^2(\Omega)$$

are uniquely solvable for $p \in L^2(\Omega)/\mathbb{R}$ and $\phi \in H^1(\Omega)/\mathbb{R}$, and $\|p\|_{L^2(\Omega)/\mathbb{R}} \leq c_1 \|\Phi_1\|_{\boldsymbol{H}^{-1}(\Omega)}$ and $\|\phi\|_{H^1(\Omega)/\mathbb{R}} \leq c_2 \|\Phi_2\|_{\boldsymbol{L}^2(\Omega)'}$ hold for some $c_1, c_2 > 0$.

Proof. See [7, Chap. I, Lem. 4.1]. ∎

LEMMA 2.3 (passing to the limit in c).

(1) *Let $\boldsymbol{u}^n \rightharpoonup \boldsymbol{u}$ in $\boldsymbol{L}^2(\Omega)$, $\boldsymbol{v}^n \to \boldsymbol{v}$ in $\boldsymbol{L}^3(\Omega)$, and $\boldsymbol{w} \in \boldsymbol{L}^6(\Omega)$. Then $c(\boldsymbol{u}^n, \boldsymbol{v}^n, \boldsymbol{w}) \to c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$.*

(2) *Let $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega)$, $\boldsymbol{v}^n \to \boldsymbol{v}$ in $\boldsymbol{L}^3(\Omega)$, and $\boldsymbol{w}^n \rightharpoonup \boldsymbol{w}$ in $\boldsymbol{L}^6(\Omega)$. Then $c(\boldsymbol{u}, \boldsymbol{v}^n, \boldsymbol{w}^n) \to c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$.*

Proof. For the first claim, we use the estimate

$$|c(\boldsymbol{u}^n, \boldsymbol{v}^n, \boldsymbol{w}) - c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})| \leq \left| \int_\Omega (\boldsymbol{u}^n \times (\boldsymbol{v}^n - \boldsymbol{v})) \cdot \boldsymbol{w} \right| + \left| \int_\Omega ((\boldsymbol{u}^n - \boldsymbol{u}) \times \boldsymbol{v}) \cdot \boldsymbol{w} \right|.$$

We apply Hölder's inequality to the first term, using the native norms of all three factors involved. It converges to zero since $\|\boldsymbol{v}^n - \boldsymbol{v}\|_{\boldsymbol{L}^3(\Omega)}$ converges to zero and $\|\boldsymbol{u}^n\|_{\boldsymbol{L}^2(\Omega)}$ is bounded by assumption (1). Hölder's inequality again shows that $\int_\Omega (\cdot \times \boldsymbol{v}) \cdot \boldsymbol{w}$ is a continuous linear functional on $\boldsymbol{L}^2(\Omega)$ so that also the second term converges to zero. The second claim follows similarly, using the splitting

$$|c(\boldsymbol{u}, \boldsymbol{v}^n, \boldsymbol{w}^n) - c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})| \leq \left| \int_\Omega (\boldsymbol{u} \times (\boldsymbol{v}^n - \boldsymbol{v})) \cdot \boldsymbol{w}^n \right| + \left| \int_\Omega (\boldsymbol{u} \times \boldsymbol{v}) \cdot (\boldsymbol{w}^n - \boldsymbol{w}) \right|. \qquad ∎$$

LEMMA 2.4 (properties of $\widetilde{\mathcal{B}}$). *The following properties hold:*

(1) *The Biot–Savart operator* (1.7) *maps any given* $\boldsymbol{J} \in \boldsymbol{L}^2(\mathbb{R}^3)$ *to* $\widetilde{\mathcal{B}}(\boldsymbol{J}) \in \boldsymbol{V}^1(\mathbb{R}^3)$. *The restriction of* $\widetilde{\mathcal{B}}(\boldsymbol{J})$ *to* Ω *lies in* $\boldsymbol{H}^1(\Omega)$.
In this sense, the Biot–Savart operator defines a continuous linear map between $\boldsymbol{L}^2(\mathbb{R}^3)$ *and* $\boldsymbol{H}^1(\Omega)$, *i.e.,*

$$\|\widetilde{\mathcal{B}}(\boldsymbol{J})\|_{\boldsymbol{H}^1(\Omega)} \le c_B \|\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)}$$

for all $\boldsymbol{J} \in \boldsymbol{L}^2(\Omega)$ *and some* $c_B > 0$.

(2) *If* $\boldsymbol{J} \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega)$, *then* $\widetilde{\mathcal{B}}(\boldsymbol{J})$ *is the unique solution of the* div-curl *system* (1.2) *vanishing at infinity.*
The operator $\widetilde{\mathcal{B}}$ *is self-adjoint in* $\boldsymbol{L}^2(\mathbb{R}^3)$.

Remark 2.5. Whenever $\boldsymbol{J}$ has compact support, as will be the case in our applications, $\widetilde{\mathcal{B}}(\boldsymbol{J})$ belongs to $\boldsymbol{H}^1(\mathbb{R}^3)$. However, it is sufficient for our purpose that the restriction of $\widetilde{\mathcal{B}}(\boldsymbol{J})$ to Ω is in $\boldsymbol{H}^1(\Omega)$, as guaranteed by the lemma.

Proof of Lemma 2.4. Consider the Newton potential

$$(\mathcal{L}_0 \boldsymbol{J})(\boldsymbol{x}) = -\frac{\mu}{4\pi} \int_{\mathbb{R}^3} \frac{1}{|\boldsymbol{x}-\boldsymbol{y}|} \boldsymbol{J}(\boldsymbol{y})\, d\boldsymbol{y},$$

which defines an isomorphism from $\boldsymbol{L}^2(\mathbb{R}^3)$ to $\boldsymbol{V}^2(\mathbb{R}^3)$ [21, 3]. One argues that $\widetilde{\mathcal{B}}(\boldsymbol{J}) = \nabla \times \mathcal{L}_0(\boldsymbol{J})$ holds for all $\boldsymbol{J} \in \boldsymbol{L}^2(\mathbb{R}^3)$. This implies that $\boldsymbol{B}(\boldsymbol{J}) \in \boldsymbol{V}^1(\mathbb{R}^3)$.

Since $\boldsymbol{V}^1(\mathbb{R}^3)$ embeds into $\boldsymbol{H}^1_{\mathrm{loc}}(\mathbb{R}^3)$ [3], claims (1) and (2) follow. Theorem 2.7 and Remark 2.8(b) in [21] imply that (3) holds. To prove self-adjointness, we multiply $\widetilde{\mathcal{B}}(\boldsymbol{J})$ by a function $\boldsymbol{C} \in \boldsymbol{L}^2(\Omega)$ and integrate over $\mathbb{R}^3$:

$$\begin{aligned}
\int_{\mathbb{R}^3} \boldsymbol{C} \cdot \widetilde{\mathcal{B}}(\boldsymbol{J}) &= -\frac{\mu}{4\pi} \int_{\mathbb{R}^3} \boldsymbol{C}(\boldsymbol{x}) \cdot \left(\int_{\mathbb{R}^3} \frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3} \times \boldsymbol{J}(\boldsymbol{y})\, d\boldsymbol{y} \right) d\boldsymbol{x} \\
&= -\frac{\mu}{4\pi} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \boldsymbol{C}(\boldsymbol{x}) \cdot \left(\frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3} \times \boldsymbol{J}(\boldsymbol{y}) \right) d\boldsymbol{x}\, d\boldsymbol{y} \\
&= -\frac{\mu}{4\pi} \int_{\mathbb{R}^3} \boldsymbol{J}(\boldsymbol{y}) \cdot \left(\int_{\mathbb{R}^3} \boldsymbol{C}(\boldsymbol{x}) \times \frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3}\, d\boldsymbol{x} \right) d\boldsymbol{y}. \\
&= -\frac{\mu}{4\pi} \int_{\mathbb{R}^3} \boldsymbol{J}(\boldsymbol{x}) \cdot \left(\int_{\mathbb{R}^3} \frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3} \times \boldsymbol{C}(\boldsymbol{y})\, d\boldsymbol{y} \right) d\boldsymbol{x}.
\end{aligned}$$

Since the left-hand side by definition equals $\int_{\mathbb{R}^3} \boldsymbol{J} \cdot \widetilde{\mathcal{B}}^\star(\boldsymbol{C})$, we have found $\widetilde{\mathcal{B}}^\star(\boldsymbol{C}) = \widetilde{\mathcal{B}}(\boldsymbol{C})$. ∎

We conclude this section by a compact embedding result whose proof can be found in [7].

LEMMA 2.6 (compact maps and embeddings). *For a bounded domain* $\Omega \subset \mathbb{R}^3$ *with Lipschitz boundary, the embeddings* $H^1(\Omega) \hookrightarrow L^{6-\varepsilon}(\Omega)$ *and* $L^2(\Omega) \hookrightarrow W^{-1,6-\varepsilon}(\Omega)$ *are compact for all* $\varepsilon > 0$. *In addition, the pointwise product map* $(u, v) \mapsto u\, v$ *is continuous from* $H^1(\Omega) \times H^1(\Omega)$ *to* $W^{1,3/2}(\Omega)$, *and the latter embeds compactly into* $L^{3-\varepsilon}(\Omega)$ *for all* $\varepsilon > 0$.

3. The optimal control problem. We analyze an optimal control problem for the stationary MHD system motivated by the applications described in [4]. A typical geometry that we have in mind is depicted in Figure 3.1. We assume that the electrically conducting fluid, e.g., a liquid metal, is contained in a vessel occupying

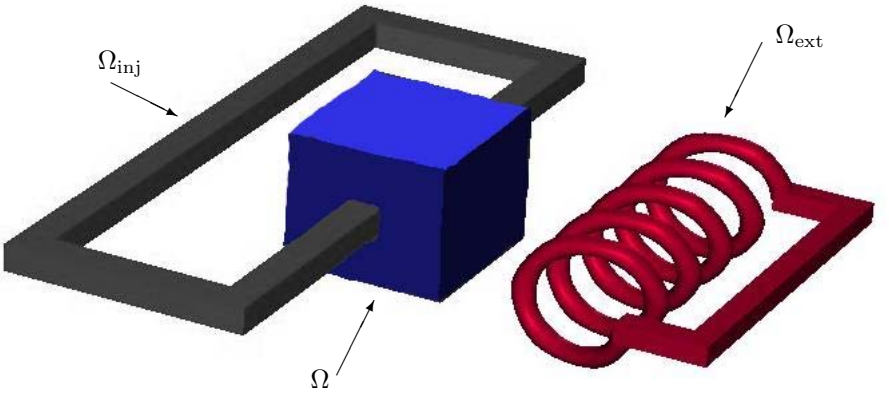


FIG. 3.1. *General setup: Fluid region Ω (cube), external conductor Ω_{inj} attached to the fluid region, and external conductor Ω_{ext} separate from the fluid region.*

the domain Ω. On part of the boundary $\partial\Omega$, an external conductor Ω_{inj} is attached. The current distribution in Ω_{inj} is assumed to be known, but its magnitude can be adjusted. The purpose of such a propulsion device is to drive the fluid in Ω in a desired way, both through the action of the magnetic field induced by the current in Ω_{inj} and through the current, which is "injected" into the fluid region Ω through the electrodes attached to its surface. The same assembly can be found in electromagnetic filtration devices. Since in some cases it may be undesirable to attach the external conductor to the surface of the fluid vessel, we included a second conductor Ω_{ext} separately from the fluid region in which, again, the current distribution is given but its magnitude can be controlled. This external conductor has an impact on the fluid motion in Ω solely through its induced magnetic field. An assembly in which a number of such coils is distributed around the fluid vessel can be found, e.g., in electromagnetic stirring devices [4, 23], albeit their magnetic field is usually amplified by yokes in the coil centers, which are currently not included in our model in view of the assumption that the permeability μ is constant. In addition, we allow for another external magnetic field $\boldsymbol{B}_{\text{ext}}$ which is subject to optimization. In practice such a field cannot be shaped at will. We consider it as originating from a permanent magnet whose field is again known except for its magnitude, which serves as an optimization parameter.

From (1.8)–(1.9) we recall the stationary MHD system in velocity-current formulation,

$$\varrho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} - \eta\Delta\boldsymbol{u} + \nabla p = \boldsymbol{J}\times\boldsymbol{B}, \qquad \nabla\cdot\boldsymbol{u} = 0 \quad \text{on } \Omega, \tag{3.1}$$

$$\sigma^{-1}\boldsymbol{J} - \boldsymbol{u}\times\boldsymbol{B} + \nabla\phi = 0, \qquad \nabla\cdot\boldsymbol{J} = 0 \quad \text{on } \Omega. \tag{3.2}$$

We will consider solutions

$$y = (\boldsymbol{u}, p, \boldsymbol{J}, \phi) \tag{3.3}$$

with

$$\boldsymbol{u} \in \boldsymbol{H}^1(\Omega)\cap\boldsymbol{L}^2_{\text{div}}(\Omega), \qquad \boldsymbol{J}\in\boldsymbol{L}^2_{\text{div}}(\Omega),$$
$$p\in L^2(\Omega)/\mathbb{R}, \qquad \phi\in H^1(\Omega)/\mathbb{R},$$

which satisfy (3.1)–(3.2) in variational form, together with the following boundary conditions:

$$\boldsymbol{J}\cdot\boldsymbol{n} = \boldsymbol{J}_{\mathrm{inj}}\cdot\boldsymbol{n} \quad \text{on } \partial\Omega_{\mathrm{inj}}\cap\partial\Omega, \tag{3.4}$$

$$\boldsymbol{J}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega\setminus\partial\Omega_{\mathrm{inj}}, \tag{3.5}$$

with the injected current $\boldsymbol{J}_{\mathrm{inj}}$ specified below. For the fluid velocity, we impose Dirichlet boundary conditions

$$\boldsymbol{u} = \boldsymbol{h} \quad \text{on } \partial\Omega. \tag{3.6}$$

For our setup the total magnetic field $\boldsymbol{B}$ is the superposition of the field $\widetilde{\mathcal{B}}(\boldsymbol{J})$ induced by the current $\boldsymbol{J}$ inside the fluid domain, the fields $\widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{ext}})$ and $\widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{inj}})$ induced by the currents in the external conductors (whether or not attached to the fluid domain), and the magnetic field $\boldsymbol{B}_{\mathrm{ext}}$ associated with the permanent magnet, i.e.,

$$\boldsymbol{B} = \widetilde{\mathcal{B}}(\boldsymbol{J}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{ext}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{inj}}) + \boldsymbol{B}_{\mathrm{ext}}. \tag{3.7}$$

We repeat that the external magnetic field $\boldsymbol{B}_{\mathrm{ext}}$ and the current fields $\boldsymbol{J}_{\mathrm{ext}}$ and $\boldsymbol{J}_{\mathrm{inj}}$ are assumed known except for their magnitude. For instance, in the case of a smooth wire, the current field simply follows its shape, and thus

$$\boldsymbol{J}_{\mathrm{ext}} = I_{\mathrm{ext}}\cdot\widetilde{\boldsymbol{J}}_{\mathrm{ext}}, \qquad \boldsymbol{J}_{\mathrm{inj}} = I_{\mathrm{inj}}\cdot\widetilde{\boldsymbol{J}}_{\mathrm{inj}}, \qquad \boldsymbol{B}_{\mathrm{ext}} = B_{\mathrm{ext}}\cdot\widetilde{\boldsymbol{B}}_{\mathrm{ext}}, \tag{3.8}$$

where

$$u = (I_{\mathrm{ext}}, I_{\mathrm{inj}}, B_{\mathrm{ext}}) \in \mathbb{R}^3 \tag{3.9}$$

is the vector of control variables. Herein, I_{ext} and I_{inj} denote the adjustable scalar current strengths and $\widetilde{\boldsymbol{J}}_{\mathrm{ext}}$ and $\widetilde{\boldsymbol{J}}_{\mathrm{inj}}$ are the given solenoidal current field distributions in the external conductors Ω_{ext} and Ω_{inj}, respectively. In practice, these currents must be maintained by an adjustable voltage source. Likewise, B_{ext} relates to the strength of the external magnetic field (associated with a permanent magnet) $\boldsymbol{B}_{\mathrm{ext}}$. The boundary conditions (3.4)–(3.5), together with Assumption 3.1(3) below, close the current loop and ensure that the total current $\boldsymbol{J}+\boldsymbol{J}_{\mathrm{ext}}+\boldsymbol{J}_{\mathrm{inj}}$ is solenoidal on $\mathbb{R}^3$.

The restriction to finite-dimensional controls is motivated by applications. Nonparametrized distributed current and magnetic fields $(\boldsymbol{J}_{\mathrm{ext}}, \boldsymbol{J}_{\mathrm{inj}}, \boldsymbol{B}_{\mathrm{ext}}) \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega)\times\big(\boldsymbol{H}^1(\Omega_{\mathrm{inj}})\cap\boldsymbol{L}^2_{\mathrm{div}}(\Omega_{\mathrm{inj}})\big)\times\boldsymbol{V}^1(\mathbb{R}^3)$ can be used as controls if the norms in the objective below are adjusted accordingly. Besides technical difficulties, the most significant change that will occur is in the necessary optimality conditions (3.28) (see Theorem 3.11 below), which will involve Poisson equations.

We are now prepared to state our optimal control problem as follows:

$$\begin{aligned}\text{Minimize } & \frac{\alpha_u}{2}\|\boldsymbol{u}-\boldsymbol{u}_{\mathrm{d}}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \frac{\alpha_B}{2}\|\boldsymbol{B}-\boldsymbol{B}_{\mathrm{d}}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} + \frac{\alpha_J}{2}\|\boldsymbol{J}-\boldsymbol{J}_{\mathrm{d}}\|^2_{L^2(\Omega_{J,\mathrm{obs}})} \\ & + \frac{\gamma_{\mathrm{ext}}}{2}|I_{\mathrm{ext}}|^2 + \frac{\gamma_{\mathrm{inj}}}{2}|I_{\mathrm{inj}}|^2 + \frac{\gamma_B}{2}|B_{\mathrm{ext}}|^2 \end{aligned} \tag{P}$$

subject to (3.1)–(3.7).

The control objective reflects the goal of steering the fluid velocities and the magnetic and current fields towards the given desired fields $\boldsymbol{u}_{\mathrm{d}}$, $\boldsymbol{B}_{\mathrm{d}}$, and $\boldsymbol{J}_{\mathrm{d}}$, possibly

only on subdomains $\Omega_{u,\mathrm{obs}} \subset \Omega$, $\Omega_{B,\mathrm{obs}} \subset \mathbb{R}^3$, and $\Omega_{J,\mathrm{obs}} \subset \Omega$ of interest. The desired fields are $\boldsymbol{L}^2$ functions in their respective domains of definition. One may choose among those goals by setting the respective weights α equal to zero. The control weights γ are assumed to be positive. Note that due to the "state times control" terms $\boldsymbol{u} \times \boldsymbol{B}$ (through Ohm's law (1.3)) and $\boldsymbol{J} \times \boldsymbol{B}$ (the Lorentz force), problem (P) is a particular type of bilinear control problem.

We shall require the following assumption for the analysis of the MHD system (3.1)–(3.7).

Assumption 3.1 (problem data).

1. Ω, Ω_{inj}, and Ω_{ext} are bounded mutually disjoint domains with $C^{0,1}$ boundary, such that Ω_{inj} and Ω have a part of their boundary of positive surface measure in common; see Figure 3.1.
2. The boundary velocity $\boldsymbol{h} \in \boldsymbol{H}^{1/2}(\partial\Omega)$ satisfies $\int_{\partial\Omega} \boldsymbol{h} \cdot \boldsymbol{n} = 0$.
3. $\widetilde{\boldsymbol{J}}_{\mathrm{ext}}$ and $\widetilde{\boldsymbol{J}}_{\mathrm{inj}}$ are (current) fields in $\boldsymbol{L}^2_{\mathrm{div}}(\Omega_{\mathrm{ext}})$ and $\boldsymbol{H}^1(\Omega_{\mathrm{inj}}) \cap \boldsymbol{L}^2_{\mathrm{div}}(\Omega_{\mathrm{inj}})$, respectively, satisfying $\int_{\partial\Omega\cap\partial\Omega_{\mathrm{inj}}} \widetilde{\boldsymbol{J}}_{\mathrm{inj}} \cdot \boldsymbol{n} = 0$.
4. $\widetilde{\boldsymbol{B}}_{\mathrm{ext}}$ is a divergence-free (magnetic) field on $\mathbb{R}^3$ such that its restriction to Ω lies in $\boldsymbol{L}^{3+\varepsilon}(\Omega)$ for some $\varepsilon > 0$.
5. The nonzero fields $\widetilde{\boldsymbol{B}}_{\mathrm{ext}}$, $\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}})$, and $\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}})$ are linearly independent.

The assumption $\widetilde{\boldsymbol{J}}_{\mathrm{inj}} \in \boldsymbol{H}^1(\Omega_{\mathrm{inj}})$ implies that the normal trace $\widetilde{\boldsymbol{J}}_{\mathrm{inj}} \cdot \boldsymbol{n}$, when restricted to the intersection $\partial\Omega \cap \partial\Omega_{\mathrm{inj}}$ and extended by zero, yields a function $j \in L^2(\partial\Omega)$; hence in particular $j \in H^{-1/2}(\partial\Omega)$ holds. The latter is needed to ensure the existence of a lifting $\boldsymbol{J}_0$ whose normal boundary values coincide with j, see Lemma 3.2 below. Note that $j \in H^{-1/2}(\partial\Omega)$ can in general not be achieved if merely $\widetilde{\boldsymbol{J}}_{\mathrm{inj}} \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega_{\mathrm{inj}})$.

In order to eliminate the boundary conditions for the velocity and current and to homogenize the problem, we introduce liftings $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$ of the given boundary data such that

$$\begin{aligned} &\boldsymbol{u}_0 \in \boldsymbol{H}^1(\Omega), && \boldsymbol{u}_0|_{\partial\Omega} = \boldsymbol{h}, && \nabla \cdot \boldsymbol{u}_0 = 0, \\ &\boldsymbol{J}_0 \in \boldsymbol{L}^2(\Omega), && \boldsymbol{J}_0 \cdot \boldsymbol{n}|_{\partial\Omega} = j, && \nabla \cdot \boldsymbol{J}_0 = 0, \end{aligned}$$

with

$$j = \boldsymbol{J}_{\mathrm{inj}} \cdot \boldsymbol{n} \text{ on } \partial\Omega \cap \partial\Omega_{\mathrm{inj}} \quad \text{and} \quad j = 0 \text{ on } \partial\Omega \setminus \partial\Omega_{\mathrm{inj}}. \tag{3.10}$$

Such a lifting exists according to the following lemma. Note that Assumption 3.1(3) implies that $\int_{\partial\Omega} j = 0$ as required in part (b) of Lemma 3.2 below.

LEMMA 3.2 (lifting). *Let β_i be the constants from the LBB condition (Lemma 2.1). Then we have the following:*

(a) *For every $\boldsymbol{h} \in \boldsymbol{H}^{1/2}(\partial\Omega)$, there exists $\boldsymbol{u}_0 \in \boldsymbol{H}^1(\Omega)$ such that $\boldsymbol{u}_0|_{\partial\Omega} = \boldsymbol{h}$ and $d_1(\boldsymbol{u}_0, q) = 0$ holds for all $q \in \boldsymbol{L}^2(\Omega)/\mathbb{R}$, i.e., $\nabla \cdot \boldsymbol{u}_0 = 0$. Moreover, the map $\boldsymbol{h} \mapsto \boldsymbol{u}_0$ can be chosen linearly and continuously, such that*

$$\|\boldsymbol{u}_0\|_{\boldsymbol{H}^1(\Omega)} \le (1 + \beta_1^{-1}) \|\boldsymbol{h}\|_{\boldsymbol{H}^{1/2}(\partial\Omega)}$$

is satisfied.

(b) *For every* $j \in H^{-1/2}(\partial\Omega)$ *which satisfies* $\langle j, 1\rangle_{\partial\Omega} = 0$, *there exists* $\boldsymbol{J}_0 \in \boldsymbol{L}^2(\Omega)$ *such that* $d_2(\boldsymbol{J}_0, \psi) = \langle j, \psi\rangle_{\partial\Omega}$ *holds for all* $\psi \in H^1(\Omega)/\mathbb{R}$, *i.e.,* $\nabla \cdot \boldsymbol{J}_0 = 0$ *and* $\boldsymbol{J}_0 \cdot \boldsymbol{n} = j$. *Moreover, the map* $j \mapsto \boldsymbol{J}_0$ *can be chosen linearly and continuously, such that*

$$\|\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)} \le \beta_2^{-1} \|j\|_{H^{-1/2}(\partial\Omega)}$$

is satisfied.

Proof. The claim is a consequence of the LBB condition; see [22] for details. Note that for functions $\boldsymbol{J}_0 \in \boldsymbol{L}^2(\Omega)$ such that $\nabla \cdot \boldsymbol{J}_0 \in \boldsymbol{L}^2(\Omega)$, the normal trace $\boldsymbol{J}_0 \cdot \boldsymbol{n}$ exists with values in $H^{-1/2}(\partial\Omega)$ [7]. ∎

As a consequence, the fluid velocity and current can be written as

$$\boldsymbol{u} = \boldsymbol{u}_0 + \widehat{\boldsymbol{u}}, \qquad \boldsymbol{J} = \boldsymbol{J}_0 + \widehat{\boldsymbol{J}},$$

where $\widehat{\boldsymbol{u}} \in \boldsymbol{H}_0^1(\Omega)$ and $\widehat{\boldsymbol{J}} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$. It is important to note that in view of the current boundary condition (3.4)–(3.5), the lifting $\boldsymbol{J}_0$ depends on the control variable I_{inj}. We emphasize this dependence whenever appropriate by writing

$$\boldsymbol{J}_0 = \Lambda(I_{\text{inj}} \widetilde{\boldsymbol{J}}_{\text{inj}} \cdot \boldsymbol{n}), \tag{3.11}$$

meaning that $\boldsymbol{J}_0$ is the lifting, according to Lemma 3.2(b), of the function j in (3.10).

We now consider the homogenized state equation to find

$$\widehat{y} = (\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi). \tag{3.12}$$

In its variational form, the homogenized system is given by

$$\begin{aligned} a_1(\widehat{\boldsymbol{u}} + \boldsymbol{u}_0, \boldsymbol{v}) - c(\widehat{\boldsymbol{J}} + \boldsymbol{J}_0, \boldsymbol{B}, \boldsymbol{v}) + b(\widehat{\boldsymbol{u}} + \boldsymbol{u}_0, \widehat{\boldsymbol{u}} + \boldsymbol{u}_0, \boldsymbol{v}) + d_1(\boldsymbol{v}, p) &= 0, \\ d_1(\widehat{\boldsymbol{u}}, q) &= 0, \\ a_2(\widehat{\boldsymbol{J}} + \boldsymbol{J}_0, \boldsymbol{K}) + c(\boldsymbol{K}, \boldsymbol{B}, \widehat{\boldsymbol{u}} + \boldsymbol{u}_0) + d_2(\boldsymbol{K}, \phi) &= 0, \\ d_2(\widehat{\boldsymbol{J}}, \psi) &= 0 \end{aligned} \tag{3.13}$$

for all $(\boldsymbol{v}, q, \boldsymbol{K}, \psi) \in \boldsymbol{H}_0^1(\Omega) \times L^2(\Omega)/\mathbb{R} \times \boldsymbol{L}^2(\Omega) \times H^1(\Omega)/\mathbb{R}$, where we set

$$\boldsymbol{B} = \widetilde{\mathcal{B}}(\Lambda(I_{\text{inj}} \widetilde{\boldsymbol{J}}_{\text{inj}} \cdot \boldsymbol{n})) + \widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\text{ext}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\text{inj}}) + \boldsymbol{B}_{\text{ext}} \tag{3.14}$$

as an abbreviation. The homogeneous solution $\widehat{y}$ is sought in the space

$$\widehat{Y} = \boldsymbol{H}_0^1(\Omega) \times L^2(\Omega)/\mathbb{R} \times \boldsymbol{L}^2(\Omega) \times H^1(\Omega)/\mathbb{R}. \tag{3.15}$$

In its strong form, (3.13) corresponds to

$$\begin{aligned} -\eta\Delta\widehat{\boldsymbol{u}} + \rho(\widehat{\boldsymbol{u}} \cdot \nabla)\widehat{\boldsymbol{u}} + \rho(\widehat{\boldsymbol{u}} \cdot \nabla)\boldsymbol{u}_0 + \rho(\boldsymbol{u}_0 \cdot \nabla)\widehat{\boldsymbol{u}} + \nabla p \\ = \eta\Delta\boldsymbol{u}_0 - \rho(\boldsymbol{u}_0 \cdot \nabla)\boldsymbol{u}_0 + \boldsymbol{J} \times \boldsymbol{B} + \boldsymbol{J}_0 \times \boldsymbol{B}, \\ \sigma^{-1}\widehat{\boldsymbol{J}} - \widehat{\boldsymbol{u}} \times \boldsymbol{B} + \nabla\phi = -\sigma^{-1}\boldsymbol{J}_0 + \boldsymbol{u}_0 \times \boldsymbol{B} \end{aligned}$$

plus the incompressibility conditions $\nabla \cdot \widehat{\boldsymbol{u}} = 0$ and $\nabla \cdot \widehat{\boldsymbol{J}} = 0$ and boundary conditions $\widehat{\boldsymbol{J}} \cdot \boldsymbol{n} = 0$ and $\widehat{\boldsymbol{u}} = 0$ on $\partial\Omega$. We note that the velocity boundary condition is incorporated into the space $\boldsymbol{H}_0^1(\Omega)$, whereas the boundary condition for the current is of variational type.

We now comment on the solvability of (3.13)–(3.14) and thus of the original system (3.1)–(3.7). As was observed in [22] for the MHD system without $\boldsymbol{J}_{\text{ext}}$ and $\boldsymbol{B}_{\text{ext}}$, the existence of a solution seems to be contingent upon the smallness of the liftings $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$, i.e., smallness of the boundary data $\boldsymbol{h}$ and $\boldsymbol{J}_{\text{inj}} \cdot \boldsymbol{n}$. Note that, while the Navier–Stokes nonlinearity $\rho(\widehat{\boldsymbol{u}} \cdot \nabla)\widehat{\boldsymbol{u}}$ is conservative in the sense that $b(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{u}}) = 0$, this is not the case for the bilinear terms $\boldsymbol{J} \times \boldsymbol{B}$ and $\widehat{\boldsymbol{u}} \times \boldsymbol{B}$. The existence proof given in [22] uses the LBB theory [7, Chap. IV, Thm. 1.2], and it is based on a limiting process of Galerkin approximations. Applying this technique to the present situation likewise yields solvability, provided that the data $\boldsymbol{h}$ and $\boldsymbol{J}_{\text{inj}} \cdot \boldsymbol{n}$ are sufficiently small. Under stronger assumptions involving also the remaining controls $\boldsymbol{J}_{\text{ext}}$ and $\boldsymbol{B}_{\text{ext}}$, uniqueness of the solution can be proved using [7, Chap. IV, Thm. 1.3]. In any case, the required bounds on the data seem not exactly tangible since they involve the embedding constants of $H^1(\Omega) \hookrightarrow L^p(\Omega)$ and the constant in the Poincaré inequality, as well as the norms of the lifting operator Λ and the Biot–Savart operator $\widetilde{\mathcal{B}}$. This is the reason that we refrain from stating the exact conditions here.

We provide here an alternative existence proof based on the Leray–Schauder fixed point theorem; see, for instance, [6, p. 222]. Let us define the operator $A : \widehat{Y} \to \widehat{Y}'$ by its components:

$$\begin{aligned}
A^1(\delta\boldsymbol{u}, \delta p, \delta\boldsymbol{J}, \delta\phi)(\boldsymbol{v}) &= a_1(\delta\boldsymbol{u}, \boldsymbol{v}) + d_1(\boldsymbol{v}, \delta p),\\
A^2(\delta\boldsymbol{u}, \delta p, \delta\boldsymbol{J}, \delta\phi)(q) &= d_1(\delta\boldsymbol{u}, q),\\
A^3(\delta\boldsymbol{u}, \delta p, \delta\boldsymbol{J}, \delta\phi)(\boldsymbol{K}) &= a_2(\delta\boldsymbol{J}, \boldsymbol{K}) + d_2(\boldsymbol{K}, \delta\phi),\\
A^4(\delta\boldsymbol{u}, \delta p, \delta\boldsymbol{J}, \delta\phi)(\psi) &= d_2(\delta\boldsymbol{J}, \psi).
\end{aligned}$$

We observe that A is an isomorphism since $(A_1, A_2)(\cdot, \cdot, 0, 0)$ defined between $\boldsymbol{H}^1_0(\Omega) \times L^2(\Omega)/\mathbb{R} \times \{0\} \times \{0\}$ and its dual is an isomorphism, and so is $(A_3, A_4)(0, 0, \cdot, \cdot)$, defined between $\{0\} \times \{0\} \times \boldsymbol{L}^2(\Omega) \times H^1(\Omega)/\mathbb{R}$ and its dual; see [22].

PROPOSITION 3.3 (state equation). *Suppose that Assumption* 3.1 *holds and that* $\|\boldsymbol{u}_0\|_{\boldsymbol{H}^1(\Omega)}$ *and* $\|\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}$ *are sufficiently small. Then the homogenized state equations* (3.13)–(3.14)*, and hence the original system* (3.1)–(3.7)*, possess at least one variational solution. Every such solution satisfies the a priori bound*

$$\begin{aligned}
\|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)} + \|\boldsymbol{J}\|^2_{\boldsymbol{L}^2(\Omega)} &\le c_1 \|\boldsymbol{J}_0\|^2_{\boldsymbol{L}^2(\Omega)} \left(1 + \|\boldsymbol{J}_0\|^2_{\boldsymbol{L}^2(\Omega)} + |u|^2\right)\\
&\quad + c_2 \|\boldsymbol{u}_0\|^2_{\boldsymbol{H}^1(\Omega)} \left(1 + \|\boldsymbol{J}_0\|^2_{\boldsymbol{L}^2(\Omega)} + |u|^2\right).
\end{aligned} \tag{3.16}$$

Moreover, we have the bound

$$\|p\|_{L^2(\Omega)/\mathbb{R}} + \|\phi\|_{H^1(\Omega)/\mathbb{R}} \le c_3 \left(\|\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)} + \|\boldsymbol{u}\|^2_{\boldsymbol{H}^1(\Omega)} + \|\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)} + \|\boldsymbol{J}\|^2_{\boldsymbol{L}^2(\Omega)}\right). \tag{3.17}$$

Proof. The proof can be found in the appendix. ∎

Remark 3.4. The constants c_1 and c_2 are of the form $\overline{c}_i(1 - \|\boldsymbol{J}_0\| - \|\boldsymbol{u}_0\|)^{-1}$. From the proof we infer that the larger the viscosity η of the fluid is and the smaller the conductivity σ is, the larger the liftings $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$ in Proposition 3.3 are allowed to become, and thus the larger the boundary data $\boldsymbol{h}$ and the control I_{inj} are allowed to become.

The variational form of the state equation (3.13) gives rise to the definition of the PDE constraint operator

$$e : \widehat{Y} \times \mathbb{R}^3 \to \widehat{Y}'. \tag{3.18}$$

The components $e^1(\widehat{y},u)(\boldsymbol{v}),\dots,e^4(\widehat{y},u)(\psi)$ are defined through the left-hand sides of (3.13). This concise form of the MHD system

$$e(\widehat{y},u)=0 \qquad \text{in } \widehat{Y}' \tag{3.19}$$

will be used below to argue existence of the Lagrange multipliers in the optimality system, based on the following results on the linearization of e whose proofs are only given as necessary.

LEMMA 3.5 (linearized PDE constraint). *The operator e is infinitely Fréchet differentiable. Its first order partial derivative with respect to the state variables in the direction of $\delta y=(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)$ is given by*

$$\begin{aligned}
e_y^1(\widehat{y},u)(\delta y)(\boldsymbol{v}) &= a_1(\delta\boldsymbol{u},\boldsymbol{v})-c(\delta\boldsymbol{J},\boldsymbol{B},\boldsymbol{v})-c(\widehat{\boldsymbol{J}}+\Lambda(I_{\mathrm{inj}}\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}),\widetilde{\mathcal{B}}(\delta\boldsymbol{J}),\boldsymbol{v})\\
&\quad+b(\delta\boldsymbol{u},\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\boldsymbol{v})+b(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\delta\boldsymbol{u},\boldsymbol{v})+d_1(\boldsymbol{v},\delta p),\\
e_y^2(\widehat{y},u)(\delta y)(q) &= d_1(\delta\boldsymbol{u},q),\\
e_y^3(\widehat{y},u)(\delta y)(\boldsymbol{K}) &= a_2(\delta\boldsymbol{J},\boldsymbol{K})+c(\boldsymbol{K},\boldsymbol{B},\delta\boldsymbol{u})+c(\boldsymbol{K},\widetilde{\mathcal{B}}(\delta\boldsymbol{J}),\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)+d_2(\boldsymbol{K},\delta\phi),\\
e_y^4(\widehat{y},u)(\delta y)(\psi) &= d_2(\delta\boldsymbol{J},\psi),
\end{aligned} \tag{3.20}$$

where we have set again $\boldsymbol{B}=\widetilde{\mathcal{B}}(\Lambda(I_{\mathrm{inj}}\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}))+\widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}})+\widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{ext}})+\widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{inj}})+\boldsymbol{B}_{\mathrm{ext}}$. As for the control variables, the first order derivative in the direction of $\delta u=(\delta I_{\mathrm{ext}},\delta I_{\mathrm{inj}},\delta B_{\mathrm{ext}})$ is

$$\begin{aligned}
e_u^1(\widehat{y},u)(\delta u)(\boldsymbol{v}) &= -c(\delta I_{\mathrm{inj}}\cdot\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}),\boldsymbol{B},\boldsymbol{v})-c(\widehat{\boldsymbol{J}}+\Lambda(I_{\mathrm{inj}}\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}),\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}))\\
&\quad+\delta I_{\mathrm{ext}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}})+\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}})+\delta B_{\mathrm{ext}}\cdot\widetilde{\boldsymbol{B}}_{\mathrm{ext}},\boldsymbol{v}),\\
e_u^2(\widehat{y},u)(\delta u)(q) &= 0,\\
e_u^3(\widehat{y},u)(\delta u)(\boldsymbol{K}) &= a_2(\delta I_{\mathrm{inj}}\cdot\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}),\boldsymbol{K}),\\
&\quad+c(\boldsymbol{K},\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}))+\delta I_{\mathrm{ext}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}})+\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}})+\delta B_{\mathrm{ext}}\cdot\widetilde{\boldsymbol{B}}_{\mathrm{ext}},\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)\\
e_u^4(\widehat{y},u)(\delta u)(\psi) &= 0.
\end{aligned} \tag{3.21}$$

We note that in its strong form, the system $e_y(\widehat{y},u)(\delta y)=(\bar{e},\bar{f},\bar{g},\bar{h})\in\widehat{Y}'$ corresponds to

$$\begin{aligned}
-\eta\Delta\delta\boldsymbol{u}+\rho(\delta\boldsymbol{u}\cdot\nabla)(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)+\rho((\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)\cdot\nabla)\delta\boldsymbol{u}+\nabla\delta p &= \delta\boldsymbol{J}\times\boldsymbol{B}+(\widehat{\boldsymbol{J}}+\boldsymbol{J}_0)\times\widetilde{\mathcal{B}}(\delta\boldsymbol{J})+\bar{e},\\
\nabla\cdot\delta\boldsymbol{u} &= \bar{f},\\
\sigma^{-1}\delta\boldsymbol{J}-\delta\boldsymbol{u}\times\boldsymbol{B}-(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)\times\widetilde{\mathcal{B}}(\delta\boldsymbol{J})+\nabla\delta\phi &= \bar{g},\\
\nabla\cdot\boldsymbol{J} &= \bar{h}.
\end{aligned}$$

The first and third components of $e_u(\widehat{y},u)\delta u$ must be read as

$$\begin{aligned}
&-\delta I_{\mathrm{inj}}\cdot\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n})\times\boldsymbol{B}-(\widehat{\boldsymbol{J}}+\Lambda(I_{\mathrm{inj}}\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}))\\
&\qquad\times\left(\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}))+\delta I_{\mathrm{ext}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}})+\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}})+\delta B_{\mathrm{ext}}\cdot\widetilde{\boldsymbol{B}}_{\mathrm{ext}}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\sigma^{-1}\delta I_{\mathrm{inj}}\cdot\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n})-(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)\\
&\qquad\times\left(\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\Lambda(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n}))+\delta I_{\mathrm{ext}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}})+\delta I_{\mathrm{inj}}\cdot\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}})+\delta B_{\mathrm{ext}}\cdot\widetilde{\boldsymbol{B}}_{\mathrm{ext}}\right),
\end{aligned}$$

respectively.

In preparation for the following proposition, let us define the operator $C : \widehat{Y} \to \widehat{Y}'$ by its components:

$$\begin{aligned}
C^1(\widehat{y},u)(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)(\boldsymbol{v}) &= -c(\delta\boldsymbol{J},\boldsymbol{B},\boldsymbol{v}) - c(\widehat{\boldsymbol{J}}+\boldsymbol{J}_0,\widetilde{\mathcal{B}}(\delta\boldsymbol{J}),\boldsymbol{v}) + b(\delta\boldsymbol{u},\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\boldsymbol{v}) \\
&\quad + b(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\delta\boldsymbol{u},\boldsymbol{v}),\\
C^2(\widehat{y},u)(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)(q) &= 0,\\
C^3(\widehat{y},u)(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)(\boldsymbol{K}) &= c(\boldsymbol{K},\boldsymbol{B},\delta\boldsymbol{u}) + c(\boldsymbol{K},\widetilde{\mathcal{B}}(\delta\boldsymbol{J}),\widehat{\boldsymbol{u}}+\boldsymbol{u}_0),\\
C^4(\widehat{y},u)(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)(\psi) &= 0.
\end{aligned}$$

The quantities $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$ are the liftings according to Lemma 3.2 for given arbitrary but fixed controls u and boundary data $\boldsymbol{h}$.

PROPOSITION 3.6 (linearized state equation). *For any given $\widehat{y} \in \widehat{Y}$, $u \in \mathbb{R}^3$, and $\boldsymbol{h} \in \boldsymbol{H}^{1/2}(\partial\Omega)$, the linearization with respect to the state variables of the operator e can be decomposed as*

$$e_y(\widehat{y},u) = A + C(\widehat{y},u), \tag{3.22}$$

where $A : \widehat{Y} \to \widehat{Y}'$ is an isomorphism, independent of $(\widehat{y},u)$, and $C(\widehat{y},u) : \widehat{Y} \to \widehat{Y}'$ is a compact linear operator.

Proof. The isomorphism property of A has been noted previously; see the definition of A immediately preceding Proposition 3.3. As for compactness of C, we recall that $\boldsymbol{B} \in \boldsymbol{L}^{3+\varepsilon}(\Omega)$ (see Lemma 2.4 and Assumption 3.1) and infer from Lemma 2.6 that $\delta\boldsymbol{J} \mapsto \delta\boldsymbol{J} \times \boldsymbol{B}$ is compact from $\boldsymbol{L}^2(\Omega)$ to $\boldsymbol{H}^{-1}(\Omega)$. Hence $\delta\boldsymbol{J} \mapsto c(\delta\boldsymbol{J},\boldsymbol{B},\cdot)$ is compact from $\boldsymbol{L}^2(\Omega)$ to $\boldsymbol{H}^{-1}(\Omega)$. Similarly, by Lemmas 2.4 and 2.6, $\delta\boldsymbol{J} \mapsto \boldsymbol{J} \times \widetilde{\mathcal{B}}(\delta\boldsymbol{J})$ is compact from $\boldsymbol{L}^2(\Omega)$ to $\boldsymbol{H}^{-1}(\Omega)$, and hence $\delta\boldsymbol{J} \mapsto c(\boldsymbol{J},\widetilde{\mathcal{B}}(\delta\boldsymbol{J}),\cdot)$ is compact from $\boldsymbol{L}^2(\Omega)$ to $\boldsymbol{H}^{-1}(\Omega)$. In addition, $\delta\boldsymbol{u} \mapsto b(\delta\boldsymbol{u},\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\cdot)$ and $\delta\boldsymbol{u} \mapsto b(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\delta\boldsymbol{u},\cdot)$ are continuous from $\boldsymbol{H}^1_0(\Omega)$ to $\boldsymbol{L}^{3/2}(\Omega)$, which embeds compactly into $\boldsymbol{W}^{-1,3-\varepsilon}(\Omega)$ for all $\varepsilon > 0$ and thus into $\boldsymbol{H}^{-1}(\Omega)$. This completes the proof of compactness for C^1. As for C^3, we let $p = 2(3+\varepsilon)/(1+\varepsilon) < 6$ and observe that $\delta\boldsymbol{u} \mapsto c(\cdot,\boldsymbol{B},\delta\boldsymbol{u})$ is continuous from $\boldsymbol{L}^p(\Omega)$ to $\boldsymbol{L}^2(\Omega)$. In view of the compact embedding $\boldsymbol{H}^1_0(\Omega) \hookrightarrow \boldsymbol{L}^p(\Omega)$, this map is compact from $\boldsymbol{H}^1_0(\Omega)$ to $\boldsymbol{L}^2(\Omega)$. Finally, since $\delta\boldsymbol{J} \mapsto \widetilde{\mathcal{B}}(\delta\boldsymbol{J})$ is continuous from $\boldsymbol{L}^2(\Omega)$ to $V^1(\Omega)$, which embeds compactly into $\boldsymbol{L}^3(\Omega)$, the map $\delta\boldsymbol{J} \mapsto c(\cdot,\widetilde{\mathcal{B}}(\delta\boldsymbol{J}),\widehat{\boldsymbol{u}}+\boldsymbol{u}_0)$ is compact from $\boldsymbol{L}^2(\Omega)$ to $\boldsymbol{L}^2(\Omega)$, which completes the proof. ∎

The previous proposition allows us to draw the following conclusions about the properties of the linearized state operator $e_y(\widehat{y},u)$.

PROPOSITION 3.7 (bounded invertibility of $e_y(\widehat{y},u)$). *Except for a countable set of (η,σ) values, the operator $e_y(\widehat{y},u)$ is an isomorphism. Moreover, $e_y(\widehat{y},u)$ is an isomorphism whenever η is sufficiently large and σ is sufficiently small.*

Proof. For $(\bar{e},\bar{f},\bar{g},\bar{h}) \in \widehat{Y}'$, consider the equation

$$\left(A + C(\widehat{y},u)\right)(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi) = (\bar{e},\bar{f},\bar{g},\bar{h}) \tag{3.23}$$

and define $\mathcal{A} : \widehat{Y} \to \widehat{Y}'$ through its coordinates

$$\begin{aligned}
\mathcal{A}^1(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)(\boldsymbol{v}) &= \int_\Omega \nabla\delta\boldsymbol{u} : \nabla\boldsymbol{v} - \int_\Omega (\nabla\cdot\boldsymbol{v})\,\mathcal{P}(\delta p),\\
\mathcal{A}^2 &= A^2,\\
\mathcal{A}^3(\delta\boldsymbol{u},\delta p,\delta\boldsymbol{J},\delta\phi)(\boldsymbol{K}) &= \int_\Omega \delta\boldsymbol{J}\cdot\boldsymbol{K} + \int_\Omega \boldsymbol{K}\cdot\nabla(\delta\phi),\\
\mathcal{A}^4 &= A^4,
\end{aligned}$$

where $(\boldsymbol{v}, \boldsymbol{K}) \in \boldsymbol{H}_0^1(\Omega) \times \boldsymbol{L}^2(\Omega)$, i.e., $\mathcal{A}$ arises from A by setting $\eta = \sigma = 1$. Multiplying the first two equations in (3.23) by η^{-1} and the last two by σ, we find that the following equation is equivalent to (3.23):

$$(\mathcal{A} + \eta^{-1}\mathcal{C}_1 + \sigma\mathcal{C}_2)(\delta\boldsymbol{u}, \widehat{\delta p}, \delta\boldsymbol{J}, \widehat{\delta\phi}) = (\eta^{-1}\bar{e}, \bar{f}, \sigma\bar{g}, \bar{h}), \tag{3.24}$$

where $\widehat{\delta p} = \eta^{-1}\delta p$ and $\widehat{\delta\phi} = \sigma\delta\phi$ and $\mathcal{C}_1 = (C_1, C_2, 0, 0)^\top$, $\mathcal{C}_2 = (0, 0, C_3, C_4)^\top$. From the proof of Proposition 3.6 we have that $\mathcal{A} : \widehat{Y} \to \widehat{Y}'$ is an isomorphism and that $\mathcal{C}_1$ and $\mathcal{C}_2$ are compact operators from $\widehat{Y}$ to $\widehat{Y}'$. Moreover, (3.24) is equivalent to

$$(I + \eta^{-1}\mathcal{K}_1 + \sigma\mathcal{K}_2)(\delta\boldsymbol{u}, \widehat{\delta p}, \delta\boldsymbol{J}, \widehat{\delta\phi}) = (e, f, g, h), \tag{3.25}$$

where $(e, f, g, h) = \mathcal{A}^{-1}(\eta^{-1}\bar{e}, \bar{f}, \sigma\bar{g}, \bar{h})$, and

$$\mathcal{K}_1 = \mathcal{A}^{-1}\mathcal{C}_1, \qquad \mathcal{K}_2 = \mathcal{A}^{-1}\mathcal{C}_2.$$

Hence, $\mathcal{K}_1$ and $\mathcal{K}_2$ are compact operators in $\widehat{Y}$. Therefore the spectrum of $\mathcal{K}_1$, denoted by $\sum(\mathcal{K}_1)$, consists of 0 and at most countably many eigenvalues, with 0 being the only possible accumulation point. For $-\eta \notin \sum(\mathcal{K}_1)$ we have

$$(I + \sigma(I + \eta^{-1}\mathcal{K}_1)^{-1}\mathcal{K}_2)(\delta\boldsymbol{u}, \widehat{\delta p}, \delta\boldsymbol{J}, \widehat{\delta\phi}) = (I + \eta^{-1}\mathcal{K}_1)^{-1}(e, f, g, h). \tag{3.26}$$

Since $(I + \eta^{-1}\mathcal{K}_1)$ has a continuous inverse, we find that (3.26), and hence (3.23) are solvable if $-\sigma^{-1} \notin \sum((I + \eta^{-1}\mathcal{K}_1)^{-1}\mathcal{K}_2)$. Since the set of points $\mathcal{S} := \{(\eta, \sigma^{-1}) : -\eta \in \Sigma(\mathcal{K}_1), -\sigma^{-1} \in \Sigma((I - \eta^{-1}\mathcal{K}_1)^{-1}\mathcal{K}_2)\}$ is countable in $\mathbb{R}^2$, the first claim follows. The second claim is a consequence of a Neumann series argument applied to (3.25). ∎

From the proof of Proposition 3.7 we conclude that $e_y(\widehat{y}, u)$, and therefore $e_x(\widehat{y}, u)$, is generically surjective. Lack of surjectivity of $e_y(\widehat{y}, u)$ occurs if and only if $(\eta, \sigma^{-1}) \in \mathcal{S}$. In that case, Fredholm's alternative provides another way to prove surjectivity of $e_x(\widehat{y}, u)$. Note that $e_u(\widehat{y}, u)\,\delta u$ can be written in the form

$$e_u(\widehat{y}, u)\,\delta u = \delta I_{\text{inj}}\,\psi_1 + \delta I_{\text{ext}}\,\psi_2 + \delta B_{\text{ext}}\,\psi_3$$

with $\psi_i \in \widehat{Y}'$. In the following proposition, $(A + C)^\star : \widehat{Y}' \to \widehat{Y}$ refers to the Hilbert space adjoint.

PROPOSITION 3.8 (surjectivity of $e_x(\widehat{y}, u)$). *If span* $\{\psi_1, \psi_2, \psi_3\} \supseteq \ker(A + C)^\star$, *then* $e_x(\widehat{y}, u)$ *is surjective.*

Proof. Since $A : \widehat{Y} \to \widehat{Y}'$ is an isomorphism and C is compact, the Fredholm alternative implies that the range of $A + C$ is closed and $R(A + C)^\perp = \ker(A + C)^\star$, with $\dim \ker(A + C)^\star =: L < \infty$. Let $\{\omega_i\}_{i=1}^L \subset \widehat{Y}'$ be a basis for $\ker(A + C)^\star$, orthonormalized such that $\langle \omega_i, \omega_j \rangle_{\widehat{Y}'} = \delta_{ij}$. For arbitrary $\widehat{f} \in \widehat{Y}'$, define $\widetilde{f} = \widehat{f} - \sum_{i=1}^L \langle \widehat{f}, \omega_i \rangle_{\widehat{Y}'} \cdot \omega_i$. Then $\widetilde{f} \in \ker(A+C)^{\star,\perp}$ and there exists $x_{\widetilde{f}}$ such that $(A+C)x_{\widetilde{f}} = \widetilde{f}$. By assumption there exist $(\delta I_{\text{inj}}, \delta I_{\text{ext}}, \delta\boldsymbol{B}_{\text{ext}})$ such that $e_u(\widehat{y}, u)\delta u = \widehat{f} - \widetilde{f}$ and the claim follows. ∎

Having established these properties for the stationary MHD system and its linearization, we now return to the optimal control problem. We recall that we aim to

$$
\begin{aligned}
&\text{minimize } f(\widehat{y},u) = \frac{\alpha_u}{2}\|\boldsymbol{u}-\boldsymbol{u}_{\mathrm{d}}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \frac{\alpha_B}{2}\|\boldsymbol{B}-\boldsymbol{B}_{\mathrm{d}}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} \\
\text{(P)}\qquad &\qquad + \frac{\alpha_J}{2}\|\boldsymbol{J}-\boldsymbol{J}_{\mathrm{d}}\|^2_{L^2(\Omega_{J,\mathrm{obs}})} + \frac{\gamma_{\mathrm{ext}}}{2}|I_{\mathrm{ext}}|^2 + \frac{\gamma_{\mathrm{inj}}}{2}|I_{\mathrm{inj}}|^2 + \frac{\gamma_B}{2}|B_{\mathrm{ext}}|^2 \\
&\qquad \text{over } \widehat{y}\in\widehat{Y} \text{ and } u\in U_{\mathrm{ad}} \\
&\qquad \text{subject to } e(\widehat{y},u)=0,
\end{aligned}
$$

where $\boldsymbol{u}=\widehat{\boldsymbol{u}}+\boldsymbol{u}_0$ and $\boldsymbol{J}=\widehat{\boldsymbol{J}}+\boldsymbol{J}_0$. $\boldsymbol{B}$ is defined in (3.14), and U_{ad} in Assumption 3.9 below. Recall also that the lifting $\boldsymbol{J}_0$, and hence $\boldsymbol{B}$, depends on the control I_{inj}. In order to ensure well-posedness of problem (P), we need the following assumption on the problem data.

Assumption 3.9 (control problem data). Assume that the boundary data

$$\|\boldsymbol{h}\|_{\boldsymbol{H}^{1/2}(\partial\Omega)}$$

is sufficiently small, and that for some fixed $r>0$ and all controls in the set

$$U_{\mathrm{ad}} = \{(I_{\mathrm{ext}}, I_{\mathrm{inj}}, B_{\mathrm{ext}})\in\mathbb{R}^3 \,:\, |I_{\mathrm{inj}}|\le r\},$$

the liftings $\boldsymbol{u}_0$ and $\boldsymbol{J}_0=\Lambda(I_{\mathrm{inj}}\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n})$ allow the existence of a solution to the stationary MHD system.

In fact, smallness of $\boldsymbol{h}$ and I_{inj} implies by Lemma 3.2 smallness of $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$, which in turn implies existence of solutions according to Proposition 3.3.

Proposition 3.10 (existence of a global minimum). *Under Assumptions* 3.1 *and* 3.9, *problem* (P) *possesses at least one global optimal solution in* $\widehat{Y}\times U_{\mathrm{ad}}$.

Proof. The proof follows along the usual lines. We set $m=\inf f(\widehat{y},u)$, where the infimum extends over all state/control pairs $(\widehat{y},u)\in\widehat{Y}\times U_{\mathrm{ad}}$ which satisfy the state equation (admissible pairs). Note that m is nonnegative and finite since f is nonnegative and the set of admissible pairs is nonempty (Assumption 3.9 and Proposition 3.3). Now if $\{(\widehat{y}^n,u^n)\}$ is a minimizing sequence, we can infer from the cost functional that the controls u^n are bounded in $\mathbb{R}^3$. By the a priori estimate (3.16), $\widehat{y}^n$ is bounded in $\widehat{Y}$. We extract weakly convergent subsequences, still denoted by index n, such that

$$
\begin{aligned}
&\widehat{\boldsymbol{u}}^n \rightharpoonup \widehat{\boldsymbol{u}} \quad \text{in } \boldsymbol{H}^1_0(\Omega), && p^n \rightharpoonup p \quad \text{in } L^2(\Omega)/\mathbb{R},\\
&\widehat{\boldsymbol{J}}^n \rightharpoonup \widehat{\boldsymbol{J}} \quad \text{in } \boldsymbol{L}^2(\Omega), && \phi^n \rightharpoonup \phi \quad \text{in } H^1(\Omega)/\mathbb{R},\\
&u^n \to u \quad \text{in } \mathbb{R}^3. &&
\end{aligned}
$$

Note that the lifting $\boldsymbol{u}_0$ is independent of n and that $\boldsymbol{J}_0^n=\Lambda(I_{\mathrm{inj}}^n\widetilde{\boldsymbol{J}}_{\mathrm{inj}}\cdot\boldsymbol{n})$ converges strongly to some $\boldsymbol{J}_0$ in $\boldsymbol{L}^2(\Omega)$. In order to pass to the limit in $e(\widehat{y}^n,u^n)$, we consider the individual terms in (3.13). For the terms involving the bilinear forms a_i and d_i, the convergence is evident. In addition, $b(\widehat{\boldsymbol{u}}^n,\widehat{\boldsymbol{u}}^n,\boldsymbol{v})\to b(\widehat{\boldsymbol{u}},\widehat{\boldsymbol{u}},\boldsymbol{v})$ is known from the theory of Navier–Stokes problems, see [7, Chap. IV, Thm. 2.1]. The convergence of all terms involving the trilinear form c follows from Lemmas 2.3 and 2.6. Consequently, the weak limit $(\widehat{y},u)$ satisfies the state equation $e(\widehat{y},u)=0$, and hence the weak limit $(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,p,\widehat{\boldsymbol{J}}+\boldsymbol{J}_0,\phi)$ satisfies our inhomogeneous MHD system. The claim now follows from the weak lower semicontinuity of the objective, by which

$$m\le f(\widehat{y},u)\le\liminf_{n\to\infty} f(\widehat{y}^n,u^n)=m. \qquad \square$$

THEOREM 3.11 (optimality system). *Let Assumptions* 3.1 *and* 3.9 *hold, and let the state* $\widehat{y} = (\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi) \in \widehat{Y}$ *and control* $u = (I_{\mathrm{ext}}, I_{\mathrm{inj}}, B_{\mathrm{ext}}) \in U_{\mathrm{ad}}$ *constitute a local optimal pair for problem* (P). *In addition, let* $e_x(\widehat{y}, u)$ *be surjective. Then there exists a unique Lagrange multiplier*

$$\lambda = (\boldsymbol{v}, q, \boldsymbol{K}, \psi) \in \widehat{Y}$$

which satisfies the adjoint equations

$$\begin{aligned} &a_1(\delta\boldsymbol{u}, \boldsymbol{v}) + b(\delta\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) + b(\boldsymbol{u}, \delta\boldsymbol{u}, \boldsymbol{v}) + d_1(\delta\boldsymbol{u}, q) + c(\boldsymbol{K}, \boldsymbol{B}, \delta\boldsymbol{u}) \\ &\qquad + \alpha_u \int_{\Omega_{u,\mathrm{obs}}} (\boldsymbol{u} - \boldsymbol{u}_{\mathrm{d}}) \cdot \delta\boldsymbol{u} = 0, \end{aligned} \tag{3.27a}$$

$$d_1(\boldsymbol{v}, \delta p) = 0, \tag{3.27b}$$

$$\begin{aligned} &-c(\delta\boldsymbol{J}, \boldsymbol{B}, \boldsymbol{v}) - c(\boldsymbol{J}, \widetilde{\mathcal{B}}(\delta\boldsymbol{J}), \boldsymbol{v}) + a_2(\delta\boldsymbol{J}, \boldsymbol{K}) + c(\boldsymbol{K}, \widetilde{\mathcal{B}}(\delta\boldsymbol{J}), \boldsymbol{u}) \\ &\quad + d_2(\delta\boldsymbol{J}, \psi) + \alpha_B \int_{\Omega_{B,\mathrm{obs}}} (\boldsymbol{B} - \boldsymbol{B}_{\mathrm{d}}) \cdot \widetilde{\mathcal{B}}(\delta\boldsymbol{J}) + \alpha_J \int_{\Omega_{J,\mathrm{obs}}} (\boldsymbol{J} - \boldsymbol{J}_{\mathrm{d}}) \cdot \delta\boldsymbol{J} = 0, \end{aligned} \tag{3.27c}$$

$$d_2(\boldsymbol{K}, \delta\phi) = 0 \tag{3.27d}$$

for all $(\delta\boldsymbol{u}, \delta p, \delta\boldsymbol{J}, \delta\phi) \in \boldsymbol{H}^1_0(\Omega) \times L^2(\Omega)/\mathbb{R} \times \boldsymbol{L}^2(\Omega) \times H^1(\Omega)/\mathbb{R}$, *and which satisfy the three scalar optimality conditions*

$$-c(\boldsymbol{J}, \widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}}), \boldsymbol{v}) + c(\boldsymbol{K}, \widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}}), \boldsymbol{u}) + \gamma_{\mathrm{ext}} I_{\mathrm{ext}} + \alpha_B \int_{\Omega_{B,\mathrm{obs}}} (\boldsymbol{B} - \boldsymbol{B}_{\mathrm{d}}) \cdot \widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{ext}}) = 0, \tag{3.28a}$$

$$\begin{aligned} &\Big[-c(\boldsymbol{J}, \widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}), \boldsymbol{v}) + c(\boldsymbol{K}, \widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}), \boldsymbol{u}) + \gamma_{\mathrm{inj}} I_{\mathrm{inj}} + \alpha_B \int_{\Omega_{B,\mathrm{obs}}} (\boldsymbol{B} - \boldsymbol{B}_{\mathrm{d}}) \cdot \widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_{\mathrm{inj}}) \\ &\quad - \big\langle \widetilde{\boldsymbol{J}}_{\mathrm{inj}} \cdot \boldsymbol{n}, \psi \big\rangle_{\partial\Omega \cap \partial\Omega_{\mathrm{inj}}} \Big] (\overline{I}_{\mathrm{inj}} - I_{\mathrm{inj}}) \geq 0 \quad \textit{for all } |\overline{I}_{\mathrm{inj}}| \leq r, \end{aligned} \tag{3.28b}$$

$$-c(\boldsymbol{J}, \widetilde{\boldsymbol{B}}_{\mathrm{ext}}, \boldsymbol{v}) + c(\boldsymbol{K}, \widetilde{\boldsymbol{B}}_{\mathrm{ext}}, \boldsymbol{u}) + \gamma_{\mathrm{B}} B_{\mathrm{ext}} + \alpha_B \int_{\Omega_{B,\mathrm{obs}}} (\boldsymbol{B} - \boldsymbol{B}_{\mathrm{d}}) \cdot \widetilde{\boldsymbol{B}}_{\mathrm{ext}} = 0. \tag{3.28c}$$

Proof. Our proof relies on a classical abstract multiplier result; see, e.g., Maurer and Zowe [20]. Since f is Fréchet differentiable, e is continuously Fréchet differentiable, and e_x is assumed surjective at $(\widehat{y}, u)$, it follows that there exists a Lagrange multiplier $\lambda \in \widehat{Y}$, which satisfies

$$\begin{aligned} &f_y(\widehat{y}, u)(\delta y) + \big\langle \lambda, e_y(\widehat{y}, u)(\delta y) \big\rangle = 0 \qquad\qquad \text{for all } \delta y \in \widehat{Y}, \\ &f_u(\widehat{y}, u)(\overline{\delta u} - \delta u) + \big\langle \lambda, e_u(\widehat{y}, u)(\overline{\delta u} - \delta u) \big\rangle \geq 0 \text{ for all } \overline{\delta u} \in U_{\mathrm{ad}}. \end{aligned} \tag{3.29}$$

Above, the duality holds in $\widehat{Y} \times \widehat{Y}'$. It is now straightforward to verify that (3.29) is nothing else but (3.27)–(3.28). ∎

The elements of the Lagrange multiplier λ are termed the adjoint velocity $\boldsymbol{v}$, the adjoint pressure q, the adjoint current density $\boldsymbol{K}$, and the adjoint potential ψ, respectively, all defined on Ω.

In order to improve our understanding of the adjoint system (3.27), we also paraphrase it in its strong form. Exploiting the self-adjointness of the linear Biot–Savart

operator $\widetilde{\mathcal{B}}$, we obtain the dual state equations on Ω,

$$\begin{aligned} \varrho(\nabla \boldsymbol{u})^\top \boldsymbol{v} - \varrho(\boldsymbol{u}\cdot\nabla)\boldsymbol{v} - \eta\Delta\boldsymbol{v} + \nabla q - \boldsymbol{B}\times\boldsymbol{K} &= -\alpha_u \chi_{\Omega_{u,\mathrm{obs}}}(\boldsymbol{u}-\boldsymbol{u}_\mathrm{d}), \qquad (3.30)\\ \sigma^{-1}\boldsymbol{K} - \widetilde{\mathcal{B}}(\boldsymbol{K}\times\boldsymbol{u} + \boldsymbol{v}\times\boldsymbol{J}) + \nabla\psi - \boldsymbol{B}\times\boldsymbol{v} &= -\alpha_J\chi_{\Omega_{J,\mathrm{obs}}}(\boldsymbol{J}-\boldsymbol{J}_\mathrm{d}) - \alpha_B\chi_{\Omega_{B,\mathrm{obs}}\cap\Omega}\widetilde{\mathcal{B}}(\boldsymbol{B}-\boldsymbol{B}_\mathrm{d}), \qquad (3.31)\end{aligned}$$

plus incompressibility conditions

$$\nabla\cdot\boldsymbol{v} = 0 \quad \text{on } \Omega, \qquad\qquad \nabla\cdot\boldsymbol{K} = 0 \quad \text{on } \Omega \tag{3.32}$$

and boundary conditions

$$\boldsymbol{v} = 0 \quad \text{on } \partial\Omega, \qquad\qquad \boldsymbol{K}\cdot\boldsymbol{n} = 0 \quad \text{on } \partial\Omega. \tag{3.33}$$

The symbol χ_A denotes the characteristic function of a set A. Note that the controls I_ext, I_inj, and B_ext appear (hidden within $\boldsymbol{B}$) in the adjoint equations (3.27a) and (3.27c), or in (3.30)–(3.31) for that matter. The reason is that the state equations (3.1)–(3.2) contain the "state times control" terms $\boldsymbol{u}\times\boldsymbol{B}$ (through Ohm's Law (1.3)) and $\boldsymbol{J}\times\boldsymbol{B}$ (the Lorentz force).

Let us now turn to a condition which ensures the strict local optimality of a given point (y,u,λ) satisfying the first order optimality condition set forth in Theorem 3.11. To this end, we define the Lagrangian function for problem (P) as

$$\begin{aligned}\mathcal{L}(\widehat{y},u,\lambda) = {}& \frac{\alpha_u}{2}\|\boldsymbol{u}-\boldsymbol{u}_\mathrm{d}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \frac{\alpha_B}{2}\|\boldsymbol{B}-\boldsymbol{B}_\mathrm{d}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} + \frac{\alpha_J}{2}\|\boldsymbol{J}-\boldsymbol{J}_\mathrm{d}\|^2_{L^2(\Omega_{J,\mathrm{obs}})}\\ &+ \frac{\gamma_\mathrm{ext}}{2}|I_\mathrm{ext}|^2 + \frac{\gamma_\mathrm{inj}}{2}|I_\mathrm{inj}|^2 + \frac{\gamma_B}{2}|B_\mathrm{ext}|^2 + a_1(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\boldsymbol{v}) - c(\widehat{\boldsymbol{J}}+\boldsymbol{J}_0,\boldsymbol{B},\boldsymbol{v})\\ &+ b(\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\widehat{\boldsymbol{u}}+\boldsymbol{u}_0,\boldsymbol{v}) + d_1(\boldsymbol{v},p) - d_1(\widehat{\boldsymbol{u}},q) + a_2(\widehat{\boldsymbol{J}}+\boldsymbol{J}_0,\boldsymbol{K})\\ &+ c(\boldsymbol{K},\boldsymbol{B},\widehat{\boldsymbol{u}}+\boldsymbol{u}_0) + d_2(\boldsymbol{K},\phi) - d_2(\widehat{\boldsymbol{J}},\psi) + \big\langle I_\mathrm{inj}\widetilde{\boldsymbol{J}}_\mathrm{inj}\cdot\boldsymbol{n},\psi\big\rangle_{\partial\Omega\cap\partial\Omega_\mathrm{inj}},\end{aligned}$$

where again $\boldsymbol{u} = \widehat{\boldsymbol{u}}+\boldsymbol{u}_0$, $\boldsymbol{J} = \widehat{\boldsymbol{J}}+\Lambda(I_\mathrm{inj}\widetilde{\boldsymbol{J}}_\mathrm{inj}\cdot\boldsymbol{n})$, and $\boldsymbol{B}$ is defined in (3.14). Moreover, $\boldsymbol{u}_0$ is the fixed lifting of the velocity boundary data $\boldsymbol{h}$ from Lemma 3.2, and $\boldsymbol{B}$ is still taken according to (3.14). It is readily checked that the following quadratic form is the Hessian of $\mathcal{L}$ with respect to the state/control pair:

$$\begin{aligned}\mathcal{L}''(\widehat{y},u,\lambda)[(\delta y,\delta u)]^2 = {}& \alpha_u\|\delta\boldsymbol{u}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \alpha_J\|\delta\boldsymbol{J}\|^2_{L^2(\Omega_{J,\mathrm{obs}})} + \alpha_B\|\delta\widetilde{\mathcal{B}}\|^2_{L^2(\Omega_{B,\mathrm{obs}})}\\ &+ \gamma_\mathrm{ext}|\delta I_\mathrm{ext}|^2 + \gamma_\mathrm{inj}|\delta I_\mathrm{inj}|^2 + \gamma_B|\delta B_\mathrm{ext}|^2 + 2b(\delta\boldsymbol{u},\delta\boldsymbol{u},\boldsymbol{v})\\ &- 2c(\delta\boldsymbol{J},\delta\widetilde{\mathcal{B}},\boldsymbol{v}) + 2c(\boldsymbol{K},\delta\widetilde{\mathcal{B}},\delta\boldsymbol{u})\end{aligned}$$

with the abbreviation

$$\delta\widetilde{\mathcal{B}} = \widetilde{\mathcal{B}}(\delta\boldsymbol{J}) + \delta I_\mathrm{inj}\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_\mathrm{inj}) + \delta I_\mathrm{ext}\widetilde{\mathcal{B}}(\widetilde{\boldsymbol{J}}_\mathrm{ext}) + \delta B_\mathrm{ext}\widetilde{\boldsymbol{B}}_\mathrm{ext}.$$

PROPOSITION 3.12 (second order sufficient conditions). *Suppose that $(\widehat{y},u,\lambda)$ satisfies the optimality system consisting of* (3.13)–(3.14), (3.27)–(3.28), *and that $e_y(\widehat{y},u)$ is boundedly invertible. If, moreover,*

$$\alpha_u\|\boldsymbol{u}-\boldsymbol{u}_\mathrm{d}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \alpha_B\|\boldsymbol{B}-\boldsymbol{B}_\mathrm{d}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} + \alpha_J\|\boldsymbol{J}-\boldsymbol{J}_\mathrm{d}\|^2_{L^2(\Omega_{J,\mathrm{obs}})}$$

is sufficiently small, then there exists a neighborhood $\mathcal{U}$ *of* $(\widehat{y},u)$ *and* $\kappa>0$ *such that*

$$f(\overline{y},\overline{u}) \geq f(\widehat{y},u) + \kappa\left(\|\overline{u}-u\|^2 + \|\overline{y}-\widehat{y}\|^2_{\widehat{Y}}\right)$$

holds for all $(\overline{y},\overline{u})\in\mathcal{U}$ *satisfying the state equation. In particular,* $(\widehat{y},u)$ *is a strict local optimum for* (P).

Proof. We shall argue that there exists $\rho>0$ such that the coercivity condition

$$\mathcal{L}''(\widehat{y},u,\lambda)[(\delta y,\delta u)]^2 \geq \rho\left(\|\delta y\|^2_{\widehat{Y}} + |\delta u|^2\right) \tag{3.34}$$

holds for all $(\delta y,\delta u)\in\widehat{Y}\times\mathbb{R}^3$, which satisfy the linear MHD system (see Lemma 3.5)

$$e_y(\widehat{y},u)\,\delta y + e_u(\widehat{y},u)\,\delta u = 0. \tag{3.35}$$

The claim then follows from a Taylor series expansion of $\mathcal{L}$ at $(\widehat{y},u,\lambda)$; see, e.g., [20, Thm. 5.6]. In fact, since $\|\delta p\|_{L^2(\Omega)/\mathbb{R}} + \|\delta\phi\|_{H^1(\Omega)/\mathbb{R}} \leq c\,(\|\delta\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)} + \|\delta\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)})$ holds (cf. Lemma 2.2), we need to verify (3.34) only for the components $\delta\boldsymbol{u}$ and $\delta\boldsymbol{J}$ of δy.

Since $e_y(\widehat{y},u)$ is surjective, $e_y^\star(\widehat{y},u):\widehat{Y}\to\widehat{Y}'$ has closed range and is continuously invertible on its range [2]. Hence in view of (3.30)–(3.33), there exists $\kappa_1>0$ such that

$$\begin{aligned}&\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)} + \|\boldsymbol{K}\|_{\boldsymbol{L}^2(\Omega)}\\ &\quad\leq \kappa_1\left(\alpha_u\|\boldsymbol{u}-\boldsymbol{u}_{\mathrm{d}}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \alpha_B\|\boldsymbol{B}-\boldsymbol{B}_{\mathrm{d}}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} + \alpha_J\|\boldsymbol{J}-\boldsymbol{J}_{\mathrm{d}}\|^2_{L^2(\Omega_{J,\mathrm{obs}})}\right)\end{aligned}$$

holds. From (3.35) and the bounded invertibility of $e_y(\widehat{y},u)$, we have

$$\|\delta\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)} + \|\delta\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)} \leq 2\|\delta y\|_{\widehat{Y}} \leq \kappa_2|\delta u| \tag{3.36}$$

for a constant $\kappa_2>0$ independent of $\delta u\in\mathbb{R}^3$. Hence

$$\begin{aligned}&|b(\delta\boldsymbol{u},\delta\boldsymbol{u},\boldsymbol{v})| \leq \kappa_3|\delta u|^2\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)}\\ &\leq \kappa_1\kappa_3|\delta u|^2\left(\alpha_u\|\boldsymbol{u}-\boldsymbol{u}_{\mathrm{d}}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \alpha_B\|\boldsymbol{B}-\boldsymbol{B}_{\mathrm{d}}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} + \alpha_J\|\boldsymbol{J}-\boldsymbol{J}_{\mathrm{d}}\|^2_{L^2(\Omega_{J,\mathrm{obs}})}\right).\end{aligned} \tag{3.37}$$

Further there exists κ_4 independent of δu such that

$$\begin{aligned}&|c(\delta\boldsymbol{J},\delta\widetilde{\mathcal{B}},\boldsymbol{v})| + |c(\boldsymbol{K},\delta\widetilde{\mathcal{B}},\delta\boldsymbol{u})|\\ &\quad\leq \kappa_4\left(\|\delta\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)} + |\delta u|\right)\left(\|\delta\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)}\|\boldsymbol{v}\|_{\boldsymbol{H}^1(\Omega)} + \|\boldsymbol{K}\|_{\boldsymbol{L}^2(\Omega)}\|\delta\boldsymbol{u}\|_{\boldsymbol{H}^1(\Omega)}\right)\\ &\quad\leq C\left(\alpha_u\|\boldsymbol{u}-\boldsymbol{u}_{\mathrm{d}}\|^2_{L^2(\Omega_{u,\mathrm{obs}})} + \alpha_B\|\boldsymbol{B}-\boldsymbol{B}_{\mathrm{d}}\|^2_{L^2(\Omega_{B,\mathrm{obs}})} + \alpha_J\|\boldsymbol{J}-\boldsymbol{J}_{\mathrm{d}}\|^2_{L^2(\Omega_{J,\mathrm{obs}})}\right),\end{aligned}$$

where $C=\kappa_1\kappa_2(\kappa_2+1)\kappa_4|\delta u|^2$. This last estimate, together with (3.36) and (3.37), implies (3.34). ∎

4. An operator splitting scheme. In this section, we address an operator splitting scheme for the numerical solution of the MHD state equations (3.13)–(3.14). Our approach is based on the hypothesis that one wants to decouple the system and

put to use existing and validated solvers for the Navier–Stokes equations and for div-curl systems. The same idea can be applied to the adjoint system. Different iterative schemes for MHD have been proposed; see, for instance [22, 5]. In [5] two decoupling algorithms for the velocity-magnetic field formulation are discussed and numerical examples are given. This work contains a stability result for a special geometry but no convergence analysis. The classical proofs (see, e.g., [8]) for general purpose decomposition techniques rely on monotonicity of the differential operators, not only in each variable separately, but also jointly in all variables simultaneously, to obtain unconditional convergence with respect to the problem data. This behavior is not possible for MHD systems, and localization naturally leads to smallness requirements.

With this in mind, we analyze the following iterative scheme to compute a solution of the MHD system for given controls $u = (I_{\mathrm{ext}}, I_{\mathrm{inj}}, B_{\mathrm{ext}}) \in U_{\mathrm{ad}}$ and given velocity boundary data $\boldsymbol{h}$. As before, $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$ denote the liftings according to Lemma 3.2.

ALGORITHM 4.1 (operator splitting scheme).

(1) *Choose an initial guess* $\widehat{\boldsymbol{J}}^0 \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega)$*; set* $n = 0$.

(2) *Solve the* div-curl *system for* $\boldsymbol{B}^{n+1} \in \boldsymbol{V}^1(\mathbb{R}^3)$,

$$\nabla \cdot \boldsymbol{B}^{n+1} = 0, \qquad \nabla \times (\mu^{-1}\boldsymbol{B}^{n+1}) = \widehat{\boldsymbol{J}}^n,$$

with the interface condition $[\boldsymbol{B}^{n+1}]_{\partial\Omega} = 0$.

(3) *Solve the Navier–Stokes system with Lorentz force for* $\widehat{\boldsymbol{u}}^{n+1} \in \boldsymbol{H}^1_0(\Omega)$ *and* $p^{n+1} \in L^2(\Omega)/\mathbb{R}$,

$$\begin{aligned} -\eta\Delta\widehat{\boldsymbol{u}}^{n+1} + \rho(\widehat{\boldsymbol{u}}^{n+1}\cdot\nabla)\widehat{\boldsymbol{u}}^{n+1} + \rho(\widehat{\boldsymbol{u}}^{n+1}\cdot\nabla)\boldsymbol{u}_0 + \rho(\boldsymbol{u}_0\cdot\nabla)\widehat{\boldsymbol{u}}^{n+1} + \nabla p^{n+1} \\ = \eta\Delta\boldsymbol{u}_0 - \rho(\boldsymbol{u}_0\cdot\nabla)\boldsymbol{u}_0 + (\widehat{\boldsymbol{J}}^n + \boldsymbol{J}_0)\times(\boldsymbol{B}^{n+1} + \boldsymbol{B}_0), \\ \nabla\cdot\widehat{\boldsymbol{u}}^{n+1} = 0 \end{aligned}$$

with homogeneous Dirichlet boundary data on $\partial\Omega$.

(4) *Solve for* $\widehat{\boldsymbol{J}}^{n+1} \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega)$ *and* $\phi^{n+1} \in H^1(\Omega)/\mathbb{R}$

$$\begin{aligned} \sigma^{-1}\widehat{\boldsymbol{J}}^{n+1} + \nabla\phi^{n+1} &= (\widehat{\boldsymbol{u}}^{n+1} + \boldsymbol{u}_0)\times(\boldsymbol{B}^{n+1} + \boldsymbol{B}_0) - \sigma^{-1}\boldsymbol{J}_0, \\ \nabla\cdot\widehat{\boldsymbol{J}}^{n+1} &= 0 \end{aligned}$$

with boundary condition $\widehat{\boldsymbol{J}}^{n+1}\cdot\boldsymbol{n} = 0$ *on* $\partial\Omega$.

(5) *Unless* $\|\widehat{\boldsymbol{J}}^{n+1} - \widehat{\boldsymbol{J}}^n\|_{\boldsymbol{L}^2(\Omega)}$ *is sufficiently small, increase* n *and go to* (2).

Note that the solution to step (2) is given by the Biot–Savart operator $\widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}}^n)$. In steps (3) and (4), $\boldsymbol{B}_0 = \widetilde{\mathcal{B}}(\boldsymbol{J}_0) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{ext}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{inj}}) + \boldsymbol{B}_{\mathrm{ext}}$ collects the constant contributions to the total magnetic field. Obviously, instead of computing the liftings $\boldsymbol{u}_0$ and $\boldsymbol{J}_0$ and repeatedly solving homogeneous problems in steps (3) and (4), one may directly address the inhomogeneous ones with unknowns $\widehat{\boldsymbol{u}}^{n+1} + \boldsymbol{u}_0$ and $\widehat{\boldsymbol{J}}^{n+1} + \boldsymbol{J}_0$. The same applies to the div-curl system in step (2), which yields $\boldsymbol{B}^{n+1} + \boldsymbol{B}_0 - \boldsymbol{B}_{\mathrm{ext}}$ if $\widehat{\boldsymbol{J}}^n$ is replaced with $\widehat{\boldsymbol{J}}^n + \boldsymbol{J}_0 + \boldsymbol{J}_{\mathrm{ext}} + \boldsymbol{J}_{\mathrm{inj}}$.

Remark 4.2 (alternative form of step (4)). Note that step (4) in Algorithm 4.1 above is equivalent to the solution of the div-curl system on Ω,

$$\begin{aligned} \nabla\cdot\widehat{\boldsymbol{J}}^{n+1} &= 0, \\ \nabla\times(\sigma^{-1}\widehat{\boldsymbol{J}}^{n+1}) &= \nabla\times\left[(\widehat{\boldsymbol{u}}^{n+1} + \boldsymbol{u}_0)\times(\boldsymbol{B}^{n+1} + \boldsymbol{B}_0)\right] - \nabla\times(\sigma^{-1}\boldsymbol{J}_0) \end{aligned}$$

with boundary condition $\widehat{\boldsymbol{J}}^{n+1}\cdot\boldsymbol{n}=0$ on $\partial\Omega$, provided that the right-hand side is in $\boldsymbol{L}^2(\Omega)$. This can be guaranteed if $\boldsymbol{J}_0$, $\boldsymbol{B}_{\rm ext}$, and $\partial\Omega$ are smooth enough.

For the proposed scheme, we have the following conditional convergence result.

PROPOSITION 4.3 (convergence of the operator splitting scheme). *Let $u\in\mathbb{R}^3$ and $\boldsymbol{h}\in\boldsymbol{H}^{1/2}(\partial\Omega)$ be a given control vector and boundary data and suppose that η is sufficiently large and σ is sufficiently small. Then there exists $\rho_J>0$ such that whenever the initial iterate $\widehat{\boldsymbol{J}}^0\in\boldsymbol{L}^2_{\rm div}(\Omega)$ satisfies $\|\widehat{\boldsymbol{J}}^0+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}<\rho_J$, then the iterates $(\widehat{\boldsymbol{J}}^n,\widehat{\boldsymbol{u}}^n)$ of Algorithm 4.1 converge in $\boldsymbol{L}^2(\Omega)\times\boldsymbol{H}^1(\Omega)$ to the necessarily unique solution of (3.13)–(3.14) which satisfies $\|\widehat{\boldsymbol{J}}+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}\le\rho_J$.*

Proof. The proof uses the Banach fixed point theorem. This choice is due to the fact that the nonlinearities in (3.13)–(3.14) are not of strictly monotone or energy preservation type, so that techniques analogous to those developed for decomposition methods, e.g., in [8], cannot be used. Let $T:\boldsymbol{L}^2_{\rm div}(\Omega)\to\boldsymbol{L}^2_{\rm div}(\Omega)$ denote the operator which assigns to $\widehat{\boldsymbol{J}}^n$ the value $\widehat{\boldsymbol{J}}^{n+1}$ defined by steps (2)–(4) of Algorithm 4.1. Let us denote by ρ_I a common bound for the inhomogeneities $\widetilde{u}=(I_{\rm ext},I_{\rm inj})$ and $\boldsymbol{h}$, i.e., $|\widetilde{u}|\le\rho_I$ and $\|\boldsymbol{h}\|_{\boldsymbol{H}^{1/2}(\partial\Omega)}\le\rho_I$. Given the solenoidal current field $\widehat{\boldsymbol{J}}^n$, we infer from Lemma 2.4 the existence of $\boldsymbol{B}^{n+1}$ satisfying the equations in step (2) and the a priori estimate

$$\|\boldsymbol{B}^{n+1}+\boldsymbol{B}_0\|_{\boldsymbol{L}^3(\Omega)}\le c_1\,\mu\left(\|\widehat{\boldsymbol{J}}^n+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}+|\widetilde{u}|\right)+c_1|\boldsymbol{B}_{\rm ext}|.$$

Here and below, the constants c_i are independent of μ, η, σ, iteration index n, and controls u. Let us further assume that $\|\widehat{\boldsymbol{J}}^n+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}\le\rho_J$. Then we have

$$\|\boldsymbol{B}^{n+1}+\boldsymbol{B}_0\|_{\boldsymbol{L}^3(\Omega)}\le c_1\big(\mu(\rho_J+\rho_I)+|B_{\rm ext}|\big). \tag{4.1}$$

Standard estimates for the Navier–Stokes equations in step (3) imply that

$$\begin{aligned}\|\widehat{\boldsymbol{u}}^{n+1}+\boldsymbol{u}_0\|_{\boldsymbol{H}^1(\Omega)}&\le\mu\eta^{-1}c_2\left(\|\widehat{\boldsymbol{J}}^n+\boldsymbol{J}_0\|^2_{\boldsymbol{L}^2(\Omega)}+|\widetilde{u}|^2\right)\\&\quad+c_2\|\widehat{\boldsymbol{J}}^n+\boldsymbol{J}_0\||B_{\rm ext}|+c_2\Big(\|\boldsymbol{h}\|_{\boldsymbol{H}^{1/2}(\partial\Omega)}+\|\boldsymbol{h}\|^2_{\boldsymbol{H}^{1/2}(\partial\Omega)}\Big)\\&\le c_2\Big(\mu\eta^{-1}(\rho_J^2+\rho_I^2)+\rho_J|B_{\rm ext}|+\rho_I+\rho_I^2\Big).\end{aligned} \tag{4.2}$$

By Lemma 2.1 and a direct computation (or by [7, Chap. I, Cor. 4.1]), the system

$$\sigma^{-1}\widehat{\boldsymbol{J}}+\nabla\phi=f\ \text{on}\ \Omega,\qquad\qquad\nabla\cdot\widehat{\boldsymbol{J}}=0\ \text{on}\ \Omega$$

with given $f\in\boldsymbol{L}^2(\Omega)$ has a unique solution $\boldsymbol{J}\in\boldsymbol{L}^2_{\rm div}(\Omega)$ and $\phi\in H^1(\Omega)/\mathbb{R}$ which satisfies $\|\boldsymbol{J}\|_{\boldsymbol{L}^2(\Omega)}\le\sigma\,\|f\|_{\boldsymbol{L}^2(\Omega)}$. Hence we conclude from step (4) and (4.1) and (4.2) that

$$\begin{aligned}\|\widehat{\boldsymbol{J}}^{n+1}+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}&\le\sigma\,c_3'\,\|\widehat{\boldsymbol{u}}^{n+1}+\boldsymbol{u}_0\|_{\boldsymbol{L}^6(\Omega)}\|\boldsymbol{B}^{n+1}+\boldsymbol{B}_0\|_{\boldsymbol{L}^3(\Omega)}\\&\le\sigma\,c_3'\Big(\mu\eta^{-1}(\rho_J^2+\rho_I^2)+\rho_J|B_{\rm ext}|+\rho_I+\rho_I^2\Big)\Big(\mu(\rho_J+\rho_I)+|B_{\rm ext}|\Big)+c_3'|I_{\rm inj}|\\&\le\sigma\,c_3\Big(\mu^2\eta^{-1}(\rho_J^3+\rho_I^3)+|B_{\rm ext}|\Big(\mu\rho_J(\rho_J+\rho_I)+\mu\eta^{-1}(\rho_J^2+\rho_I^2)+\rho_J|B_{\rm ext}|\Big)\\&\qquad+(\rho_I+\rho_I^2)\big(\mu(\rho_J+\rho_I)+|B_{\rm ext}|\big)\Big)+c_3|I_{\rm inj}|.\end{aligned} \tag{4.3}$$

Concerning the initialization, note that $\widehat{\boldsymbol{J}}^0$ can be taken as zero. Then $\|\widehat{\boldsymbol{J}}^0+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}$ is bounded by $c_3|I_{\mathrm{inj}}|$. Choosing $\rho_J := 2c_3|I_{\mathrm{inj}}|$ and assuming that $\|\widehat{\boldsymbol{J}}^n+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)} \le \rho_J$ by induction, we obtain from (4.3) that $\|\widehat{\boldsymbol{J}}^{n+1}+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)} \le \rho_J$, provided that, for instance, σ is sufficiently small, or μ and $|B_{\mathrm{ext}}|$ are sufficiently small. From (4.1) and (4.2) follows the existence of constants ρ_B and ρ_u independent of n such that

$$\|\boldsymbol{B}^n+\boldsymbol{B}_0\|_{\boldsymbol{L}^3(\Omega)} \le \rho_B \quad \text{and} \quad \|\widehat{\boldsymbol{u}}^n+\boldsymbol{u}_0\|_{\boldsymbol{H}^1(\Omega)} \le \rho_u \tag{4.4}$$

for all n. To prove that T is a contraction, let $\widehat{\boldsymbol{J}}_i \in \boldsymbol{L}^2_{\mathrm{div}}(\Omega)$, $i = 1, 2$, and let $\boldsymbol{K}_i$ be their images under T. Further let $\boldsymbol{B}_i$ and $\widehat{\boldsymbol{u}}_i$ denote the associated magnetic and velocity fields according to Algorithm 4.1. Then

$$\|\boldsymbol{B}_1 - \boldsymbol{B}_2\|_{\boldsymbol{H}^1(\Omega)} \le \mu\, c_4 \|\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2\|_{\boldsymbol{L}^2(\Omega)}. \tag{4.5}$$

Here and below, the constants are also independent of $\widehat{\boldsymbol{J}}_1$ and $\widehat{\boldsymbol{J}}_2$. Moreover, $U = \widehat{\boldsymbol{u}}_1 - \widehat{\boldsymbol{u}}_2 \in \boldsymbol{H}^1_0(\Omega)$ satisfies

$$\begin{aligned} -\eta\Delta U + \rho(U\cdot\nabla)(\widehat{\boldsymbol{u}}_1+\boldsymbol{u}_0) + \rho((\widehat{\boldsymbol{u}}_2+\boldsymbol{u}_0)\cdot\nabla)U + \nabla P \\ = (\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2)\times(\boldsymbol{B}_1+\boldsymbol{B}_0) + (\widehat{\boldsymbol{J}}_2+\boldsymbol{J}_0)\times(\boldsymbol{B}_1-\boldsymbol{B}_2) \end{aligned}$$

for some $P \in L^2(\Omega)/\mathbb{R}$. This implies that

$$\begin{aligned} \eta\|\nabla U\|^2_{\boldsymbol{L}^2(\Omega)} &\le c_5\Big(\rho_u\|\nabla U\|^2_{\boldsymbol{L}^2(\Omega)} + \rho_B\|\nabla U\|_{\boldsymbol{L}^2(\Omega)}\|\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2\|_{\boldsymbol{L}^2(\Omega)} \\ &\qquad + \rho_J\|\nabla U\|_{\boldsymbol{L}^2(\Omega)}\|\boldsymbol{B}_1-\boldsymbol{B}_2\|_{\boldsymbol{L}^3(\Omega)}\Big) \\ &\le c_5\Big(\rho_u\|\nabla U\|^2_{\boldsymbol{L}^2(\Omega)} + \big(\rho_B + \mu\, c_4\,\rho_J\big)\|\nabla U\|_{\boldsymbol{L}^2(\Omega)}\|\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2\|_{\boldsymbol{L}^2(\Omega)}\Big) \\ &\le c_5\Big(\rho_u\|\nabla U\|^2_{\boldsymbol{L}^2(\Omega)} + \eta/2\|\nabla U\|^2_{\boldsymbol{L}^2(\Omega)} + \big(\rho_B + \mu\, c_4\,\rho_J\big)^2/(2\eta)\|\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2\|^2_{\boldsymbol{L}^2(\Omega)}\Big). \end{aligned}$$

Hence if η is sufficiently large, or ρ_u is sufficiently small (which can be achieved by μ, $|B_{\mathrm{ext}}|$, and ρ_I sufficiently small), we have

$$\|\nabla U\|_{\boldsymbol{L}^2(\Omega)} \le c_6\,\eta^{-1}(\rho_B + \mu\, c_4\,\rho_J)\|\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2\|_{\boldsymbol{L}^2(\Omega)}. \tag{4.6}$$

Finally, we obtain from step (4)

$$\begin{aligned} \|\boldsymbol{K}_1 - \boldsymbol{K}_2\| &\le \sigma\, c_7'\Big(\|U\|_{\boldsymbol{H}^1(\Omega)}\|\boldsymbol{B}_1+\boldsymbol{B}_0\|_{\boldsymbol{L}^3(\Omega)} + \|\widehat{\boldsymbol{u}}_2+\boldsymbol{u}_0\|_{\boldsymbol{H}^1(\Omega)}\|\boldsymbol{B}_1-\boldsymbol{B}_2\|_{\boldsymbol{L}^3(\Omega)}\Big) \\ &\le \sigma\, c_7\Big(\eta^{-1}\rho_B(\rho_B + \mu\,\rho_J) + \mu\, c_4\,\rho_u\Big)\|\widehat{\boldsymbol{J}}_1 - \widehat{\boldsymbol{J}}_2\|_{\boldsymbol{L}^2(\Omega)}. \end{aligned} \tag{4.7}$$

Hence we conclude that if σ is sufficiently small, or if μ is sufficiently small and η sufficiently large, then T is a contraction on the ball $\{\boldsymbol{J} : \|\widehat{\boldsymbol{J}}+\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)} < \rho_J\}$. ∎

5. Conclusion and outlook. In this paper, we have presented and analyzed an optimal control problem for the stationary MHD system. We derived necessary and sufficient conditions for local optimal solutions. In addition, we analyzed an iterative scheme for the numerical solution of the MHD state equations which is tailored to make use of existing Navier–Stokes and div-curl solvers. We believe that in the face of industrial MHD applications, there is ample room to extend our results in several directions. Of particular interest are the cases of instationary MHD flows,

unknown current densities in external conductors, flows with thermal coupling and Ohmic heating, and the case of material-dependent magnetic permeability. All of the above present additional technical difficulties, which are the subject of future investigations. Finally, devising an efficient numerical algorithm to solve optimal control problems involving MHD flows presents another challenging task.

Appendix. Proof of Proposition 3.3. Let us define $T : \widehat{Y} \to \widehat{Y}$ according to

$$(\delta \boldsymbol{u}, \delta p, \delta \boldsymbol{J}, \delta \phi) = T(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$$

if and only if

$$A(\delta \boldsymbol{u}, \delta p, \delta \boldsymbol{J}, \delta \phi) = R(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi) \tag{A.1}$$

holds in $\widehat{Y}'$. That is, T is the solution operator of a linear PDE, which depends nonlinearly on the data $(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$. Defining the components of R as

$$\begin{aligned}
R^1(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)(\boldsymbol{v}) &= c(\widehat{\boldsymbol{J}} + \boldsymbol{J}_0, \widetilde{\mathcal{B}}(\boldsymbol{J}_0) + \widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{ext}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{inj}}) + \boldsymbol{B}_{\mathrm{ext}}, \boldsymbol{v}) \\
&\quad - a_1(\boldsymbol{u}_0, \boldsymbol{v}) - b(\widehat{\boldsymbol{u}} + \boldsymbol{u}_0, \widehat{\boldsymbol{u}} + \boldsymbol{u}_0, \boldsymbol{v}), \\
R^2(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)(q) &= 0, \\
R^3(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)(\boldsymbol{K}) &= -c(\boldsymbol{K}, \widetilde{\mathcal{B}}(\boldsymbol{J}_0) + \widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{ext}}) + \widetilde{\mathcal{B}}(\boldsymbol{J}_{\mathrm{inj}}) + \boldsymbol{B}_{\mathrm{ext}}, \widehat{\boldsymbol{u}} + \boldsymbol{u}_0) \\
&\quad - a_2(\boldsymbol{J}_0, \boldsymbol{K}), \\
R^4(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)(\psi) &= 0,
\end{aligned}$$

we easily verify that the solutions to the homogenized problem (3.13)–(3.14) are exactly the fixed points of T. In view of Proposition 3.6 below, $A : \widehat{Y} \to \widehat{Y}'$ is an isomorphism, and hence T is well defined from $\widehat{Y}$ to itself. We now confirm that T is compact. To this end, we consider a bounded and weakly convergent sequence $(\widehat{\boldsymbol{u}}^n, p^n, \widehat{\boldsymbol{J}}^n, \phi^n) \rightharpoonup (\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$ in $\widehat{Y}$. Since the norm in $\widehat{Y}'$ of the right-hand side in (A.1) is a quadratic polynomial in the norms of $\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{J}}$, the sequence $(\delta \boldsymbol{u}^n, \delta p^n, \delta \boldsymbol{J}^n, \delta \phi^n) := T(\widehat{\boldsymbol{u}}^n, p^n, \widehat{\boldsymbol{J}}^n, \phi^n)$ is bounded in $\widehat{Y}$ and thus possesses a weakly convergent subsequence in $\widehat{Y}$, i.e., $(\delta \boldsymbol{u}^n, \delta p^n, \delta \boldsymbol{J}^n, \delta \phi^n) \rightharpoonup (\delta \boldsymbol{u}, \delta p, \delta \boldsymbol{J}, \delta \phi)$. Using Lemma 2.3, one confirms that the weak limit $(\delta \boldsymbol{u}, \delta p, \delta \boldsymbol{J}, \delta \phi)$ satisfies $(\delta \boldsymbol{u}, \delta p, \delta \boldsymbol{J}, \delta \phi) = T(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$; i.e., T is weakly continuous. The difference $(\delta \boldsymbol{u}^n, \delta p^n, \delta \boldsymbol{J}^n, \delta \phi^n) - (\delta \boldsymbol{u}, \delta p, \delta \boldsymbol{J}, \delta \phi)$ satisfies (A.1) with right-hand side

$$R(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi) - R(\widehat{\boldsymbol{u}}^n, p^n, \widehat{\boldsymbol{J}}^n, \phi^n),$$

which converges to zero strongly in $\widehat{Y}'$, as a straightforward application of Lemmas 2.3, 2.4, and 2.6 shows. Hence T is indeed compact.

Now let $(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$ be a fixed point of $s \cdot T$ for any $s \in [0, 1]$; i.e., $(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$ satisfies (A.1) with the right-hand side multiplied by s. Testing this system with $(\widehat{\boldsymbol{u}}, p, \widehat{\boldsymbol{J}}, \phi)$, we obtain

$$\begin{aligned}
\eta \, \|\nabla \widehat{\boldsymbol{u}}\|^2_{\boldsymbol{L}^2(\Omega)} + \sigma^{-1} \|\widehat{\boldsymbol{J}}\|^2_{\boldsymbol{L}^2(\Omega)} = s\big(&c(\boldsymbol{J}_0, \boldsymbol{B}, \widehat{\boldsymbol{u}}) - c(\widehat{\boldsymbol{J}}, \boldsymbol{B}, \boldsymbol{u}_0) - b(\widehat{\boldsymbol{u}}, \boldsymbol{u}_0, \widehat{\boldsymbol{u}}) \\
&- a_1(\boldsymbol{u}_0, \widehat{\boldsymbol{u}}) - a_2(\boldsymbol{J}_0, \widehat{\boldsymbol{J}})\big)
\end{aligned}$$

with $\boldsymbol{B}$ according to (3.14). By Poincaré's inequality $\|\widehat{\boldsymbol{u}}\|_{\boldsymbol{L}^2(\Omega)} \le c_P \|\nabla \widehat{\boldsymbol{u}}\|_{\boldsymbol{L}^2(\Omega)}$, one obtains

$$\frac{\eta}{1+c_P^2}\|\widehat{\boldsymbol{u}}\|^2_{\boldsymbol{H}^1(\Omega)} + \sigma^{-1}\|\widehat{\boldsymbol{J}}\|^2_{\boldsymbol{L}^2(\Omega)} \le s\big(c(\boldsymbol{J}_0,\boldsymbol{B},\widehat{\boldsymbol{u}}) - c(\widehat{\boldsymbol{J}},\boldsymbol{B},\boldsymbol{u}_0) - b(\widehat{\boldsymbol{u}},\boldsymbol{u}_0,\widehat{\boldsymbol{u}}) - a_1(\boldsymbol{u}_0,\widehat{\boldsymbol{u}}) - a_2(\boldsymbol{J}_0,\widehat{\boldsymbol{J}})\big). \tag{A.2}$$

The application of the Leray–Schauder fixed point theorem requires that the left-hand side be a priori bounded uniformly in $s \in [0,1]$. The bound may depend on the controls $(I_{\mathrm{ext}}, I_{\mathrm{inj}}, B_{\mathrm{ext}})$ and the boundary data $\boldsymbol{h}$. We observe that the right-hand side in (A.2) is bounded above by

$$\|\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}\|\widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}})\|_{\boldsymbol{L}^3(\Omega)}\|\widehat{\boldsymbol{u}}\|_{\boldsymbol{L}^6(\Omega)} + \|\widehat{\boldsymbol{J}}\|_{\boldsymbol{L}^2(\Omega)}\|\widetilde{\mathcal{B}}(\widehat{\boldsymbol{J}})\|_{\boldsymbol{L}^3(\Omega)}\|\boldsymbol{u}_0\|_{\boldsymbol{L}^6(\Omega)} + \|\nabla \boldsymbol{u}_0\|_{\boldsymbol{L}^2(\Omega)}\|\widehat{\boldsymbol{u}}\|^2_{\boldsymbol{L}^4(\Omega)} \tag{A.3}$$

plus a number of terms which are at most linear in $\|\widehat{\boldsymbol{u}}\|$ and $\|\boldsymbol{J}\|$. The latter can be treated using Young's inequality according to the pattern $c\,\|\widehat{\boldsymbol{u}}\| \le \varepsilon\,\|\widehat{\boldsymbol{u}}\|^2 + c/(4\varepsilon)$, and $\varepsilon\,\|\widehat{\boldsymbol{u}}\|^2$ can then be absorbed into the left-hand side of (A.2) for sufficiently small $\varepsilon > 0$. However, in order for the terms in (A.3) to be likewise absorbed in the left-hand side of (A.2), the coefficients $\|\boldsymbol{J}_0\|_{\boldsymbol{L}^2(\Omega)}$, $\|\boldsymbol{u}_0\|_{\boldsymbol{L}^6(\Omega)}$, and $\|\nabla \boldsymbol{u}_0\|_{\boldsymbol{L}^2(\Omega)}$ must be sufficiently small. In this case, $\|\widehat{\boldsymbol{u}}\|_{\boldsymbol{H}^1(\Omega)}$ and $\|\widehat{\boldsymbol{J}}\|_{\boldsymbol{L}^2(\Omega)}$ are indeed a priori bounded by the right-hand side in (3.16). In view of Lemma 3.2, the same bound holds for the inhomogeneous solution $\boldsymbol{u}$ and $\boldsymbol{J}$. Finally, the bounds for the pressure p and potential ϕ follow from Lemma 2.2. Hence we conclude the applicability of the Leray–Schauder theorem which yields the existence of a fixed point of T.

Acknowledgments. We gratefully acknowledge our invitation to the workshop "Flow Control by Tailored Magnetic Fields" by M. Hinze, held in April 2004, which stimulated our interest in the subject of MHD optimal control.

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] H. Brézis, *Analyse Fonctionelle—Théorie et Applications*, Masson, Paris, 1983.
[3] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 4: Integral Equations and Numerical Methods, Springer-Verlag, Berlin, 1990.
[4] P. A. Davidson, *An Introduction to Magnetohydrodynamics*, Cambridge University Press, Cambridge, UK, 2001.
[5] J.-F. Gerbeau, *Problèmes Mathématiques et Numériques Posés par la Modélisation de L'électrolyse de L'aluminium*. Ph.D. thesis, Ecole Nationale des Ponts et Chaussées, Marne la Vallée, France, 1998.
[6] D. Gilbarg and N. S. Trudinger, *Elliptic Differential Equations of Second Order*, Springer, New York, 1977.
[7] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[8] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer Ser. Comput. Phys., Springer-Verlag, New York, 1984.
[9] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.
[10] M. D. Gunzburger and S. Manservisi, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed controls*, SIAM J. Numer. Anal., 37 (2000), pp. 1481–1512.

[11] M. D. Gunzburger and S. Manservisi, *The velocity tracking problem for Navier–Stokes flows with boundary control*, SIAM J. Control Optim., 39 (2000), pp. 594–634.
[12] M. D. Gunzburger, A. J. Meir, and J. S. Peterson, *On the existence, uniqueness, and finite element approximation of solutions of the equations of stationary, incompressible magnetohydrodynamics*, Math. Comp., 56 (1991), pp. 523–563.
[13] M. Gunzburger and C. Trenchea, *Analysis and discretization of an optimal control problem for the time-periodic MHD equations*, J. Math. Anal. Appl., 308 (2005), pp. 440–466.
[14] M. Hinze and K. Kunisch, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim., 40 (2001), pp. 925–946.
[15] M. Hinze and O. Pfeiffer, *Active Closed Loop Control of Weakly Conductive Fluids*, Preprint 20–2004, SFB 609, TU Dresden, Dresden, Germany, 2004; submitted.
[16] L. S. Hou and A. J. Meir, *Boundary optimal control of MHD flows*, Appl. Math. Optim., 32 (1995), pp. 143–162.
[17] L. S. Hou and J. Peterson, *Boundary optimal control for an electrically conducting fluid using boundary electrical potential controls*, Nonlinear Anal., 24 (1995), pp. 857–784.
[18] L. S. Hou and S. S. Ravindran, *Finite element approximation for optimal control of electrically conducting fluid flows*, in Mathematics of Computation 1943–1993: A Half-Century of Computational Mathematics (Vancouver, BC, 1993), Proc. Sympos. Appl. Math. 48, AMS, Providence, 1994, pp. 305–309.
[19] L. S. Hou and S. S. Ravindran, *Computations of boundary optimal control problems for an electrically conducting fluid*, J. Comput. Phys., 128 (1996), pp. 319–330.
[20] H. Maurer and J. Zowe, *First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
[21] A. J. Meir and P. G. Schmidt, *Variational methods for stationary MHD flow under natural interface conditions*, Nonlinear Anal., 26 (1996), pp. 659–689.
[22] A. J. Meir and P. G. Schmidt, *Analysis and numerical approximation of a stationary MHD flow problem with nonideal boundary*, SIAM J. Numer. Anal., 36 (1999), pp. 1304–1332.
[23] A. J. Meir, P. G. Schmidt, S. I. Bakhtiyarov, and R. A. Overfelt, *Velocity, potential, and temperature distributions in molten metals during electromagnetic stirring, Part* ii: *Numerical simulations*, in Proceedings of a Symposium at the ASME International Mechanical Engineering Congress and Exhibition, Nashville, 1999, pp. 87–96.
[24] S. S. Ravindran, *Real-time computational algorithm for optimal control of an MHD flow system*, SIAM J. Sci. Comput., 26 (2005), pp. 1369–1388.
[25] J. A. Shercliff, *A Textbook of Magnetohydrodynamics*, Pergamon Press, Oxford, UK, 1965.
[26] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, North-Holland, Amsterdam, 1984.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1846–1874

BARRIER OPTION HEDGING UNDER CONSTRAINTS: A VISCOSITY APPROACH*

IMEN BENTAHAR† AND BRUNO BOUCHARD‡

Abstract. We study the problem of finding the minimal initial capital needed in order to hedge without risk a barrier option when the vector of proportions of wealth invested in each risky asset is constrained to lie in a closed convex domain. In the context of a Brownian diffusion model, we provide a PDE characterization of the superhedging price. This extends the result of Broadie, Cvitanič, and Soner [Rev. Financial Stud., 11 (1998), pp. 59–79] and Cvitanič, Pham, and Touzi [J. Appl. Probab., 36 (1999), pp. 523–545] which was obtained for plain vanilla options and provides a natural numerical procedure for computing the corresponding superhedging price. As a by-product, we obtain a comparison theorem for a class of parabolic PDEs with relaxed Dirichlet conditions involving a constraint on the gradient.

Key words. superreplication, barrier options, portfolio constraints, viscosity solutions

AMS subject classifications. 91B28, 49L25, 35B05

DOI. 10.1137/06065324X

1. Introduction.

The problem of superhedging under portfolio constraints has attracted a lot of attention since the seminal work of Cvitanič and Karatzas [6]. One of the original motivations came from the hedging of plain vanilla options with discontinuous payoffs, such as digital options. For such options the delta and gamma may take very large values when the remaining maturity is small, which makes them difficult to delta-hedge.

Within diffusion models, the remarkable result of Broadie, Cvitanič, and Soner [4] shows that the optimal hedge under constraints is obtained by considering the Black–Scholes-type hedging strategy of some modified payoff. Thus, hedging the original claim under constraints corresponds to hedging a modified one without constraints. This is the "face-lifting" procedure. Within the Black–Scholes model, this allows one to find explicit optimal hedge. In more general Markov diffusion models, an explicit solution may not be available but the superhedging price can still be characterized as the solution of some Hamilton–Jacobi–Bellman equation; see Cvitanič, Pham, and Touzi [7] and the review paper Soner and Touzi [16]. In the general semi-martingale case, no explicit solution is available but a general dual formulation was obtained by Föllmer and Kramkov [8].

Similar problems may appear for path-dependent options such as barrier options. For instance, the delta of knock-out barrier options may explode when the maturity is small and the underlying asset is close to the barrier. This more difficult issue was recently considered by Shreve, Schmock, and Wystup [12]. In this paper, the authors solve the problem of hedging a knock-out call option in a one-dimensional Black–Scholes model under a constraint on the short position, i.e., the proportion of wealth invested in the risky asset is bounded from below. This result is obtained

*Received by the editors February 28, 2006; accepted for publication (in revised form) August 4, 2006; published electronically December 11, 2006. This work was partially supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk."
http://www.siam.org/journals/sicon/45-5/65324.html
†Institut für Mathematik, Technische Universität Berlin, Berlin, D-10623, Germany (imen@ceremade.dauphine.fr).
‡Université Paris VI, LPMA, Paris 75252 Cedex 05, France (bouchard@ccr.jussieu.fr).

by extending the dual formulation of Cvitanič and Karatzas [6] and by solving the associated stochastic control problem.

The aim of the present paper is to provide a PDE characterization of the superhedging price of barrier-type options. Our model is more general than the one studied in Shreve, Schmock, and Wystup [12] in two aspects. First, we consider general payoffs of the form $g(\tau, X_\tau)$, where τ is the first exit time of a d-dimensional price process X from a given domain $\mathcal{O}$. Second, our constraints on the proportions of wealth invested in the risky assets are described by a rather general closed convex set.

Our derivation of the associated PDE relies on the dual formulation of Cvitanič and Karatzas [6] as in Cvitanič, Pham, and Touzi [7]. Here, the main difficulty comes from the boundary condition on $\partial\mathcal{O}$ before maturity, a problem which does not appear in [7]. As in the vanilla option case, we have to consider as a boundary condition a "face-lifted" payoff, but in the case of barrier options this is not sufficient. Indeed, the example considered in Shreve, Schmock, and Wystup [12] shows that the boundary condition on $[0,T)\times\partial\mathcal{O}$ may not be assumed continuously by the value function, even when the payoff is "face-lifted" (in their case $g = 0$ before T). This implies that this boundary condition has to be considered in a weak sense.

In this paper, we give an appropriate sense to the boundary condition and show that the superhedging price is a (discontinuous) viscosity solution of the corresponding Hamilton–Jacobi–Bellman equation. We also show that it can actually be further characterized as its smallest viscosity supersolution. Finally, under mild additional assumptions, we prove a comparison theorem for the associated PDE which ensures uniqueness of the solution and opens the door to the implementation of a numerical scheme. Here, the difficulty comes from the constraint on the gradient of the value function which also appears in the relaxed boundary condition. To the best of our knowledge, this is the first time that such an equation has been considered.

To conclude this section, we observe that our results could be alternatively obtained by using the direct dynamic principle of [14], which was first used in [15] to derive the PDEs associated to general stochastic target problems with constraints on the control process in the case $\mathcal{O} = \mathbb{R}^d$; see also [3] for the case of diffusions with jumps.

The rest of the paper is organized as follows. The superhedging problem and its dual formulation are presented in section 2. In section 3, we describe the associated PDE and state our main results. A numerical application is presented in section 4. The remaining sections contain the proofs.

Notation. All elements $x = (x^i)_{i\le d}$ of $\mathbb{R}^d$ are identified with column vectors with Euclidean norm $|\cdot|$ and transposed vector x'. The positive orthant of $\mathbb{R}^d$ is denoted by $\mathbb{R}^d_+$ and the set of $d\times d$ matrices by $\mathbb{M}^d$. We write $\operatorname{diag}[x]$ to denote the diagonal matrix of $\mathbb{M}^d$ whose ith diagonal element is x^i. If $y\in\mathbb{R}^d$, we write xy for $(x^iy^i)_{i\le d}$, x^y for $\prod_{i\le d}(x^i)^{y^i}$, and xe^y for $(x^ie^{y^i})_{i\le d}$, whenever it is well defined. The trace of $M\in\mathbb{M}^d$ is denoted by $\operatorname{Tr}[M]$, and $|M|$ denotes its Euclidean norm when viewed as an element of $\mathbb{R}^{d^2}$. Given a family $(a^{ij})_{i,j\le d}$ of real numbers, we denote by $[a^{ij}]_{i,j}$ the matrix A whose component (i,j) is given by a^{ij}. The closure of a set $E\subset\mathbb{R}^d$ is denoted by $\bar E$, ∂E stands for its boundary, and $\operatorname{int}(E)$ for its interior. Given $\eta>0$, $B(x,\eta)$ denotes the open ball of radius η centered on x.

Given a smooth function $(t,x)\in[0,T]\times\mathbb{R}^d\mapsto\varphi(t,x)\in\mathbb{R}$, we denote by $D\varphi$ its (partial) Jacobian matrix with respect to x and by $D^2\varphi$ its (partial) Hessian matrix with respect to x. All inequalities involving random variables have to be understood in the $\mathbb{P}$-a.s. sense.

2. The superhedging price under constraints and its dual formulation. Throughout this paper, $T > 0$ is a finite time horizon and $W = (W_t)_{t \le T}$ is a d-dimensional Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the $\mathbb{P}$-augmented filtration generated by W, $\mathbb{F} = (\mathcal{F}_t)_{t \le T}$ satisfies $\mathcal{F}_0 = \{\Omega, \emptyset\}$ and $\mathcal{F}_T = \mathcal{F}$.

2.1. The barrier option hedging problem. The financial market is composed of a nonrisky asset B with price process normalized to unity, i.e., $B_t = 1$ for all $t \le T$, and d risky assets $X = (X^1, \dots, X^d)$ whose dynamics are given by the stochastic differential equation

$$X(t) = X_0 + \int_0^t \operatorname{diag}[X(s)]\, \sigma(s, X(s)) dW_s, \quad t \le T\,, \tag{2.1}$$

for some $X_0 \in (0, \infty)^d$. Here, $\sigma : [0,T] \times \mathbb{R}^d_+ \mapsto \mathbb{M}^d$ is assumed to satisfy the following.

$$\left\{ \begin{array}{ll} \text{(i)} & \sigma \text{ is continuous, bounded, and invertible with bounded inverse.} \\ \text{(ii)} & \text{The map } (t,x) \in [0,T] \times \mathbb{R}^d_+ \;\mapsto\; \operatorname{diag}[x]\, \sigma(t,x) \\ & \text{is Lipschitz continuous in } x, \text{ uniformly in } t. \end{array} \right. \tag{2.2}$$

Remark 2.1. As usual there is no loss of generality in assuming that X is a local martingale since, under mild assumptions on the original dynamics, we can always reduce to this case by passing to an equivalent probability measure. The normalization $B = 1$ means that we consider discounted processes, i.e., we take B as a numéraire.

A *financial strategy* is described by a d-dimensional predictable process $\pi = (\pi^1, \dots, \pi^d)$ satisfying the integrability condition

$$\int_0^T |\pi_t|^2 dt \;<\; \infty \quad \mathbb{P}\text{-a.s.}, \tag{2.3}$$

where π^i_t is the proportion of wealth invested at time t in the risky asset X^i. To an initial capital $y \in \mathbb{R}$ and a financial strategy π, we associate the induced wealth process Y^π_y defined as the solution on $[0,T]$ of

$$Y(t) = y + \int_0^t Y(s) \pi_s' \operatorname{diag}[X(s)]^{-1} dX(s) = y + \int_0^t Y(s) \pi_s' \sigma(s, X(s)) dW_s, \tag{2.4}$$

where $'$ stands for transposition.

Remark 2.2. Since in our model the financial strategies are described by the proportions of total wealth invested in each risky asset, the no-bankruptcy condition always holds provided that the initial wealth is nonnegative. Indeed, it is clear from (2.4) that for $y \ge 0$, the induced wealth process satisfies $Y^\pi_y(t) \ge 0$ for all $t \in [0,T]$ almost surely.

The constraints on the portfolio strategy are described by a closed convex set $K \subset \mathbb{R}^d$. We say that a *financial strategy* π is *admissible* if it satisfies, in addition to the condition (2.3), the constraint

$$\pi \in K \;\; dt \times d \quad \mathbb{P}\text{-a.e.}, \tag{2.5}$$

and we denote by $\mathcal{K}$ the set of *admissible financial strategies*. Throughout this paper, we shall assume that

$$0 \in K \neq \mathbb{R}^d\,. \tag{2.6}$$

The left-hand side condition just means that $0 \in \mathcal{K}$, while the inequality is natural since otherwise there would be no constraint on the portfolio.

The *barrier option* is described by a map g defined on $[0,T] \times \mathbb{R}^d_+$ and an open domain $\mathcal{O}$ of $\mathbb{R}^d$ such that

$$g \geq 0 \ \text{ on } \bar{\mathcal{O}} \cap \mathbb{R}^d_+ \qquad \text{and} \qquad g = 0 \ \text{ on } [0,T] \times \bar{\mathcal{O}}^c \,, \tag{2.7}$$

where $\bar{\mathcal{O}}^c := (0,\infty)^d \setminus \bar{\mathcal{O}}$. The buyer of the option receives the payment $g(\tau, X(\tau))$ at the (stopping-)time τ defined as the first time when X exits $\mathcal{O}$ if this occurs before T and T otherwise:

$$\tau := \inf\{t \in [0,T] \ : \ X(t) \notin \mathcal{O}\} \wedge T \,,$$

with the usual convention $\inf \emptyset = \infty$. The superreplication cost under constraint of the claim $g(\tau, X(\tau))$ is thus defined as

$$v(0,X_0) := \inf\left\{y \in \mathbb{R}_+ \ : \ Y^\pi_y(\tau) \geq g(\tau, X(\tau)) \quad \text{for some } \pi \in \mathcal{K}\right\} \,. \tag{2.8}$$

Remark 2.3. The condition $g = 0$ on $[0,T] \times \bar{\mathcal{O}}^c$ can be seen as a convention. Indeed, it is clear that $v(0,X_0)$ does not depend on the value of g on this set when $X_0 \in \bar{\mathcal{O}}$, while for $X_0 \in \bar{\mathcal{O}}^c$ the problem has no interest.

Hereafter we present examples of barrier options which fit into our framework.

Example 2.1. *Up-and-out call.* Let $d = 1$. The payoff of an *up-and-out call* on a single asset X^1, with strike price κ and *knock-out barrier* B, is equal to

$$\left(X^1(T) - \kappa\right)^+ \mathbf{1}_{\left\{\max_{0\leq t\leq T} X^1(t) < B\right\}} \,.$$

In our framework this corresponds to $\mathcal{O} = (-\infty, B)$ and $g(t,x) = t(x-\kappa)^+ 1_{\{t=T, x<B\}}$.

Example 2.2. *Down-and-out basket put option.* A basket option is an option whose payoff depends on a weighted average of a set of underlying values. Letting $d = 2$, we consider the *down-and-out* barrier option whose payoff is given by

$$\left(\kappa - \frac{X^1(T) + X^2(T)}{2}\right)^+ \mathbf{1}_{\left\{min_{0\leq t\leq T} X^1(t) + X^2(t) > 2B\right\}} \,.$$

In our framework this corresponds to $\mathcal{O} = \left\{x \in (0,\infty)^2 \,,\ x^1 + x^2 > 2B\right\}$ and $g(t,x) = (\kappa - \frac{x^1+x^2}{2})^+ 1_{\{t=T, x^1+x^2>2B\}}$.

2.2. The dual formulation. The dual formulation for hedging problems under general convex constraint was first established by Cvitanič and Karatzas [6] in the diffusion case and then extended to the semimartingale case by Föllmer and Kramkov [8]; see also Karatzas and Shreve [10] and the review paper by Soner and Touzi [16].

To state the dual formulation, we first need the characterization of the closed convex set K in terms of its support function δ. For $\rho \in R^d$ set

$$\delta(\rho) = \sup_{\gamma \in K} \gamma' \rho \geq 0 \,, \tag{2.9}$$

where the last inequality follows from the left-hand side of (2.6), and define

$$\tilde{K} := \{\rho \in \mathbb{R}^d \ : \ \delta(\rho) < \infty\} \,,$$

the domain of δ. Observe that the right-hand side of (2.6) implies that $\tilde{K} \neq \{0\}$. Moreover, it is a standard result of convex analysis (see, e.g., [11]) that K can be characterized in terms of

$$\tilde{K}_1 := \{\rho \in \tilde{K} \ : \ |\rho| = 1\}$$

by

$$\gamma \in K \ \Leftrightarrow \ H(1,\gamma) \geq 0 \quad \text{and} \quad \gamma \in \text{int}(K) \ \Leftrightarrow \ H(1,\gamma) > 0, \tag{2.10}$$

where

$$H(u,p) \ := \ \inf\{\delta(\rho)u - \rho' p, \ \rho \in \tilde{K}_1\} \text{ for } (u,p) \in \mathbb{R} \times \mathbb{R}^d \ .$$

Remark 2.4. Assume for a while that $0 \in \text{int}(K)$. Then there is $c_K > 0$ such that $B(0, c_K) \subset K$. Thus, for all $\rho \in \tilde{K}_1$, $c_K \rho \in K$ and therefore

$$\delta(\rho) \geq c_K > 0 \ .$$

The dual formulation is constructed as follows. Let us denote by $\tilde{\mathcal{K}}$ the set of bounded adapted processes ϑ taking values in $\tilde{K}$. To such a process, we associate the martingale M^ϑ defined on $[0,T]$ as the solution of

$$M_t := 1 + \int_0^t M_s \left(\sigma(s, X(s))^{-1} \vartheta_s\right)' dW_s \ ;$$

recall (2.2). We then define the $\mathbb{P}$-equivalent probability measure $\mathbb{Q}^\vartheta$ by

$$\frac{\mathrm{d}\mathbb{Q}^\vartheta}{\mathrm{d}\mathbb{P}} = M_T^\vartheta \ .$$

It follows from Girsanov's theorem that the process W^ϑ defined by

$$W_t^\vartheta = W_t - \int_0^t \sigma(s, X(s))^{-1} \vartheta_s ds, \ \ t \leq T \ ,$$

is a Brownian motion under $\mathbb{Q}^\vartheta$. In the following, we shall denote by $\mathbb{E}^\vartheta$ the expectation operator associated to $\mathbb{Q}^\vartheta$.

To $\vartheta \in \tilde{\mathcal{K}}$, we finally associate the process $\mathcal{E}^\vartheta$ defined by

$$\mathcal{E}_t^\vartheta := e^{-\int_0^t \delta(\vartheta_s) ds}, \quad t \leq T \ .$$

THEOREM 2.1. *The following holds:*

$$v(0, X_0) = \sup_{\vartheta \in \tilde{\mathcal{K}}} \mathbb{E}^\vartheta \left[\mathcal{E}_\tau^\vartheta g\left(\tau, X(\tau)\right)\right] \ . \tag{2.11}$$

Proof. The above result is a direct consequence of Theorem 6.2 and Remark 6.11 in [10], as follows:

1. First observe that

$$v(0, X_0) = \inf\left\{y \in \mathbb{R}_+ \ : \ Y_y^\pi(T) \geq g(\tau, X(\tau)) \quad \text{for some } \pi \in \mathcal{K}\right\} \ .$$

Indeed, it follows from (2.4) and condition (2.3) that, for all $y \in \mathbb{R}_+$ and $\pi \in \mathcal{K}$, the process Y_y^π is a nonnegative local $\mathbb{P}$-martingale on $[t,T]$. Hence it is a

supermartingale and, by taking conditional expectation, $Y_y^\pi(T) \geq g(\tau, X(\tau))$ implies $Y_y^\pi(\tau) \geq g(\tau, X(\tau))$. From this we deduce the first inequality:

$$v(0, X_0) \leq \inf\left\{y \in \mathbb{R}_+ \ : \ Y_y^\pi(T) \geq g(\tau, X(\tau)) \quad \text{for some } \pi \in \mathcal{K}\right\} .$$

For the converse inequality, notice that if $Y_y^\pi(\tau) \geq g(\tau, X(\tau))$, then $Y_T^{y,\tilde{\pi}} \geq g(\tau, X(\tau))$, where $\tilde{\pi} = \pi \mathbf{1}_{[s,\tau]}$ belongs to $\mathcal{K}$.

2. Since $g \geq 0$ (see (2.7)), it follows from Theorem 6.2 and Remark 6.11 in [10] that

$$v(0, X_0) = \sup_{\vartheta \in \tilde{\mathcal{K}}} \mathbb{E}^\vartheta \left[\mathcal{E}_T^\vartheta g\left(\tau, X(\tau)\right)\right] .$$

Observe that the process $\mathcal{E}^\vartheta$ is positive, nonincreasing in time, and recall that $g \geq 0$; then the last equality leads to (2.11). □

In order to derive the PDE characterization of the superhedging price, we shall use a standard dynamic programming principle for the dual formulation of Theorem 2.1. Before stating it, we need to extend the definition of v to general initial conditions $(t, x) \in [0, T] \times (0, \infty)^d$. For $(t, x) \in [0, T] \times (0, \infty)^d$, $y \in \mathbb{R}_+$, and $\pi \in \mathcal{K}$, we define $(X_{t,x}, Y_{t,x,y}^\pi)$ as the solution of (2.1)–(2.4) on $[t, T]$ with initial condition $(X_{t,x}(t), Y_{t,x,y}^\pi(t)) = (x, y)$.

The value function v is then defined on $[0, T] \times (0, \infty)^d$ by

$$v(t, x) := \inf\left\{y \in \mathbb{R}_+ \ : \ Y_{t,x,y}^\pi(\tau_{t,x}) \geq g(\tau_{t,x}, X_{t,x}(\tau_{t,x})) \quad \text{for some } \pi \in \mathcal{K}\right\}, \tag{2.12}$$

where

$$\tau_{t,x} := \inf\{s \in [t, T] \ : \ X_{t,x}(s) \notin \mathcal{O}\} \wedge T .$$

Remark 2.5. Observe that for $(t, x) \in ([0, T] \times \partial\mathcal{O}) \cup (\{T\} \times \bar{\mathcal{O}})$, we have $v(t, x) = g(t, x)$ by construction.

In what follows, we shall denote by $\mathcal{T}_{t,T}$ the set of all stopping times with values in $[t, T]$. Given $\vartheta \in \tilde{K}$ and $t < T$, we also set

$$\mathcal{E}_s^{t,\vartheta} := \mathcal{E}_s^\vartheta / \mathcal{E}_t^\vartheta \quad \text{for } s \geq t .$$

The following result is a consequence of Proposition 6.5 in [10].

PROPOSITION 2.1. *For all $(t, x) \in [0, T) \times \mathcal{O}$ and $\theta \in \mathcal{T}_{t,T}$,*

$$v(t, x) = \sup_{\vartheta \in \tilde{\mathcal{K}}} \mathbb{E}^\vartheta \left[\mathcal{E}_\theta^{t,\vartheta} v\left(\theta, X_{t,x}(\theta)\right) \mathbf{1}_{\theta < \tau_{t,x}} + \mathcal{E}_{\tau_{t,x}}^{t,\vartheta} g\left(\tau_{t,x}, X_{t,x}(\tau_{t,x})\right) \mathbf{1}_{\theta \geq \tau_{t,x}}\right] . \tag{2.13}$$

Proof. It follows from Proposition 6.5 in [10] that

$$v(t, x) = \sup_{\vartheta \in \tilde{\mathcal{K}}} \mathbb{E}^\vartheta \left[\mathcal{E}_{\theta \wedge \tau_{t,x}}^{t,\vartheta} v\left(\theta \wedge \tau_{t,x}, X_{t,x}(\theta \wedge \tau_{t,x})\right)\right] ,$$

where by definition of v (see Remark 2.5) $v\left(\tau_{t,x}, X_{t,x}(\tau_{t,x})\right) = g(\tau_{t,x}, X_{t,x}(\tau_{t,x}))$. This provides the required result. □

3. The PDE characterization. Our main result consists of a PDE characterization of the value function v. Before stating it, we describe the PDE associated to v and explain in which sense it has to be considered.

3.1. The associated PDE. Set $\mathcal{O}^* = \mathcal{O} \cap (0,\infty)^d$. In view of [7] and [17], it is natural to expect that the value function v is a viscosity solution on

$$D := [0,T) \times \mathcal{O}^*$$

of the PDE

$$\min\{-\mathcal{L}v\ ,\ \mathcal{H}v\} = 0\ , \tag{3.1}$$

where for a smooth function φ on $[0,T] \times \mathbb{R}^d_+$ we set

$$\mathcal{H}\varphi(t,x) = \inf\left\{\delta(\rho)\varphi(t,x) - \rho' \mathrm{diag}\,[x]\, D\varphi(t,x),\ \rho \in \tilde{K}_1\right\}\ ,$$
$$\mathcal{L}\varphi(t,x) = \frac{\partial}{\partial t}\varphi(t,x) + \frac{1}{2}\mathrm{Tr}\left[a(t,x)D^2\varphi(t,x)\right]\ ,$$

with a defined on $[0,T] \times \mathbb{R}^d_+$ by

$$a(t,x) := \mathrm{diag}\,[x]\, \sigma(t,x)\sigma(t,x)' \mathrm{diag}\,[x]\ .$$

The first part of the equation corresponds to the usual Black–Scholes equation, while the second part is due to the portfolio constraint. Indeed, assuming that v is smooth and positive, and writing formally that the hedging portfolio satisfies $Y^\pi_y(t) = v(t, X(t))$, we deduce from Itô's lemma that π_t must coincide with $\mathrm{diag}\,[X(t)]\, Dv(t,X(t))/v(t, X(t))$. Since it has to belong to K, the characterization of K given by (2.10) implies that $H(1, \mathrm{diag}\,[X(t)]\, Dv(t,X(t))/\ v(t,X(t)))$, or equivalently $H(v(t,X(t)), \mathrm{diag}\,[X(t)]\ Dv(t,X(t)))$, must be nonnegative.

In order to provide a full characterization of v, it remains to define the boundary conditions on $\partial_x D^* := [0,T) \times \partial\mathcal{O}^*$ and $\partial_T D^* := \{T\} \times \bar{\mathcal{O}}^*$, where

$$\partial\mathcal{O}^* := \partial\mathcal{O} \cap (0,\infty)^d \qquad \text{and} \qquad \bar{\mathcal{O}}^* := \bar{\mathcal{O}} \cap (0,\infty)^d\ .$$

It is known from [7] (see also [16] and [17]) that the boundary condition on $\partial_T D^*$ has to be written as

$$v = \hat{g}, \tag{3.2}$$

where, for $x \in (0,\infty)^d$,

$$\hat{g}(T,x) = \sup_{\rho \in \tilde{K}} e^{-\delta(\rho)} g(T, xe^\rho)\ .$$

This corresponds to the "face-lifting" procedure, which was already observed by [4]. This "face-lifting" is due to the portfolio constraint, $\hat{g}$ being the smallest function above g which, in a sense, satisfies $D\hat{g}/\hat{g} \in K$.

Remark 3.1. Observe that (2.7) allows us to define $\hat{g}(T,\cdot)$ on $(0,\infty)^d$ as

$$\hat{g}(T,x) = \sup_{\rho \in \tilde{K}(x,\bar{\mathcal{O}})} e^{-\delta(\rho)} g(T, xe^\rho)\ , \tag{3.3}$$

with the convention $\sup\emptyset = 0$ and

$$\tilde{K}(x,E) := \left\{\rho \in \tilde{K}\ :\ xe^\rho \in E\right\} \quad \text{for } E \subset \bar{\mathcal{O}}\ . \tag{3.4}$$

The fact that v satisfy (3.1)–(3.2) in the viscosity sense can be shown by following the arguments of [7] and is not difficult.

The difficulty comes from the boundary condition on $\partial_x D^*$. In this paper, we shall show that g has also to be modified on $\partial_x D^*$, i.e., replaced by $\hat{g}$ defined on $[0,T) \times (0,\infty)^d$ by

$$\hat{g}(t,x) = \sup_{\rho \in \tilde{K}(x,\partial\mathcal{O})} e^{-\delta(\rho)} g(t, xe^{\rho}) \ , \tag{3.5}$$

with the convention $\sup \emptyset = 0$. This result is expected and will be obtained under a smoothness condition on $\mathcal{O}$; see $\mathbf{H}_{\mathcal{O}}$ below.

But this is only a first step in the derivation of the appropriate boundary condition. Actually, [12] provides an example of a superhedging price for an up-and-out call option for which $\hat{g}(t,x) = 0$ for $t < T$ and $v(t',x')$ does not converge to 0 when $(t',x') \in D$ goes to $(t,x) \in \partial_x D^*$. This shows that the constraint on the portfolio may prevent the value function from assuming the boundary condition continuously and leads to the natural formulation of a relaxed boundary condition on $\partial_x D^*$:

$$\min\{v - \hat{g} \ , \ \mathcal{H}v\} = 0 \ . \tag{3.6}$$

However, we shall see in Remark 6.1 below that the above equation has to be corrected in order to admit a viscosity supersolution and therefore makes sense. Given a smooth function φ, we therefore define

$$\mathcal{H}_{\rm d}\varphi(t,x) = \inf\Big\{\delta(\rho)\varphi(t,x) - \rho' \mathrm{diag}\,[x]\, D\varphi(t,x), \ \rho \in \tilde{K}_1(x,\bar{\mathcal{O}})\Big\} \ ,$$

where, for $x \in E \subset \bar{\mathcal{O}}$,

$$\tilde{K}_1(x,E) := \Big\{\rho \in \tilde{K}_1 \ : \ \exists\, \lambda_0 > 0 \text{ s.t. } \lambda\rho \in \tilde{K}(x,E) \text{ for all } \lambda \in [0,\lambda_0]\Big\} \ . \tag{3.7}$$

To sum up, we introduce the operators

$$\mathcal{B}\varphi := \begin{cases} \min\{-\mathcal{L}\varphi \ , \ \mathcal{H}\varphi\} & \text{on} \quad D \ , \\ \min\{\varphi - \hat{g} \ , \ \mathcal{H}\varphi\} & \text{on} \quad \partial_x D^* \ , \\ \varphi - \hat{g} & \text{on} \quad \partial_T D^* \ , \end{cases}$$

$$\mathcal{B}_{\rm d}\varphi := \begin{cases} \mathcal{B}\varphi & \text{on} \quad D \cup \partial_T D^* \ , \\ \min\{\varphi - \hat{g} \ , \ \mathcal{H}_{\rm d}\varphi\} & \text{on} \quad \partial_x D^* \ , \end{cases}$$

and we say that a locally bounded function w on D is a *discontinuous viscosity solution* of

$$\mathcal{B}_{\rm d}\varphi = 0 \tag{3.8}$$

on $\bar{D}^* := \bar{D} \cap ([0,T] \times (0,\infty)^d)$ if w_* and w^* defined on $\bar{D}$ as

$$w_*(t,x) \ := \ \liminf_{(\tilde{t},\tilde{x}) \in D, \ (\tilde{t},\tilde{x}) \to (t,x)} w(\tilde{t},\tilde{x}) \qquad \text{and} \qquad w^*(t,x) \ := \ \limsup_{(\tilde{t},\tilde{x}) \in D, \ (\tilde{t},\tilde{x}) \to (t,x)} w(\tilde{t},\tilde{x})$$

are, respectively, viscosity super- and subsolutions of $\mathcal{B}_{\rm d}\varphi = 0$ and $\mathcal{B}\varphi = 0$ on $\bar{D}^*$. More generally, we shall say that w is a (discontinuous) viscosity supersolution (resp., subsolution) of $\mathcal{B}\varphi = 0$ on $\bar{D}^*$ if w_* is a supersolution of $\mathcal{B}_{\rm d}\varphi = 0$ (resp., subsolution of $\mathcal{B}\varphi = 0$) on $\bar{D}^*$.

Remark 3.2. Assume that the conditions of Theorem 3.1 below hold. Let us write $\mathcal{B}\varphi(t,x)$ as $\mathcal{B}(t,x,\varphi(t,x),\frac{\partial}{\partial t}\varphi(t,x),D\varphi(t,x),D^2\varphi(t,x))$ and $\mathcal{B}_{\rm d}\varphi(t,x)$ similarly. Then one easily checks that the upper-semicontinuous envelope of $\mathcal{B}_{\rm d}$ as a map on $\bar{D}^*\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}^d\times\mathbb{M}^d$ is given by

$$(\mathcal{B}_{\rm d})_+(t,x,\varphi(t,x),\tfrac{\partial}{\partial t}\varphi(t,x),D\varphi(t,x),D^2\varphi(t,x))) = \max\{\mathcal{B}_{\rm d}\varphi(t,x), \min\{-\mathcal{L}\varphi(t,x),\mathcal{H}\varphi(t,x)\}\},$$

and that its lower-semicontinuous envelope is

$$(\mathcal{B}_{\rm d})_-(t,x,\varphi(t,x),\frac{\partial}{\partial t}\varphi(t,x),D\varphi(t,x),D^2\varphi(t,x))) = \min\{\mathcal{B}\varphi(t,x),-\mathcal{L}\varphi(t,x)\}.$$

From the arguments of the proof of Propositions 6.3 and 6.6 below, we deduce that $(\mathcal{B}_{\rm d})_+\varphi=0$ (resp., $(\mathcal{B}_{\rm d})_-\varphi=0$) has the same supersolutions as $\mathcal{B}_{\rm d}\varphi=0$ (resp., subsolutions as $\mathcal{B}\varphi=0$) on $D\cup\partial_x D^*$ for the terminal condition $\varphi=\hat{g}$ at T. In other words, $\mathcal{B}_{\rm d}$ can be viewed as being upper-semicontinuous with lower-semicontinuous envelope given by $\mathcal{B}$. This justifies the above definition of a viscosity solution of $\mathcal{B}_{\rm d}\varphi=0$, and shows that it is in accordance with Definition 7.4 in [5]. This remak will be used in the example section to prove the convergence of the discretization scheme we shall consider for a particular example.

3.2. Main results. In order to establish that v is a discontinuous viscosity solution of (3.8), we shall appeal to the following additional assumptions.

Our first condition concerns the convex set K describing the portfolio constraints. It is stated in terms of $\tilde{K}(x,\mathcal{O})$; recall (3.4).

$\mathbf{H}_{\tilde{K}}$: (i) For all $x\in\mathcal{O}$, $\rho\in\tilde{K}(x,\mathcal{O})$ implies $\lambda\rho\in\tilde{K}(x,\mathcal{O})$ for all $\lambda\in[0,1)$.
(ii) For all $x\in\mathcal{O}$, the closure of $\tilde{K}(x,\mathcal{O})$ is equal to $\tilde{K}(x,\bar{\mathcal{O}})$.
(iii) If $(x_n)_n$ is a sequence in $\mathcal{O}$ such that $x_n\to x\in\partial\mathcal{O}$ and $\rho\in\tilde{K}(x,\bar{\mathcal{O}})$, then there exists a sequence $\rho_n\to\rho$ such that, up to a subsequence, $\rho_n\in\tilde{K}(x_n,\bar{\mathcal{O}})$ for all $n\geq 1$.

Remark 3.3. The conditions (i) and (ii) of $\mathbf{H}_{\tilde{K}}$ are automatically satisfied whenever the set $\ln(\mathcal{O})=\{(\ln(x^i))_{i\leq d},\ x\in\mathcal{O}\}$ is convex. Indeed, we easily check that in this case, for all $x\in\mathcal{O}$, $\tilde{K}(x,\mathcal{O})$ is convex, and since $0\in\tilde{K}(x,\mathcal{O})$, this provides (i). The convexity of $\ln(\mathcal{O})$ also implies that if $\rho\in\tilde{K}(x,\mathcal{O})$ and $\bar{\rho}\in\tilde{K}(x,\bar{\mathcal{O}})$, then $\lambda\rho+(1-\lambda)\bar{\rho}\in\tilde{K}(x,\mathcal{O})$ for all $\lambda\in(0,1)$. Since $0\in\tilde{K}$, this shows that for all $x\in\mathcal{O}$ the closure of $\tilde{K}(x,\mathcal{O})$ contains $\tilde{K}(x,\bar{\mathcal{O}})$, while the converse inclusion is obvious.

We shall also impose some regularity assumptions on g.

$\mathbf{H}_g$: (i) g is lower-semicontinuous on $[0,T]\times\partial\mathcal{O}^*$ and on $\{T\}\times\bar{\mathcal{O}}^*$.
(ii) There exist $C_g>0$ and $\bar{\gamma}\in K\cap\mathbb{R}^d_+$ s.t. $|g(\cdot,x)|\leq C_g(1+x^{\bar{\gamma}})$ for all $x\in\bar{\mathcal{O}}^*$.
(iii) $\hat{g}$ is upper-semicontinuous on $[0,T]\times(0,\infty)^d$ and has linear growth.

Under $\mathbf{H}_{\tilde{K}}$ and (i)–(ii) of $\mathbf{H}_g$, one can already derive the following qualitative properties of v.

PROPOSITION 3.1. *Assume that* $\mathbf{H}_{\tilde{K}}$ *and* (i)–(ii) *of* $\mathbf{H}_g$ *hold. Then, for all* $(t,x)\in D$, *we have*

$$v(t,x)\geq 0\,, \tag{3.9}$$

and there is a constant $C > 0$, independent of (t, x), such that

$$|v(t,x)| \le C\left(1 + x^{\bar\gamma}\right) . \tag{3.10}$$

Moreover, for all $(t, x) \in \bar D$,

$$v_*(t,x) = \sup_{\rho \in \tilde K(x,\bar{\mathcal{O}})} e^{-\delta(\rho)} v_*(t, xe^{\rho}) . \tag{3.11}$$

The proof will be provided in section 5.

In order to derive the appropriate boundary condition on $\partial_x D^*$, we shall also need some regularity on the domain $\mathcal{O}$.

$\mathbf{H}_{\mathcal{O}}$: There exists a map $d : (0,\infty)^d \mapsto \mathbb{R}$ such that
- (i) $\{x \in (0,\infty)^d \ : \ d(x) > 0\} = \mathcal{O}^*$.
- (ii) $\{x \in (0,\infty)^d \ : \ d(x) = 0\} = \partial\mathcal{O}^*$.
- (iii) For all $x \in \partial\mathcal{O}^*$, there exists $r > 0$ s.t. $d \in C^2(B(x,r))$.

This essentially amounts to saying that $\mathcal{O}$ is C^2; see [9].

Using $\mathbf{H}_{\tilde K}$, $\mathbf{H}_g$, and $\mathbf{H}_{\mathcal{O}}$, we can already characterize v not only as a discontinuous solution of (3.8) but also as its smallest supersolution.

THEOREM 3.1. *Assume that $\mathbf{H}_{\tilde K}$, $\mathbf{H}_g$, and $\mathbf{H}_{\mathcal{O}}$ hold. Then*
- (i) *v is a discontinuous viscosity solution of* (3.8)*;*
- (ii) *v is lower-continuous on D;*
- (iii) *v_* is the smallest supersolution of* (3.8) *in the class of locally bounded functions satisfying* (3.10).

Finally, under the following additional assumptions we will be able in section 7 to provide a comparison theorem for (3.8).

$\mathbf{H}'$:
- (i) Either $\bar{\mathcal{O}}$ is bounded or there exists $\underline{\varrho} > 1$ s.t. $\underline{\varrho}\bar\gamma \in K \cap (0,\infty)^d$.
- (ii) $\mathrm{int}(K) \neq \emptyset$ and either $0 \in \mathrm{int}(K)$ or $\bar{\mathcal{O}} \cap \partial\mathbb{R}^d_+ = \emptyset$.
- (iii) For all $x \in \partial\mathcal{O}^*$ there exists $\rho \in \tilde K_1$ s.t. $Dd(x)' \mathrm{diag}\,[x]\,\rho > 0$.

This will imply our last result, which characterizes v as the unique solution of (3.8) in a suitable class of functions.

THEOREM 3.2. *Let the conditions of Theorem* 3.1 *hold and assume further that $\mathbf{H}'$ is satisfied. Then*
- (i) *$v_* = v^*$ on $\bar D^*$;*
- (ii) *v is continuous on D;*
- (iii) *v is the unique discontinuous viscosity solution of* (3.8) *in the class of locally bounded function satisfying* (3.10).

Remark 3.4. Recall the examples of barrier options of section 2.
1. If we hedge the up-and-out call of the Example 2.1 with short-sale constraints, i.e., $K = [-\alpha, +\infty)$, with $\alpha > 0$, then it is easy to verify that all of the conditions $\mathbf{H}_{\tilde K}$, $\mathbf{H}_g$, $\mathbf{H}_{\mathcal{O}}$, and $\mathbf{H}'$ hold true.
2. These conditions are also satisfied when we hedge the down-and-out basket put of the Example 2.2 with a bounded portfolio, i.e., $K = \prod_{i=1}^2 [-\underline{\alpha}^i, \bar\alpha^i]$, $\underline{\alpha}^i, \bar\alpha^i > 0$ for $i = 1, 2$.

Remark 3.5. In [13], the authors consider example 1 of Remark 3.4 within a one-dimensional Black–Scholes model and use the PDE characterization of Theorem 3.2 to solve the problem explicitly; see [13, section 4.2.1]. They work in a case where

v is expected to be strictly positive while $g = 0$ for $t < T$. They therefore guess that the condition $\mathcal{H}w = 0$ should always hold at the boundary. Then they observe that, in their particular setting, $w := \mathcal{H}v$ actually solves $\mathcal{L}\varphi = 0$ if v does. Since by construction $w = 0$ on the boundary, it is easy to guess its explicit form; this is the price of a barrier option without constraint and a modified terminal payoff. The equation $\mathcal{H}v = w$ is then solved in v as an ordinary differential equation in x. They finally verify that this provides a solution to (3.8). This general methodology can be applied in various particular cases.

Remark 3.6. To conclude this section, let us comment on assumption $\mathbf{H}'$. As already mentioned, Theorem 3.2 is based on a comparison result for (3.8) stated in Theorem 7.1 below. A first difficulty in proving this theorem comes from the growth condition (3.10), which is nonstandard. In the case where $\bar{\mathcal{O}}$ is not bounded, the second assumption in (i) is used to construct a suitable penalty function which allows us to reduce to a bounded domain. The second difficulty comes from the term $\mathcal{H}\varphi$ appearing in $\mathcal{B}\varphi$. It is handled by using the first assertion of (ii) under which we can construct a strict supersolution of $\mathcal{H}\varphi = 0$. A third difficulty is due to the fact that the equation in written only on $\bar{\mathcal{O}} \cap (0,\infty)^d$. In the case where $\bar{\mathcal{O}} \cap \partial\mathbb{R}^d_+ \neq \emptyset$, we need to introduce another penalty function which permits us to reduce the analysis to $(0,\infty)^d$. We then appeal to the second assertion of (ii). Finally, a major difficulty comes from the boundary condition on $\partial_x D^*$, which is written in a weak sense. It is treated by using condition (iii), which allows us to "avoid" this boundary. We refer to step 4 of the proof of Theorem 7.1 below for a more detailed discussion of assumptions (i) and (ii).

4. A numerical application. In this section, we study a numerical scheme for the resolution of $\mathcal{B}_{\rm d}\varphi = 0$ in the simple example considered in [12]: superhedging of a knock-out call with a short-sale constraint. The general case will be discussed in the companion paper [2].

The model corresponds to our general framework with $d = 1$, $\sigma(t,x) = \sigma > 0$, a fixed constant, $\mathcal{O} = (-\infty, B)$, $K = [-\alpha, \infty)$, and $g(t,x) = [x-\kappa]^+ \mathbf{1}_{\{t=T,\, x<B\}}$, with $\alpha > 0$, $B > \kappa > 0$. In this case $\tilde{K} = (-\infty, 0]$, the function $\hat{g}(t,x)$ is equal to

$$\hat{g}(t,x) \;=\; e^{-\alpha\theta(x)}[xe^{-\theta(x)} - K]^+ \mathbf{1}_{t=T} \text{ with } \theta(x) = [-\ln(B/x)]^+ \,,$$

and all the assumptions of Theorem 3.2 are satisfied.

In order to solve numerically the equation $\mathcal{B}_{\rm d}\varphi = 0$, we propose the following discretization. We fix a regular grid $\pi_h = \{t_i := (ir_h) \wedge T,\ 0 \le i \le I_h\}$ of $[0,T]$ and $\Xi_h := \{x_i := (ih) \wedge B,\ 0 \le i \le N_h\}$ of $[0,B]$. Here, $h > 0$ is a fixed parameter, $N_h := \inf\{k \in \mathbb{N} \,:\, k \ge B/h\}$, and $I_h := \inf\{k \in \mathbb{N} \,:\, k \ge T/r_h\}$ with $r_h = h^2$. The approximation v^h of v is defined as follows.

1. For $i = I_h$, we use the boundary condition at $t = T$ to set $v^h(t_{I_h}, x_j) = \hat{g}(t_{I_h}, x_j)$, $j = 0, \dots, N_h$.
2. For $i = I_h - 1, \dots, 0$, we use the following procedure:
   (a) We initialize: $v^h(t_i, 0) = 0$.
   (b) Then we solve on $j = 1, \dots, N_h$ the system

$$v^h(t_i,x_j) = \begin{cases} \max\left\{A^h(v^h,i,j)\,;\ B^h(v^h,i,j)\right\} & \text{if } j \neq N_h\,, \\ \max\left\{0\,;\ B^h(v^h,i,j)\right\} & \text{otherwise}\,, \end{cases}$$

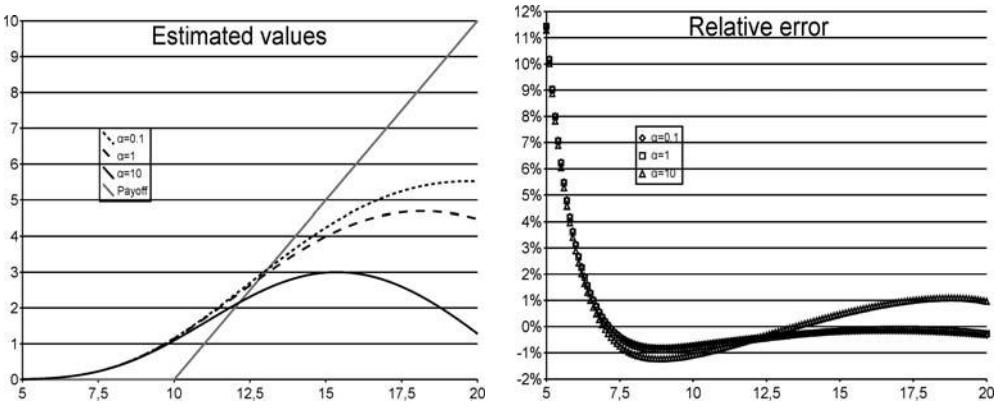


FIG. 1.

with

$$A^h(v^h,i,j) := \frac{(r_h)^{-1}v^h(t_{i+1},x_j) + (2h^2)^{-1}\sigma^2 x_j^2 (v^h(t_i,x_{j+1}) + v^h(t_i,x_{j-1}))}{(r_h)^{-1} + (h^2)^{-1}\sigma^2 x_j^2},$$

$$B^h(v^h,i,j) := \frac{x_j\, h^{-1} v^h(t_i,x_{j-1})}{\alpha + x_j\, h^{-1}}\ ,$$

The initialization of step 2(a) is justified by the continuity of v at 0, which is easily checked in this simple model by using the dual formulation of Theorem 2.1. The system given in 2(b) follows from the approximation of $\mathcal{H} = \mathcal{H}_d$ and $\mathcal{L}$ by

$$\mathcal{H}^h(t_i,x_j,v^h(t_i,x_j),v^h) = \alpha v^h(t_i,x_j) + x_j \frac{v^h(t_i,x_j) - v^h(t_i,x_{j-1})}{h}\ ,$$

$$\mathcal{L}^h(t_i,x_j,v^h(t_i,x_j),v^h) = \frac{v^h(t_{i+1},x_j) - v^h(t_i,x_j)}{r_h} + \frac{1}{2}\sigma^2 x_j^2 \frac{v^h(t_i,x_{j+1}) + v^h(t_i,x_{j-1}) - 2v^h(t_i,x_j)}{h^2}\ .$$

Observing that v^h is nonnegative and uniformly bounded from above by B, the convergence of the scheme easily follows from Remark 3.2, Theorem 3.2, Remark 7.1 below, and [1].

In Figure 1, we plot the estimation of v obtained with this scheme for $\sigma = 0.3$, $\kappa = 10$, $B = 20$, $T = 1$ and for $\alpha \in \{0.1, 1, 10\}$. The relative error is computed by using the closed-form solution obtained in [12]. We took $N_h = 200$. We observe that the estimation is very sharp with a relative error less than 1% in absolute value, except for small values of X_0, for which v is almost equal to 0.

5. Growth and monotonicity properties. In this section, we provide the proof of Proposition 3.1.

Proof of (3.9)–(3.10). The lower bound of (3.9) is an immediate consequence of the assumption $g \geq 0$ and the dual formulation (2.11). We now prove (3.10). Let $\pi \in \mathcal{K}$ be defined by $\pi_t = \bar{\gamma}$ for all $t \leq T$. Since σ is bounded (see (2.2)), one easily checks from the dynamics of the processes $X_{t,x}$ and $Y^{\pi}_{t,x,1}$ that

$$1 + \prod_{i=1}^{d} (X^i_{t,x}(u))^{\bar{\gamma}^i} \ \leq\ C\left(1 + \prod_{i=1}^{d} (x^i)^{\bar{\gamma}^i}\right) Y^{\pi}_{t,x,1}(u) \quad \text{for all } \ u \in [0,T]\ \mathbb{P}\text{-a.s.}\ ,$$

where $C > 0$ depends only on $|\bar\gamma|$ and the bound on $|\sigma|$. Then, after possibly changing the value of the constant C, $\mathbf{H}_g$-(ii) implies

$$g(u, X_{t,x}(u)) \;\le\; C\left(1 + x^{\bar\gamma}\right) Y^{\pi}_{t,x,1}(u) \quad \text{for all } \; u \in [0,T], \; \mathbb{P}\text{-a.s.}\,,$$

and since $yY^{\pi}_{t,x,1} = Y^{\pi}_{t,x,y}$ for $y > 0$, we deduce from the last inequality that $v(t,x) \le C\left(1 + x^{\bar\gamma}\right)$. □

Proof of (3.11). Since $0 \in \tilde K(x, \bar{\mathcal{O}})$, we only have to show that

$$v_*(t,x) \ge \sup_{\rho \in \tilde K(x,\bar{\mathcal{O}})} e^{-\delta(\rho)} v_*(t, xe^{\rho})\,.$$

1. We first consider the case where $(t,x) \in D$. Since by lower-semicontinuity of v_* and (ii) of $\mathbf{H}_{\tilde K}$

$$\sup_{\rho \in \tilde K(x,\mathcal{O})} e^{-\delta(\rho)} v_*(t, xe^{\rho}) = \sup_{\rho \in \tilde K(x,\bar{\mathcal{O}})} e^{-\delta(\rho)} v_*(t, xe^{\rho})\,,$$

it suffices to show that

$$v_*(t,x) \ge \sup_{\rho \in \tilde K(x,\mathcal{O})} e^{-\delta(\rho)} v_*(t, xe^{\rho})\,. \tag{5.1}$$

Fix $\rho \in \tilde K(x,\mathcal{O})$ and consider the sequence of processes ϑ^n in $\tilde{\mathcal{K}}$ defined on $[t,T]$ by $\vartheta^n := n\rho \mathbf{1}_{[t,t_n]}$ with $t_n := t + n^{-1}$ for n large enough so that $t_n < T$. By Proposition 2.1

$$\begin{aligned} v(t,x) \ge \mathbb{E}^{\vartheta^n}\Big[e^{-n\delta(\rho)(n^{-1}\wedge(\tau_{t,x}-t))} \Big(& v\left(t_n, X_{t,x}(t_n)\right) \mathbf{1}_{t_n < \tau_{t,x}} \\ & + g\left(\tau_{t,x}, X_{t,x}(\tau_{t,x})\right) \mathbf{1}_{t_n \ge \tau_{t,x}} \Big) \Big]\,. \end{aligned} \tag{5.2}$$

Let X^n be the solution on $[t,T]$ of

$$X^n(s) = x + \int_t^s \operatorname{diag}\left[X^n(r)\right] \vartheta^n_r dr + \int_t^s \operatorname{diag}\left[X^n(r)\right] \sigma(r, X^n(r)) dW_r$$

so that $X^n(s) = \beta^n_s H^n_s$ with

$$(H^n_s)^i \;:=\; \mathcal{E}\left(\sum_{j=1}^d \int_t^s \sigma^{ij}(r, X^n(r)) dW^j_r \right) \quad \text{and} \quad \beta^n_s \;:=\; xe^{\int_t^s \vartheta^n_r dr}\,,$$

where $\mathcal{E}$ denotes the Doleans–Dade exponential. By Girsanov's theorem, (5.2) can be rewritten as

$$\begin{aligned} v(t,x) \ge \mathbb{E}\Big[e^{-n\delta(\rho)(n^{-1}\wedge(\tau_n - t))} & \left(v\left(t_n, X^n(t_n)\right) \mathbf{1}_{t_n < \tau_n}\right. \\ & \left. + g\left(\tau_n, X^n(\tau_n)\right) \mathbf{1}_{t_n \ge \tau_n}\right) \Big]\,, \end{aligned} \tag{5.3}$$

where

$$\tau_n := \inf\left\{ s \in [t,T] \;:\; X^n(s) \notin \mathcal{O} \right\} \wedge T\,.$$

Since σ is bounded (see (2.2)), $H^n_{\tau_n \wedge t_n} \to (1, \dots, 1)$ $\mathbb{P}$-a.s., after possibly passing to a subsequence. Also observe that

$$\beta^n_{\tau_n \wedge t_n} = xe^{\rho[(n(\tau_n - t))\wedge 1]}\,.$$

By $\mathbf{H}_{\tilde{K}}$ and the assumption $\rho \in \tilde{K}(x, \mathcal{O})$, it follows that, $\mathbb{P}$-a.s., $X^n_{\tau_n \wedge t_n} \in \mathcal{O}$ and therefore $t_n < \tau_n$ for large values of n. In particular,

$$(X^n(t_n), \mathbf{1}_{t_n < \tau^n}) \longrightarrow (xe^{\rho}, 1) \ \ \mathbb{P}\text{-a.s.}$$

Thus, passing to the limit in (5.3) and applying Fatou's lemma shows the required result; recall (3.9).

2. We now consider the case where $(t, x) \in \partial D$. Let $(t_n, x_n)_n$ be a sequence in D that converges to (t, x) such that $v(t_n, x_n) \to v_*(t, x)$. Fix $\rho \in \tilde{K}(x, \bar{\mathcal{O}})$. By $\mathbf{H}_{\tilde{K}}$, there is a sequence $(\rho_n)_n$ with values in $\tilde{K}(x_n, \bar{\mathcal{O}})$ such that $\rho_n \to \rho$. Using Step 1 of this proof, we deduce that $v(t_n, x_n) \geq e^{-\delta(\rho_n)} v_*(t_n, x_n e^{\rho_n})$. Passing to the limit shows that $v_*(t, x) \geq e^{-\delta(\rho)} v_*(t, xe^{\rho})$ by lower-semicontinuity of v_*. □

Remark 5.1. Fix $(t, x) \in \bar{D}^*$ and assume that $(\lambda_0, \rho_0) \in \mathbb{R}_+ \times \tilde{K}$ are such that $xe^{\lambda_0 \rho_0} \in \bar{\mathcal{O}}$. By $\mathbf{H}_{\tilde{K}}$-(i), the map $\lambda \in [0, \lambda_0] \mapsto e^{-\lambda \delta(\rho_0)} v_*(t, xe^{\lambda \rho_0})$ is well defined and it follows from (3.11) that it is nonincreasing.

6. The viscosity solution property. In this section, we provide the proof of Theorem 3.1. We start with the supersolution and subsolution properties. Then we use an approximation argument combined with a comparison theorem to prove that v_* is the smallest supersolution of (3.8).

6.1. Supersolution property. In this section, we show that v_* is a supersolution of (3.8) on $\bar{D}^*$. This is a consequence of Propositions 6.1, 6.2, 6.3, and 6.4 below.

PROPOSITION 6.1. *Assume that* $\mathbf{H}_{\tilde{K}}$ *and* $\mathbf{H}_g$ *hold. Let* $(t_0, x_0) \in \bar{D}^*$ *and* $\varphi \in C^2(\bar{D}^*)$ *be such that* (t_0, x_0) *is a local minimum on* $\bar{D}^*$ *of* $v_* - \varphi$ *satisfying* $(v_* - \varphi)(t_0, x_0) = 0$. *Then*

$$\mathcal{H}_{\mathrm{d}} \varphi(t_0, x_0) \geq 0 \ .$$

Proof. By (3.11), for all $\rho \in \tilde{K}_1(x, \bar{\mathcal{O}})$ and $\lambda > 0$ small enough, we must have

$$\varphi(t_0, x_0) = v_*(t_0, x_0) \geq e^{-\lambda \delta(\rho)} v_*(t_0, x_0 e^{\lambda \rho}) \geq e^{-\lambda \delta(\rho)} \varphi(t_0, x_0 e^{\lambda \rho}) \ .$$

Thus, dividing by λ and sending λ to 0 leads to the required result. □

Remark 6.1. Assume that $\mathbf{H}_{\mathcal{O}}$ holds and that for all $(t_0, x_0) \in \bar{D}^*$ and φ as in Proposition 6.1, we have

$$\mathcal{H} \varphi(t_0, x_0) \geq 0 \ .$$

Let (t_0, x_0) and φ be as in Proposition 6.1 with $x_0 \in \partial \mathcal{O}^*$. Recall from $\mathbf{H}_{\mathcal{O}}$ the definition of the function d and observe that (t_0, x_0) is also a local minimum of $(v_* - \varphi)(t, x) + \varepsilon^{-1} d(x)$ on $\bar{D}^*$ for all $\varepsilon > 0$. Thus, if the above assertion is true, $\varphi - \varepsilon^{-1} d$ must satisfy

$$\delta(\rho) v_*(t_0, x_0) - \rho' \mathrm{diag}\,[x_0] \left(D\varphi(t_0, x_0) - \varepsilon^{-1} Dd(x_0) \right) \geq 0 \quad \text{for all} \ \ \rho \in \tilde{K}_1 \ .$$

Now observe that for $\rho \in \tilde{K}_1 \setminus \tilde{K}_1(x_0, \bar{\mathcal{O}})$, there is a sequence of positive parameters $\lambda_n \to 0$ such that $d(x_0 e^{\lambda_n \rho}) < 0 = d(x_0)$ for all n; recall (3.7). This implies that $\rho' \mathrm{diag}\,[x_0]\, Dd(x_0) < 0$. Hence, sending $\varepsilon \to 0$ in the above inequality leads to a contradiction if $\tilde{K}_1 \setminus \tilde{K}_1(x_0, \bar{\mathcal{O}}) \neq \emptyset$.

PROPOSITION 6.2. *Let $(t_0, x_0) \in D$ and $\varphi \in C^2(\bar{D}^*)$ be such that (t_0, x_0) is a local minimum on $\bar{D}^*$ of $v_* - \varphi$ satisfying $(v_* - \varphi)(t_0, x_0) = 0$. Then*

$$-\mathcal{L}\varphi(t_0, x_0) \geq 0 \,. \tag{6.1}$$

Proof. The proof is standard. Let V be a bounded open neighborhood of (t_0, x_0) such that (t_0, x_0) is a minimum on $\bar{V} \cap \bar{D}^*$ of $v_* - \varphi$ and let $(t_n, x_n)_n$ be a sequence in $V \cap D$ such that $(t_n, x_n) \to (t_0, x_0)$ and $v(t_n, x_n) \to v_*(t_0, x_0)$. For ease of notation we write $(\tau_n, X^n) = (\tau_{t_n,x_n}, X_{t_n,x_n})$. Given a sequence $(\eta_n)_n$ of positive numbers such that $t_n + \eta_n < T$ for all n, we set

$$\theta_n := \inf\{s \in [t_n, T] \ : \ (s, X^n(s)) \notin V \cap D\} \wedge (t_n + \eta_n) \,.$$

Since $0 \in \tilde{K}$, (2.13), the assumption $g \geq 0$ (see (2.7)) and the inequality $v_* \geq \varphi$ on V imply that

$$v(t_n, x_n) \geq \mathbb{E}\left[\varphi\left(\theta_n, X^n_{\theta_n}\right) \mathbf{1}_{\theta_n < \tau_n}\right] \,.$$

Set $\epsilon_n := v(t_n, x_n) - \varphi(t_n, x_n)$ and observe that ϵ_n converges to 0 as n goes to infinity. Moreover, it follows from Itô's lemma that

$$\epsilon_n \geq \mathbb{E}\left[\int_{t_n}^{\theta_n} \mathcal{L}\varphi(s, X^n(s)) ds \; \mathbf{1}_{\theta_n < \tau_n}\right] \,. \tag{6.2}$$

Using standard estimates, we then observe that

$$\liminf_{n\to\infty} \mathbb{E}\left[\eta_n^{-1} \int_{t_n}^{\theta_n} \mathcal{L}\varphi(s, X^n(s)) ds \; \mathbf{1}_{\theta_n < \tau_n}\right] \geq \mathcal{L}\varphi(t_0, x_0)$$

whenever $\eta_n \to 0$. Thus, choosing $(\eta_n)_n$ such that $\epsilon_n/\eta_n \to 0$ and using (6.2) leads to the required result. □

PROPOSITION 6.3. *Assume that $\mathbf{H}_{\tilde{K}}$ and $\mathbf{H}_g$ hold. Then $v_* \geq \hat{g}$ on $\partial_x D^*$.*

Proof.

1. We first prove that for all $(t_0, x_0) \in \partial_x D^*$ and $\varphi \in C^2(\bar{D}^*)$ such that

$$0 = (v_* - \varphi)(t_0, x_0) \ = \ \min_{\bar{D}^*} (\text{strict})(v_* - \varphi) \,,$$

we have

$$\max\{v_*(t_0, x_0) - g(t_0, x_0) \ ; \ -\mathcal{L}\varphi(t_0, x_0)\} \geq 0 \,. \tag{6.3}$$

Assume to the contrary that

$$\max\{\varphi(t_0, x_0) - g(t_0, x_0) \ ; \ -\mathcal{L}\varphi(t_0, x_0)\} \leq -2\varepsilon \tag{6.4}$$

for some $\varepsilon > 0$. Let $(t_n, x_n)_n$ be a sequence in D converging to (t_0, x_0) such that

$$v(t_n, x_n) \to v_*(t_0, x_0) \,.$$

By (2.2) and $\mathbf{H}_g$, there is an open ball B centered on (t_0, x_0) such that

$$-\mathcal{L}\varphi \leq 0 \ \text{ on } B \cap \bar{D}^* \qquad \text{and} \qquad \varphi - g \leq -\varepsilon \ \text{ on } B \cap \partial_x D^* \,. \tag{6.5}$$

Obviously, we can assume that $(t_n, x_n) \in B$. Set $(\tau_n, X^n) = (\tau_{t_n,x_n} X_{t_n,x_n})$ and let θ_n be the first exit time of $(X^n(s))_{s\geq t_n}$ from B. Observing that $\xi := \min_{\partial B\cap\bar{D}}(v_* - \varphi) > 0$, using Itô's lemma and (6.5) one obtains

$$\begin{aligned}\varphi(t_n, x_n) &\leq \mathbb{E}\left[\varphi\left(\tau_n \wedge \theta_n, X^n(\tau^n \wedge \theta^n)\right)\right] \\ &\leq -(\varepsilon \wedge \xi) + \mathbb{E}\left[g\left(\tau_n, X^n(\tau^n)\right)\mathbf{1}_{\tau_n\leq\theta_n} + v\left(\theta_n, X^n(\theta^n)\right)\mathbf{1}_{\tau_n>\theta_n}\right] .\end{aligned}$$

Since $(\varphi(t_n, x_n) - v(t_n, x_n)) \to 0$ and $0 \in \tilde{K}$, this leads to a contradiction of (2.13) for n large enough.

2. We now prove that $v_*(t_0, x_0) \geq g(t_0, x_0)$ for all $(t_0, x_0) \in \partial_x D^*$. To see this, we assume to the contrary that

$$v_*(t_0, x_0) < g(t_0, x_0) \tag{6.6}$$

for some $(t_0, x_0) \in \partial_x D^*$ and work toward a contradiction of (6.3). Let $\varphi \in C^2(\bar{D}^*)$ be such that

$$0 = (v_* - \varphi)(t_0, x_0) \;=\; \min_{\bar{D}^*}(\text{strict})(v_* - \varphi) .$$

For $\varepsilon > 0$, define ϕ^ε on $\bar{D}^*$ by

$$\phi^\varepsilon(t, x) = \varphi(t, x) - \left(d(x) - \frac{d(x)^2}{\varepsilon}\right) ,$$

where d is defined in $\mathbf{H}_{\mathcal{O}}$. Since $d(x) - \frac{d(x)^2}{\varepsilon} > 0$ when $0 < d(x) < \varepsilon$, it follows that (t_0, x_0) is a strict local minimum of $(v_* - \phi^\varepsilon)$ for each $\varepsilon > 0$. By (6.3) and (6.6), we must therefore have

$$-\mathcal{L}\varphi(t_0, x_0) + \mathrm{Tr}\left[a(t_0, x_0)D^2 d(x_0)\right] - \frac{1}{\varepsilon}\mathrm{Tr}\left[a(t_0, x_0)Dd(x_0)Dd(x_0)'\right] \geq 0 ,$$

which leads to a contradiction of (2.2) when ε tends to 0.

3. In view of step 2 and the definition of $\hat{g}$ in (3.5), (3.11) concludes the proof. □

PROPOSITION 6.4. *Assume that* $\mathbf{H}_{\tilde{K}}$ *and* $\mathbf{H}_g$ *hold. Then* $v_*(T, \cdot) \geq \hat{g}(T, \cdot)$ *on* $\bar{\mathcal{O}}^*$.

Proof. Fix $x_0 \in \bar{\mathcal{O}}^*$ and let $(t_n, x_n)_n$ be a sequence in D converging to (T, x_0) such that $v(t_n, x_n) \to v_*(T, x_0)$. Set $(\tau_n, X^n) = (\tau_{t_n,x_n}, X_{t_n,x_n})$. Since σ is bounded (see (i) of (2.2)), one easily checks that $(\tau_n, X^n(\tau_n)) \to (T, x_0)$ $\mathbb{P}$-a.s., after possibly passing to a subsequence. In view of $\mathbf{H}_g$, it follows that

$$\liminf_{n\to\infty}\left(g(\tau_n, X^n(\tau_n))\mathbf{1}_{\tau_n<T} + g(T, X^n(T))\mathbf{1}_{\tau_n=T}\right) \geq g(T, x_0) .$$

Since $g \geq 0$ by assumption and $0 \in \tilde{\mathcal{K}}$, it follows from Fatou's lemma and (2.11) that $v_*(T, x_0) \geq g(T, x_0)$. The proof is concluded by using (3.11) and recalling the definition of $\hat{g}(T, \cdot)$ in (3.3). □

6.2. Subsolution property. In view of Propositions 6.1, 6.2, 6.3, and 6.4, we already know that v_* is a supersolution of $\mathcal{B}_{\mathrm{d}}\varphi = 0$ on $\bar{D}^*$. To conclude the proof of (i) of Theorem 3.1, it remains to show that v^* is a subsolution of $\mathcal{B}\varphi = 0$ on $\bar{D}^*$. This is a consequence of Propositions 6.5, 6.6, and 6.7 below.

PROPOSITION 6.5. *Let* $(t_0, x_0) \in D$ *and* $\varphi \in C^2(\bar{D}^*)$ *be such that* (t_0, x_0) *is a local maximum on* $\bar{D}^*$ *of* $v^* - \varphi$ *satisfying* $(v^* - \varphi)(t_0, x_0) = 0$. *Then*

$$\min\left\{-\mathcal{L}\varphi(t_0, x_0)\ ;\ \mathcal{H}\varphi(t_0, x_0)\right\} \leq 0 . \tag{6.7}$$

Proof. The proof is standard. We assume that

$$G(t_0,x_0) := \inf_{\rho\in\tilde{K}} \{-\mathcal{L}\varphi(t_0,x_0) + \delta(\rho)\varphi(t_0,x_0) - \rho' \mathrm{diag}\,[x_0]\, D\varphi(t_0,x_0)\} > 0, \tag{6.8}$$

and work towards a contradiction. If (6.8) holds, then it follows from (i) of (2.2) that there exists some $\alpha > 0$ such that

$$G(t,x) > 0 \quad \text{for all } (t,x) \in B_0 := B(t_0,\alpha) \times B(x_0,\alpha) \subset D\ . \tag{6.9}$$

Let $(t_n,x_n)_{n\geq 0}$ be a sequence in B_0 such that $(t_n,x_n) \to (t_0,x_0)$ and $v(t_n,x_n) \to v^*(t_0,x_0)$. Observe that $\beta_n := \varphi(t_n,x_n) - v(t_n,x_n) \to 0$. Set $X^n = X_{t_n,x_n}$ and define the stopping times

$$\theta_n := T \wedge \inf\{s \in [t_n,T] \ :\ (s,X^n(s)) \notin B_0\}\ .$$

Let $\partial_p B_0 = [t_0,t_0+\alpha]\times\partial B(x_0,\alpha) \cup \{t_0+\alpha\}\times\bar{B}(x_0,\alpha)$ denote the parabolic boundary of B_0 and observe that

$$0 \ > \ -\zeta := \sup_{(t,x)\in\partial_p B_0} (v^* - \varphi)(t,x)\ .$$

Then we deduce from Itô's lemma applied on φ, (6.9), Girsanov's theorem (see the discussion in section 2.2), and the above assertion that

$$v(t_n,x_n) + \beta_n \geq \zeta + \sup_{\vartheta\in\tilde{\mathcal{K}}} \mathbb{E}^\vartheta \left[\mathcal{E}^\vartheta_{\theta_n} v(\theta_n, X^n(\theta_n))\right]\ .$$

Since by construction $\theta_n < \tau_{t_n,x_n}$ and $\beta_n \to 0$, we obtain a contradiction of (2.13). □

Proposition 6.6. *Assume that* $\mathbf{H}_{\mathcal{O}}$ *and* $\mathbf{H}_g$ *hold. Let* $(t_0,x_0) \in \partial_x D^*$ *and* $\varphi \in C^2(\bar{D}^*)$ *be such that* (t_0,x_0) *is a local maximum on* $\bar{D}^*$ *of* $v_* - \varphi$ *satisfying* $(v_* - \varphi)(t_0,x_0) = 0$. *Then*

$$\min\{v^*(t_0,x_0) - \hat{g}(t_0,x_0)\ ;\ \mathcal{H}\varphi(t_0,x_0)\} \leq 0\ .$$

Proof.

1. By using similar arguments as in the proof of Proposition 6.5, we first obtain that

$$\min\{v^*(t_0,x_0) - g(t_0,x_0)\ ;\ -\mathcal{L}\varphi(t_0,x_0)\ ;\ \mathcal{H}\varphi(t_0,x_0)\} \leq 0\ . \tag{6.10}$$

2. We now proceed by contradiction as in step 2 of the proof of Proposition 6.3 to show that

$$\min\{v^*(t_0,x_0) - g(t_0,x_0)\ ;\ \mathcal{H}\varphi(t_0,x_0)\} \leq 0\ .$$

As usual, we can assume that (t_0,x_0) is a strict local maximum of $v^* - \varphi$ on $\bar{D}^*$. Assume that for some $\eta > 0$,

$$\min\left\{v^*(t_0,x_0) - g(t_0,x_0)\ ;\ \inf_{\rho\in\tilde{K}_1} \delta(\rho)v^*(t_0,x_0) - \rho' \mathrm{diag}\,[x_0]\, D\varphi(t_0,x_0)\right\} > \eta.$$

Let $\lambda > 0$ be a fixed parameter to be chosen later and for $\varepsilon > 0$ set on $\bar{D}^*$

$$\phi^\varepsilon(t,x) = \varphi(t,x) + \lambda d(x) - \frac{d^2(x)}{\varepsilon}\ ,$$

where d is defined in $\mathbf{H}_{\mathcal{O}}$. For $x \in \bar{O}^*$ such that $d(x) < \varepsilon\lambda$ we have $\lambda d(x) - \frac{d^2(x)}{\varepsilon} \geq 0$. It follows that (t_0, x_0) is a local maximum of $v^* - \phi^\varepsilon$. Moreover,

$$\min\left\{ v^*(t_0,x_0) - g(t_0,x_0) \; ; \; \inf_{\rho\in\tilde{K}_1} \delta(\rho)v^*(t_0,x_0) - \rho' \mathrm{diag}\,[x_0]\, D\phi^\varepsilon(t_0,x_0) \right\} > 0,$$

for $\lambda > 0$ small enough since $d(x_0) = 0$ and therefore $D\phi^\varepsilon(t_0,x_0) = D\varphi(t_0,x_0) + \lambda Dd(x_0) - 2Dd(x_0)d(x_0)/\varepsilon = D\varphi(t_0,x_0) + \lambda Dd(x_0)$. Thus, it follows from step 1 that we must have

$$-\mathcal{L}(\varphi(t_0,x_0) + \lambda d(x_0)) + \frac{1}{\varepsilon}\mathrm{Tr}\left[a(t_0,x_0)Dd(x_0)Dd(x_0)'\right] \leq 0 \; .$$

Sending $\varepsilon \to 0$ leads to a contradiction of (i) of (2.2). □

PROPOSITION 6.7. *Assume that* $\mathbf{H}_g$ *holds. Then* $v^*(T,\cdot) \leq \hat{g}(T,\cdot)$ *on* $\bar{\mathcal{O}}^*$.

Proof.

1. Let $(t_n, x_n)_n$ be a sequence in D which converges to (T, x_0) and such that $v(t_n,x_n) \to v^*(T,x_0)$. Set $(\tau_n, X^n) = (\tau_{t_n,x_n}, X_{t_n,x_n})$. By the dual formulation (2.11), there is some $\vartheta^n \in \tilde{\mathcal{K}}$ such that

$$v(t_n,x_n) \leq \mathbb{E}^{\vartheta^n}\left[e^{-\int_{t_n}^{\tau_n}\delta(\vartheta_s^n)ds} g(\tau_n, X^n(\tau_n)) \right] + n^{-1} \; .$$

Since $\tilde{K}$ is a convex cone, δ is sublinear, and $g \geq 0$, it follows that

$$\begin{aligned} e^{-\int_{t_n}^{\tau_n}\delta(\vartheta_s^n)ds} g(\tau_n, X^n(\tau_n)) &\leq e^{-\delta(\int_{t_n}^{\tau_n}\vartheta_s^n ds)} g(\tau_n, X^n(\tau_n)) \\ &\leq \sup_{t_n\leq t\leq T} \hat{g}\left(t, X^n(t) e^{-\int_{t_n}^t \vartheta_s^n ds}\right) \end{aligned}$$

by definition of $\hat{g}$ in (3.3)–(3.5). In view of the above inequalities and the definition of (t_n, x_n), it remains to show that

$$\limsup_{n\to\infty} \mathbb{E}^{\vartheta^n}\left[\sup_{t_n\leq t\leq T} \hat{g}\left(t, Z^n(t)\right) \right] \leq \hat{g}(T,x_0) \; , \tag{6.11}$$

where $Z^n := X^n e^{-\int_{t_n}^{\cdot}\vartheta_s^n ds}$ solves on $[t_n, T]$

$$dZ^n(t) = \mathrm{diag}\,[Z^n(t)]\,\sigma(t,X^n(t))dW_t^n \; , \; Z^n(t_n) = x_n \; ,$$

and W^n is a standard Brownian motion under $\mathbb{Q}^{\vartheta^n}$, recall the discussion of section 2.2. Using the boundedness assumption on σ (see (2.2)), we deduce from standard arguments that there is a constant $C > 0$ independent of n such that

$$\mathbb{E}^{\vartheta^n}\left[\sup_{t_n\leq t\leq T} |Z^n(t) - x_0| \right] \leq C\left(|x_n - x_0| + (T-t_n)^{1/2} \right) \; .$$

We shall prove in step 2 that for each $\varepsilon > 0$ there is a Lipschitz function Ψ_ε such that $|\hat{g}(T,x_0) - \Psi_\varepsilon(T,x_0)| \leq \varepsilon$ and $\Psi_\varepsilon \geq \hat{g}$. It follows that, for each ε, we can find some finite $K_\varepsilon > 0$ such that

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{E}^{\vartheta^n}\left[\sup_{t_n\leq t\leq T} \hat{g}\left(t, Z^n(t)\right) \right] &\leq \limsup_{n\to\infty} \mathbb{E}^{\vartheta^n}\left[\sup_{t_n\leq t\leq T} \Psi_\varepsilon\left(t, Z^n(t)\right) \right] \\ &\leq \Psi_\varepsilon(T,x_0) + \limsup_{n\to\infty} K_\varepsilon\left(|x_n - x_0| + (T-t_n)^{1/2} \right) \\ &= \Psi_\varepsilon(T,x_0) \; . \end{aligned}$$

By definition of Ψ_ε this implies that

$$\limsup_{n\to\infty} \mathbb{E}^{\vartheta^n}\left[\sup_{t_n\le t\le T} \hat{g}\left(t, Z^n(t)\right)\right] \le \hat{g}(T,x_0)+\varepsilon\ ,$$

and the proof of (6.11) is concluded by sending ε to 0.

2. We conclude this proof by constructing the sequence of functions $(\Psi_\varepsilon)_{\varepsilon>0}$. For $(t,x)\in[0,T]\times(0,\infty)^d$, we define

$$G_k(t,x) = \sup_{(s,z)\in[0,T]\times(0,\infty)^d} \left[\hat{g}(s,z) - k\left(|s-t|+|z-x|\right)\right],\ k\ge 1\ .$$

Clearly, $G_k\ge\hat{g}$ and G_k is k-Lipschitz. Moreover, taking k large enough, it follows from the linear growth and upper-semicontinuity assumptions on $\hat{g}$ (see $\mathbf{H}_g$) that for all $(t,x)\in[0,T]\times(0,\infty)^d$, the maximum is attained in the above definition by some $(t_k(t,x),x_k(t,x))$. In particular,

$$G_k(t,x) = \hat{g}(t_k(t,x),x_k(t,x)) - k\left(|t_k(t,x)-t|+|x_k(t,x)-x|\right) \ge \hat{g}(t,x)\ .$$

Using the linear growth of $\hat{g}$ again, we deduce that $(t_k(t,x),x_k(t,x))\to(t,x)$ as $k\to\infty$ after possibly passing to a subsequence. Since $\hat{g}$ is upper-semicontinuous, this also implies that

$$\hat{g}(T,x_0)\ge \limsup_{k\to\infty}\hat{g}(t_k(T,x_0),x_k(T,x_0)) \ \ge\ \limsup_{k\to\infty} G_k(T,x_0)\ge \hat{g}(T,x_0)\ .$$

We can then choose k_ε such that $|G_{k_\varepsilon}(T,x_0)-\hat{g}(T,x_0)|\le\varepsilon$ and set $\Psi_\varepsilon := G_{k_\varepsilon}$. □

6.3. Characterization of v_* as the smallest supersolution. In this section, we prove that $v=v_*$ on D and that v_* is the smallest supersolution of (3.8).

To this end, we introduce a sequence of approximating control problems as follows. For all $\eta\ge1$, we define $\tilde{K}_\eta$ as the set of elements $\rho\in\tilde{K}$ such that $|\rho|\le\eta$ and $\tilde{\mathcal{K}}_\eta$ as the set of elements $\vartheta\in\tilde{\mathcal{K}}$ that take values in $\tilde{K}_\eta$. We then define on $\bar{D}^*$

$$w_\eta(t,x) = \sup_{\vartheta\in\tilde{\mathcal{K}}_\eta} \mathbb{E}^\vartheta\left[\mathcal{E}^\vartheta_\tau g\left(\tau_{t,x}, X_{t,x}(\tau_{t,x})\right)\right]\ . \tag{6.12}$$

It is clear that w_η is a nondecreasing sequence, and it follows directly from Theorem 2.1 and the definition of $\tilde{\mathcal{K}}$ that

$$\lim_{\eta\to\infty}\uparrow w_\eta(t,x) = v(t,x) \quad \text{for all}\ \ (t,x)\in\bar{D}^*\ . \tag{6.13}$$

For $\eta\ge1$, let us introduce the operator $\mathcal{G}_\eta$ defined for smooth functions by

$$\mathcal{G}_\eta\varphi(t,x) := \min_{\rho\in\tilde{K}_\eta}\left\{-\mathcal{L}\varphi(t,x)+\delta(\rho)\varphi(t,x)-\rho'\mathrm{diag}\,[x]\,D\varphi(t,x)\right\}\ .$$

PROPOSITION 6.8. *Let the conditions $\mathbf{H}_g$ and $\mathbf{H}_\mathcal{O}$ hold. Then, for all $\eta\ge1$, w^*_η is a viscosity subsolution on D of*

$$\mathcal{G}_\eta\varphi(t_0,x_0)=0\ . \tag{6.14}$$

Moreover, $w^*_\eta\le\hat{g}$ on $\partial_x D^*\cup\partial_T D^*$.

Proof. The proof is standard. Set $\varphi\in C^2(D)$ and let (t_0,x_0) be a strict global maximizer of $w^*_\eta-\varphi$ on $\bar{D}^*$ such that $(w^*_\eta-\varphi)(t_0,x_0)=0$.

1. If $(t_0, x_0) \in D$, then the result follows from the same arguments as in the proof of Proposition 6.5.
2. Arguing as in Proposition 6.5 again, we deduce that

$$\min\left\{w_\eta^*(t_0,x_0) - \hat{g}(t_0,x_0)\ ,\ \mathcal{G}_\eta\varphi(t_0,x_0)\right\} \le 0$$

if $(t_0, x_0) \in \partial_x D^*$. The required result is then obtained by arguing as in step 2 of the proof of Proposition 6.6.
3. Since $w_\eta^* \le v^*$, the inequality $w_\eta^*(T,\cdot) \le \hat{g}(T,\cdot)$ follows from Proposition 6.7. □

PROPOSITION 6.9. *Assume that* $\mathbf{H}_g$ *holds. Let* u *(resp.,* w*) be a viscosity subsolution (resp., supersolution) of* (6.14) *on* D *satisfying the growth condition* (3.10)*. If* $u \le w$ *on* $\partial_x D^* \cup \partial_T D^*$*, then* $u \le w$ *on* $\bar{D}^*$*.*

Proof.

1. Given $\kappa > 0$, we set $\tilde{u}(t,x) = e^{\kappa t}u(t,x)$ and $\tilde{w}(t,x) = e^{\kappa t}w(t,x)$ so that $\tilde{u}$ and $\tilde{v}$ are, respectively, sub- and supersolutions of

$$\tilde{\mathcal{G}}_\eta\varphi(t,x) := \min_{\rho\in\tilde{K}_\eta}\left\{(\kappa+\delta(\rho))\varphi(t,x) - \mathcal{L}^\rho\varphi(t,x)\right\} \ =\ 0\ ,$$

where for $\rho \in \tilde{K}$

$$\mathcal{L}^\rho\varphi(t,x) := \mathcal{L}\varphi(t,x) + \rho'\mathrm{diag}\,[x]\,D\varphi(t,x)\ .$$

Recall the definition of $\bar{\gamma}$ in $\mathbf{H}_g$ and set

$$\gamma = 2\bar{\gamma} \in \mathbb{R}^d_+, \tilde{\gamma} = (2,\dots,2) \in (0,\infty)^d\ . \tag{6.15}$$

Define on $\bar{D}^*$

$$\beta(t,x) := e^{\tau(T-t)}\left(1 + x^\gamma + x^{\tilde{\gamma}}\right).$$

Observing that

$$\frac{\partial}{\partial t}\beta(t,x) = -\tau\beta(t,x), \mathrm{diag}\,[x]\,D\beta(t,x) = e^{\tau(T-t)}\left(x^\gamma\gamma + x^{\tilde{\gamma}}\tilde{\gamma}\right)\ ,$$
$$\mathrm{Tr}\left[a(t,x)D^2\beta(t,x)\right] = e^{\tau(T-t)}\left(x^\gamma\mathrm{Tr}\left[\sigma\sigma'(t,x)M\right] + x^{\tilde{\gamma}}\mathrm{Tr}\left[\sigma\sigma'(t,x)\tilde{M}\right]\right)$$

with $M := [\gamma^i(\gamma^i - 1)\mathbf{1}_{i=j} + \gamma^i\gamma^j\mathbf{1}_{i\neq j}]_{ij}$ and $\tilde{M}$ defined similarly, it follows from (i) of (2.2) and the compactness of $\tilde{K}_\eta$ that we can find $\tau > 0$ such that

$$\tilde{\mathcal{G}}_\eta\beta(t,x) \ge 0 \quad \text{on } \bar{D}^*\ . \tag{6.16}$$

2. We now argue by contradiction and assume that

$$\sup_{\bar{D}^*}(\tilde{u} - \tilde{w}) > 0\ .$$

2.1. In view of the growth condition on $\tilde{u}$, $\tilde{w}$ and (6.15), we then have

$$0 < 2m := \sup_{\bar{D}^*}(\tilde{u} - \tilde{w} - 2\alpha\beta) < \infty \tag{6.17}$$

for $\alpha > 0$ small enough. For $x \in \bar{D}^*$, set

$$f(x) = \sum_{i=1}^{d} (x^i)^{-2} \,. \tag{6.18}$$

Combining the growth condition on $\tilde{u}$, $\tilde{w}$ with (6.15) and the definition of f implies that, for each $\varepsilon > 0$, the upper-semicontinuous function

$$\Phi^\varepsilon := \tilde{u} - \tilde{w} - 2(\alpha\beta + \varepsilon f)$$

admits a maximum $(t_\varepsilon, x_\varepsilon)$ on $\bar{D}^*$. By (6.17), we can choose ε small enough so that

$$\Phi^\varepsilon(t_\varepsilon, x_\varepsilon) \geq m > 0 \,. \tag{6.19}$$

Let $(t_0^\varepsilon, x_0^\varepsilon)$ be a sequence in D such that $\Phi^\varepsilon(t_0^\varepsilon, x_0^\varepsilon) \to 2m$. By (6.17) and the definition of $(t_\varepsilon, x_\varepsilon)$, we have

$$\begin{aligned}
\liminf_{\varepsilon\to 0} \; (2m - 2\varepsilon f(x_\varepsilon)) &\geq \liminf_{\varepsilon\to 0} \; (\tilde{u} - \tilde{w} - 2(\alpha\beta + \varepsilon f))(t_\varepsilon, x_\varepsilon) \\
&\geq \lim_{\varepsilon\to 0} \; (\tilde{u} - \tilde{w} - 2(\alpha\beta + \varepsilon f))(t_0^\varepsilon, x_0^\varepsilon) \\
&= 2m \,.
\end{aligned}$$

This shows that

$$\limsup_{\varepsilon\to 0} \; \varepsilon f(x_\varepsilon) = \limsup_{\varepsilon\to 0} \; \varepsilon \sum_{i=1}^{d} (x_\varepsilon^i)^{-2} = 0 \,,$$

which, by (i) of (2.2) and the compactness of $\tilde{K}_\eta$, implies

$$\limsup_{\varepsilon\to 0} \; \sup_{\rho \in \tilde{K}_\eta} \; \varepsilon \left(|f(x_\varepsilon)| + |\mathcal{L}^\rho f(x_\varepsilon)| \right) = 0 \,. \tag{6.20}$$

2.2. For $(t, x) \in [0, T] \times \mathbb{R}^d$, set $G^\varepsilon(t, x) = |t - t_\varepsilon|^4 + |x - x_\varepsilon|^4$. Given $n > 0$, it follows from similar arguments to those above that the map

$$\begin{aligned}
\Phi_n^\varepsilon(t, x, y) := \tilde{u}(t, x) - \tilde{w}(t, y) - \frac{n}{2}|x - y|^2 - \alpha \left(\beta(t, x) + \beta(t, y) \right) \\
- \; \varepsilon \left(f(x) + f(y) + G^\varepsilon(t, x) \right)
\end{aligned}$$

also admits a maximum point $(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) \in \bar{D}^*$ which necessarily satisfies

$$\Phi_n^\varepsilon(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) \geq \Phi_n^\varepsilon(t_\varepsilon, x_\varepsilon, x_\varepsilon) = \Phi^\varepsilon(t_\varepsilon, x_\varepsilon) \geq m > 0 \,. \tag{6.21}$$

Using the growth assumption on u and w, (6.15), and the definition of f again, one easily checks that this implies that the sequence $(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon)_n$ is bounded and therefore converges, after possibly passing to a subsequence. Moreover, (6.21) implies that $n|x_n^\varepsilon - y_n^\varepsilon|^2 + \varepsilon f(x_n^\varepsilon)$ is bounded. Thus, there is $(\bar{t}_\varepsilon, \bar{x}_\varepsilon) \in \bar{D}^*$ such that $(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) \to (\bar{t}_\varepsilon, \bar{x}_\varepsilon)$ and, by the definition of $(t_\varepsilon, x_\varepsilon)$ and (6.21), we must have

$$\begin{aligned}
\Phi^\varepsilon(t_\varepsilon, x_\varepsilon) &\geq \Phi^\varepsilon(\bar{t}_\varepsilon, \bar{x}_\varepsilon) \\
&\geq \limsup_{n\to\infty} \left(\Phi^\varepsilon(\bar{t}_\varepsilon, \bar{x}_\varepsilon) - \frac{n}{2}|x_n^\varepsilon - y_n^\varepsilon|^2 - \varepsilon G^\varepsilon(t_n^\varepsilon, x_n^\varepsilon) \right) \\
&\geq \Phi^\varepsilon(t_\varepsilon, x_\varepsilon) \,.
\end{aligned}$$

This shows that, up to a subsequence,

$$(t_n^\varepsilon, x_n^\varepsilon) \to (t_\varepsilon, x_\varepsilon) \in \bar{D}^* \ , \ \Phi_n^\varepsilon(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) \to \Phi^\varepsilon(t_\varepsilon, x_\varepsilon) \quad \text{and} \quad n|x_n^\varepsilon - y_n^\varepsilon|^2 \to 0 \tag{6.22}$$

as $n \to \infty$.

3. Since the upper-semicontinuous function $u - w$ is nonpositive on $\partial_x D^* \cup \partial_T D^*$, it follows from (6.19) that $(t_\varepsilon, x_\varepsilon) \in D$ and that we may assume that $(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) \in [0,T) \times \mathcal{O}^2$ for each $n > 0$. Using Ishii's lemma and following standard arguments (see [5, Theorem 8.3] and the discussion after [5, Theorem 3.2]) we deduce from the viscosity property of $\tilde{u}$ and $\tilde{w}$ that for some $\hat{\rho}_n^\varepsilon$ in the compact set $\tilde{K}_\eta$

$$\begin{aligned} 0 \le\ & (\kappa + \delta(\hat{\rho}_n^\varepsilon))(\tilde{w}(t_n^\varepsilon, y_n^\varepsilon) - \tilde{u}(t_n^\varepsilon, x_n^\varepsilon)) + \frac{1}{2}\mathrm{Tr}\,[a(t_n^\varepsilon, x_n^\varepsilon)A_n^\varepsilon - a(t_n^\varepsilon, y_n^\varepsilon)B_n^\varepsilon] \\ & + (\hat{\rho}_n^\varepsilon)'\mathrm{diag}\,[x_n^\varepsilon - y_n^\varepsilon]\, q_n^\varepsilon \\ & + \left\{ \mathcal{L}^{\hat{\rho}_n^\varepsilon}(\alpha\beta + \varepsilon[f + G^\varepsilon])(t_n^\varepsilon, x_n^\varepsilon) + \mathcal{L}^{\hat{\rho}_n^\varepsilon}(\alpha\beta + \varepsilon f)(t_n^\varepsilon, y_n^\varepsilon) \right\} , \end{aligned}$$

where

$$q_n^\varepsilon := n(x_n^\varepsilon - y_n^\varepsilon) \tag{6.23}$$

and A_n^ε, B_n^ε are two symmetric matrices satisfying

$$-3n \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix} \le \begin{pmatrix} A_n^\varepsilon & 0 \\ 0 & -B_n^\varepsilon \end{pmatrix} \le 3n \begin{pmatrix} I_d & -I_d \\ -I_d & I_d \end{pmatrix} . \tag{6.24}$$

Using (6.16), (6.21), (6.22), (6.23), (6.24), and (i) of (2.2), we then deduce that

$$\begin{aligned} 0 \le\ & -m(\kappa + \delta(\hat{\rho}_n^\varepsilon)) + C\, n|x_n^\varepsilon - y_n^\varepsilon|^2 - \varepsilon(\kappa + \delta(\hat{\rho}_n^\varepsilon))\, \{(f + G^\varepsilon)(t_n^\varepsilon, x_n^\varepsilon) \\ & +\ f(y_n^\varepsilon)\} + \varepsilon \left\{ \mathcal{L}^{\hat{\rho}_n^\varepsilon}(f + G^\varepsilon)(t_n^\varepsilon, x_n^\varepsilon) + \mathcal{L}^{\hat{\rho}_n^\varepsilon} f(y_n^\varepsilon) \right\} \end{aligned}$$

for some $C > 0$ independent of n. Sending n to ∞, it follows from the compactness of $\tilde{K}_\eta$ and (6.22) that

$$0 \le -m(\kappa + \delta(\hat{\rho}^\varepsilon)) + 2\varepsilon \left\{ \mathcal{L}^{\hat{\rho}^\varepsilon} f(x_\varepsilon) - (\kappa + \delta(\hat{\rho}^\varepsilon)) f(x_\varepsilon) \right\}$$

for some $\hat{\rho}^\varepsilon \in \tilde{K}_\eta$. Sending ε to 0 and using (6.20) finally leads to a contradiction since $\kappa, m > 0$ and $\delta \ge 0$ by (2.9). □

We now conclude the proof of Theorem 3.1.

Proof of (ii) *and* (iii) *of Theorem* 3.1. Observe that a supersolution u of $\mathcal{B}_{\mathrm{d}}\varphi = 0$ on $\bar{D}^*$ is also a supersolution of (6.14) on D, and, by Proposition 6.8, satisfies $u \ge w_\eta$ on $\partial_x D^* \cup \partial_T D^*$ for all $\eta \ge 1$. In view of Proposition 6.9 and (6.13), it follows that $u \ge \lim_{\eta\to\infty} \uparrow w_\eta = v$ on D whenever u satisfies (3.10). In particular, since v_* is a supersolution of (3.8) satisfying (3.10) (see Proposition 3.1) we have $v_* \ge v$ so that $v_* = v \le u$ on D and $v_* \le u_*$ on $\bar{D}^*$. □

7. A uniqueness result. We now proceed with the proof of Theorem 3.2. It is an immediate consequence of Proposition 3.1, Theorem 3.1, and the following comparison result.

THEOREM 7.1. *Assume that the conditions of Theorem* 3.2 *hold. Let u be an upper-semicontinuous viscosity subsolution of $\mathcal{B}\varphi = 0$ on $\bar{D}^*$. Assume furthermore that u satisfies the growth condition* (3.10). *Then $u \le v_*$ on $\bar{D}^*$.*

Remark 7.1.

1. It will be clear from the proof that the above theorem can be stated as follows. Let u and w be, respectively, the sub- and supersolution of $\mathcal{B}\varphi = 0$ and $\mathcal{B}_{\rm d}\varphi = 0$ on $D \cup \partial_x D^*$ satisfying the growth condition (3.10). Assume further that w satisfies

   C: For all $(t,x) \in D \cup \partial_x D^*$ and $\rho \in \tilde{K}_1(x, \bar{\mathcal{O}})$ there exists $\lambda_0 > 0$ s.t. $\lambda \in [0, \lambda_0] \mapsto w(t, xe^{\lambda\rho})e^{-\lambda\delta(\rho)}$

   is nonincreasing.

   Then $u \le w$ on $\partial_T D^*$ implies $u \le w$ on $\bar{D}^*$.
2. One can actually show that any supersolution of $\mathcal{H}_{\rm d}\varphi = 0$ satisfies condition C. Since it is not useful for our main result, we do not provide the proof, which is rather long.
3. Combining the above assertions provides a general comparison result for super- and subsolutions of, respectively, $\mathcal{B}_{\rm d}\varphi = 0$ and $\mathcal{B}\varphi = 0$ on $D \cup \partial_x D^*$.

In order to prove Theorem 7.1, we need the following intermediate lemma.

LEMMA 7.1. *Assume that $\mathbf{H}_{\mathcal{O}}$ holds. Fix $x_0 \in \partial\mathcal{O}^*$. If $\rho \in \tilde{K}_1$ satisfies*

$$Dd(x_0)' \operatorname{diag}[x_0]\, \rho \;>\; 0\,,$$

then there exists some positive parameters r_0 and λ_0 such that $xe^{\lambda\rho} \in \mathcal{O}$ for all $x \in B(x_0, r_0) \cap \bar{\mathcal{O}}$ and $\lambda \in (0, \lambda_0)$.

Proof. Recall from $\mathbf{H}_{\mathcal{O}}$ that the function d is C^2 on a neighborhood of x_0. Thus, $Dd(x_0)' \operatorname{diag}[x_0]\, \rho > 0$ implies that for some $\delta_0, r_0 > 0$

$$Dd(\bar{x})' \operatorname{diag}[x]\, \rho \;\ge\; \delta_0 \quad \text{for all } \bar{x}, x \in B(x_0, r_0)\,. \tag{7.1}$$

Given that $xe^{\lambda\rho} - x = \lambda \operatorname{diag}[x]\, \rho + o(\lambda)$, we can fix some $\lambda_0 > 0$ such that, for all $x \in B(x_0, r_0/2)$ and $\lambda \in (0, \lambda_0)$,

$$[x, xe^{\lambda\rho}] \subset B(x_0, r_0) \quad \text{and} \quad |xe^{\lambda\rho} - x - \lambda \operatorname{diag}[x]\, \rho| < \frac{\lambda\delta_0/2}{1 + \max\limits_{x \in \bar{B}(x_0, r_0)} |Dd(x)|}. \tag{7.2}$$

Let x be in $B(x_0, r_0/2) \cap \bar{\mathcal{O}}$, so that $d(x) \ge 0$. Since d is C^1, for each $\lambda \in (0, \lambda_0)$ there exists $\bar{x} \in [x, xe^{\lambda\rho}] \subset B(x_0, r_0)$ such that

$$\begin{aligned} d(xe^{\lambda\rho}) &= d(x) + Dd(\bar{x})' \left(xe^{\lambda\rho} - x\right) \\ &= d(x) + \lambda Dd(\bar{x})' \operatorname{diag}[x]\, \rho + Dd(\bar{x})' \left[xe^{\lambda\rho} - x - \lambda \operatorname{diag}[x]\, \rho\right] \\ &\ge d(x) + \lambda \frac{\delta_0}{2} \;>\; 0\,, \end{aligned}$$

where the last inequality follows from (7.1) and (7.2). This shows that $xe^{\lambda\rho} \in \mathcal{O}$. □

Proof of Theorem 7.1. In order to avoid too many complications, we make the proof under the following assumption in place of $\mathbf{H}'$:

$\mathbf{H}''$: (i) There exists $\varrho > 1$ s.t. $\varrho\bar{\gamma} \in K \cap (0, \infty)^d$.
(ii) $0 \in \operatorname{int}(K)$.
(iii) For all $x \in \partial\mathcal{O}^*$ there exists $\rho \in \tilde{K}_1$ s.t. $Dd(x)' \operatorname{diag}[x]\, \rho > 0$.

We shall explain in the last step how to adapt this proof when $\bar{\mathcal{O}}$ is bounded but (i) of $\mathbf{H}''$ does not hold, or $0 \notin \mathrm{int}(K)$ but $\bar{\mathcal{O}} \cap \partial\mathbb{R}^d_+ = \emptyset$ and $\mathrm{int}(K) \neq \emptyset$.

1. Given some positive parameter κ, we introduce the functions $\tilde{u}(t,x) := e^{\kappa t} u(t,x)$, $\tilde{v}(t,x) := e^{\kappa t} v_*(t,x)$, and $\tilde{g}(t,x) := e^{\kappa t}\hat{g}(t,x)$. One easily checks that the function $\tilde{u}$ (resp., $\tilde{v}$) is a viscosity subsolution (resp., supersolution) of $\tilde{\mathcal{B}}\varphi = 0$ (resp., $\tilde{\mathcal{B}}_{\mathrm{d}}\varphi = 0$), where for a smooth function φ

$$\tilde{\mathcal{B}}\varphi = \begin{cases} \min\left\{\tilde{\mathcal{L}}\varphi\ ,\ \mathcal{H}\varphi\right\} & \text{on} \quad D\ , \\ \min\left\{\varphi - \tilde{g}\ ,\ \mathcal{H}\varphi\right\} & \text{on} \quad \partial_x D^*\ , \\ \varphi - \tilde{g} & \text{on} \quad \partial_T D^*\ , \end{cases}$$

$$\tilde{\mathcal{B}}_{\mathrm{d}}\varphi = \begin{cases} \tilde{\mathcal{B}}\varphi & \text{on} \quad D \cup \partial_T D^*\ , \\ \min\left\{\varphi - \tilde{g}\ ,\ \mathcal{H}_{\mathrm{d}}\varphi\right\} & \text{on} \quad \partial_x D^*\ , \end{cases}$$

and

$$\tilde{\mathcal{L}}\varphi := \kappa\varphi - \mathcal{L}\varphi\ .$$

Let $\varrho \in \mathbb{R}$ be as in $\mathbf{H}''$, i.e.,

$$\gamma := \varrho\bar{\gamma} \in K \cap (0,\infty)^d \ \text{ and } \ \varrho > 1\ . \tag{7.3}$$

Since $0 \in \mathrm{int}(K)$ by $\mathbf{H}''$, it follows from (2.9) and Remark 2.4 that the map defined by $\beta(t,x) = e^{\tau(T-t)}(1 + x^\gamma)$ on $\bar{D}^*$ satisfies

$$H\left(\beta(t,x), \mathrm{diag}\,[x]\, D\beta(t,x)\right) \geq c_K > 0 \quad \text{for all } (t,x) \in \bar{D}^*\ . \tag{7.4}$$

Moreover, by the same computations as in the proof of Proposition 6.9, one easily checks that we can choose τ large enough so that

$$\tilde{\mathcal{L}}\beta \geq 0 \quad \text{on } \bar{D}^*\ . \tag{7.5}$$

2. We argue by contradiction. We assume that

$$\sup_{\bar{D}^*}(u - v_*) > 0$$

and work towards a contradiction.

2.1. In this step, we follow the same construction as in the proof of Proposition 6.9. By the growth condition on $\tilde{u}$, $\tilde{v}$ and (7.3), we deduce that

$$0 < 2m := \sup_{\bar{D}^*}(\tilde{u} - \tilde{v} - 2\alpha\beta) < \infty \tag{7.6}$$

for $\alpha > 0$ small enough. Fix $\varepsilon > 0$ and let f be defined as in (6.18). Arguing as in the proof of Proposition 6.9, we obtain that

$$\Phi^\varepsilon := \tilde{u} - \tilde{v} - 2(\alpha\beta + \varepsilon f)$$

admits a maximum $(t_\varepsilon, x_\varepsilon)$ on $\bar{D}^*$, which, for $\varepsilon > 0$ small enough, satisfies

$$\Phi^\varepsilon(t_\varepsilon, x_\varepsilon) \geq m > 0\ . \tag{7.7}$$

Moreover, using the same arguments as in step 2.1 of the proof of Proposition 6.9, we obtain that

$$\limsup_{\varepsilon \to 0} \sup_{\rho \in \tilde{K}_1} \varepsilon\left(|f(x_\varepsilon)| + |\mathrm{diag}\,[x_\varepsilon]\, Df(x_\varepsilon)| + |\mathcal{L}f(x_\varepsilon)|\right) = 0\ . \tag{7.8}$$

Finally, since $\beta, f \geq 0$ and $v_*(T,\cdot) \geq u(T,\cdot)$, (7.7) implies that $t_\varepsilon < T$, i.e.,

$$(t_\varepsilon, x_\varepsilon) \in [0,T) \times \bar{\mathcal{O}}^* \,. \tag{7.9}$$

2.2. In the following, we fix $\rho_\varepsilon \in \tilde{K}_1$ such that

$$\begin{array}{rll} \rho_\varepsilon := 0 & \text{if} & x_\varepsilon \in \mathcal{O}\,, \\ Dd(x_\varepsilon)' \operatorname{diag}[x_\varepsilon]\, \rho_\varepsilon > 0 & \text{if} & x_\varepsilon \in \partial\mathcal{O}^*\,; \end{array} \tag{7.10}$$

see (iii) of $\mathbf{H}''$. By Lemma 7.1 and (3.11), we can fix r_ε, $\lambda_\varepsilon > 0$, such that

$$x e^{\lambda \rho_\varepsilon} \in \mathcal{O} \qquad \text{and} \qquad e^{-\lambda \delta(\rho_\varepsilon)} \tilde{v}(t, x e^{\lambda \rho_\varepsilon}) \leq \tilde{v}(t,x) \tag{7.11}$$

for all $t \in (t_\varepsilon - r_\varepsilon, t_\varepsilon + r_\varepsilon) \cap [0,T)$, $x \in B_\varepsilon := B(x_\varepsilon, r_\varepsilon) \cap \bar{\mathcal{O}}$, and $\lambda \in (0, \lambda_\varepsilon)$. For $n \geq 1$ and $\zeta \in (0,1)$, we then define the function $\Psi^\varepsilon_{n,\zeta}$ on $[0,T] \times (\bar{\mathcal{O}}^*)^2$ by

$$\begin{aligned} \Psi^\varepsilon_{n,\zeta}(t,x,y) := \Theta(t,x,y) - \varepsilon(f(x) + f(y)) - \zeta(|x - x_\varepsilon|^2 + |t - t_\varepsilon|^2) \\ - n^2 |x e^{\frac{\zeta}{n}\rho_\varepsilon} - y|^2, \end{aligned}$$

where

$$\Theta(t,x,y) := \tilde{u}(t,x) - \tilde{v}(t,y) - \alpha\left(\beta(t,x) + \beta(t,y)\right) \,.$$

It follows from (7.3) and the growth condition (3.10) satisfied by $\tilde{v}$ and $\tilde{u}$ that $\Psi^\varepsilon_{n,\zeta}$ attains its maximum at some $(t^\varepsilon_n,\, x^\varepsilon_n,\, y^\varepsilon_n) \in [0,T] \times (\bar{\mathcal{O}}^*)^2$. The inequality $\Psi^\varepsilon_{n,\zeta}(t^\varepsilon_n, x^\varepsilon_n, y^\varepsilon_n) \geq \Psi^\varepsilon_{n,\zeta}(t_\varepsilon, x_\varepsilon, x_\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon})$ implies that

$$\begin{aligned} \Theta(t^\varepsilon_n, x^\varepsilon_n, y^\varepsilon_n) \geq \Theta(t_\varepsilon, x_\varepsilon, x_\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}) - \varepsilon\left(f(x_\varepsilon) + f(x_\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon})\right) \\ + n^2 |x^\varepsilon_n e^{\frac{\zeta}{n}\rho_\varepsilon} - y^\varepsilon_n|^2 + \zeta\left(|x^\varepsilon_n - x^\varepsilon|^2 + |t^\varepsilon_n - t_\varepsilon|^2\right) \\ + \varepsilon\left(f(x^\varepsilon_n) + f(y^\varepsilon_n)\right), \end{aligned}$$

which combined with the growth condition (3.10) and (7.3) shows that $n^2 |x^\varepsilon_n$ $e^{\frac{\zeta}{n}\rho_\varepsilon} - y^\varepsilon_n|^2 + f(x^\varepsilon_n)$ is bounded in n so that, up to a subsequence,

$$\text{(i)}\ x^\varepsilon_n e^{\frac{\zeta}{n}\rho_\varepsilon}, x^\varepsilon_n, y^\varepsilon_n \xrightarrow[n\to\infty]{} \bar{x}^\varepsilon \in \bar{\mathcal{O}}^* \qquad \text{and} \qquad t^\varepsilon_n \xrightarrow[n\to\infty]{} \bar{t}^\varepsilon \in [0,T]\,.$$

Let n be large enough so that $\frac{\zeta}{n} < \lambda_\varepsilon$. Recall from (7.11) that this implies that $\tilde{v}(t_\varepsilon, x_\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}) \leq \tilde{v}(t_\varepsilon, x_\varepsilon) e^{\frac{\zeta}{n}\delta(\rho_\varepsilon)}$, which combined with the previous inequality yields

$$\begin{aligned} \Theta(t^\varepsilon_n, x^\varepsilon_n, y^\varepsilon_n) \geq \tilde{u}(t_\varepsilon, x_\varepsilon) - \tilde{v}(t_\varepsilon, x_\varepsilon) e^{\frac{\zeta}{n}\delta(\rho_\varepsilon)} - \alpha\left(\beta(t_\varepsilon, x_\varepsilon) + \beta(t_\varepsilon, x_\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon})\right) \\ - \varepsilon\left(f(x_\varepsilon) + f(x_\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon})\right) + n^2 |x^\varepsilon_n e^{\frac{\zeta}{n}\rho_\varepsilon} - y^\varepsilon_n|^2 \\ + \zeta\left(|x^\varepsilon_n - x_\varepsilon|^2 + |t^\varepsilon_n - t_\varepsilon|^2\right) + \varepsilon\left(f(x^\varepsilon_n) + f(y^\varepsilon_n)\right) \,. \end{aligned}$$

Sending $n \to \infty$ and using the maximum property of $(t_\varepsilon, x_\varepsilon)$, we get

$$\begin{aligned} 0 &\geq \Phi^\varepsilon(\bar{t}^\varepsilon, \bar{x}^\varepsilon) - \Phi^\varepsilon(t_\varepsilon, x_\varepsilon) \\ &\geq \limsup_{n\to\infty} \left(n^2 |x^\varepsilon_n e^{\frac{\zeta}{n}\rho_\varepsilon} - y^\varepsilon_n|^2 + \zeta\left(|x^\varepsilon_n - x_\varepsilon|^2 + |t^\varepsilon_n - t_\varepsilon|^2\right)\right) \,. \end{aligned}$$

Recalling (7.7) and (7.9), this shows that

(ii) $n^2|x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon|^2 + \zeta\left(|x_n^\varepsilon - x_\varepsilon|^2 + |t_n^\varepsilon - t_\varepsilon|^2\right) \xrightarrow[n\to\infty]{} 0$,

(iii) $\tilde{u}(t_n^\varepsilon, x_n^\varepsilon) - \tilde{v}(t_n^\varepsilon, y_n^\varepsilon) \xrightarrow[n\to\infty]{} (\tilde{u} - \tilde{v})(t_\varepsilon, x_\varepsilon) \geq m + 2\alpha\beta(t_\varepsilon, x_\varepsilon) + 2\varepsilon f(x_\varepsilon)$,

(iv) $(t_n^\varepsilon, x_n^\varepsilon) \in [0,T) \times \bar{\mathcal{O}}^*$ for n large enough.

3. From Theorem 8.3 in [5], we deduce that for each $\eta > 0$ there are real coefficients $b_{1,n}^\varepsilon$, $b_{2,n}^\varepsilon$ and symmetric matrices $\mathcal{X}_n^{\varepsilon,\eta}$ and $\mathcal{Y}_n^{\varepsilon,\eta}$ such that

$$\left(b_{1,n}^\varepsilon, p_n^\varepsilon, \mathcal{X}_n^{\varepsilon,\eta}\right) \in \bar{\mathcal{P}}_{\bar{\mathcal{O}}}^+ \tilde{u}(t_n^\varepsilon, x_n^\varepsilon) \text{ and } \left(-b_{2,n}^\varepsilon, q_n^\varepsilon, \mathcal{Y}_n^{\varepsilon,\eta}\right) \in \bar{\mathcal{P}}_{\bar{\mathcal{O}}}^- \tilde{v}(t_n^\varepsilon, y_n^\varepsilon)$$

—see [5] for the standard notations $\bar{\mathcal{P}}_{\bar{\mathcal{O}}}^+$ and $\bar{\mathcal{P}}_{\bar{\mathcal{O}}}^-$, where

$$p_n^\varepsilon := 2n^2(x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon)e^{\frac{\zeta}{n}\rho_\varepsilon} + 2\zeta(x_n^\varepsilon - x_\varepsilon) + \alpha\, , D\beta(t_n^\varepsilon, x_n^\varepsilon) + \varepsilon Df(x_n^\varepsilon)\, ,$$
$$q_n^\varepsilon := 2n^2(x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon) - \alpha D\beta(t_n^\varepsilon, y_n^\varepsilon) - \varepsilon Df(y_n^\varepsilon)$$

—and such that $b_{1,n}^\varepsilon$, $b_{2,n}^\varepsilon$, $\mathcal{X}_n^{\varepsilon,\eta}$, and $\mathcal{Y}_n^{\varepsilon,\eta}$ satisfy

$$\text{(7.12)} \quad \left\{ \begin{array}{c} b_{1,n}^\varepsilon + b_{2,n}^\varepsilon = 2\zeta(t_n^\varepsilon - t_\varepsilon) - \alpha\tau\left(\beta(t_n^\varepsilon, x_n^\varepsilon) + \beta(t_n^\varepsilon, y_n^\varepsilon)\right)\, , \\ \begin{pmatrix} \mathcal{X}_n^{\varepsilon,\eta} & 0 \\ 0 & -\mathcal{Y}_n^{\varepsilon,\eta} \end{pmatrix} \leq (A_n^\varepsilon + B_n^\varepsilon) + \eta(A_n^\varepsilon + B_n^\varepsilon)^2\, , \end{array} \right.$$

with

$$A_n^\varepsilon := \begin{pmatrix} 2n^2\, \mathrm{diag}[e^{2\frac{\zeta}{n}\rho_\varepsilon}] + 2\zeta I_d & -2n^2\, \mathrm{diag}[e^{\frac{\zeta}{n}\rho_\varepsilon}] \\ -2n^2\, \mathrm{diag}[e^{\frac{\zeta}{n}\rho_\varepsilon}] & 2n^2 I_d \end{pmatrix}\, ,$$
$$B_n^\varepsilon := \begin{pmatrix} \alpha D^2\beta(t_n^\varepsilon, x_n^\varepsilon) + \varepsilon D^2 f(x_n^\varepsilon) & 0 \\ 0 & \alpha D^2\beta(t_n^\varepsilon, y_n^\varepsilon) + \varepsilon D^2 f(y_n^\varepsilon) \end{pmatrix}\, .$$

3.1. We now show that, up to a subsequence,

$$y_n^\varepsilon \in \mathcal{O}\, . \tag{7.13}$$

In view of (ii), this is clearly true when $x_\varepsilon \in \mathcal{O}$. In the case $x_\varepsilon \in \partial\mathcal{O}$, we deduce from (ii) that

$$y_n^\varepsilon = x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} + o(n^{-1}) \;=\; x_n^\varepsilon + \frac{\zeta}{n}\mathrm{diag}\,[x_n^\varepsilon]\,\rho_\varepsilon + o(n^{-1})\, .$$

This implies that, for some $\epsilon_n \to 0$,

$$d(y_n^\varepsilon) = d(x_n^\varepsilon) + \frac{\zeta}{n}\left(Dd(x_n^\varepsilon)'\mathrm{diag}\,[x_n^\varepsilon]\,\rho_\varepsilon + \epsilon_n\right)$$

so that (7.13) is a consequence of (7.10), the continuity of Dd, and (ii).

3.2. In this step, we show that there is a subsequence of $(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon)$ such that

$$\text{(7.14)}\ x_n^\varepsilon \in \mathcal{O} \qquad \text{and} \qquad \kappa\tilde{u}(t_n^\varepsilon, x_n^\varepsilon) - b_{1,n}^\varepsilon - \frac{1}{2}\mathrm{Tr}\left[a(t_n^\varepsilon, x_n^\varepsilon)\mathcal{X}_n^{\varepsilon,\eta}\right] \leq 0\, .$$

First observe that we cannot have $x_n^\varepsilon \in \partial\mathcal{O}^*$ and $\tilde{u}(t_n^\varepsilon, x_n^\varepsilon) \leq \tilde{g}(t_n^\varepsilon, x_n^\varepsilon)$ for all n. In view of (ii), this is obvious if $x_\varepsilon \in \mathcal{O}$. If $x_\varepsilon \in \partial\mathcal{O}^*$, it follows from (7.9)

and the viscosity property of $\tilde v$ that $\tilde v(t_\varepsilon, x_\varepsilon) \ge \tilde g(t_\varepsilon, x_\varepsilon)$. Since $\tilde g$ is upper-semicontinuous (see $\mathbf{H}_g$), this would imply that $\tilde u(t_n^\varepsilon, x_n^\varepsilon) \le \tilde v(t_n^\varepsilon, y_n^\varepsilon) + m/2$ for n large enough (see (ii)), thus leading to a contradiction to (iii) since $\beta, f \ge 0$. By (iv) and the viscosity subsolution property of $\tilde u$, we then deduce that either (7.14) holds or

$$H\left(\tilde u(t_n^\varepsilon, x_n^\varepsilon), \mathrm{diag}\,[x_n^\varepsilon]\, p_n^\varepsilon\right) \le 0\,. \tag{7.15}$$

Thus, it remains to prove that the above inequality leads to a contradiction. Using the supersolution property of $\tilde v$, (7.13), (ii)–(iii), and (2.9), we observe that (7.15) implies

$$\begin{aligned}
0 &\ge H\left(\tilde u(t_n^\varepsilon, x_n^\varepsilon), \mathrm{diag}\,[x_n^\varepsilon]\, p_n^\varepsilon\right) - H\left(\tilde v(t_n^\varepsilon, y_n^\varepsilon), \mathrm{diag}\,[y_n^\varepsilon]\, q_n^\varepsilon\right)\\
&\ge \alpha\left\{H\left(\beta(t_n^\varepsilon, x_n^\varepsilon), \mathrm{diag}\,[x_n^\varepsilon]\, D\beta(t_n^\varepsilon, x_n^\varepsilon)\right) + H\left(\beta(t_n^\varepsilon, y_n^\varepsilon), \mathrm{diag}\,[y_n^\varepsilon]\, D\beta(t_n^\varepsilon, y_n^\varepsilon)\right)\right\}\\
&+ \varepsilon\left\{H\left(f(x_n^\varepsilon), \mathrm{diag}\,[x_n^\varepsilon]\, Df(x_n^\varepsilon)\right) + H\left(f(y_n^\varepsilon), \mathrm{diag}\,[y_n^\varepsilon]\, Df(y_n^\varepsilon)\right)\right\}\\
&+ \inf_{\rho\in\tilde K_1} \delta(\rho)\left[\Theta(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) - \varepsilon\left(f(x_n^\varepsilon) + f(y_n^\varepsilon)\right)\right]\\
&- \sup_{\rho\in\tilde K_1} \left[2n^2\rho'\mathrm{diag}\left[x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon\right]\left(x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon\right) + 2\zeta\rho'\mathrm{diag}\,[x_n^\varepsilon]\,(x_n^\varepsilon - x_\varepsilon)\right]\\
&\ge \inf_{\rho\in\tilde K_1} \delta(\rho)(m/2) + 2\alpha H(\beta(t_\varepsilon, x_\varepsilon), \mathrm{diag}\,[x_\varepsilon]\, D\beta(t_\varepsilon, x_\varepsilon))\\
&+ \epsilon_n + \varepsilon\left\{H\left(f(x_\varepsilon), \mathrm{diag}\,[x_\varepsilon]\, Df(x_\varepsilon)\right) + H\left(f(y_\varepsilon), \mathrm{diag}\,[y_\varepsilon]\, Df(y_\varepsilon)\right)\right\}\,,
\end{aligned}$$

where $\epsilon_n \to 0$ when $n \to \infty$, but depend on ε. Recalling (7.4) and (7.8), we get a contradiction for ε small and n large enough. This concludes the proof of (7.14).

3.3. We can now provide the required contradiction and conclude the proof. Let $\tilde\sigma$ be defined on $\bar D$ by $\tilde\sigma(t,x) = \mathrm{diag}\,[x]\,\sigma(t,x)$. By the viscosity supersolution property of $\tilde v$, (7.13), (7.14), and (7.12), $(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon)$ must satisfy

$$\begin{aligned}
\kappa\left(\tilde u(t_n^\varepsilon, x_n^\varepsilon) - \tilde v(t_n^\varepsilon, y_n^\varepsilon)\right) &\le b_{1,n}^\varepsilon + b_{2,n}^\varepsilon + \frac{1}{2}\mathrm{Tr}\left[a(t_n^\varepsilon, x_n^\varepsilon)\mathcal{X}_n^{\varepsilon,\eta} - a(t_n^\varepsilon, y_n^\varepsilon)\mathcal{Y}_n^{\varepsilon,\eta}\right]\\
&\le 2\zeta(t_n^\varepsilon - t_\varepsilon) - \alpha\tau\left(\beta(t_n^\varepsilon, x_n^\varepsilon) + \beta(t_n^\varepsilon, y_n^\varepsilon)\right)\\
&+ \frac{1}{2}\mathrm{Tr}\left[\Xi(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon)\left(A_n^\varepsilon + B_n^\varepsilon + \eta(A_n^\varepsilon + B_n^\varepsilon)^2\right)\right]\,,
\end{aligned}$$

where

$$\Xi(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon) := \begin{pmatrix} \tilde\sigma(t_n^\varepsilon, x_n^\varepsilon)\tilde\sigma'(t_n^\varepsilon, x_n^\varepsilon) & \tilde\sigma(t_n^\varepsilon, y_n^\varepsilon)\tilde\sigma'(t_n^\varepsilon, x_n^\varepsilon) \\ \tilde\sigma(t_n^\varepsilon, x_n^\varepsilon)\tilde\sigma'(t_n^\varepsilon, y_n^\varepsilon) & \tilde\sigma(t_n^\varepsilon, y_n^\varepsilon)\tilde\sigma'(t_n^\varepsilon, y_n^\varepsilon) \end{pmatrix}$$

is a nonnegative symmetric matrix. Using (ii)–(iii), (7.5), and (7.8), it follows that for ε small and n large enough

$$\begin{aligned}
\kappa\, m/2 &\le \kappa\left(\tilde u(t_n^\varepsilon, x_n^\varepsilon) - \tilde v(t_n^\varepsilon, y_n^\varepsilon) - (\alpha\beta + \varepsilon f)(t_n^\varepsilon, x_n^\varepsilon) - (\alpha\beta + \varepsilon f)(t_n^\varepsilon, y_n^\varepsilon)\right)\\
&\le 2\zeta(t_n^\varepsilon - t_\varepsilon) + \frac{1}{2}\mathrm{Tr}\left[\Xi(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon)\left(A_n^\varepsilon + \eta(A_n^\varepsilon + B_n^\varepsilon)^2\right)\right] + \theta(\varepsilon, n)\,,
\end{aligned}$$

where $\theta(\varepsilon, n)$ is independent of (η, ζ) and satisfies

$$\limsup_{\varepsilon\to 0}\limsup_{n\to\infty} |\theta(\varepsilon, n)| = 0\,. \tag{7.16}$$

Sending $\eta \to 0$ in the previous inequality provides

$$\kappa\, m/2 \le 2\zeta(t_n^\varepsilon - t_\varepsilon) + \frac{1}{2}\mathrm{Tr}\left[\Xi(t_n^\varepsilon, x_n^\varepsilon, y_n^\varepsilon)A_n^\varepsilon\right] + \theta(\varepsilon, n)\ ,$$

so that

$$\begin{aligned}\kappa\, m/2 \le{}& 2\zeta(t_n^\varepsilon - t_\varepsilon) + \zeta\mathrm{Tr}\left[\tilde\sigma(t_n^\varepsilon, x_n^\varepsilon)\tilde\sigma'(t_n^\varepsilon, x_n^\varepsilon)\right]\\ &+ n^2\left|\mathrm{diag}\left[x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}\right]\sigma(t_n^\varepsilon, x_n^\varepsilon) - \mathrm{diag}\left[y_n^\varepsilon\right]\sigma(t_n^\varepsilon, y_n^\varepsilon)\right|^2 + \theta(\varepsilon, n)\ .\end{aligned}$$

Using (2.2), we now observe that

$$\begin{aligned}&\left|\mathrm{diag}\left[x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}\right]\left(\sigma(t_n^\varepsilon, x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}) - \sigma(t_n^\varepsilon, x_n^\varepsilon)\right)\right|\\ &\le \left|\mathrm{diag}\left[x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}\right]\sigma(t_n^\varepsilon, x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}) - \mathrm{diag}\left[x_n^\varepsilon\right]\sigma(t_n^\varepsilon, x_n^\varepsilon)\right|\\ &\quad + \left|\mathrm{diag}\left[x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - x_n^\varepsilon\right]\sigma(t_n^\varepsilon, x_n^\varepsilon)\right|\\ &\le C_\varepsilon\ \left|x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - x_n^\varepsilon\right|\\ &\le \zeta C_\varepsilon\ n^{-1}\ ,\end{aligned}$$

and

$$\left|\mathrm{diag}\left[x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}\right]\sigma(t_n^\varepsilon, x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon}) - \mathrm{diag}\left[y_n^\varepsilon\right]\sigma(t_n^\varepsilon, y_n^\varepsilon)\right|^2 \le C_\varepsilon\ \left|x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon\right|^2\ ,$$

where $C_\varepsilon > 0$ denotes a generic constant independent of n and ζ. Plugging this into the previous inequality implies that there is some $C_\varepsilon > 0$ independent of n and ζ for which

$$\begin{aligned}\kappa\, m/2 \le{}& 2\zeta(t_n^\varepsilon - t_\varepsilon) + \zeta\mathrm{Tr}\left[\tilde\sigma'(t_n^\varepsilon, x_n^\varepsilon)\tilde\sigma(t_n^\varepsilon, x_n^\varepsilon)\right]\\ &+ C_\varepsilon\left(\zeta + n^2|x_n^\varepsilon e^{\frac{\zeta}{n}\rho_\varepsilon} - y_n^\varepsilon|^2\right) + \theta(\varepsilon, n).\end{aligned}$$

Finally, using (ii) and sending n to ∞ and then ζ to 0 in the last inequality implies

$$\kappa\, m/2 \le \limsup_{n\to\infty}\theta(\varepsilon, n)\ ,$$

which by (7.16) provides the required contradiction and concludes the proof.

4. We now explain how to adapt this proof to the alternative assumptions of $\mathbf{H}'$.

4.1. Observe that the penalty function β is introduced in order to obtain a finite supremum for $\tilde u - \tilde v - 2\alpha\beta$ and existence of an optimum for Φ^ε and $\Psi^\varepsilon_{n,\zeta}$. If $\bar{\mathcal{O}}$ is bounded, the introduction of such a penalty function is not required and we can reproduce the same proof with $\beta \equiv 0$ whenever $0 \in \mathrm{int}(K)$. Indeed, by Remark 2.4, $\inf_{\rho\in\tilde K_1}\delta(\rho) > 0$, so that we still obtain a contradiction at the end of step 3.2. The arguments of step 3.3 also work with $\beta \equiv 0$. The case where $0 \notin \mathrm{int}(K)$ is discussed below.

4.2. Similarly, the map f is introduced only to prevent the different maxima from taking values outside $\bar{\mathcal{O}}^*$. If $\bar{\mathcal{O}} \cap \partial\mathbb{R}^d_+ = \emptyset$, this penalty function is useless

and can be fixed to $f \equiv 0$. In this case, one can also fix some $\underline{\gamma} \in \text{int}(K)$, if nonempty, and add the term $e^{\tau(T-t)}x^{\underline{\gamma}}$ in the definition of β. Thus, β becomes $e^{\tau(T-t)}\,(1 + x^{\gamma} + x^{\underline{\gamma}})$ or $e^{\tau(T-t)}\,(1 + x^{\underline{\gamma}})$, depending on whether $\bar{\mathcal{O}}$ is bounded or not; see step 4.1. For fixed $\varepsilon > 0$, we deduce from Remark 2.4 and the fact that $\underline{\gamma} \in \text{int}(K)$ that $H(\beta(t,x), \text{diag}\,[x]\,D\beta(t,x)) > 0$. Since $f = 0$, there is no ε to send to 0 at the end of steps 3.2 and 3.3, and we obtain the same contradictions by simply sending n to ∞ and ζ to 0. □

REFERENCES

[1] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.

[2] I. Bentahar, *Numerical resolution of the barrier option pricing problem under constraints*, forthcoming.

[3] B. Bouchard, *Stochastic targets with mixed diffusion processes*, Stochastic Process. Appl., 101 (2002), pp. 273–302.

[4] M. Broadie, J. Cvitanić, and M. Soner, *Optimal replication of contingent claims under portfolio constraints*, Rev. Financial Stud., 11 (1998), pp. 59–79.

[5] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Amer. Math. Soc., 27 (1992), pp. 1–67.

[6] J. Cvitanić and I. Karatzas, *Hedging contingent claims with constrained portfolios*, Ann. Appl. Probab., 3 (1993), pp. 652–681.

[7] J. Cvitanić, H. Pham, and N. Touzi, *Super-replication in stochastic volatility models with portfolio constraints*, J. Appl. Probab., 36 (1999), pp. 523–545.

[8] H. Föllmer and D. Kramkov, *Optional decomposition under constraints*, Probab. Theory Related Fields, 109 (1997), pp. 1–25.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[10] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[11] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[12] S. E. Shreve, U. Schmock, and U. Wystup, *Valuation of exotic options under shortselling constraints*, Finance Stoch., 6 (2002), pp. 143–172.

[13] S. E. Shreve, U. Schmock, and U. Wystup, *Dealing with dangerous digitals*, in Foreign Exchange Risk, Risk Publications, London, 2002, pp. 327–348.

[14] H. M. Soner and N. Touzi, *Dynamic programming for stochastic target problems and geometric flows*, J. Eur. Math. Soc. (JEMS), 4 (2002), pp. 201–236.

[15] H. M. Soner and N. Touzi, *Stochastic target problems, dynamic programming, and viscosity solutions*, SIAM J. Control Optim., 41 (2002), pp. 404–424.

[16] H. M. Soner and N. Touzi, *The problem of super-replication under constraints*, in Paris-Princeton Lectures on Mathematical Finance, Lecture Notes in Math. 1814, Springer-Verlag, Berlin, 2004, pp. 133–172.

[17] N. Touzi, *Direct characterization of the value of super-replication under stochastic volatility and portfolio constraints*, Stochastic Process. Appl., 88 (2000), pp. 305–328.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1875–1897

QUASI-OPTIMAL ROBUST STABILIZATION OF CONTROL SYSTEMS*

CHRISTOPHE PRIEUR† AND EMMANUEL TRÉLAT‡

Abstract. In this paper, we investigate the problem of semiglobal minimal time robust stabilization of analytic control systems with controls entering linearly, by means of a hybrid state feedback law. It is shown that in the absence of minimal time singular trajectories, the solutions of the closed-loop system converge to the origin in quasi-minimal time (for a given bound on the controller) with a robustness property with respect to small measurement noise, external disturbances, and actuator noise.

Key words. hybrid feedback, robust stabilization, measurement errors, actuator noise, external disturbances, optimal control, singular trajectory, sub-Riemannian geometry

AMS subject classifications. 93B52, 93D15

DOI. 10.1137/050642629

1. Introduction.

Let m and n be two positive integers. Consider on $\mathbb{R}^n$ the control system

$$\dot{x}(t) = \sum_{i=1}^{m} u_i(t) f_i(x(t)), \tag{1.1}$$

where $f_1, \ldots, f_m$ are analytic vector fields on $\mathbb{R}^n$, and where the control function $u(\cdot) = (u_1(\cdot), \ldots, u_m(\cdot))$ satisfies the constraint

$$\sum_{i=1}^{m} u_i(t)^2 \leqslant 1. \tag{1.2}$$

All results of this paper still hold on a Riemannian analytic manifold of dimension n, which is connected and complete. However, for the sake of simplicity, our results are stated in $\mathbb{R}^n$. Let $\bar{x} \in \mathbb{R}^n$. The system (1.1), together with the constraint (1.2), is said to be *globally asymptotically stabilizable* at the point $\bar{x}$ if, for each point $x \in \mathbb{R}^n$, there exists a control law satisfying the constraint (1.2) such that the solution of (1.1) associated to this control law with the initial condition x tends to $\bar{x}$ as t tends to $+\infty$.

This asymptotic stabilization problem has a long history and has been widely investigated. Note that, due to *Brockett's condition* [16, Theorem 1 (iii)], if $m < n$, and if the vector fields $f_1, \ldots, f_m$ are independent, then there does not exist any continuous stabilizing feedback law for (1.1). However several control laws have been derived for such control systems (see, for instance, [8, 29] and the references therein).

The *robust asymptotic stabilization problem* is under current and active research. Many notions of controllers have been introduced to treat this problem, such as discontinuous sampling feedbacks [19, 45], time varying control laws [20, 21, 33, 34],

*Received by the editors October 13, 2005; accepted for publication (in revised form) August 8, 2006; published electronically December 11, 2006.
http://www.siam.org/journals/sicon/45-5/64262.html

†LAAS-CNRS, 7 avenue du Colonel Roche, 31077 Toulouse, France (christophe.prieur@laas.fr). The research of this author was done partly in the framework of the HYCON Network of Excellence, contract FP6-IST-511368.

‡Mathématiques, Laboratoire MAPMO, Université d'Orléans, UMR 6628, Route de Chartres, B.P. 6759, 45067 Orléans Cedex 2, France (emmanuel.trelat@univ-orleans.fr).

patchy feedbacks (as in [5]), and SRS feedbacks [43], enjoying different robustness properties depending on the errors under consideration.

In the present paper, we consider feedback laws having both discrete and continuous components, which generate closed-loop systems with *hybrid* terms (see, for instance, [11, 49]). Such feedback appeared first in [37] to stabilize nonlinear systems having a priori no discrete state. They consist in defining a *switching strategy* between several smooth control laws defined on a partition of the state space. Many results on the stabilization problem of nonlinear systems by means of hybrid controllers have been recently established (see, for instance, [14, 22, 25, 26, 30, 32, 35, 53]). The notion of solution, connected with the robustness problem, is by now well defined in the hybrid context (see [25, 39], among others). Specific conditions for the optimization can be found in the literature (see, e.g., [9, 24]).

The strategy of our paper is to combine a minimal time controller that is smooth on a part of the state space, and other controllers defined on the complement of this part, so as to provide a *quasi-minimal time hybrid controller* by defining a switching strategy between all control laws. The resulting hybrid law enjoys a quasi-minimal time property, and robustness with respect to (small) measurement noise, actuator errors, and external disturbances.

More precisely, in a first step, we consider the minimal time problem for the system (1.1) with the constraint (1.2) of steering a point $x \in \mathbb{R}^n$ to the point $\bar{x}$. Note that this problem is solvable as soon as the *Lie algebra rank condition* holds for the m-tuple of vector fields $(f_1, \dots, f_m)$. Of course, in general, it is impossible to compute explicitly the minimal time feedback controllers for this problem. Moreover, Brockett's condition implies that such control laws are not smooth whenever $m < n$ and the vector fields $f_1, \dots, f_m$ are independent. This raises the problem of the regularity of optimal feedback laws. The literature on this subject is immense. In an analytic setting, the problem of determining the analytic regularity of the minimal time function has been investigated in among others, [47]. For systems of the form (1.1), it follows from [1, 2, 50] that the minimal time function to $\bar{x}$ is *subanalytic*, provided there does not exist any nontrivial singular minimal time trajectory starting from $\bar{x}$ (see [27, 28] for a general definition of subanalytic sets). This assumption holds generically for systems (1.1), whenever $m \geqslant 3$ (see [18]). In particular, this function is analytic outside a stratified submanifold $\mathcal{S}$ of $\mathbb{R}^n$ of codimension greater than or equal to 1 (see [48]). As a consequence, outside this submanifold it is possible to provide an analytic minimal time feedback controller for the system (1.1), (1.2). This optimal controller gives rise to trajectories never crossing again the singular set $\mathcal{S}$.

Note that the analytic context is used so as to ensure stratification properties, which do not hold a priori if the system is smooth only. These properties are related to the notion of an *o-minimal category* (see [23]).

In a neighborhood of $\mathcal{S}$, we prove the existence of a set of controllers steering the system (1.1), (1.2) outside of this neighborhood in small time.

Then, in order to achieve a minimal time robust stabilization procedure, using a hybrid feedback law, we define a suitable switching strategy (more precisely, a hysteresis) between the minimal time feedback controller and other controllers defined in a neighborhood of $\mathcal{S}$. The resulting hybrid system has the following property: If the state is close to the singular submanifold $\mathcal{S}$, the feedback controller will push the state far enough from $\mathcal{S}$, in small time; if the state is not too close to $\mathcal{S}$, then the feedback controller will steer the system to $\bar{x}$ in minimal time. Hence, the stabilization is quasi-optimal and is proved to enjoy robustness properties.

Note that we thus give an alternative solution, in the context of hybrid systems using hysteresis, to a conjecture of [15, Conjecture 1, p. 101] concerning patchy feedbacks for smooth control systems.[1]

In a previous paper [41], this program was achieved on the so-called Brockett system, for which $n = 3$, $m = 2$, and, denoting $x = (x_1, x_2, x_3)$,

$$f_1 = \frac{\partial}{\partial x_1} + x_2 \frac{\partial}{\partial x_3}, \quad f_2 = \frac{\partial}{\partial x_2} - x_1 \frac{\partial}{\partial x_3}.$$

In this case, there does not exist any nontrivial singular trajectory, and the manifold $\mathcal{S}$ coincides with the axis $(0x_3)$. A simple explicit hybrid strategy was described. In contrast, in the present paper, we derive a general result that requires a countable number of components in the definition of the hysteresis hybrid feedback law.

The paper is organized as follows. In section 2, we first recall some facts about the minimal time problem for the system (1.1) with (1.2) and recall the definition of a singular trajectory. Then we give a notion of solution adapted to hybrid feedback laws and define the concept of stabilization via a minimal time hybrid feedback law. The main result, Theorem 2.10 in section 2.3, states that if there does not exist any nontrivial singular minimal time trajectory of (1.1), (1.2), starting from $\bar{x}$, then there exists a minimal time hybrid feedback law stabilizing semiglobally the point $\bar{x}$ for the system (1.1), (1.2). Section 2.4 describes the main ideas of the proof of the main result and, in particular, contains two key lemmas. Section 3 is then devoted to the detailed proof of Theorem 2.10 and gathers all technical aspects needed to deal with hybrid systems: the components of the hybrid feedback law and a switching strategy between both components are defined, and properties of the closed-loop system are investigated.

The results in this work were announced in [42].

2. Definitions and main result.

2.1. The minimal time problem. Consider the minimal time problem for the system (1.1) with the constraint (1.2).

Throughout the paper, we assume that the *Lie algebra rank condition* holds, that is, the Lie algebra spanned by the vector fields $f_1, \ldots, f_m$ is equal to $\mathbb{R}^n$, at every point x of $\mathbb{R}^n$.

It is well known that, under this condition, any two points of $\mathbb{R}^n$ can be joined by a minimal time trajectory of (1.1), (1.2).

Let $\bar{x} \in \mathbb{R}^n$ be fixed. We denote by $T_{\bar{x}}(x)$ the minimal time needed to steer the system (1.1) with the constraint (1.2) from a point $x \in \mathbb{R}^n$ to the point $\bar{x}$.

Remark 2.1. Obviously, the control function u associated to a minimal time trajectory of (1.1), (1.2), actually satisfies $\sum_{i=1}^{m} u_i^2 = 1$.

For $T > 0$, let $\mathcal{U}_T$ denote the (open) subset of $u(\cdot)$ in $L^\infty([0,T], \mathbb{R}^m)$ such that the solution of (1.1), starting from $\bar{x}$ and associated to a control $u(\cdot) \in \mathcal{U}_T$, is well defined on $[0,T]$. The mapping

$$\begin{array}{rccc} E_{\bar{x},T}: & \mathcal{U}_T & \longrightarrow & \mathbb{R}^n \\ & u(\cdot) & \longmapsto & x(T), \end{array}$$

[1] This conjecture on patchy feedbacks has been recently considered in [6]. In this preprint, written during the review process of the present work, the authors prove, using a penalization method, a general result on stabilization by means of patchy feedback of nonlinear control systems in quasi-minimal time.

which to a control $u(\cdot)$ associates the end-point $x(T)$ of the corresponding solution $x(\cdot)$ of (1.1) starting at $\bar{x}$, is called *end-point mapping* at the point $\bar{x}$, in time T; it is a smooth mapping.

DEFINITION 2.2. *A trajectory $x(\cdot)$ of (1.1), so that $x(0) = \bar{x}$, is said to be singular on $[0,T]$ if its associated control $u(\cdot)$ is a singular point of the end-point mapping $E_{\bar{x},T}$ (i.e., if the Fréchet derivative of $E_{\bar{x},T}$ at $u(\cdot)$ is not onto). The control $u(\cdot)$ is said to be singular.*

Remark 2.3. If $x(\cdot)$ is singular on $[0,T]$, then it is singular on $[t_0,t_1]$ for all $t_0,t_1 \in [0,T]$ such that $t_0 < t_1$.

Remark 2.4. It is a standard fact that the minimal time control problem for the system (1.1) with the constraint (1.2) is equivalent to the sub-Riemannian problem associated to the m-tuple of vector fields $(f_1,\dots,f_m)$ (see [10] for a general definition of a sub-Riemannian distance). In this context, there holds $T_{\bar{x}}(x) = d_{SR}(\bar{x},x)$, where d_{SR} is the sub-Riemannian distance. This implies that the functions $T_{\bar{x}}(\cdot)$ and $d_{SR}(\bar{x},\cdot)$ share the same regularity properties. In particular, the function $T_{\bar{x}}(\cdot)$ is continuous.

2.2. Class of controllers and notion of hybrid solution. Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be defined by $f(x,u) = \sum_{i=1}^m u_i f_i(x)$. The system (1.1) is written as

$$\dot{x}(t) = f(x(t),u(t)). \tag{2.1}$$

Let $\bar{x} \in \mathbb{R}^n$ be fixed.

The controllers under consideration in this paper depend on the continuous state $x \in \mathbb{R}^n$ and also on a discrete variable $s_d \in \mathcal{N}$, where $\mathcal{N}$ is a nonempty subset of $\mathbb{N}$. According to the concept of a hybrid system of [25], we introduce the following definition.

DEFINITION 2.5. *A hybrid feedback is a 4-tuple (C,D,k,k_d), where*

- *C and D are subsets of $\mathbb{R}^n \times \mathcal{N}$;*
- *$k : \mathbb{R}^n \times \mathcal{N} \to \mathbb{R}^m$ is a function;*
- *$k_d : \mathbb{R}^n \times \mathcal{N} \to \mathcal{N}$ is a function.*

The sets C and D are, respectively, called the controlled continuous evolution set *and the* controlled discrete evolution set.

We next recall the notion of robustness to small noise (see [46]). Consider two functions e and d satisfying the following *regularity assumptions*:

$$\begin{aligned} &e(\cdot,\cdot), d(\cdot,\cdot) \in L^\infty_{loc}(\mathbb{R}^n \times [0,+\infty);\mathbb{R}^n),\\ &e(\cdot,t), d(\cdot,t) \in C^0(\mathbb{R}^n,\mathbb{R}^n)\ \forall t \in [0,+\infty). \end{aligned} \tag{2.2}$$

We introduce these functions as a measurement noise e and an external disturbance d.

Formally, the k-component of a hybrid feedback (k,k_d,C,D) governs the differential equation

$$\dot{x} = f(x,k(x+e)) + d \ \ \forall (x,s_d) \in C,$$

whereas the k_d-component governs the jump equation

$$s_d^+ = k_d(x,s_d) \ \ \forall (x,s_d) \in D.$$

The set C indicates where in the state space flow may occur, while the set D indicates where in the state space jumps may occur. The combination of this flow equation and this jump equation, under the perturbations e and d, is a perturbed hybrid system

$\mathcal{H}_{(e,d)}$, as considered, e.g., in [26]. We next provide a precise definition of the notion of solutions considered here.

This concept is well studied in the literature (see, e.g., [11, 14, 31, 38, 39, 49]). Here, we consider the notion of solution given in [25, 26].

DEFINITION 2.6. *Let* $S = \bigcup_{j=0}^{J-1} [t_j, t_{j+1}] \times \{j\}$, *where* $J \in \mathbb{N} \cup \{+\infty\}$ *and* $(x_0, s_0) \in \mathbb{R}^n \times \mathcal{N}$. *The domain* S *is said to be a hybrid time domain. A map* $(x, s_d) : S \to \mathbb{R}^n \times \mathcal{N}$ *is said to be a solution of* $\mathcal{H}_{(e,d)}$ *with the initial condition* (x_0, s_0) *if*

- *the map* x *is continuous on* S;
- *for every* j, $0 \leqslant j \leqslant J-1$, *the map* $x : t \in (t_j, t_{j+1}) \mapsto x(t, j)$ *is absolutely continuous;*
- *for every* j, $0 \leqslant j \leqslant J-1$, *and almost every* $t \geqslant 0$, $(t, j) \in S$, *we have*

$$(x(t,j) + e(x(t,j), t), s_d(t,j)) \in C \tag{2.3}$$

and

$$\dot{x}(t,j) = f(x(t), k(x(t,j) + e(x(t,j), t), s_d(t,j))) + d(x(t,j), t), \tag{2.4}$$

$$\dot{s}_d(t,j) = 0 \tag{2.5}$$

(where the dot stands for the derivative with respect to the time variable t*);*

- *for every* $(t, j) \in S$, $(t, j+1) \in S$, *we have*

$$(x(t,j) + e(x(t,j), t), s_d(t,j)) \in D \tag{2.6}$$

and

$$x(t, j+1) = x(t, j), \tag{2.7}$$

$$s_d(t, j+1) = k_d(x(t,j) + e(x(t,j), t), s_d(t,j)); \tag{2.8}$$

- $(x(0,0), s_d(0,0)) = (x_0, s_0)$.

In this context, we next define the concept of stabilization of (2.1) by a minimal time hybrid feedback law sharing a robustness property with respect to measurement noise and external disturbances. The usual Euclidean norm in $\mathbb{R}^n$ is denoted by $|\cdot|$, and the open ball centered at $\bar{x}$ with radius R is denoted $B(\bar{x}, R)$. Recall that a function of class $\mathcal{K}_\infty$ is a function δ: $[0, +\infty) \to [0, +\infty)$ which is continuous, increasing, and satisfies $\delta(0) = 0$ and $\lim_{R \to +\infty} \delta(R) = +\infty$.

As usual, the system is said to be complete if all solutions are maximally defined in $[0, +\infty)$ (see, e.g., [7]). More precisely, we have the following definition.

DEFINITION 2.7. *Let* $\rho : \mathbb{R}^n \to \mathbb{R}$ *be a continuous function satisfying*

$$\rho(x) > 0 \ \forall x \neq \bar{x}. \tag{2.9}$$

We say that the completeness assumption *for* ρ *holds if, for all* (e, d) *satisfying the regularity assumptions* (2.2), *and such that*

$$\sup_{[0,+\infty)} |e(x, \cdot)| \leqslant \rho(x), \ \operatorname{esssup}_{[0,+\infty)} |d(x, \cdot)| \leqslant \rho(x) \ \forall x \in \mathbb{R}^n, \tag{2.10}$$

for every $(x_0, s_0) \in \mathbb{R}^n \times \mathcal{N}$, *there exists a maximal solution on* $[0, +\infty)$ *of* $\mathcal{H}_{(e,d)}$ *starting from* (x_0, s_0).

Roughly speaking, the finite time convergence property means that all solutions reach $\bar{x}$ within finite time. A precise definition of this concept follows.

DEFINITION 2.8. *We say that the* uniform finite time convergence property *holds if there exists a continuous function* $\rho : \mathbb{R}^n \to \mathbb{R}$ *satisfying* (2.9)*, such that the completeness assumption for* ρ *holds, and if there exists a function* $\delta : [0, +\infty) \to [0, +\infty)$ *of class* $\mathcal{K}_\infty$ *such that, for every* $R > 0$*, there exists* $\tau = \tau(R) > 0$*, so that for all functions* e, d *satisfying the regularity assumptions* (2.2) *and inequalities* (2.10) *for this function* ρ*, for every* $x_0 \in B(\bar{x}, R)$*, and every* $s_0 \in \mathcal{N}$*, the maximal solution* (x, s_d) *of* $\mathcal{H}_{(e,d)}$ *starting from* (x_0, s_0) *satisfies*

$$|x(t, j) - \bar{x}| \leqslant \delta(R) \ \forall t \geqslant 0, \ (t, j) \in S, \tag{2.11}$$

and

$$x(t, j) = \bar{x} \ \forall t \geqslant \tau, \ (t, j) \in S. \tag{2.12}$$

We are now in a position to introduce our main definition. It deals with closed-loop systems whose trajectories converge to the equilibrium within quasi-minimal time and with a robustness property with respect to measurement noise and external disturbances.

DEFINITION 2.9. *The point* $\bar{x}$ *is said to be a* semiglobally quasi-minimal time robustly stabilizable equilibrium *for the system* (2.1) *if, for every* $\varepsilon > 0$ *and every compact subset* $K \subset \mathbb{R}^n$*, there exists a hybrid feedback law* (C, D, k, k_d) *satisfying the constraint*

$$\|k(x, s_d)\| \leqslant 1, \tag{2.13}$$

where $\|\cdot\|$ *stands for the Euclidean norm in* $\mathbb{R}^m$*, such that*

- *the uniform finite time convergence property holds;*
- *there exists a continuous function* $\rho_{\varepsilon,K} : \mathbb{R}^n \to \mathbb{R}$ *satisfying* (2.9) *for* $\rho = \rho_{\varepsilon,K}$*, such that, for every* $(x_0, s_0) \in K \times \mathcal{N}$*, all functions* e, d *satisfying the regularity assumptions* (2.2) *and inequalities* (2.10) *for* $\rho = \rho_{\varepsilon,K}$*, the maximal solution of* $\mathcal{H}_{(e,d)}$ *starting from* (x_0, s_0) *reaches* $\bar{x}$ *within time* $T_{\bar{x}}(x_0) + \varepsilon$*, where* $T_{\bar{x}}(x_0)$ *denotes the minimal time to steer the system* (2.1) *from* x_0 *to* $\bar{x}$*, under the constraint* $\|u\| \leqslant 1$*.*

2.3. Main result. The main result of this article is the following.

THEOREM 2.10. *Let* $\bar{x} \in \mathbb{R}^n$*. If there exists no nontrivial minimal time singular trajectory of* (1.1)*,* (1.2)*, starting from* $\bar{x}$*, then* $\bar{x}$ *is a semiglobally quasi-minimal time robustly stabilizable equilibrium for the system* (1.1)*, under the constraint* (1.2)*.*

Remark 2.11. The problem of global quasi-minimal time robust stabilization (i.e., $K = \mathbb{R}^n$ in Definition 2.9) cannot be achieved a priori because measurement noise may then accumulate and slow down the solution reaching $\bar{x}$ (compare with [15]).

Remark 2.12. The assumption of the absence of nontrivial singular minimizing trajectory is crucial. Notice the following facts, which show the relevance of this assumption:

- If $m \geqslant n$ and if the vector fields $f_1, \ldots, f_m$, are everywhere linearly independent, then there exists no singular trajectory. In this case, we are actually in the framework of Riemannian geometry (see Remark 2.4).
- Let $\mathcal{F}_m$ be the set of m-tuples of linearly independent vector fields $(f_1, \ldots, f_m)$, endowed with the C^∞ Whitney topology. If $m \geqslant 3$, there exists an open dense subset of $\mathcal{F}_m$, such that any control system of the form (1.1), associated to a m-tuple of this subset, admits no nontrivial singular minimizing trajectory

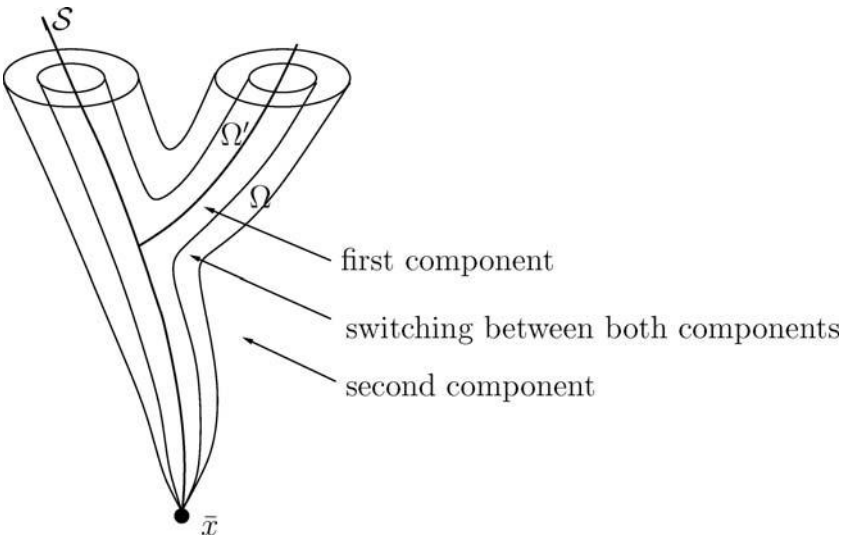


FIG. 2.1. *Switching strategy.*

(see [17, 18]; see also [2] for the existence of a dense set only). Hence generically the conclusion of Theorem 2.10 holds without assuming the absence of nontrivial singular minimizing trajectories.

- If there exist singular minimizing trajectories, then the conclusion on subanalyticity of the function T may fail (see [13, 50]), and we cannot a priori prove that the set $\mathcal{S}$ of singularities of T is a stratifiable manifold, which is the crucial fact in order to define a hybrid strategy.

2.4. Short description of the proof. The strategy of the proof of Theorem 2.10 is the following.

Under the assumption of the absence of nontrivial singular minimal time trajectory, the minimal time function $T_{\bar{x}}$ associated to the system (1.1), (1.2) is subanalytic, and hence is analytic outside a stratified submanifold $\mathcal{S}$ of $\mathbb{R}^n$, of codimension greater than or equal to one. Therefore, the corresponding minimal time feedback controller (further precisely defined in section 3.2.3) is continuous (even analytic) on $\mathbb{R}^n \setminus \mathcal{S}$ (see Figure 2.1). In a neighborhood of $\mathcal{S}$, it is therefore necessary to use other controllers, and then to define an adequate switching strategy.

More precisely, the proof of Theorem 2.10 relies on both of the following key lemmas.

LEMMA 2.13. *For every $\varepsilon > 0$, there exists a neighborhood Ω of $\mathcal{S}$ such that, for every stratum*[2] *M_i of $\mathcal{S}$, there exist a nonempty subset $\mathcal{N}_i$ of $\mathbb{N}$, a locally finite family $(\Omega_{i,p})_{p\in\mathcal{N}_i}$ of open subsets of Ω, a sequence of smooth controllers $u_{i,p}$ defined in a neighborhood of $\Omega_{i,p}$, satisfying $\|u_{i,p}\| \leqslant 1$, and there exists a continuous function $\rho_{i,p} : \mathbb{R}^n \to [0, +\infty)$ satisfying $\rho_{i,p}(x) > 0$ whenever $x \neq \bar{x}$, such that every solution of*

$$\dot{x}(t) = f(x(t), u_{i,p}(x(t) + e(x(t), t))) + d(x(t), t), \tag{2.14}$$

[2]Since $\mathcal{S}$ is a stratified submanifold of $\mathbb{R}^n$ of codimension greater than or equal to one, there exists a partition $(M_i)_{i\in\mathbb{N}}$ of $\mathcal{S}$, where M_i is a stratum, i.e., a locally closed submanifold of $\mathbb{R}^n$. Recall that if $M_i \cap \partial M_j \neq \emptyset$, then $M_i \subset M_j$ and $\dim(M_i) < \dim(M_j)$.

where $e, d : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$ *are two functions satisfying the regularity assumptions* (2.2) *and*

$$\sup_{[0,+\infty)} |e(x,\cdot)| \leqslant \rho_{i,p}(x), \quad \operatorname{esssup}_{[0,+\infty)} |d(x,\cdot)| \leqslant \rho_{i,p}(x), \tag{2.15}$$

starting from $\Omega_{i,p}$ *and maximally defined on* $[0, T)$*, leaves* Ω *within time* ε*; moreover, there exists a function* $\delta_{i,p}$ *of class* $\mathcal{K}_\infty$ *such that, for every* $R > 0$*, every such solution starting from* $\Omega_{i,p} \cap B(\bar{x}, R)$ *satisfies*

$$|x(t) - \bar{x}| \leqslant \delta_{i,p}(R) \quad \forall t \in [0, T). \tag{2.16}$$

According to this lemma, in a neighborhood Ω of $\mathcal{S}$, there exist controllers steering the system outside Ω. Moreover, since this neighborhood can be chosen arbitrarily thin, the time ε needed for its traversing is arbitrarily small.

Outside Ω, the optimal controller is analytic. The following lemma shows that this controller shares an invariance property; in brief, it gives rise to trajectories never again crossing the singular set $\mathcal{S}$.

LEMMA 2.14. *For every neighborhood* Ω *of* $\mathcal{S} \setminus \{\bar{x}\}$ *in* $\mathbb{R}^n$*, there exists a neighborhood* Ω' *of* $\mathcal{S} \setminus \{\bar{x}\}$ *in* $\mathbb{R}^n$*, satisfying*

$$\Omega' \subsetneq \operatorname{clos}(\Omega') \subsetneq \Omega, \tag{2.17}$$

such that every trajectory of the closed-loop system (1.1) *with the optimal controller, starting from any point* $x \in \mathbb{R}^n \setminus \Omega$*, reaches* $\bar{x}$ *in minimal time, and is contained in* $\mathbb{R}^n \setminus \Omega'$*.*

Finally, our hybrid strategy is the following. For every $\varepsilon > 0$, there exists a neighborhood Ω of the singular set $\mathcal{S}$, and there exist controllers which steer the system outside this neighborhood in time less than ε. Outside Ω, there exists a continuous controller giving rise to trajectories never again crossing $\mathcal{S}$ and joining $\bar{x}$ in minimal time.

It is therefore necessary to define an adequate switching strategy connecting both controllers (see Figure 2.1). This is achieved in the context of hybrid systems, using a hysteresis strategy. The first component consists of controllers which are defined in Ω, and whose existence is stated in Lemma 2.13. The second component of the hysteresis is defined by the optimal controller, outside Ω. Both components are united using an hysteresis, by adding a dynamical discrete variable s_d and using a hybrid feedback law. With this resulting hybrid controller, the time needed to join $\bar{x}$, from any point x_0 of $\mathbb{R}^n$, is less than $T_{\bar{x}}(x_0) + \varepsilon$.

The next section, devoted to the detailed proof of Theorem 2.10, is organized as follows.

The first component of the hysteresis is defined in section 3.1, and Lemma 2.13 is proved.

Section 3.2 concerns the definition and properties of the second component of the hysteresis, defined by the minimal time controller. In section 3.2.1, we recall how to compute minimal time trajectories of the system (1.1), (1.2), using the Pontryagin maximum principle. We then provide in section 3.2.2 a crucial result on the cut locus (Proposition 3.6). The optimal feedback controller is defined in section 3.2.3; basic facts on subanalytic functions are recalled, permitting us to define the singular set $\mathcal{S}$. Invariance properties of this optimal controller are then investigated: Lemma 2.14 is proved in section 3.2.4; robustness properties are given and proved in section 3.2.5.

The hybrid controller is then defined in section 3.3. A definition of a hybrid control system, and properties of solutions, are given in sections 3.3.1 and 3.3.2. A

precise description of the switching strategy is provided in section 3.3.3. Theorem 2.10 is proved in section 3.4.

3. Proof of Theorem 2.10. In what follows, let $\bar{x} \in \mathbb{R}^n$ be fixed.

3.1. The first component of the hysteresis. The first component of the hysteresis consists of a set of controllers, defined in a neighborhood of $\mathcal{S}$, whose existence is stated in Lemma 2.13. Hereafter, we provide a proof of this lemma.

Proof of Lemma 2.13. First of all, recall that, on the one hand, the minimal time function coincides with the sub-Riemannian distance associated to the m-tuple $(f_1, \dots, f_m)$ (see Remark 2.4); on the other, since the Lie algebra rank condition holds, the topology defined by the sub-Riemannian distance d_{SR} coincides with the Euclidean topology of $\mathbb{R}^n$, and, since $\mathbb{R}^n$ is complete, any two points of $\mathbb{R}^n$ can be joined by a minimizing path (see [10]).

Let $\varepsilon > 0$ fixed. Since $\mathcal{S}$ is a stratified submanifold of $\mathbb{R}^n$, there exists a neighborhood Ω of $\mathcal{S}$ satisfying the following property: for every $y \in \mathcal{S}$, there exists $z \in \mathbb{R}^n \setminus \mathrm{clos}(\Omega)$ such that $d_{SR}(y, z) < \varepsilon$.

Consider a stratum M_i of $\mathcal{S}$. For every $y \in M_i$, let $z \in \mathbb{R}^n \setminus \mathrm{clos}(\Omega)$ such that $d_{SR}(y, z) < \varepsilon$. The Lie algebra rank condition implies that there exists an open-loop control $t \mapsto u_y(t)$, defined on $[0, T)$ for a $T > \varepsilon$, satisfying the constraint $\|u_y\| \leqslant 1$, such that the associated trajectory $x_y(\cdot)$ (which can be assumed to be one-to-one), solution of (1.1), starting from y, reaches z (and thus leaves $\mathrm{clos}(\Omega)$) within time ε. Using a density argument, the control u_y can, moreover, be chosen as a smooth function (see [10, Theorem 2.8, p. 21] for the proof of this statement). Since the trajectory is one-to-one, the open-loop control $t \mapsto u_y(t)$ can be considered as a feedback $t \mapsto u_y(x_y(t))$ along $x_y(\cdot)$. Consider a smooth extension of u_y on $\mathbb{R}^n$, still denoted u_y, satisfying the constraint $\|u_y(x)\| \leqslant 1$ for every $x \in \mathbb{R}^n$. By continuity, there exists a neighborhood Ω_y of y, and positive real numbers δ_y and ρ_y, such that every solution of

$$\dot{x}(t) = f(x(t), u_y(x(t) + e(x(t), t))) + d(x(t), t), \tag{3.1}$$

where $e, d : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$ are two functions satisfying the regularity assumptions (2.2) and

$$\sup_{[0,+\infty)} |e(x, \cdot)| \leqslant \rho_y, \quad \operatorname{esssup}_{[0,+\infty)} |d(x, \cdot)| \leqslant \rho_y,$$

starting from Ω_y and maximally defined on $[0, T)$, leaves Ω within time ε; moreover,

$$|x(t) - \bar{x}| \leqslant \delta_y \ \forall t \in [0, T).$$

Repeat this construction for each $y \in M_i$.

Now, on the one hand, let $(y_p)_{p \in \mathcal{N}_i}$ be a sequence of points of M_i such that the family $(\Omega_{y_p})_{p \in \mathcal{N}_i}$ is a locally finite covering of M_i, where $\mathcal{N}_i$ is a subset of $\mathbb{N}$. Define $\Omega_{i,p} = \Omega_{y_p}$ and $u_{i,p} = u_{y_p}$.

On the other hand, the existence of a continuous function $\rho_{i,p} : \mathbb{R}^n \to [0, +\infty)$, satisfying $\rho_{i,p}(x) > 0$ whenever $x \neq x$, follows for the continuity of solutions with respect to disturbances. The existence of a function $\delta_{i,p}$ of class $\mathcal{K}_\infty$ such that (2.16) holds is obvious.

Repeat this construction for every stratum M_i of $\mathcal{S}$. Then the statement of the lemma follows. ∎

3.2. The second component of the hysteresis.

3.2.1. Computation of minimal time trajectories. Let $x_1 \in \mathbb{R}^n$, and let $x(\cdot)$ be a minimal time trajectory, associated to a control $u(\cdot)$, steering the system (1.1), (1.2), from $\bar{x}$ to x_1, in time $T = T_{\bar{x}}(x_1)$. According to Pontryagin's maximum principle (see [36]), the trajectory $x(\cdot)$ is the projection of an *extremal*, i.e., a triple $(x(\cdot), p(\cdot), u(\cdot))$ solution of the constrained Hamiltonian system

$$\dot{x}(t) = \frac{\partial H}{\partial p}(x(t), p(t), u(t)), \quad \dot{p}(t) = -\frac{\partial H}{\partial x}(x(t), p(t), u(t)),$$

$$H(x(t), p(t), p^0, u(t)) = \max_{\|v\| \leqslant 1} H(x(t), p(t), p^0, v),$$

almost everywhere on $[0, T]$, where

$$H(x, p, u) = \sum_{i=1}^{m} u_i \langle p, f_i(x) \rangle$$

is the *Hamiltonian* of the optimal control problem, and $p(\cdot)$ (called the *adjoint vector*) is an absolutely continuous mapping on $[0, T]$ such that $p(t) \in \mathbb{R}^n \setminus \{0\}$. Moreover, the function $t \mapsto \max_{\|v\| \leqslant 1} H(x(t), p(t), p^0, v)$ is Lipschitzian and everywhere constant on $[0, T]$. If this constant is not equal to zero, then the extremal is said to be *normal*; otherwise it is *abnormal*.

Remark 3.1. Any singular trajectory is the projection of an abnormal extremal, and conversely.

Controls associated to normal extremals can be computed as

$$u_i(t) = \frac{\langle p(t), f_i(x(t)) \rangle}{\sqrt{\sum_{j=1}^{m} \langle p(t), f_j(x(t)) \rangle^2}}, \ i = 1, \ldots, m. \tag{3.2}$$

Indeed, note that, by definition of normal extremals, the denominator of (3.2) cannot vanish. It follows that normal extremals are solutions of the Hamiltonian system

$$\dot{x}(t) = \frac{\partial H_1}{\partial p}(x(t), p(t)), \quad \dot{p}(t) = -\frac{\partial H_1}{\partial x}(x(t), p(t)), \tag{3.3}$$

where

$$H_1(x, p) = \sqrt{\sum_{i=1}^{m} \langle p, f_i(x) \rangle^2}.$$

Notice that $H_1(x(t), p(t))$ is constant, nonzero, along each normal extremal. Since $p(0)$ is defined up to a multiplicative scalar, it is usual to normalize it so that $H_1(x(t), p(t)) = 1$. Hence, we introduce the set

$$X = \{p_0 \in \mathbb{R}^n \mid H_1(\bar{x}, p_0) = 1\}.$$

It is a submanifold of $\mathbb{R}^n$ of codimension one, since $\frac{\partial H_1}{\partial p}(\bar{x}, p_0) = \dot{x}(0) \neq 0$ (see [12] for a similar construction in the general case). There exists a connected open subset

U of $[0,+\infty)\times X$ such that, for every $(t^*,p_0)\in X$, the differential system (3.3) has a well-defined smooth solution on $[0,t^*]$ such that $x(0)=\bar{x}$ and $p(0)=p_0$.

DEFINITION 3.2. *The smooth mapping*

$$\begin{array}{rccc}\exp_{\bar{x}}: & U & \longrightarrow & \mathbb{R}^n\\ & (t,p_0) & \longmapsto & x(t),\end{array}$$

where $(x(\cdot),p(\cdot))$ is the solution of the system (3.3) *such that $x(0)=\bar{x}$ and $p(0)=p_0\in X$, is called* exponential mapping *at the point $\bar{x}$.*

The exponential mapping parameterizes normal extremals. Note that the domain of $\exp_{\bar{x}}$ is a subset of $\mathbb{R}\times X$ which is locally diffeomorphic to $\mathbb{R}^n$ (since we are in the normal case).

DEFINITION 3.3. *A point $x\in\exp_{\bar{x}}(U)$ is said to be* conjugate *to $\bar{x}$ if it is a critical value of the mapping $\exp_{\bar{x}}$, i.e., if there exists $(t_c,p_0)\in U$ such that $x=\exp_{\bar{x}}(t_c,p_0)$ and the differential $d\exp_{\bar{x}}(t_c,p_0)$ is not onto. The* conjugate locus *of $\bar{x}$, denoted by $\mathcal{C}(\bar{x})$, is defined as the set of all points conjugate to $\bar{x}$.*

With the previous notation, define $\mathcal{C}_{min}(\bar{x})$ as the set of points $x\in\mathcal{C}(\bar{x})$ such that the trajectory $t\mapsto\exp_{\bar{x}}(t,p_0)$ is minimizing between $\bar{x}$ and x.

3.2.2. The cut locus. A standard definition is the following.

DEFINITION 3.4. *A point $x\in\mathbb{R}^n$ is not a* cut point *with respect to $\bar{x}$ if there exists a minimal time trajectory of* (1.1), (1.2), *joining $\bar{x}$ to x, which is the strict restriction of a minimal time trajectory starting from $\bar{x}$. The* cut locus *of $\bar{x}$, denoted by $\mathcal{L}(\bar{x})$, is defined as the set of all cut points with respect to $\bar{x}$.*

In other words, a cut point is a point at which a minimal time trajectory ceases to be optimal.

Remark 3.5. In the analytic case, it follows from the theory of conjugate points that every nonsingular minimal time trajectory ceases to be minimizing beyond its first conjugate point (see, for instance, [3, 13]). Hence, if there exists no singular minimal time trajectory starting from $\bar{x}$, then $\mathcal{C}_{min}(\bar{x})\subset\mathcal{L}(\bar{x})$.

The following result on the cut locus is crucial for the proof of Theorem 2.10.

PROPOSITION 3.6. *Assume that the vector fields $f_1,\dots,f_m$ are analytic, and that there exists no singular minimal time trajectory starting from $\bar{x}$. Then the set of points of $\mathbb{R}^n$ where the function $T_{\bar{x}}(\cdot)$ is not analytic is equal to the cut locus of $\bar{x}$, that is,*

$$\operatorname{Sing} T_{\bar{x}}(\cdot)=\mathcal{L}(\bar{x}). \tag{3.4}$$

Remark 3.7. Under the previous assumptions, one can prove that the set of points of $\mathbb{R}^n$ where $T_{\bar{x}}(\cdot)$ is analytic is equal to the set of points where $T_{\bar{x}}(\cdot)$ is of class C^1.

Proof. Let $x\in\mathbb{R}^n$ so that $T_{\bar{x}}(\cdot)$ is analytic at x. Then there exists a neighborhood V of x in $\mathbb{R}^n$ such that $T_{\bar{x}}(\cdot)$ is analytic on V. Let us prove that $x\notin\mathcal{L}(\bar{x})$. It follows from the maximum principle and the Hamilton–Jacobi theory (see [36]) that, for every $y\in V$, there exists a unique minimal time trajectory joining $\bar{x}$ to y having, moreover, a normal extremal lift $(x(\cdot),p(\cdot),u(\cdot))$ satisfying

$$p(T_{\bar{x}}(y))=\nabla T_{\bar{x}}(y)$$

(compare with [44, Proposition 2.3]). Set $U_1=\exp_{\bar{x}}^{-1}(V)$. It follows easily from the Cauchy–Lipschitz theorem that the mapping $\exp_{\bar{x}}$ is an analytic diffeomorphism from U_1 into V. Hence, obviously, the point x is not in the cut locus of $\bar{x}$.

Conversely, let $x \notin \mathcal{L}(\bar{x})$. To prove that $T_{\bar{x}}(\cdot)$ is analytic at x, we need the following two lemmas.

LEMMA 3.8. *The point x is not conjugate to $\bar{x}$ and is joined from $\bar{x}$ by a unique minimal time trajectory.*

Proof of Lemma 3.8. From the assumption of the absence of a singular minimal time trajectory, there exists a nonsingular minimal time trajectory joining $\bar{x}$ to x. From Remark 3.5, the point x is not conjugate to $\bar{x}$.

By contradiction, suppose that x is joined from $\bar{x}$ by at least two minimal time trajectories. By assumption, these two trajectories must admit normal extremal lifts. Since the structure is analytic, their junction at the point x is necessarily not smooth. This implies that both trajectories lose their optimality at the point x (indeed if not, there would exist a nonsmooth normal extremal, which is absurd), and thus $x \in \mathcal{L}(\bar{x})$. This is a contradiction. □

LEMMA 3.9. *There exists a neighborhood V of x in $\mathbb{R}^n$ such that every point $y \in V$ is not conjugate to $\bar{x}$, and there exists a unique (nonsingular) minimal time trajectory joining $\bar{x}$ to y.*

Proof of Lemma 3.9. Let $p_0 \in X$ so that $x = \exp_{\bar{x}}(T_{\bar{x}}(x), p_0)$. Since x is not conjugate to $\bar{x}$, the exponential mapping $\exp_{\bar{x}}$ is a diffeomorphism from a neighborhood U_1 of $(T_{\bar{x}}(x), p_0)$ in U into a neighborhood V of x in $\mathbb{R}^n$. Set $U_2 = \exp_{\bar{x}}^{-1}(V)$.

Let us prove that $\exp_{\bar{x}}$ is proper from U_2 into V. We argue by contradiction and suppose that there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of points of V converging towards x, such that for each integer n there exists $p_n \in X$, satisfying $(T_{\bar{x}}(x_n), p_n) \in U_2$ and $x_n = \exp_{\bar{x}}(T_{\bar{x}}(x_n), p_n)$, such that $(p_n)_{n\in\mathbb{N}}$ is not bounded. It then follows from [51, Lemmas 4.8 and 4.9] (see also [52, Fact 1, p. 378]) that x is joined from $\bar{x}$ by a singular control u. In particular, x is conjugate to $\bar{x}$; this is a contradiction.

Therefore, the set $\{p \mid \exp_{\bar{x}}(T_{\bar{x}}(x), p) = x\}$ is compact in U_2. Moreover, since x is not conjugate to $\bar{x}$, this set has no cluster point, and thus is finite. As a consequence, up to reducing V, we assume that V is a connected open subset of $\exp_{\bar{x}}(U_2)$, and that U_2 is a finite union of disjoint connected open sets, all of which are diffeomorphic to V by the mapping $\exp_{\bar{x}}$. We infer that every point $y \in V$ is not conjugate to $\bar{x}$. Hence, the mapping $\exp_{\bar{x}}$ is a proper submersion from U_2 into V, and thus is a fibration with finite degree. Since, from Lemma 3.8, there exists a unique minimal time trajectory joining $\bar{x}$ to x, this degree is equal to one, that is, $\exp_{\bar{x}}$ is a diffeomorphism from U_2 into V. The conclusion follows. □

It follows from the previous lemma that $(T_{\bar{x}}(y), p_0) = \exp_{\bar{x}}^{-1}(y)$ for every $y \in V$, and hence $T_{\bar{x}}(\cdot)$ is analytic on V. □

3.2.3. Definition of the optimal controller. By assumption, there does not exist any nontrivial singular minimal time trajectory starting from $\bar{x}$. Under these assumptions, the function $T_{\bar{x}}(\cdot)$ is *subanalytic* outside $\bar{x}$ (see [1, 2, 50], combined with Remark 2.4).

For the sake of completeness, we recall below the definition of a subanalytic function (see [27, 28]) and some properties that are used in a crucial way in the present paper (see [48]).

Let M be a real analytic finite-dimensional manifold. A subset A of M is said to be *semianalytic* if and only if, for every $x \in M$, there exists a neighborhood U of x in M and $2pq$ analytic functions g_{ij}, h_{ij} $(1 \leqslant i \leqslant p$ and $1 \leqslant j \leqslant q)$, such that

$$A \cap U = \bigcup_{i=1}^{p} \{y \in U \mid g_{ij}(y) = 0 \text{ and } h_{ij}(y) > 0,\ j = 1, \dots, q\}.$$

Let SEM(M) denote the set of semianalytic subsets of M. The image of a semianalytic subset by a proper analytic mapping is not in general semianalytic, and thus this class has to be enlarged.

A subset A of M is said to be *subanalytic* if and only if, for every $x \in M$, there exist a neighborhood U of x in M and $2p$ couples $(\Phi_i^\delta, A_i^\delta)$ $(1 \leqslant i \leqslant p$ and $\delta = 1, 2)$, where $A_i^\delta \in SEM(M_i^\delta)$, and where the mappings $\Phi_i^\delta : M_i^\delta \to M$ are proper analytic, for real analytic manifolds M_i^δ, such that

$$A \cap U = \bigcup_{i=1}^{p} (\Phi_i^1(A_i^1) \backslash \Phi_i^2(A_i^2)).$$

Let SUB(M) denote the set of subanalytic subsets of M.

The subanalytic class is closed by union, intersection, complementary, inverse image by an analytic mapping, and image by a proper analytic mapping. In brief, the subanalytic class is *o-minimal* (see [23]). Moreover subanalytic sets are *stratifiable* in the following sense. A *stratum* of a differentiable manifold M is a locally closed submanifold of M. A locally finite partition $\mathcal{S}$ of M is a *stratification* of M if any $S \in \mathcal{S}$ is a stratum such that

$$\forall T \in \mathcal{S} \quad T \cap \partial S \neq \emptyset \Rightarrow T \subset \partial S \quad \text{and} \quad \dim T < \dim S.$$

Finally, a mapping $f : M \to N$ between two analytic manifolds is said to be *subanalytic* if its graph is a subanalytic subset of $M \times N$.

Let M be an analytic manifold, and let f be a subanalytic function on M. The *analytic singular support* of f is defined as the complement of the set of points x in M such that the restriction of f to some neighborhood of x is analytic. The following property is of great interest in the present paper (see [48]): the analytic singular support of f is subanalytic (and thus, in particular, is stratifiable). If f is, moreover, locally bounded on M, then it is of codimension greater than or equal to one.

Let us turn back to our problem. The function $T_{\bar{x}}(\cdot)$ is subanalytic outside $\bar{x}$, and, hence, its singular set $\mathcal{S} = \text{Sing } T_{\bar{x}}(\cdot)$ (i.e., the analytic singular support of $T_{\bar{x}}(\cdot)$) is a stratified submanifold of $\mathbb{R}^n$, of codimension greater than or equal to 1.

Remark 3.10. Note that the point $\bar{x}$ belongs to the adherence of $\mathcal{S}$ (see [1]).

Outside the singular set $\mathcal{S}$, it follows from the dynamic programming principle (see [36]) that the minimal time controllers steering a point $x \in \mathbb{R}^n \setminus \mathcal{S}$ to $\bar{x}$ are given by the closed-loop formula

$$u_i(x) = -\frac{\langle \nabla T_{\bar{x}}(x), f_i(x) \rangle}{\sqrt{\sum_{j=1}^m \langle \nabla T_{\bar{x}}(x), f_j(x) \rangle^2}}, \ i = 1, \ldots, m. \tag{3.5}$$

The objective is to construct neighborhoods of $\mathcal{S} \setminus \{\bar{x}\}$ in $\mathbb{R}^n$ whose complements share invariance properties for the optimal flow. This is the contents of Lemma 2.14, proved next.

3.2.4. Proof of Lemma 2.14. It suffices to prove that, for every compact subset K of $\mathbb{R}^n$ and for every neighborhood Ω of $\mathcal{S} \setminus \{\bar{x}\}$ in $\mathbb{R}^n$, there exists a neighborhood Ω' of $\mathcal{S} \setminus \{\bar{x}\}$ in $\mathbb{R}^n$, satisfying (2.17), such that every trajectory of the closed-loop system (1.1) with the optimal controller, joining a point $x \in (\mathbb{R}^n \setminus \Omega) \cap K$ to $\bar{x}$, is contained in $\mathbb{R}^n \setminus \Omega'$.

By definition of the cut locus, and using Proposition 3.6, every optimal trajectory joining a point $x \in (\mathbb{R}^n \setminus \Omega) \cap K$ to $\bar{x}$ does not intersect $\mathcal{S}$, and thus has a positive

distance to the set $\mathcal{S}$. Using the assumption of the absence of nontrivial singular minimizing trajectories starting from $\bar{x}$, a reasoning similar to the proof of Lemma 3.9 proves that the optimal flow joining points of the compact set $(\mathbb{R}^n \setminus \Omega) \cap K$ to $\bar{x}$ is parameterized by a compact set. Hence, there exists a positive real number $\delta > 0$ so that every optimal trajectory joining a point $x \in (\mathbb{R}^n \setminus \Omega) \cap K$ to $\bar{x}$ has a distance to the set $\mathcal{S}$ which is greater than or equal to δ. The existence of Ω' follows.

3.2.5. Robustness properties of the optimal controller. In this section, we prove robustness properties of the Carathéodory solutions of system (2.1) in closed-loop with this feedback optimal controller. Given $e, d : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$, the perturbed system in closed-loop with the optimal controller (denoted u_{opt}) is

$$\dot{x}(t) = f(x(t), u_{opt}(x(t) + e(x(t), t))) + d(x(t), t). \tag{3.6}$$

Since the optimal controller is continuous outside the singular set $\mathcal{S}$, it enjoys a natural robustness property, stated below. In the next result, the notation $d(x, \mathcal{S})$ stands for the Euclidean distance from x to $\mathcal{S}$.

LEMMA 3.11. *There exist a continuous function $\rho_{opt} : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\rho_{opt}(\xi) > 0 \ \forall \xi \neq 0 \tag{3.7}$$

and a continuous function $\delta_{opt} : [0, +\infty) \to [0, +\infty)$ of class $\mathcal{K}_\infty$ such that the following three properties hold:

- Robust Stability.
  For every neighborhood Ω of $\mathcal{S}$, there exists a neighborhood $\Omega' \subset \Omega$ of $\mathcal{S}$, such that, for all $e, d : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$ satisfying the regularity assumptions (2.2) *and, for every $x \in \mathbb{R}^n$,*

  $$\sup_{[0,+\infty)} |e(x, \cdot)| \leqslant \rho_{opt}(d(x, \mathcal{S})), \ \operatorname{esssup}_{[0,+\infty)} |d(x, \cdot)| \leqslant \rho_{opt}(d(x, \mathcal{S})), \tag{3.8}$$

  and for every $x_0 \in \mathbb{R}^n \setminus \Omega$, there exists a unique Carathéodory solution $x(\cdot)$ of (3.6) *starting from x_0, maximally defined on $[0, +\infty)$, and satisfying $x(t) \in \mathbb{R}^n \setminus \Omega'$ for every $t > 0$.*
- Finite time convergence.
  For every $R > 0$, there exists $\tau_{opt} = \tau_{opt}(R) > 0$ such that, for all $e, d : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$ satisfying the regularity assumptions (2.2) *and* (3.8)*, for every $x_0 \in \mathbb{R}^n$ with $|x_0 - \bar{x}| \leqslant R$, and every maximal solution $x(\cdot)$ of* (3.6) *starting from x_0, one has*

  $$|x(t) - \bar{x}| \leqslant \delta_{opt}(R) \ \forall t \geqslant 0, \tag{3.9}$$

  $$x(t) = \bar{x} \ \forall t \geqslant \tau_{opt}, \tag{3.10}$$

  and

  $$\|u_{opt}(x(t))\| \leqslant 1 \ \forall t \geqslant 0. \tag{3.11}$$

- Optimality.
  For every neighborhood Ω of $\mathcal{S}$, every $\varepsilon > 0$, and every compact subset K of $\mathbb{R}^n$, there exists a continuous function $\rho_{\varepsilon,K} : \mathbb{R}^n \to \mathbb{R}$ satisfying (2.9) *such*

that, for all $e, d : \mathbb{R}^n \times [0, +\infty) \to \mathbb{R}^n$ *satisfying the regularity assumptions* (2.2) *and*

$$\begin{aligned} &\sup_{[0,+\infty)}|e(x,\cdot)| \leqslant \min(\rho_{opt}(d(x,\mathcal{S})), \rho_{\varepsilon,K}(x)), \\ &\operatorname{esssup}_{[0,+\infty)}|d(x,\cdot)| \leqslant \min(\rho_{opt}(d(x,\mathcal{S}), \rho_{\varepsilon,K}(x)) \ \forall x \in \mathbb{R}^n, \end{aligned} \tag{3.12}$$

and for every $x_0 \in K \cap (\mathbb{R}^n \setminus \Omega)$*, the solution of* (3.6)*, starting from* x_0*, reaches* $\bar{x}$ *within time* $T_{\bar{x}}(x_0) + \varepsilon$*.*

Proof. Since Carathéodory conditions hold for the system (3.6), the existence of a unique Carathéodory solution of (3.6), for every initial condition, is ensured. Inequality (3.11) follows from the constraint (1.2). Since the optimal controller u_{opt} defined by (3.5) is continuous on $\mathbb{R}^n \setminus \mathcal{S}$, Lemma 2.14 implies the existence of $\rho_{opt} : \mathbb{R}^n \to [0, +\infty)$ so that the *robust stability* and the *finite time convergence* properties hold.

The so-called *optimality* property follows from the definition of u_{opt}, from the continuity of solutions with respect to disturbances, and from the compactness of the set of all solutions starting from $K \cup (\mathbb{R}^n \setminus \Omega)$. ∎

3.3. Definition of the hybrid feedback law. A switching strategy must be defined in order to connect the first component (optimal controller), and the second component (consisting of a set of controllers, stated in Lemma 2.13). The switching strategy is achieved by adding a dynamical discrete variable s_d and using a hybrid feedback law, described next.

3.3.1. A class of hybrid feedbacks. Let $\mathcal{F} = \{1, \dots, 6\}$, and let $\mathcal{N}$ be a countable set. In what follows, Greek letters refer to elements of $\mathcal{N}$. Fix ω as an element of $\mathcal{N}$. We emphasize that we do not introduce any order in $\mathcal{N}$.

Given a set-valued map $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, we define the solutions $x(\cdot)$ of the differential inclusion $\dot{x} \in F(x)$ as all absolutely continuous functions satisfying $\dot{x}(t) \in F(x(t))$ almost everywhere.

DEFINITION 3.12. *The family* $(\mathbb{R}^n \setminus \{\bar{x}\}, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{N}})$ *is said to satisfy the property* $(\mathcal{P})$ *if*

1. *for every* $(\alpha, l) \in \mathcal{N} \times \mathcal{F}$*, the set* $\Omega_{\alpha,l}$ *is an open subset of* $\mathbb{R}^n$*;*
2. *for every* $\alpha \in \mathcal{N}$ *and every* $m > l \in \mathcal{F}$*,*

$$\Omega_{\alpha,l} \subsetneq \operatorname{clos}(\Omega_{\alpha,l}) \subsetneq \Omega_{\alpha,m}; \tag{3.13}$$

3. *for every* α *in* $\mathcal{N}$*,* g_α *is a smooth vector field, defined in a neighborhood of* $\operatorname{clos}(\Omega_{\alpha,6})$*, taking values in* $\mathbb{R}^n$*;*
4. *for every* $(\alpha, l) \in \mathcal{N} \times \mathcal{F}$*,* $l \leqslant 5$*, there exists a continuous function* $\rho_{\alpha,l} : \mathbb{R}^n \to [0, +\infty)$ *satisfying* $\rho_{\alpha,l}(x) \neq 0$ *whenever* $x \neq \bar{x}$ *such that every maximal solution* $x(\cdot)$ *of*

$$\dot{x} \in g_\alpha(x) + B(0, \rho_{\alpha,l}(x)), \tag{3.14}$$

defined on $[0, T)$ *and starting from* $\partial\Omega_{\alpha,l}$*, is such that*

$$x(t) \in \operatorname{clos}(\Omega_{\alpha,l+1}) \ \forall t \in [0, T);$$

5. *for every* $l \in \mathcal{F}$*, the sets* $(\Omega_{\alpha,l})_{\alpha\in\mathcal{N}}$ *form a locally finite covering of* $\mathbb{R}^n \setminus \{\bar{x}\}$*.*

Remark 3.13. Some observations are in order.

- First note that this notion is close to the notion of a family of nested patchy vector fields defined in [38]. However, note that, in general, the sets $(\Omega_{\alpha,l}, g_\alpha)$ may not be a patch as defined in [4, 38]. Indeed, due to property 4, the set $\Omega_{\alpha,l}$ may not be invariant for the system (3.14). Since the notion of a patch is one of the main ingredients of the proofs of [40], we cannot apply [40] directly, even though some notions are similar (see in particular Definition 3.14 below).
- To state our main result, we need a family of six nested patchy vector fields (see Remark 3.21 for comments on the necessity of six families). The patches 1, 2, 4, and 6 define the dynamics of the discrete component of our hybrid controller (see Definition 3.14 below). The patches 3 and 5 are used for technical reasons to handle the measurement noise.

We next define a class of hybrid controllers as those considered in section 2 (see also [40]).

DEFINITION 3.14. *Let $(\mathbb{R}^n \setminus \{\bar{x}\}, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{N}})$ satisfy the property $(\mathcal{P})$ as in Definition* 3.12. *Assume that, for every α in $\mathcal{N}$, there exists a smooth function k_α defined in a neighborhood of $\Omega_{\alpha,6}$ and taking values in $\mathbb{R}^m$, such that, for every x in a neighborhood of $\Omega_{\alpha,6}$,*

$$g_\alpha(x) = f(x, k_\alpha(x)). \tag{3.15}$$

Set

$$D_1 = \Omega_{\omega,2}, \tag{3.16}$$

$$D_{\alpha,2} = \mathbb{R}^n \setminus \Omega_{\alpha,6}. \tag{3.17}$$

Let (C, D, k, k_d) be the hybrid feedback defined by

$$C = \{(x,\alpha) \mid x \in (\mathrm{clos}(\Omega_{\alpha,4}) \setminus \Omega_{\omega,1})\}, \tag{3.18}$$

$$D = \{(x,\alpha) \mid x \in D_1 \cup D_{\alpha,2}\}, \tag{3.19}$$

$$\begin{array}{rccll} k : \mathbb{R}^n \times \mathcal{N} & \to & \mathbb{R}^m & & \\ (x,\alpha) & \mapsto & k_\alpha(x) & \text{if} & x \in \Omega_{\alpha,6}, \\ & & 0 & \text{else}, & \end{array} \tag{3.20}$$

and

$$\begin{array}{rccll} k_d : \mathbb{R}^n \times \mathcal{N} & \rightrightarrows & \mathcal{N} & & \\ (x,\alpha) & \mapsto & \omega & \text{if } x \in \mathrm{clos}(\Omega_{\omega,1} \cap D_1) \text{ and if } x \notin D_{\alpha,2}, \\ & & \alpha' & \text{if } x \in \mathrm{clos}(\Omega_{\alpha',1} \cap D_{\alpha,2}). & \end{array} \tag{3.21}$$

The 4-tuple (C, D, k, k_d) is a hybrid feedback law on $\mathbb{R}^n$ as considered in section 2.2. *We denote by $\mathcal{H}_{(e,d)}$ the system* (2.1) *in closed-loop with such feedback with the perturbations e and d as measurement noise and external disturbance, respectively.*

In this definition, we do not use any order in $\mathcal{N}$. However, the element ω has a particular role in what follows, since it will refer to the optimal controller in the hybrid feedback law.

3.3.2. Properties of solutions. In this section, we investigate some properties of the solutions of the system in closed-loop with the hybrid feedback law defined above.

DEFINITION 3.15. *Let $\chi : \mathbb{R}^n \to \mathbb{R}$ be a continuous map such that $\chi(x) > 0$ for every $x \neq \bar{x}$.*

- *We say that χ is an admissible radius for the measurement noise if, for every $x \in \mathbb{R}^n$ and every $\alpha \in \mathcal{N}$, such that $x \in \Omega_{\alpha,6}$,*

$$\chi(x) < \frac{1}{2} \min_{l \in \{1,\dots,5\}} d(\mathbb{R}^n \setminus \Omega_{\alpha,l+1}, \Omega_{\alpha,l}). \tag{3.22}$$

- *We say that χ is an admissible radius for the external disturbances if, for every $x \in \mathbb{R}^n$, we have*

$$\chi(x) \leqslant \max_{(\alpha,l),\ x \in \Omega_{\alpha,l}} \rho_{\alpha,l}(x).$$

There exists an admissible radius for the measurement noise and for the external disturbances (note that, from (3.13), the right-hand side of inequality (3.22) is positive).

Consider an admissible radius χ for the measurement noise and the external disturbances. Let e and d be a measurement noise and an external disturbance respectively, such that, for all $(x,t) \in \mathbb{R}^n \times [0,+\infty)$,

$$e(x,t) \leqslant \chi(x), \quad d(x,t) \leqslant \chi(x). \tag{3.23}$$

The properties of the solutions of the system in closed-loop with the hybrid feedback law defined in Definition 3.14 are similar to those of [40]. Hence, we skip the proof of the following three lemmas, which do not use statement 4 of Definition 3.12 but only the definition of the hybrid feedback law.

LEMMA 3.16. *For all $(x_0, s_0) \in \mathbb{R}^n \times \mathcal{N}$, there exists a solution of $\mathcal{H}_{(e,d)}$ starting from (x_0, s_0).*

Recall that a Zeno solution is a complete solution whose domain of definition is bounded in the t-direction. A solution (x, s_d), defined on a hybrid domain S, is an instantaneous Zeno solution if there exist $t \geqslant 0$ and an infinite number of $j \in \mathbb{N}$ such that $(t,j) \in S$.

The Zeno solutions do not require a special treatment.

LEMMA 3.17. *There do not exist instantaneous Zeno solutions, although a finite number of switches may occur at the same time.*

We note, as usual, that maximal solutions of $\mathcal{H}_{(e,d)}$ blow up if their domain of definition is bounded. Since Zeno solutions are avoided, the blow-up phenomenon only concerns the t-direction of the domain of definition, and we get the following result (see also [25, Proposition 2.1]).

LEMMA 3.18. *Let (x, s_d) be a maximal solution of $\mathcal{H}_{(e,d)}$ defined on a hybrid time S. Suppose that the supremum T of S in the t-direction is finite. Then*

$$\limsup_{t \to T,\,(t,l) \in S} |x(t,l)| = +\infty.$$

We conclude this series of technical lemmas by studying the behavior of solutions between two jumps. For every $\alpha \in \mathcal{N}$, set

$$\begin{aligned} \tau_\alpha = \sup\big\{ T \mid x \text{ is a Carathéodory solution of } \dot{x} = f(x,k_\alpha) + B(0,\chi(x)) \\ \text{with } x(t) \in \Omega_{\alpha,6}\ \forall t \in [0,T) \big\}. \end{aligned} \tag{3.24}$$

Note that, at this stage, there may exist $\alpha \in \mathcal{N}$ such that $\tau_\alpha = +\infty$.

LEMMA 3.19. *Let (x, s_d) be a solution of $\mathcal{H}_{(e,d)}$ defined on a hybrid time domain S and starting in $(\mathbb{R}^n \setminus \{\bar{x}\}) \times \mathcal{N}$. Let T be the supremum in the t-direction of S. Then one of the two following cases may occur:*

- *Either there exists no positive jump time—more precisely, there exists* $\alpha \in \mathcal{N}$ *such that*
  1. *for almost every* $t \in (0,T)$ *and for every* l *such that* $(t,l) \in S$*, one has* $k(s_d(t,l)) = k_\alpha$*;*
  2. *for every* $t \in (0,T)$ *and every* l *such that* $(t,l) \in S$*, one has* $x(t,l) + e(x(t,l),t) \in \mathrm{clos}(\Omega_{\alpha,4}) \setminus \Omega_{\omega,1}$*;*
  3. *for all* $(t,l) \in S$*,* $t > 0$*, one has* $x(t,l) + e(x(t,l),t) \notin D$*, where* D *is defined by* (3.19)*;*
  4. *the inequality* $T < \tau_\alpha$ *holds;*
- *or there exists a unique positive jump time—more precisely, there exist* $\alpha \in \mathcal{N} \setminus \{\omega\}$ *and* $t_1 \in (0,T)$ *such that, letting* $t_0 = 0$*,* $t_2 = T$*,* $\alpha_0 = \alpha$*,* $\alpha_1 = \omega$*, for every* $j = 0,1$*, the following properties hold:*
  5. *for almost every* $t \in (t_j, t_{j+1})$ *and for every* l *such that* $(t,l) \in S$*, one has* $k(s_d(t,l)) = k_{\alpha_j}$*;*
  6. *for every* $t \in (t_0,t_1)$ *and every* l *such that* $(t,l) \in S$*, one has* $x(t,l) + e(x(t,l),t) \in \mathrm{clos}(\Omega_{\alpha,4}) \setminus \Omega_{\omega,1}$*;*
  7. *for every* t *in* (t_j, t_{j+1}) *and every* l *such that* $(t,l) \in S$*, one has* $x(t,l) + e(x(t,l),t) \notin D_{\alpha_j,2}$*, where* $D_{\alpha_j,2}$ *is defined by* (3.17)*;*
  8. *the inequality* $t_1 < \tau_{\alpha_j}$ *holds.*

Proof. Consider the sequence $(t_j)_{j \in \overline{m}}$ of jump times, i.e., the times such that $t_0 = 0$ and, for every $j \in \overline{m}$, $j \leqslant m-1$,

$$t_j \leqslant t_{j+1}, \tag{3.25}$$

$$(x(t_{j+1},j) + e(x(t_{j+1},j),t_{j+1}), s_d(t_{j+1},j)) \in D, \tag{3.26}$$

and

$$(x(t_{j+1},j+n_j) + e(x(t_{j+1},j),t_{j+1}), s_d(t_{j+1},j+n_j)) \in C, \tag{3.27}$$

where n_j is the finite number of instantaneous switches[3] (see Lemma 3.17). Let $\sigma : \mathbb{N} \to \mathbb{N}$ be an increasing function such that $t_{\sigma(j)} < t_{\sigma(j+1)}$.

Between two jumps, $s_d(t)$ is constant, and thus there exists a sequence (α_j) in $\mathcal{N}$ such that, for every $t \in (t_{\sigma(j)}, t_{\sigma(j+1)})$, except for a finite number of t, we have

$$s_d(t,\sigma(j)) = \alpha_j, \tag{3.28}$$

$$x \text{ is a Carathéodory solution of } \dot{x} = f(x,k_{\alpha_j}) + d \text{ on } (t_{\sigma(j)}, t_{\sigma(j+1)}), \tag{3.29}$$

and

$$k(s_d(t,\sigma(j))) = k_{\alpha_j}. \tag{3.30}$$

From (3.18), (3.27), and (3.28), we have, for every $t \in [t_{\sigma(j)}, t_{\sigma(j+1)}]$,

$$x(t,j) + e(x(t,j),t) \in \mathrm{clos}(\Omega_{\alpha_j,4}) \setminus \Omega_{\omega,1}. \tag{3.31}$$

Note that, from (3.30), (3.31), and statement 4 of Definition 3.12, for every $t > 0$ such that $t \in [t_{\sigma(j)}, t_{\sigma(j+1)}]$, one has

$$x(t,\sigma(j)) \notin D_{\alpha_j,2}. \tag{3.32}$$

[3]We emphasize that we may have jumps at time $t = 0$ (see also Remark 3.21).

Therefore, the positive jump time may occur only at time t_j where the point $x(t_j, l) + e(x(t_j, l), t_j)$ belongs to D_1. Thus, there exists at most one positive jump time. From (3.29) and (3.30), Statements 1 and 5 hold. Statements 2 and 6 are deduced from (3.31). Equation (3.32) implies Statements 3 and 7 Finally, Statements 4 and 8 are a consequence of (3.24) and (3.29). □

3.3.3. Definition of the hybrid feedback law, and switching strategy. We next define our hybrid feedback law. Let $\varepsilon > 0$ and K be a compact subset of $\mathbb{R}^n$. Let Ω be the neighborhood of $\mathcal{S}$ given by Lemma 2.13. For this neighborhood Ω, let $\Omega' \subset \Omega$ be the neighborhood of $\mathcal{S}$ yielded by Lemma 2.14.

Let $\mathcal{N}$ be the countable set defined by

$$\mathcal{N} = \{(i, p),\ i \in \mathbb{N},\ p \in \mathcal{N}_i\} \cup \{\omega\},$$

where ω is an element of $\mathbb{N} \times \mathbb{N}$, distinct from every (i, p), $i \in \mathbb{N}$, $p \in \mathcal{N}_i$.

We proceed in two steps. We first define k_α and $\Omega_{\alpha,l}$, where $\alpha \in \mathcal{N} \setminus \{\omega\}$ and $l \in \mathcal{F}$. Then we define k_ω and $\Omega_{\omega,l}$, where $l \in \mathcal{F}$.

1. Let $i \in \mathbb{N}$. Lemma 2.13, applied with the stratum M_i, implies the existence of a family of smooth controllers $(k_{i,p})_{p\in\mathcal{N}_i}$ satisfying the constraint (1.2) and of a family of neighborhoods $(\Omega_{i,p,7})_{p\in\mathcal{N}_i}$. The existence of the families $(\Omega_{i,p,1})_{p\in\mathcal{N}_i}, \dots, (\Omega_{i,p,6})_{p\in\mathcal{N}_i}$, satisfying

$$\Omega_{i,p,l} \subsetneq \operatorname{clos}(\Omega_{i,p,l}) \subsetneq \Omega_{i,p,m}$$

for every $m > l \in \mathcal{F}$, follows from a finite induction argument, using Lemma 2.13.

We have thus defined $k_{i,p}$ and $\Omega_{i,p,l}$, where $(i, p) \in \mathcal{N} \setminus \{\omega\}$ and $l \in \mathcal{F}$.

Remark 3.20. It follows from [1] that, near the point $\bar{x}$, the cut locus $\mathcal{S}$ is contained in a conic neighborhood $\mathcal{C}$ centered at $\bar{x}$ (as illustrated in Figure 2.1), the axis of the cone being transversal to the subspace $\operatorname{Span}\{f_1(\bar{x}), \dots, f_m(\bar{x})\}$. Moreover, we may modify slightly the previous construction, and assume that, near $\bar{x}$, the set $\bigcup_{\alpha\in\mathcal{N}\setminus\{\omega\},\ l\in\mathcal{F}} \Omega_{\alpha,l}$ is contained in this conic neighborhood.

2. It remains to define the sets $\Omega_{\omega,l}$, where $l \in \mathcal{F}$, and the controller k_ω. Let $\Omega_{\omega,1}$ be an open set of $\mathbb{R}^n$ containing $\mathbb{R}^n \setminus \bigcup_{\alpha\in\mathcal{N}\setminus\{\omega\}} \Omega_{\alpha,1}$ and contained in $\mathbb{R}^n \setminus \mathcal{S}$. From the previous remark, the point $\bar{x}$ belongs to $\operatorname{clos}(\Omega_{\omega,1})$. Lemma 2.14, applied with $\Omega = \mathbb{R}^n \setminus \operatorname{clos}(\Omega_{\omega,1})$, allows to define k_ω as k_{opt}, and Ω' a closed subset of $\mathbb{R}^n$ such that

$$\Omega' \subsetneq \Omega, \tag{3.33}$$

and such that Ω' is a neighborhood of $\mathcal{S}$. Set $\Omega_{\omega,2} = \mathbb{R}^n \setminus \Omega'$; it is an open subset of $\mathbb{R}^n$, contained in $\mathbb{R}^n \setminus \mathcal{S}$. Moreover, from (3.33), $\Omega_{\omega,1} \subsetneq \operatorname{clos}(\Omega_{\omega,1}) \subsetneq \Omega_{\omega,2}$. The existence of the sets $\Omega_{\omega,3}, \dots, \Omega_{\omega,6}$ follows from a finite induction argument, using Lemma 2.14. Moreover, from Lemma 3.11, we have the following property: for every $l \in \{1, \dots, 6\}$, for every $x_0 \in \Omega_{\omega,l}$, the unique Carathéodory solution $x(\cdot)$ of (3.6), with $x(0) = x_0$, satisfies $x(t) \in \Omega_{\omega,l+1}$ for every $t \geqslant 0$.

Therefore, $(\mathbb{R}^n \setminus \{\bar{x}\}, ((\Omega_{\alpha,l})_{l\in\mathcal{F}}, g_\alpha)_{\alpha\in\mathcal{N}})$ is a family satisfying the property $(\mathcal{P})$ as in Definition 3.12, where g_α is a function defined in a neighborhood of $\Omega_{\alpha,6}$ by $g_\alpha(x) = f(x, k_\alpha)$. The hybrid feedback law (C, D, k, k_d) is then defined according to Definition 3.14.

3.4. Proof of Theorem 2.10. Let $\varepsilon > 0$, and let K be a compact subset of $\mathbb{R}^n$. Consider the hybrid feedback law (C, D, k, k_d) defined previously. Let χ be an admissible radius for the external disturbances and the measurement noise (see Definition 3.15). We may reduce this function, and assume that, for every $\alpha \in \mathcal{N} \setminus \{\omega\}$,

$$\chi(x) \leqslant \rho_{opt}(d(x, \mathcal{S})) \quad \forall x \in \Omega_{\omega,6},$$
$$\chi(x) \leqslant \rho_\alpha(x) \quad \forall x \in \Omega_{\alpha,6}.$$

Note that, from the choice of the components of the hybrid feedback law, and from Lemmas 2.13 and 3.19, for every $\alpha \in \mathcal{N} \setminus \{\omega\}$, the constant τ_α defined by (3.24) is such that $\tau_\alpha < \varepsilon$.

Let us prove that the point $\bar{x}$ is a semiglobal quasi-minimal time robust stable equilibrium for the system (1.1) in closed-loop with the hybrid feedback law (C, D, k, k_d) as stated in Theorem 2.10.

Step 1: *Completeness and global stability.* Let $R > 0$ and $\delta : [0, +\infty) \to [0, +\infty)$ of class $\mathcal{K}_\infty$ be such that, for every $\alpha \in \mathcal{N}$ and every x, with $|x - \bar{x}| < R$ and $x \in \Omega_{\alpha,6}$, we have

$$\delta(R) \leqslant \delta_{opt}(\delta_\alpha(R)),$$

where the functions δ_α are defined in Lemma 2.13. Let e, d be two functions satisfying the regularity assumptions and (3.23). Let (x, s_d) be a maximal solution of $\mathcal{H}_{(e,d)}$ on a hybrid domain S with the initial condition (x_0, s_0), where $|x_0 - \bar{x}| < R$. From Lemmas 2.13, 3.11, and 3.19, we have, for every $(t, l) \in S$,

$$|x(t, l) - \bar{x}| \leqslant \delta(R). \tag{3.34}$$

Therefore, the conclusion of Lemma 3.18 cannot hold (since $\limsup_{t \to T, (t,l) \in S} |x(t, l)| \neq +\infty$), and thus the supremum T of S in the t-direction is infinite, and the maximality property follows. The stability property follows from (3.34).

Step 2: *Uniform finite time convergence property.* Let $x_0 \in B(\bar{x}, R)$, and $s_0 \in \mathcal{N}$. Let (x, s_d) denote the solution of $\mathcal{H}_{(e,d)}$ with the initial condition (x_0, s_0). If $x_0 = \bar{x}$, then, using (3.20) and the fact that $\chi(\bar{x}) = 0$, the solution remains at the point $\bar{x}$. We next assume that $x_0 \neq \bar{x}$. Let $\alpha_0 \in \mathcal{N}$ such that $x(\cdot)$ is a solution of $\dot{x} = f(x, k_{\alpha_0}(x)) + d$ on $(0, t_1)$ for a $t_1 > 0$ given by Lemma 3.19.

If $\alpha_0 = \omega$, then the feedback law under consideration coincides with the optimal controller and, from statement 3 of Lemma 3.19, there does not exist any switching time $t > 0$. Then, from Lemma 3.11, $x(\cdot)$ reaches $\bar{x}$ within time $T_{\bar{x}}(x_0) + \varepsilon$.

If $\alpha_0 \neq \omega$, then, from Lemmas 2.13 and 3.19, the solution $x(\cdot)$ leaves $\Omega_{\alpha_0,6}$ within time ε and then enters the set $\Omega_{\omega,4}$. Therefore, since $\tau_\alpha < \varepsilon$, $x(\cdot)$ reaches $\bar{x}$ within time $T_{\bar{x}}(x_1) + \varepsilon$, where x_1 denotes the point of $x(\cdot)$ when entering $\Omega_{\omega,4}$.

Let $\tau(R) = \max_{x \in \delta(R)} T(x) + \varepsilon$. With (3.34), we get (2.12) and the uniform finite time property. Note that, from Lemma 2.13, the constraint (2.13) is satisfied.

Step 3: *Quasi-optimality.* Let K be a compact subset of $\mathbb{R}^n$, and $(x_0, s_0) \in K \times \mathcal{N}$. Let $R > 0$ such that $K \subset B(0, R)$. From the previous arguments, two cases occur:

- The solution starting from (x_0, s_0) reaches $\bar{x}$ within time $T_{\bar{x}}(x_0) + \varepsilon$ whenever $\alpha_0 = \omega$.
- The solution starting from (x_0, s_0) reaches $\bar{x}$ within time $T_{\bar{x}}(x_1) + \varepsilon$, whenever $\alpha_0 \neq \omega$, where x_1 denotes the point of $x(\cdot)$ when entering $\Omega_{\omega,6}$. We may reduce the neighborhoods $\Omega_{\alpha,l}$, one has $|T_{\bar{x}}(x_0) - T_{\bar{x}}(x_1)| \leqslant \varepsilon$. Indeed, from Remark 2.4, the function $T_{\bar{x}}(\cdot)$ is uniformly continuous on the compact K.

Hence, the maximal solution starting from (x_0, s_0) reaches $\bar{x}$ within time $T_{\bar{x}}(x_0) + 2\varepsilon$. This is the quasi-optimality property.

Theorem 2.10 is proved.

Remark 3.21. We deduce from Lemma 3.19 and from the proof of Theorem 2.10 that the following two cases may occur:

- If $x_0 \in \Omega_{\omega,2}$, then the solutions (x, s_d) of $\mathcal{H}_{(e,d)}$ defined on the hybrid time domain S and with the initial condition (x_0, ω) satisfy $x(t,l) + e(x(t,l),t) \in \Omega_{\omega,4}$ for all $(t,l) \in S$.
- If $x_0 \in \Omega_{\alpha,2}$ for $\alpha \neq \omega$, then there exists a jump time $\bar{t} \geqslant 0$, $(\bar{t}, \bar{l}) \in S$ such that $x(\bar{t}, \bar{l}) + e(x(\bar{t}, \bar{l}), \bar{t}) \in \Omega_{\omega,2}$, and for all $(t,l) \in S$, $t \geqslant \bar{t}$, we have $x(t, \bar{l}) + e(x(t,l), t) \in \Omega_{\omega,4}$.

Hence, for every $x_0 \in \mathbb{R}^n \setminus \{\bar{x}\}$, by considering only the initial conditions (x_0, s_0) with $x_0 \in \Omega_{s_0,2}$, the solutions of $\mathcal{H}_{(e,d)}$ depend only on the sets $\Omega_{\alpha,2}$, $\Omega_{\alpha,3}$, and $\Omega_{\alpha,4}$ for some $\alpha \in \mathcal{N}$. In other words, for such initial conditions, we only need three nested families of open sets to state a quasi-optimality property. However, six nested families of open sets $(\Omega_{\alpha,l})_{\alpha\in\mathcal{N},\, l\in\mathcal{F}}$ are required to establish the quasi-optimality property stated in Theorem 2.10 *for all initial conditions* $(x_0, s_0) \in \mathbb{R}^n \setminus \{\bar{x}\} \times \mathcal{N}$, and *not only* for initial conditions $\{(x_0, s_0),\ x_0 \in \Omega_{s_0,2}\}$.

Since there is no restriction on the s_0 variable, a jump may occur at time $t = 0$ and the second claim of (3.21) is active for some initial conditions. We emphasize that Lemma 3.19 asserts that there exists at most one positive jump time, but another jump may occur at time $t = 0$.

Intuitively, two patches are required for defining both kinds of jumps: from the controllers built in Lemma 2.13 to the controller considered in Lemma 2.14, and conversely. The other four patches are used to define neighborhoods of these jump sets and to bound the admissible measurement noise (see Definition 3.15). This explains the necessity of considering at least six patches to establish our main result.

Acknowledgments. The authors are grateful to Ludovic Rifford for constructive comments, and to the reviewers for helpful suggestions about the presentation of this paper.

REFERENCES

[1] A. Agrachev, *Compactness for sub-Riemannian length-minimizers and subanalyticity*, Rend. Sem. Mat. Univ. Politec. Torino, 56 (1998), pp. 1–12.

[2] A. Agrachev and J. P. Gauthier, *On subanalyticity of Carnot-Carathéodory distances*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 359–382.

[3] A. Agrachev and Yu. Sachkov, *Control Theory from the Geometric Viewpoint*, Springer-Verlag, Berlin, 2004.

[4] F. Ancona and A. Bressan, *Patchy vector fields and asymptotic stabilization*, ESAIM Cont. Optim. Calc. Var., 4 (1999), pp. 445–471.

[5] F. Ancona and A. Bressan, *Flow stability of patchy vector fields and robust feedback stabilization*, SIAM J. Control Optim., 41 (2003), pp. 1455–1476.

[6] F. Ancona and A. Bressan, *Nearly Time Optimal Stabilizing Patchy Feedbacks*, Preprint, 2005, http://arxiv.org/abs/math/0512531.

[7] V. Andrian, A. Bacciotti, and G. Beccari, *Global stability and external stability of dynamical systems*, Nonlinear Anal., 28 (1997), pp. 1167–1185.

[8] A. Astolfi, *Discontinuous control of the Brockett integrator*, European J. Control, 4 (1998), pp. 49–53.

[9] A. Bacciotti and F. Ceragioli, *Nonsmooth optimal regulation and discontinuous stabilization*, Abstr. Appl. Anal., 20 (2003), pp. 1159–1195.

[10] A. Bellaïche, *The tangent space in sub-Riemannian geometry*, in Sub-Riemannian Geometry, Birkhäuser, Basel, 1996, pp. 1–78.

[11] A. Bensoussan and J. L. Menaldi, *Hybrid control and dynamic programming*, Dynam. Contin. Discrete Impuls. Systems, 3 (1997), pp. 395–442.
[12] B. Bonnard, J.-B. Caillau, and E. Trélat, *Second order optimality conditions in the smooth case and applications in optimal control*, ESAIM Control Optim. Calc. Var. (2006), to appear.
[13] B. Bonnard and M. Chyba, *The Role of Singular Trajectories in Control Theory*, Math. Monograph, Springer-Verlag, Berlin, 2003.
[14] M. S. Branicky, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.
[15] A. Bressan, *Singularities of stabilizing feedbacks*, Rend. Sem. Mat. Univ. Politec. Torino, 56 (1998), pp. 87–104.
[16] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.
[17] Y. Chitour, F. Jean, and E. Trélat, *Propriétés génériques des trajectoires singulières*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 49–52.
[18] Y. Chitour, F. Jean, and E. Trélat, *Genericity properties for singular trajectories*, J. Differential Geom., 73 (2006), pp. 45–73.
[19] F. H. Clarke, Yu. S. Ledyaev, L. Rifford, and R. J. Stern, *Feedback stabilization and Lyapunov functions*, SIAM J. Control Optim, 39 (2000), pp. 25–48.
[20] R. T. Closkey and R. M. Murray, *Exponential stabilization of driftless nonlinear control systems using homogeneous feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 614–628.
[21] J.-M. Coron, *Global asymptotic stabilization for controllable systems without drift*, Math. Control Signals Systems, 5 (1992), pp. 295–312.
[22] C. D'Apice, M. Garavello, R. Manzo, and B. Piccoli, *Hybrid optimal control: Case study of a car with gears*, Internat. J. Control, 76 (2003), pp. 1272–1284.
[23] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J., 84 (1996), pp. 497–540.
[24] M. Garavello and B. Piccoli, *Hybrid necessary principle*, SIAM J. Control Optim., 43 (2005), pp. 1867–1887.
[25] R. Goebel and A. R. Teel, *Solutions to hybrid inclusions via set and graphical convergence with stability theory applications*, Automatica, 42 (2006), pp. 513–696.
[26] R. Goebel, J. Hespanha, A. R. Teel, C. Cai, and R. Sanfelice, *Hybrid systems: Generalized solutions and robust stability*, in Proceedings of the IFAC Symposium on Nonlinear Control Systems, Stuttgart, Germany, 2004.
[27] R. M. Hardt, *Stratification of real analytic mappings and images*, Invent. Math., 28 (1975), pp. 193–208.
[28] H. Hironaka, *Subanalytic sets*, in Number Theory, Algebraic Geometry and Commutative Algebra, in honor of Y. Akizuki, Kinokuniya, Tokyo, 1973, pp. 453–493.
[29] I. Kolmanovsky and N. H. McClamroch, *Developments in nonholonomic control problems*, IEEE Control Systems, 15 (1995), pp. 20–36.
[30] D. Liberzon, *Switching in Systems and Control*, Systems Control Found. Appl., Birkhäuser, Boston, 2003.
[31] E. Litsyn, Y. V. Nepomnyashchikh, and A. Ponosov, *Stabilization of linear differential systems via hybrid feedback controls*, SIAM J. Control Optim., 38 (2000), pp. 1468–1480.
[32] J. Lygeros, K. H. Johansson, S. N. Simic, J. Zhang, and S. S. Sastry, *Dynamical properties of hybrid automata*, IEEE Trans. Automat. Control, 48 (2003), pp. 2–17.
[33] P. Morin and C. Samson, *A characterization of the Lie algebra rank condition by transverse periodic functions*, SIAM J. Control Optim., 40 (2001), pp. 1227–1249.
[34] P. Morin and C. Samson, *Exponential stabilization of nonlinear driftless systems with robustness to unmodeled dynamics*, ESAIM Cont. Optim. Calc. Var., 4 (1999), pp. 1–35.
[35] B. Piccoli, F. M. Pait, and A. Bittar, *A hybrid controller for a nonholonomic system*, SBA Cont. Aut., 9 (1998), pp. 85–89.
[36] L. S. Pontryagin, V. G. Boltyanskij, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience Publishers, John Wiley & Sons, New York, 1962.
[37] C. Prieur, *Uniting local and global controllers with robustness to vanishing noise*, Math. Control Signals Systems, 14 (2001), pp. 143–172.
[38] C. Prieur, *Asymptotic controllability and robust asymptotic stabilizability*, SIAM J. Control Optim., 43 (2005), pp. 1888–1912.
[39] C. Prieur and A. Astolfi, *Robust stabilization of chained systems via hybrid control*, IEEE Trans. Automat. Control, 48 (2003), pp. 1768–1772.

[40] C. Prieur, R. Goebel, and A. Teel, *Results on robust stabilization of asymptotically controllable systems by hybrid feedback*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05), Seville, Spain, 2005, pp. 2598–2603.

[41] C. Prieur and E. Trélat, *Robust optimal stabilization of the Brockett integrator via a hybrid feedback*, Math. Control Signals Systems, 17 (2005), pp. 201–216.

[42] C. Prieur and E. Trélat, *Semi-global minimal time hybrid robust stabilization of analytic driftless control-affine systems*, in Proceedings of the 44th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC'05), Seville, Spain, 2005, pp. 5438–5443.

[43] L. Rifford, *The stabilization problem: AGAS and SRS feedbacks*, in Optimal Control, Stabilization, and Nonsmooth Analysis, Lectures Notes in Control and Inform. Sci. 301, Springer-Verlag, Heidelberg, 2004, pp. 173–184.

[44] L. Rifford and E. Trélat, *Morse-Sard type results in sub-Riemannian geometry*, Math. Ann., 332 (2005), pp. 145–159.

[45] E. D. Sontag, *Clocks and insensitivity to small measurement errors*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 537–557.

[46] E. D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations, and Control, Proc. NATO Advanced Study Institute, Montreal, F. H. Clarke and R. J. Stern, eds., Kluwer Academic, Dordrecht, The Netherlands, 1999, pp. 551–598.

[47] H. J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.

[48] M. Tamm, *Subanalytic sets in the calculus of variation*, Acta Math., 146 (1981), pp. 167–199.

[49] L. Tavernini, *Differential automata and their discrete simulators*, Nonlinear Anal., 11 (1987), pp. 665–683.

[50] E. Trélat, *Etude asymptotique et transcendance de la fonction valeur en contrôle optimal; catégorie log-exp en géométrie sous-riemannienne dans le cas Martinet*, Ph.D. Thesis, Univ. Dijon, France, 2000.

[51] E. Trélat, *Some properties of the value function and its level sets for affine control systems with quadratic cost*, J. Dyn. Control Syst., 6 (2000), pp. 511–541.

[52] E. Trélat, *Global subanalytic solutions of Hamilton-Jacobi type equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 23 (2006), pp. 363–387.

[53] H. Ye, A. N. Michel, and L. Hou, *Stability theory for hybrid dynamical systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 461–474.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1898–1914

OUTPUT FEEDBACK POLE ASSIGNMENT FOR TRANSFER FUNCTIONS WITH SYMMETRIES*

UWE HELMKE†, JOACHIM ROSENTHAL‡, AND XIAOCHANG ALEX WANG§

Abstract. This paper studies the problem of pole assignment for symmetric and Hamiltonian transfer functions. A necessary and sufficient condition for pole assignment by complex symmetric output feedback transformations is given. Moreover, in the case where the McMillan degree coincides with the number of parameters appearing in the symmetric feedback transformations, we derive an explicit combinatorial formula for the number of pole assigning symmetric feedback gains. The proof uses intersection theory in projective space as well as a formula for the degree of the complex Lagrangian Grassmann manifold.

Key words. output feedback, pole placement, inverse eigenvalue problems, Lagrangian Grassmannian, symmetric or Hamiltonian realizations, degree of a projective variety

AMS subject classifications. 14M15, 15A18, 70H14, 70S05, 93B55, 93B60

DOI. 10.1137/050644276

1. Introduction.

One of the best known inverse eigenvalue problems from linear system theory is that of pole assignment, i.e., finding a static output feedback gain for a given linear system such that the closed loop poles of the system coincide with a specified subset of the complex plane. Moreover, in the case of finitely many solutions, a formula for the number of pole assigning feedback transformations is desirable. Early contributions on the subject were obtained by, e.g., Davison and Wang [7] and Kimura [20], who derived sufficient conditions for the solvability. However, these conditions were far from being necessary as well. In a series of pioneering papers [16, 24, 25], Hermann and Martin applied tools from algebraic geometry to obtain necessary and sufficient conditions, valid for a generic class of systems and for *complex* feedback transformations. Their approach was based on the dominant morphism theorem [2, Chapter AG, section 17, Theorem 17.3] from complex algebraic geometry. A second breakthrough was subsequently made by Brockett and Byrnes [3], who used intersection theoretic arguments and the Schubert calculus on Grassmann manifolds to count the number of pole assigning complex feedback transformations. By refining these algebraic-geometric approaches of Hermann and Martin and Brockett and Byrnes, a number of fundamental contributions on the subject were made that finally led to a solution of the problem in the *real* case, with important contributions due to [8, 21, 29, 37, 38]. For an excellent survey paper on this subject, written from a control-theoretic point of view, see, e.g., [4]. More recently, various intersection theoretic tasks related to the Schubert calculus have been studied in the algebraic geometry literature; see, e.g., [12, 18, 34]. The focus of most of the investigations so far has been on the unstructured case, where no underlying symmetries

*Received by the editors November 4, 2005; accepted for publication (in revised form) June 16, 2006; published electronically December 18, 2006.
http://www.siam.org/journals/sicon/45-5/64427.html

†Department of Mathematics, University of Würzburg, Am Hubland, 97074 Würzburg, Germany (helmke@mathematik.uni-wuerzburg.de).

‡Department of Mathematics, University of Zürich, Winterthurerstr 190, 8057 Zürich, Switzerland (rosenthal@math.unizh.ch, www.math.unizh.ch/aa/). This author's research was supported in part by NSF grant DMS-00-72383 and Swiss National Science Foundation grant 107887.

§Department of Mathematics, Texas Tech University, Lubbock, TX 79409-1024 (alex.wang@ttu.edu).

for the involved transfer function or for the associated feedback transformations are imposed. However, transfer functions with symmetries occur naturally in various application areas, such as in network theory or mechanics. For example, the transfer functions $G(s)$ of linear RLC-circuits, consisting solely of resistors, capacitors, and inductive elements, are symmetric; i.e., they satisfy $G(s)^\top = G(s)$. In mechanics, the transfer functions of linear Hamiltonian systems are characterized by the symmetry relation $G(-s)^\top = G(s)$, while second order mechanical systems of the form

$$M\ddot{x} = Nx + Bu, \quad y = B^\top x$$

yield symmetric Hamiltonian transfer functions, satisfying

$$G(s) = H(s^2), \quad H(s) = H(s)^\top;$$

see, e.g., [1, 5, 6, 9]. For such structured systems it is reasonable to restrict the class of admissible feedback transformations to those that preserve the symmetry properties of the transfer functions. Therefore the known results on pole placement on unstructured systems do not apply in these cases and instead require a new approach.

In this paper we start an investigation of the pole placement problem for $n \times n$ symmetric transfer functions $G(s) = G(s)^\top$, arising in electrical network theory, and Hamiltonian transfer functions. For both types of systems the natural class of admissible output feedback transformations are the symmetric ones $F = F^\top$, yielding a symmetric closed loop transfer function

$$G_F(s) := (I_n - G(s)F)^{-1}G(s).$$

As the number of free parameters occurring in the symmetric feedback matrices F is $n(n+1)/2$, a necessary condition for generic solvability of this output feedback problem is that the McMillan degree δ of the transfer function G satisfies $\delta \geq \binom{n+1}{2}$ in the symmetric case, and $\delta \geq n(n+1)$ in the Hamiltonian case. In fact, we show that generically for *complex* symmetric output feedback transformations this condition is also sufficient. Moreover, for the limit case $\delta = \binom{n+1}{2}$ (or $\delta = n(n+1)$), we derive an explicit combinatorial formula for the number of complex symmetric output feedback gains that place the poles at given points. Our formula coincides with that of the degree for the complex Lagrangian manifold, given in [36].

In the real case such complete results cannot be expected. In fact, the symmetry of the transfer functions then imposes a priori limitations for the possible pole locations of such systems. This has been observed in [23], where it is shown for symmetric transfer functions that—in the special case that the Cauchy index of G coincides with the McMillan degree—generically real symmetric output feedback pole assignability holds if and only if $n \geq \delta$. Of course, in most applications we have $n \leq \delta$, and therefore the description of the set of poles that can be achieved by real symmetric output feedback becomes a complicated and nontrivial task.

2. Complex symmetric and Hamiltonian realizations. In this section we recall some basic facts concerning complex symmetric and Hamiltonian transfer functions and associated signature symmetric and Hamiltonian realizations. Let $\mathbb{C}$ denote the field of complex numbers. A complex rational transfer function $G(s) \in \mathbb{C}(s)^{n\times n}$ of McMillan degree δ is called *symmetric or Hamiltonian, respectively,* if

$$G(s) = G(s)^\top \quad \text{or} \quad G(s) = G(-s)^\top, \quad \text{respectively,}$$

holds for all $s \in \mathbb{C}$. A *complex symmetric realization* is a linear system of the form

$$\dot{x} = Ax + Bu, \quad y = B^\top x,$$

where $A \in \mathbb{C}^{\delta\times\delta}$ is symmetric, i.e., $A^\top = A$, and $B \in \mathbb{C}^{\delta\times n}$. Similarly, a *Hamiltonian realization* is a linear system

$$\dot{x} = Ax + Bu, \quad y = Cx,$$

where $A \in \mathbb{C}^{\delta\times\delta}, B \in \mathbb{C}^{\delta\times n}, C \in \mathbb{C}^{n\times\delta}$ satisfies

$$AJ = (AJ)^\top, \quad C^\top = JB$$

and

$$\begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$$

denotes the standard symplectic form on $\mathbb{C}^{\delta\times\delta}$. In particular, Hamiltonian systems always have even McMillan degree δ.

Complex symmetric realizations are the natural class of realizations for complex symmetric transfer functions. In fact, they are the proper analogue of signature symmetric realizations of real rational transfer functions, appearing in network theory. Here a signature symmetric realization (A, B, C) is one that is symmetric relative to the bilinear form $I_{pq} := \mathrm{diag}(\mathrm{I_p}, -\mathrm{I_q})$; i.e., (A, B, C) satisfies $(AI_{pq})^\top = AI_{pq}, C^\top = I_{pq}B$. Such realizations always exist for real symmetric transfer functions G, and the integer $p - q$ coincides with the so-called Cauchy–Maslov index of G. If $q = 0$, then this definition coincides with the one above for symmetric realizations. In particularly, over $\mathbb{R}$, real symmetric realizations correspond to linear models of *RC*-networks, constructed entirely using capacitors and resistors. The real symmetric transfer functions defined by them are characterized by the property that the Cauchy–Maslov index coincides with the McMillan degree [1, 9, 14].

The following variant of the Kalman realization theorem is well known; see, e.g., [9, 10]. Recall that the complex orthogonal group $O(\delta, \mathbb{C})$ is the matrix group consisting of all complex $\delta \times \delta$ matrices S, satisfying $SS^\top = I_\delta$. Given any complex realization (A, B, C) of a symmetric transfer function $G(s) = C(sI - A)^{-1}B$, note that $(A^\top, C^\top, B^\top)$ is also a realization.

PROPOSITION 2.1. *Let $G(s) = G(s)^\top$ be an $n\times n$ strictly proper, complex rational transfer function of McMillan degree δ. Then the following hold:*

(1) *$G(s)$ has a controllable and observable complex symmetric realization*

$$(A, B, C) = (A^\top, C^\top, B^\top).$$

(2) *If (A_i, B_i, C_i), i=1, 2, are two controllable and observable complex symmetric realizations of $G(s)$, then there exists a unique complex orthogonal transformation $S \in O(\delta, \mathbb{C})$ such that $(A_2, B_2, C_2) = (SA_1S^{-1}, SB_1, C_1S^{-1})$.*

In the literature usually only the real case of the above result is proved, where the statement is actually slightly different due to the presence of signature symmetric realizations. In the complex case the result simplifies to the one given here. For the sake of completeness we include the proof; see also [10].

Proof. If (A, B, C) is a minimal realization of $G(s)$, then, by symmetry of G, $(A^\top, C^\top, B^\top)$ also is a minimal realization. Applying Kalman's realization theorem implies the existence of a unique invertible complex $\delta \times \delta$ matrix S with

$$(A^\top, C^\top, B^\top) = (SAS^{-1}, SB, CS^{-1}).$$

By transposing this equation and using the uniqueness of S we conclude that $S = S^\top$. It is a well-known fact from linear algebra that every complex symmetric invertible matrix has a representation $S = XX^\top$ by a complex invertible matrix X. Moreover, X is uniquely determined up to right factors XT, where $T \in O(\delta, \mathbb{C})$. Then (XAX^{-1}, XB, CX^{-1}) is a complex symmetric realization, which completes the proof. □

There is a similar realization theorem for Hamiltonian systems, for which we refer to the literature; see, e.g., [6, 9]. Static linear output feedback can be meaningfully defined for such systems only through symmetric gain matrices. Note that

$$\operatorname{adj}(sI - A) = s^{n-1}I + s^{n-2}(A - \operatorname{trace}(A)I) + \text{lower order terms.}$$

Thus, for symmetric A, the rational matrix

$$B^\top(sI - A - BFB^\top)^{-1}B = B^\top \frac{1}{\det(sI - A - BFB^\top)} \operatorname{adj}(sI - A - BFB^\top)B$$

is symmetric only if $B^\top BFB^\top B$ is symmetric. Therefore, if B has full column rank, an output feedback transformation

$$u = Fy + v$$

with the closed system

$$\dot{x} = (A + BFB^\top)x + Bu, \quad y = B^\top x$$

preserves the complex symmetry of the realizations if and only if $F = F^\top$. Thus we define two complex symmetric realizations (A_i, B_i, C_i), $i = 1, 2$, to be *symmetric output feedback equivalent* if and only if there exist $S \in O(\delta, \mathbb{C})$, $F = F^\top \in \mathbb{C}^{n\times n}$ with

$$(A_2, B_2, C_2) = (S(A_1 + B_1FB_1^\top)S^{-1}, SB_1, C_1S^{-1}).$$

Equivalently, they are symmetric output feedback equivalent if and only if the associated transfer functions satisfy

$$G_2(s) = (I_n - G_1(s)F)^{-1}G_1(s).$$

Similarly, output feedback for Hamiltonian systems

$$\dot{x} = (A + BFC)x + Bu, \quad y = Cx$$

preserves the Hamiltonian properties of the realization if and only if $F = F^\top$. Thus in both cases we have to focus on symmetric output feedback.

We note some elementary geometric properties of the set of complex symmetric transfer functions that will be important in the subsequent development; see, e.g., [6] for some of the details for the proof of the subsequent theorem. We omit a full proof as it would take us too far from the subject.

PROPOSITION 2.2. *Let $SRat_{\delta,n}(\mathbb{C})$ and $Ham_{\delta,n}(\mathbb{C})$, respectively, denote the sets of strictly proper, complex symmetric and Hamiltonian $n \times n$ transfer functions of McMillan degree δ. Then $SRat_{\delta,n}(\mathbb{C})$ and $Ham_{\delta,n}(\mathbb{C})$ are, respectively, a smooth complex manifold of complex dimension $\delta(n + 1)$ and dimension δn. Moreover, they are nonsingular irreducible quasi-affine varieties.*

In particular, there is a canonical notion of "genericity" for complex symmetric or Hamiltonian transfer functions. Explicitly, a property E of complex symmetric transfer functions is called generic if the set defined by E,

$$\{G \in SRat_{\delta,n}(\mathbb{C}) \mid G \text{ has property } E\},$$

is a Zariski open subset of $SRat_{\delta,m}(\mathbb{C})$. Equivalently, this also can be expressed in terms of complex symmetric realizations.

3. Main result. After these preliminaries we can now formulate and prove the main technical results of this paper. Let $G(s)$ be an $n \times n$ complex symmetric or Hamiltonian transfer function, i.e., $G(s)^\top = G(s)$ or $G(-s)^\top = G(s)$, respectively. Assume that $G(s)$ is strictly proper and has McMillan degree δ. The complex symmetric eigenvalue assignment problem then asks the following question.

PROBLEM 3.1. *For a given arbitrary monic polynomial $\varphi(s) \in \mathbb{C}[s]$ of degree δ ($\varphi(s) = \varphi(-s)$ is assumed to be* even *in the Hamiltonian case), is there an $n \times n$ complex symmetric matrix F such that the closed loop transfer function*

$$G_F(s) := (I_n - G(s)F)^{-1}G(s)$$

has the characteristic polynomial $\varphi(s)$, i.e., such that the poles of $G_F(s)$ are the zeros of $\varphi(s)$?

If for a particular symmetric (Hamiltonian) transfer function $G(s)$ Problem 3.1 has an affirmative answer, we will say that $G(s)$ is *pole assignable in the class of complex symmetric (Hamiltonian) feedback compensators.* We say that $G(s)$ is *generically pole assignable* if the problem is solvable for a generic choice of admissible polynomials $\varphi(s)$.

Similar to the situation of the static pole placement problem [3, 37] and the dynamic pole placement problem [28], Problem 3.1 turns out to be highly nonlinear, and techniques from algebraic geometry will be required to study the problem. The first main result is in the spirit of Hermann and Martin [16], by deriving a generic necessary and sufficient condition via the dominant morphism theorem.

We prove some lemmas first. Let $\pi(A) = (a_{11}, \ldots, a_{\delta\delta})$ be the projection onto the diagonal entries of an $\delta \times \delta$ matrix A. In what follows we will identify $\mathbb{C}^\delta$ with the complex vector space of row vectors. For any symmetric matrix L, define $\theta^L : O(\delta, \mathbb{C}) \to \mathbb{C}^\delta$ through

$$\theta^L(S) = \pi(SLS^{-1}).$$

As $O(\delta, \mathbb{C})$ is a Lie group, its tangent space at the identity matrix I is given by the Lie algebra of complex skew-symmetric matrices

$$so(\delta, \mathbb{C}) = \{X \in \mathbb{C}^{\delta\times\delta} \mid X + X^\top = 0\}.$$

Moreover, the Jacobian $d\theta_I^L$ of θ^L at I is given by

$$d\theta_I^L : so(\delta, \mathbb{C}) \to V, \qquad d\theta_I^L(X) = \pi(XL - LX),$$

where

$$V = \left\{ (x_1, \ldots, x_\delta) \in \mathbb{C}^\delta \mid \sum_1^\delta x_i = 0 \right\}.$$

For any $\delta \times \delta$ matrix L, the graph $\mathcal{G}(L)$ of L is defined as a graph with δ vertices such that there is a path from vertex i to vertex j if and only if the ijth entry of L is nonzero. It is a well-known fact from linear algebra that the graph $\mathcal{G}(L)$ is connected if and only if L is irreducible, i.e., if and only if there exists no permutation matrix P such that PLP^{-1} is block diagonal. We use this fact together with an idea developed in [15, Lemma 2.5] to prove the following equivalent characterization.

LEMMA 3.2. *The Jacobian $d\theta_I^L$ is surjective if and only if the associated graph $\mathcal{G}(L)$ is connected.*

Proof. By inspection, the derivative $d\theta_I^L$ is not surjective if and only if there exists a nonzero diagonal matrix Z of trace zero such that for all $X \in so(\delta, \mathbb{C})$

$$\text{trace}(Z(XL - LX)) = \text{trace}((LZ - ZL)X) = 0.$$

By symmetry of L, Z we have $LZ - ZL \in so(\delta, \mathbb{C})$. Since the trace function defines a nondegenerate bilinear form on $so(\delta, \mathbb{C})$, the condition $\text{trace}((LZ - ZL)X) = 0$ is equivalent to $LZ = ZL$. Since Z is a nonzero diagonal matrix of trace zero, there is a permutation matrix P such that $\hat{Z} := PZP^{-1} = \text{block diag}(a_1 I_1, \ldots, a_k I_k)$ with $k \geq 2$ and a_i's distinct. Let $\hat{L} = PLP^{-1}$. Then $LZ = ZL$ is equivalent to $\hat{L}\hat{Z} = \hat{Z}\hat{L}$, which is equivalent to $\hat{L}$ being block diagonal. But from the above remark this is equivalent to the graph $\mathcal{G}(L)$ being disconnected. The result follows. ∎

LEMMA 3.3. *Let L be a nonzero complex symmetric matrix such that $\pi(L) = 0$. Then there is a family of orthogonal matrices $S(\epsilon) \in O(\delta, \mathbb{C})$, $\epsilon \geq 0$, with $S(0) = I$ such that the matrix $\hat{L}(\epsilon) := S(\epsilon)LS(\epsilon)^{-1}$ has the properties that $\pi(\hat{L}(\epsilon)) = 0$ and $d\varphi_I^{\hat{L}(\epsilon)}$ is surjective for all $\epsilon \in (0, \pi/2)$.*

Proof. If $\mathcal{G}(L)$ is connected, then by the previous lemma the choice $S(\epsilon) := I$ does the job. Thus it suffices to prove that $\mathcal{G}(L)$ not connected implies that then we can find a family of transformations $S(\epsilon)$ such that $\pi(S(\epsilon)LS(\epsilon)^{-1}) = 0$ and the largest connected subgraph of $\mathcal{G}(S(\epsilon)LS(\epsilon)^{-1})$ contains more vertices than that of $\mathcal{G}(L)$ for all $0 < \epsilon < \pi/2$.

Note that $\pi(L) = 0$ and $L \neq 0$ imply that the largest connected subgraph of $\mathcal{G}(L)$ must contain at least 2 vertices. Assume that the largest connected subgraph of $\mathcal{G}(L)$ contains k vertices, $2 \leq k < \delta$. Without loss of generality, assume

$$L = \begin{bmatrix} L_1 & 0 \\ 0 & L_2 \end{bmatrix},$$

where the graph of the $k \times k$ submatrix L_1 is connected. Write

$$L_1 = \begin{bmatrix} L_{11} & \alpha \\ \alpha^\top & 0 \end{bmatrix} \text{ and } L_2 = \begin{bmatrix} 0 & \beta^\top \\ \beta & L_{22} \end{bmatrix},$$

where L_{11} and L_{22} are sizes $(k-1)\times(k-1)$ and $(\delta-k-1)\times(\delta-k-1)$, respectively. By irreducibility of L_1 we have $\alpha \neq 0$. Thus L has the form

$$\begin{bmatrix} L_{11} & \alpha & 0 & 0 \\ \alpha^\top & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta^\top \\ 0 & 0 & \beta & L_{22} \end{bmatrix}.$$

Let

$$S(\epsilon) = \begin{bmatrix} I_{(k-1)\times(k-1)} & 0 & 0 & 0 \\ 0 & \cos\epsilon & -\sin\epsilon & 0 \\ 0 & \sin\epsilon & \cos\epsilon & 0 \\ 0 & 0 & 0 & I_{(\delta-k-1)\times(\delta-k-1)} \end{bmatrix}.$$

Then

$$S(\epsilon)LS^{-1}(\epsilon) = \begin{bmatrix} L_{11} & (\cos\epsilon)\alpha & (\sin\epsilon)\alpha & 0 \\ (\cos\epsilon)\alpha^\top & 0 & 0 & (-\sin\epsilon)\beta^\top \\ (\sin\epsilon)\alpha^\top & 0 & 0 & (\cos\epsilon)\beta^\top \\ 0 & (-\sin\epsilon)\beta & (\cos\epsilon)\beta & L_{22} \end{bmatrix}.$$

For the graph $\mathcal{G}(S(\epsilon)LS^{-1}(\epsilon))$ with $0 < \epsilon < \pi/2$, the vertices $\{1,\dots,k\}$ are still connected and the vertex $k+1$ is symmetrically connected to at least one of the first k vertices. Thus the vertices $\{1,\dots,k+1\}$ are connected. □

LEMMA 3.4. *Let $\mathcal{L}$ be a linear subspace of complex symmetric matrices of dimension δ, and $\mathcal{L} \not\subset sl(\delta,\mathbb{C})$. Then there exists an orthogonal matrix $S \in O(\delta,\mathbb{C})$ such that $\pi\mid_{S\mathcal{L}S^{-1}}$ is one to one, and onto.*

Proof. The proof goes by recursively constructing a basis $\{L_1,\dots,L_\delta\}$ of $\mathcal{L}$ such that $\{\pi(SL_1S^{-1}),\dots,\pi(SL_\delta S^{-1})\}$ are linearly independent for a suitable complex orthogonal matrix $S \in O(\delta,\mathbb{C})$. First note that we can modify any basis of $\mathcal{L}$ into a basis $\mathcal{L}^{(1)} := \{L_1,\dots,L_\delta\}$ of $\mathcal{L}$ such that $L_1 \notin sl(\delta,\mathbb{C})$ and $L_i \in sl(\delta,\mathbb{C})$ for $i = 2,\dots,\delta$. In fact, if $\{K_1,\dots,K_\delta\}$ denotes any basis of $\mathcal{L}$ with $\operatorname{trace}(K_1) \neq 0$, then $\{L_1 := K_1, L_2 := K_2 - c_2K_1,\dots,L_\delta := K_\delta - c_\delta K_1\}$, $c_i := \operatorname{trace}(K_i)/\operatorname{trace}(K_1)$, is as desired. By construction of L_1, we have $\pi(L_1) \neq 0$.

Let $\{L_1,\dots,L_\delta\}$ be a basis of $\mathcal{L}$ such that $L_1 \notin sl(\delta,\mathbb{C})$ and $L_i \in sl(\delta,\mathbb{C})$ for $i = 2,\dots,\delta$. Then $\dim \operatorname{span}\{\pi(L_1),\dots,\pi(L_\delta)\} := k \geq 1$. If $k < \delta$, then by reordering the indices we can assume that $\{\pi(L_1),\dots,\pi(L_k)\}$ are linearly independent and

$$\pi(L_j) = \sum_{i=1}^{k} c_{ij}\pi(L_i) \text{ for } j = k+1,\dots,\delta.$$

By replacing L_j with $L_j - \sum_{i=1}^{k} c_{ij}L_i$, we can further assume that $\pi(L_j) = 0$ for $j = k+1,\dots,\delta$. It is thus sufficient to show that if there is an orthogonal matrix $\hat{S}$ such that the matrices $\{M_j := \hat{S}L_j\hat{S}^{-1}, j = 1,\dots,\delta\}$ have the property that $\{\pi(M_1),\dots,\pi(M_k)\}$ are linearly independent and $\pi(M_j) = 0$, $j = k+1,\dots,\delta$, for some $k < n$, then we can find an orthogonal S such that

$$\{\pi(SM_1S^{-1}),\dots,\pi(SM_kS^{-1}),\pi(SM_{k+1}S^{-1})\}$$

are linearly independent.

By Lemma 3.3, there exists $S_\epsilon \in O(\delta,\mathbb{C})$ arbitrarily close to the identity matrix such that $\pi(S_\epsilon M_1 S_\epsilon^{-1}),\dots,\pi(S_\epsilon M_k S_\epsilon^{-1})$ are linearly independent and the graph $\mathcal{G}(S_\epsilon M_{k+1} S_\epsilon^{-1})$ is connected. By replacing M_i with $S_\epsilon M_i S_\epsilon^{-1}$, we can assume further that $d\theta_I^{M_{k+1}}$ is onto V. Then there exists a skew-symmetric matrix X such that

$$\pi(XM_{k+1} - M_{k+1}X) \notin \operatorname{span}\{\pi(M_1),\dots,\pi(M_k)\}.$$

Let

$$S(\epsilon) = \exp(\epsilon X).$$

Then $S(\epsilon)$ is orthogonal for all ϵ, and

$$S(\epsilon) = I + \epsilon X + \text{higher order terms}.$$

The Taylor series expansions of $\{\pi(S(\epsilon)M_iS(\epsilon)^{-1})\}$ have the forms

$$\pi(S(\epsilon)M_iS(\epsilon)^{-1}) = \pi(M_i) + \beta_i(\epsilon), \quad i = 1, \ldots, k,$$

and

$$\pi(S(\epsilon)M_{k+1}S(\epsilon)^{-1}) = \epsilon\left(\pi(XM_{k+1} - M_{k+1}X) + \beta_{k+1}(\epsilon)\right),$$

where $\beta_i(\epsilon)$ are continuous with respect to ϵ and $\beta_i(\epsilon) \to 0$ as $\epsilon \to 0$ for $i = 1, \ldots, k+1$. Since $\{\pi(M_1), \ldots, \pi(M_k), \pi(XM_{k+1} - M_{k+1}X)\}$ are linearly independent, for sufficient small $\epsilon > 0$, $\{\pi(M_1) + \beta_1(\epsilon), \ldots, \pi(M_k) + \beta_k(\epsilon), \pi(XM_{k+1} - M_{k+1}X) + \beta_{k+1}(\epsilon)\}$ are also linearly independent; i.e.,

$$\{\pi(S(\epsilon)M_1S(\epsilon)^{-1}, \ldots, \pi(S(\epsilon)M_kS(\epsilon)^{-1}, \pi(S(\epsilon)M_{k+1}S(\epsilon)^{-1}\}$$

are linearly independent. □

THEOREM 3.5. *If $G(s)$ is a symmetric (or Hamiltonian) transfer function of McMillan degree $\delta > \binom{n+1}{2}$ (or $\delta > n(n+1)$), then $G(s)$ is not pole assignable in the class of (real or) complex symmetric feedback compensators.*

When $\delta \leq \binom{n+1}{2}$ (or $\delta \leq n(n+1)$), then there is a generic set of $n \times n$ symmetric (or Hamiltonian) transfer functions of degree δ which are generically pole assignable via complex symmetric feedback compensators.

Proof. We give only a sketch of the proof, as the arguments based on the dominant morphism theorem are well known from [16, 24]. Note, however, that there is a serious gap in the proof of [24] for the pole placement result on Hamiltonian systems because it is not proved that the set of generically pole assignable Hamiltonian systems is nonempty. In fact, a construction of such an example is not completely trivial and depends on our previous lemmas.

The first claim follows immediately from a standard dimension argument, as the vector space $\mathrm{Sym}(n)$ of complex $n \times n$ symmetric matrices has dimension $\binom{n+1}{2}$. For the second claim we note that the set of generically pole assignable systems is a Zariski open subset of the nonsingular, irreducible quasi-affine variety of symmetric or Hamiltonian transfer functions, respectively. Therefore we need only show that this Zariski open subset is nonempty. By the dominant morphism theorem, it suffices to find one system whose Jacobian of the pole placement map at one point is onto.

Note, by the Newton formula, that the coefficients of the characteristic polynomial $\det(sI - A) = s^\delta + \alpha_{\delta-1}s^{\delta-1} + \cdots + \alpha_1 s + \alpha_0$ are related to the traces of powers of A as follows:

$$\begin{aligned}
\alpha_{\delta-1} &= -\operatorname{trace}(A),\\
\alpha_{\delta-2} &= -\frac{1}{2}(\operatorname{trace}(A^2) + \alpha_{\delta-1}\operatorname{trace}(A)),\\
&\vdots\\
\alpha_0 &= -\frac{1}{\delta}(\operatorname{trace}(A^\delta) + \alpha_{\delta-1}\operatorname{trace}(A^{\delta-1}) + \cdots + \alpha_1\operatorname{trace}(A)).
\end{aligned}$$

Therefore for the case of symmetric transfer functions, the pole placement map is equivalent to the map

$$\begin{array}{rccc}
\phi: & \mathrm{Sym}(n) & \longrightarrow & \mathbb{C}^\delta\\
& F & \longmapsto & (\operatorname{trace}(A + BFB^\top), \ldots, \operatorname{trace}(A + BFB^\top)^\delta),
\end{array} \tag{3.1}$$

and its Jacobian at 0 is given by

$$d\phi_0(F) = (\operatorname{trace}(BFB^\top), 2\operatorname{trace}(ABFB^\top), \ldots, \delta \operatorname{trace}(A^{\delta-1}BFB^\top)).$$

For the case of Hamiltonian transfer functions, since $JAJ = A^\top$ and $J^2 = -I$, we have $(-1)^{k-1}JA^kJ = (A^k)^\top$ for $k = 1, 2, \ldots$, which implies that the characteristic polynomial of A is even and

$$\operatorname{trace}(A^k) = 0 \text{ holds for all odd } k\text{'s}.$$

Therefore the pole placement map is equivalent to the map

(3.2)

$$\begin{array}{rccc} \psi: & \operatorname{Sym}(n) & \longrightarrow & \mathbb{C}^{\delta/2} \\ & F & \longmapsto & (\operatorname{trace}(A+BFC)^2, \operatorname{trace}(A+BFC)^4, \ldots, \operatorname{trace}(A+BFC)^\delta), \end{array}$$

and its Jacobian at 0 is given by

$$d\psi_0(F) = (2\operatorname{trace}(ABFC), 4\operatorname{trace}(A^3BFC) \ldots, \delta \operatorname{trace}(A^{\delta-1}BFC)).$$

We first consider the case of symmetric transfer functions. Let B be any real nonzero matrix and $\mathcal{L} = \{BFB^\top \mid F \in \operatorname{Sym}(n)\}$. Then $\mathcal{L} \not\subset sl(\delta, \mathbb{C})$ and $\dim\mathcal{L} \geq \delta$. By Lemma 3.4 there exists an orthogonal matrix $S \in O(\delta, \mathbb{C})$ such that $\pi|_{S\mathcal{L}S^{-1}}$ is surjective. Let $D = \operatorname{diag}(1, 2, \ldots, \delta)$ and $A = S^{-1}DS$. Then

$$\begin{aligned} d\phi_0(F) &= (\operatorname{trace}(BFB^\top), 2\operatorname{trace}(ABFB^\top), \ldots, \delta \operatorname{trace}(A^{\delta-1}BFB^\top)) \\ &= (\operatorname{trace}(SBFB^\top S^{-1}), 2\operatorname{trace}(DSBFB^\top S^{-1}), \ldots, \delta \operatorname{trace}(D^{\delta-1}SBFB^\top S^{-1})) \\ &= \pi(SBFB^\top S^{-1})V, \end{aligned}$$

where

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2^{\delta-1} \\ \vdots & \vdots & & \vdots \\ 1 & \delta & \cdots & \delta^{\delta-1} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \delta \end{bmatrix}.$$

Since $\pi|_{S\mathcal{L}S^{-1}}$ is surjective and V is nonsingular, $d\phi_0$ is onto.

For the case of Hamiltonian transfer functions, let

$$B = \begin{bmatrix} 0 \\ B_1 \end{bmatrix} \text{ and } C = \begin{bmatrix} B_1^\top & 0 \end{bmatrix},$$

where B_1 is any real nonzero $\frac{\delta}{2} \times n$ matrix and $\mathcal{L} = \{B_1FB_1^\top \mid F \in \operatorname{Sym}(n)\}$. Then $\mathcal{L} \not\subset sl(\delta/2, \mathbb{C})$. By Lemma 3.4 there exists an orthogonal matrix $S_1 \in O(\delta/2, \mathbb{C})$ such that $\pi : S_1\mathcal{L}S_1^{-1} \mapsto \mathbb{C}^{\frac{\delta}{2}}$ is surjective. Let $D_1 = \operatorname{diag}(1, 2, \ldots, \delta/2)$,

$$S = \begin{bmatrix} S_1 & 0 \\ 0 & S_1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & D_1 \\ D_1 & 0 \end{bmatrix},$$

and $A = S^{-1}DS$. Note that D and S are Hamiltonian and symplectic matrices, respectively. In particular, A is Hamiltonian. Then

$$\begin{aligned} d\psi_0(F) &= (2\operatorname{trace}(ABFC), 4\operatorname{trace}(A^3BFC), \ldots, \delta \operatorname{trace}(A^{\delta-1}BFC)) \\ &= (2\operatorname{trace}(DSBFCS^{-1}), 4\operatorname{trace}(D^3SBFCS^{-1}), \ldots, \delta \operatorname{trace}(D^{\delta-1}SBFCS^{-1})) \\ &= (2\operatorname{trace}(D_1S_1B_1FB_1^\top S_1^{-1}), \ldots, \delta \operatorname{trace}(D_1^{\delta-1}S_1B_1FB_1^\top S_1^{-1})) \\ &= \pi(S_1B_1FB_1^\top S_1^{-1})U, \end{aligned}$$

where

$$U = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 2^3 & \cdots & 2^{\delta-1} \\ \vdots & \vdots & & \vdots \\ \frac{\delta}{2} & \left(\frac{\delta}{2}\right)^3 & \cdots & \left(\frac{\delta}{2}\right)^{\delta-1} \end{bmatrix} \begin{bmatrix} 2 & 0 & \cdots & 0 \\ 0 & 4 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \delta \end{bmatrix}.$$

Since $\pi|_{S_1 B_1 F B_1^\top S_1^{-1}}$ is surjective and U is nonsingular, $d\psi_0$ is onto. □

The second main theorem in this paper deals with the limit case $\delta = \binom{n+1}{2}$, where we can prove a more precise statement.

THEOREM 3.6. *Let $\delta = \binom{n+1}{2}$ in the symmetric case, and $\delta = n(n+1)$ for Hamiltonian systems. Then for a generic set of $n \times n$ symmetric (or Hamiltonian) transfer functions of degree δ, the number of pole assigning complex symmetric feedback compensators is finite, and when counted with multiplicities there are exactly*

$$d(n) := 2^{\binom{n}{2}} \frac{\binom{n+1}{2}!\ 1!\ 2!\ \cdots\ (n-1)!}{1!\ 3!\ \cdots\ (2n-1)!} = \frac{\binom{n+1}{2}!}{\prod_{i=0}^{n-1} (2\,i+1)^{n-i}} \tag{3.3}$$

many symmetric compensators as solution.

One immediately computes $d(1) = 1$, $d(2) = 2$, $d(3) = 2^4$, $d(4) = 3 \cdot 2^8$, $d(5) = 11 \cdot 13 \cdot 2^{11}$, and $d(6) = 13 \cdot 17 \cdot 19 \cdot 2^{18}$. The integer sequence $d(n)$ is sequence A005118 in Sloane's data bank of integer sequences [33]. The sequence has several combinatorial and geometric interpretations. For the context of this paper it will be important that $d(n)$ is equal to the degree of the Lagrangian Grassmannian, the projective variety of all maximal isotropic subspaces in a complex vector space of dimension $2n$.

As for the Grassmann variety, classical Schubert calculus [30] (see [11, Chapter 14] for a modern treatment of Schubert calculus) provides the tools to compute the degree of the Lagrangian Grassmannian. An explicit formula for the integers $d(n)$ was probably first given by Hiller [17], who computed

$$d(n) = \begin{cases} 2^{\binom{n}{2}} \frac{\binom{n+1}{2}!\ 2!\ 4!\ \cdots\ (n-2)!}{(n+1)1!\ (n+3)!\ \cdots\ (2n-1)!} & \text{if } n \text{ is even,} \\ 2^{\binom{n}{2}} \frac{\binom{n+1}{2}!\ 2!\ 4!\ \cdots\ (n-2)!}{(n+1)1!\ (n+3)!\ \cdots\ (2n-1)!} & \text{if } n \text{ is odd.} \end{cases} \tag{3.4}$$

The hard combinatorial work to derive formula (3.4) is actually due to Schur [31]. It has been pointed out by Totaro [36] that $d(n)$ is equal to the Kostka number $K_{\lambda,(1^N)}$, where λ is the partition $\lambda = (n, n-1, \ldots, 1)$ and $N = \binom{n+1}{2}$. Totaro derived an explicit formula for more general Kostka numbers, and Totaro's formula specializes to formula (3.3). Readers interested in combinatorial aspects of Kostka numbers should consult the book by Macdonald [22].

It can be seen from the formulas that $d(n)$ is always even, except for $n = 1$. This is related to the fact that the symmetric output feedback pole placement problem is *not* generically solvable over the reals. Actually, more can be said about $d(n)$. The sequence

$$\tilde{d}(n) := d(n) 2^{-\binom{n}{2}}$$

is the degree of the spinor variety, the complex projective variety $SO(2n+1)/U(n)$ [17]; in particular, $\tilde{d}(n)$ represents an integer sequence again. The sequence $\tilde{d}(n)$ appears under the number A003121 in Sloane's data bank [33].

The proof of Theorem 3.6 will occupy the rest of this section. The proof will necessitate a geometric reformulation and several technical lemmas.

First we will describe the closed loop characteristic equation in a slightly more convenient way. Consider a left coprime factorization $D^{-1}(s)N(s) = G(s)$ of the symmetric or Hamiltonian transfer function $G(s)$. Let $F \in \mathrm{Sym}(n)$ be an $n \times n$ complex symmetric matrix. When the feedback law $y = -Fu + v$ is applied, then up to a constant factor the characteristic polynomial $\varphi(s)$ is also equal to

$$\det \begin{bmatrix} D(s) & N(s) \\ F & I_n \end{bmatrix}. \tag{3.5}$$

The vector space $\mathrm{Sym}(n)$ describing the set of $n \times n$ complex symmetric matrices is not very well suited to invoking strong theorems from algebraic geometry and intersection theory [11], as these usually require compactness assumptions on the underlying spaces. A similar difficulty exists for the static output pole placement problem. Brockett and Byrnes showed in [3] how to translate the static pole placement problem into a geometric problem. This then resulted in an intersection problem on a compact Grassmann variety, and methods from classical Schubert calculus [30, 35] could be invoked.

We will follow this compactification strategy for Problem 3.1 as well. This will lead us to an intersection problem on some projective variety. In order to do so we therefore need a good compactification of $\mathrm{Sym}(n)$. For this identify the row span $\mathrm{rowsp}\,[F\ I_n]$ of any symmetric matrix F with an element of the Grassmann variety $\mathrm{Grass}(n, \mathbb{C}^{2n})$. Using the Plücker embedding

$$\mathrm{Grass}(n, \mathbb{C}^{2n}) \longrightarrow \mathbb{P}\left(\wedge^n \mathbb{C}^{2n}\right) = \mathbb{P}^N, \quad N = \binom{2n}{n} - 1,$$

we can then identify $\mathrm{Sym}(n)$ with a quasi-projective subset of the complex projective variety $\mathbb{P}^N$.

Definition 3.7. The algebraic closure of the set

$$\{\mathrm{rowsp}\,[F\ I_n] \mid F \in \mathrm{Sym}(n)\}$$

is called the *complex Lagrangian Grassmann manifold*. It will be denoted by $\mathbb{L}G(n)$.

It is well known that $\mathbb{L}G(n)$ is a smooth projective variety of dimension $\binom{n+1}{2}$, the dimension of $\mathrm{Sym}(n)$. Note that every element in $\mathbb{L}G(n)$ can be simply represented by a subspace of the form $\mathrm{rowsp}\,[F_1\ F_2]$, where $F_1(F_2)^\top$ is a symmetric matrix, i.e., $F_1(F_2)^\top = F_2(F_1)^\top$. The elements of $\mathbb{L}G(n)$ are thus exactly the Lagrangian subspaces of $\mathbb{C}^{2n}$. The subspace $\mathrm{rowsp}\,[F_1\ F_2]$ coincides with the subspace $\mathrm{rowsp}\,[S\ I_n]$ associated with an element S of $\mathrm{Sym}(n)$ if and only if F_2 is invertible. Moreover, then $S = (F_2)^{-1}F_1$. When F_2 is singular one can still define a characteristic polynomial through

$$\varphi(s) := \det \begin{bmatrix} D(s) & N(s) \\ F_1 & F_2 \end{bmatrix}. \tag{3.6}$$

Note that in the Hamiltonian case, $\varphi(s)$ is necessarily even; i.e.,

$$\varphi(s) = \varphi(-s).$$

Let $f_i,\ i = 0, \ldots, N$, be the Plücker coordinates of $\operatorname{rowsp}[F_1\ F_2]$. In terms of the Plücker coordinates the characteristic equation can then be written as

$$\det \begin{bmatrix} D(s) & N(s) \\ F_1 & F_2 \end{bmatrix} = \sum_{i=0}^{N} p_i(s) f_i, \tag{3.7}$$

where $p_i(s)$ is the cofactor of f_i in the determinant (3.7).

Let $\mathcal{Z} \subset \mathbb{P}^N$ be the linear subspace defined by

$$\mathcal{Z} = \left\{ z \in \mathbb{P}^N \,|\, \sum_{i=0}^{N} p_i(s) z_i = 0 \right\}. \tag{3.8}$$

Following [19, 27, 28, 37] we identify a closed loop characteristic polynomial $\varphi(s)$ with a point in $\mathbb{P}^\delta$. In analogy to the situation of the static pole placement problem considered in [3, 37] (compare also with [28, section 5]), one has a well-defined characteristic map

$$\begin{array}{rccc} \chi: & \mathbb{L}G(n) - \mathcal{Z} & \longrightarrow & \mathbb{P}^\delta \\ & \operatorname{rowsp}[F_1\ F_2] & \longmapsto & \sum_{i=0}^{N} f_i p_i(s) \end{array} \tag{3.9}$$

in the complex symmetric case and

$$\begin{array}{rccc} \chi': & \mathbb{L}G(n) - \mathcal{Z} & \longrightarrow & \mathbb{P}^{\delta/2} \\ & \operatorname{rowsp}[F_1\ F_2] & \longmapsto & \text{even part of } \sum_{i=0}^{N} f_i p_i(s) \end{array} \tag{3.10}$$

in the Hamiltonian case. In the latter case the reduction in dimension of the projective space arises due to the evenness of the closed loop characteristic polynomial, so that in the second map only the coefficients of the even terms of $\sum_{i=0}^{N} f_i p_i(s)$ appear.

Recall the notion of degree of a variety [13, Chapter I, section 7] and the notion of a central projection (see [32, Chapter I, section 4]). The geometric properties of the map χ are as follows.

THEOREM 3.8. *The maps χ, χ' define central projections. In particular, if $\mathcal{Z} \cap \mathbb{L}G(n) = \emptyset$ and $\dim \mathbb{L}G(n) = \binom{n+1}{2} = \delta$, then χ is surjective, and there are $\deg \mathbb{L}G(n)$ many preimages (counted with multiplicity) for each point in $\mathbb{P}^\delta$, where $\deg \mathbb{L}G(n)$ is the degree of the Lagrangian manifold $\mathbb{L}G(n)$ in $\mathbb{P}^N$. Similarly, if $\dim \mathbb{L}G(n) = \binom{n+1}{2} = \delta/2$, then χ' is surjective with exactly $\deg \mathbb{L}G(n)$ many preimage points in each fiber.*

Proof. By definition (see, e.g., [26, 32]), χ represents a central projection of $\mathbb{L}G(n)$ *from the center* $\mathcal{Z}$ *to* $\mathbb{P}^\delta$. When $\mathcal{Z} \cap \mathbb{L}G(n) = \emptyset$ and $\dim \mathbb{L}G(n) = \binom{n+1}{2} = \delta$, then χ is a finite morphism [32, Chapter I, section 5, Theorem 7] and onto of degree $\deg \mathbb{L}G(n)$ [26, Corollary 5.6]. Similar argument can be applied to χ'. ∎

The set $\mathcal{Z} \cap \mathbb{L}G(n)$ is sometimes referred to as the *base locus*. The interesting part of the theorem occurs when the base locus $\mathcal{Z} \cap \mathbb{L}G(n) = \emptyset$ since in this situation very specific information on the number of solutions is provided. If $\mathcal{Z} \cap \mathbb{L}G(n) = \emptyset$ and $\binom{n+1}{2} = \delta$ (or $n(n+1) = \delta$), then one says that χ (or χ') describes a *finite morphism* from the projective variety $\mathbb{L}G(n)$ onto the projective space $\mathbb{P}^\delta$ (or $\mathbb{P}^{\delta/2}$).

This last situation is most desirable, and this motivates the following definition.

DEFINITION 3.9. A particular symmetric transfer function $G(s)$ is called *nondegenerate* if $\mathcal{Z} \cap \mathbb{L}G(n) = \emptyset$. A system which is not nondegenerate will be called degenerate.

In terms of matrices a symmetric transfer function $G(s) = D(s)^{-1}N(s)$ is degenerate as soon as there is a Lagrangian subspace $\operatorname{rowsp}[F_1\ F_2] \in \mathbb{L}G(n)$, such that

$$\det \begin{bmatrix} D(s) & N(s) \\ F_1 & F_2 \end{bmatrix} = 0.$$

In more geometric language this means that the Hermann–Martin curve [25] defined by $\operatorname{rowsp}[D(s)\ N(s)]$ is fully contained in a Lagrangian hyperplane defined by $\operatorname{rowsp}[F_1\ F_2]$. In the study of the static pole placement problem [3] and the dynamic pole placement problem [28], definitions analogous to Definition 3.9 played an important role.

The next lemmas give specific information as to under which conditions $\mathcal{Z} \cap \mathbb{L}G(n) = \emptyset$, i.e., under which conditions a symmetric transfer function is nondegenerate. Similar results were crucial in proving the pole placement results in [3, 19, 28].

LEMMA 3.10. *If $\delta < \binom{n+1}{2} = \dim \mathbb{L}G(n)$, then every $n \times n$ symmetric transfer function of McMillan degree δ is degenerate. Similarly, any $n \times n$ Hamiltonian transfer function of McMillan degree δ is degenerate if $\delta < n(n+1) = 2 \dim \mathbb{L}G(n)$.*

Proof. $\dim \mathcal{Z} \geq N - \delta - 1$ as $\mathcal{Z}$ is defined by $\delta + 1$ linear equations ($\delta/2 + 1$ in the Hamiltonian case). If $\dim \mathbb{L}G(n) > \delta$ (or $\dim \mathbb{L}G(n) > \delta/2$ in the Hamiltonian case), then $\mathcal{Z} \cap \mathbb{L}G(n)$ is nonempty by the (projective) dimension theorem (see, e.g., [13, Chapter I, Theorem 7.2]). ∎

LEMMA 3.11. *If $\delta = \binom{n+1}{2} = \dim \mathbb{L}G(n)$ (or $\delta = n(n+1)$), then a generic set of $n \times n$ symmetric (or Hamiltonian) transfer functions of McMillan degree δ is nondegenerate.*

Proof. Let $\mathcal{Q}$ be the set of all $n \times n$ symmetric transfer functions of McMillan degree n. $\mathcal{Q}$ can be given the structure of a quasi-projective variety. For this, recall the definition of the projective variety $K^{\delta}_{n,n}$ that was introduced in [28] and which compactifies the set of all $n \times n$ transfer functions of McMillan degree δ. An element (Hermann–Martin curve) $\operatorname{rowsp}[D(s)\ N(s)] \in K^{\delta}_{n,n}$ describes an element of $\mathcal{Q}$ as soon as $\deg \det D(s) = \delta$ and $D(s)N(s)^{\top} = N(s)D(s)^{\top}$. The last condition translates into some linear conditions to be satisfied among the Plücker coordinates of $K^{\delta}_{n,n}$. The resulting subvariety of $K^{\delta}_{n,n}$ constitutes a natural compactification of $\mathcal{Q}$, and $\mathcal{Q}$ itself is a quasi-projective subset.

Consider now the coincidence set

$$\mathcal{S} := \left\{ (D(s)^{-1}N(s);\ F_1, F_2) \in \mathcal{Q} \times \mathbb{L}G(n) \mid \det \begin{bmatrix} D(s) & N(s) \\ F_1 & F_2 \end{bmatrix} = 0 \right\}.$$

Since $\mathbb{L}G(n)$ is projective, the projection onto $\mathcal{Q}$ is an algebraic set by the main theorem of elimination theory (see, e.g., [26]). The set of nondegenerate systems therefore forms a Zariski open subset of $\mathcal{Q}$. We have shown the result if we can exhibit one $n \times n$ transfer function of McMillan degree $\binom{n+1}{2}$ which is nondegenerate. The next lemma gives such an example, and the claim therefore follows. Note that the previous arguments run in a completely similar manner for the Hamiltonian case, and it therefore remains to construct one example as well. However, the symmetric Hamiltonian transfer function $G(s^2)$ is exactly such an example. ∎

LEMMA 3.12. *The symmetric transfer function*

$$G(s) := \begin{bmatrix} \frac{1}{s} & & & \\ & \frac{1}{s^2} & & \\ & & \ddots & \\ & & & \frac{1}{s^n} \end{bmatrix}$$

is nondegenerate.

Proof. First it is clear that $G(s)$ has McMillan degree $\delta = \binom{n+1}{2}$ and that

$$[D(s)\ N(s)] = \left[\begin{array}{cccc|cccc} s & & & & 1 & & & \\ & s^2 & & & & 1 & & \\ & & \ddots & & & & \ddots & \\ & & & s^n & & & & 1 \end{array}\right]$$

forms a left coprime factorization of $G(s)$. Let

$$R := \begin{bmatrix} & & & 1 \\ & & \iddots & \\ & \iddots & & \\ 1 & & & \end{bmatrix}$$

and assume by contradiction that $G(s)$ is degenerate. Then there exists rowsp $[F_1\ F_2] \in \mathbb{L}G(n)$, such that

$$0 = \det \begin{bmatrix} D(s) & N(s) \\ F_1 & F_2 \end{bmatrix} = \det \begin{bmatrix} D(s) & N(s)R \\ F_1 & F_2R \end{bmatrix}. \tag{3.11}$$

Let $S \in Gl_n$ be the matrix which transforms the $n \times 2n$ matrix $[F_1\ F_2R]$ into row reduced echelon form, i.e.,

$$[(SF_1)\ \ (SF_2R)] = \begin{bmatrix} * & \cdots & * & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ * & \cdots & * & 0 & * & \cdots & * & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ * & \cdots & * & 0 & * & \cdots & * & 0 & \cdots & * & \cdots & * & 1 & 0 & \cdots & 0 \end{bmatrix} =: [\tilde{F}_1\ \ \tilde{F}_2]. \tag{3.12}$$

Let

$$i_1 < \cdots i_k \le n < i_{k+1} < \cdots i_n \le 2n$$

be the pivot indices. We claim that the first k pivot indices determine the last $n-k$ pivot indices uniquely. For this let $\hat{i}_1 < \cdots < \hat{i}_{n-k}$ be the complementary indices of the indices $\{i_1, \ldots i_k\}$ inside the set $\{1, \ldots, n\}$. Then we claim that

$$\begin{aligned} i_{k+1} &= 2n - \hat{i}_{n-k} + 1 \\ &\vdots \\ i_n &= 2n - \hat{i}_1 + 1. \end{aligned}$$

Indeed, if this is not the case, then it follows that $\tilde{F}_1 R (\tilde{F}_2)^\top$ cannot be symmetric for any choice of values in the row reduced echelon form (3.12). On the other hand, the matrix $\tilde{F}_1 R (\tilde{F}_2)^\top$ has to be symmetric since by assumption $F_1 (F_2)^\top$ is symmetric.

The indices $i_1, \ldots, i_n$ describe the maximal Plücker coordinate (with regard to the Bruhat order) of rowsp $[F_1\ \ F_2R]$ which is nonzero, and the corresponding cofactor of $[D(s)\ \ N(s)R]$ is computed as $\pm s^\alpha$, where $\alpha = \sum_{\ell=1}^{n-k} \hat{i}_\ell$. In general there are other full-size minors (Plücker coordinates) of $[D(s)\ \ N(s)R]$ which have the form $\pm s^\alpha$. All

other Plücker coordinates with this value, however, are not comparable with regard to the Bruhat order, and since $i_1, \ldots, i_n$ was the maximal nonzero Plücker coordinate of rowsp $[F_1 \;\; F_2 R]$, it follows that the determinant expansion in (3.11) cannot be zero. This is a contradiction, and it follows that $G(s)$ is nondegenerate. □

Remark 3.13. For the static pole placement problem Brockett and Byrnes [3] showed that the osculating normal curve

$$\text{rowsp} \begin{bmatrix} 1 & s & s^2 & \ldots & \ldots & s^{m+p-1} \\ 0 & 1 & 2s & \ldots & \ldots & \binom{m+p-1}{1} s^{m+p-2} \\ \vdots & & \ddots & & & \vdots \\ 0 & \ldots & 0 & 1 & \ldots & \binom{m+p-1}{m-1} s^p \end{bmatrix} \in \text{Grass}(m, \mathbb{K}^{m+p})$$

is nondegenerate. Also in this situation the Plücker coordinates have the simple form $\pm s^\beta$, where $\beta = \sum_{\ell=1}^{m} i_\ell - \ell$ and there are no two Plücker coordinates which are comparable in the Bruhat order and give rise to the same monomial s^β.

We have now all pieces together in order to prove the main result.

Proof of Theorem 3.6. Without loss of generality we focus on the case of symmetric transfer functions. The arguments for the Hamiltonian case run in a completely similar manner. Note, however, that the closed loop characteristic polynomial of a Hamiltonian system is always an *even* polynomial. Therefore our definition of generic pole assignability for Hamiltonian systems restricts to the space of even polynomials. Since the dimension of the space of even monic polynomials of degree δ is $\delta/2$, the appropriate condition for Hamiltonian systems is $\delta/2 \leq \binom{n+1}{2}$. With these comments in mind, we return to the proof for symmetric transfer functions.

When $\delta > \binom{n+1}{2}$, then a simple dimension argument shows that the image of the characteristic map χ described in (3.9) has dimension at most $\binom{n+1}{2}$ and therefore there is a Zariski open set in $\mathbb{P}^\delta$ not in the image of χ.

When $\delta = \binom{n+1}{2}$, then Lemmas 3.11 and 3.12 show that there is a generic set of $n \times n$ symmetric transfer functions of McMillan degree δ which are nondegenerate. The characteristic map (3.9) therefore has no base locus, and every point in the image of χ has $\deg \mathbb{L}G(n)$ preimage points when counted with multiplicities. The degree of the variety $\mathbb{L}G(n)$ was recently computed by Totaro [36] and resulted in the number (3.3).

A priori the geometric formulation only predicts $\deg \mathbb{L}G(n)$ many solutions inside $\mathbb{L}G(n)$, and it is not clear if all these solutions correspond to regular feedback laws of the form $u = -Fy + v$. If $G(s)$ is a strictly proper symmetric transfer function, then this is indeed the case and the same argument applies as in [3]. □

Acknowledgment. This work was started during the Special Year in Control 2003 at the Mittag-Leffler Institute, Stockholm, whose support is gratefully acknowledged.

REFERENCES

[1] R. R. Bitmead and B. D. O. Anderson, *The matrix Cauchy index: Properties and applications*, SIAM J. Appl. Math., 33 (1977), pp. 655–672.

[2] A. Borel, *Linear Algebraic Groups*, 2nd ed., Grad. Texts in Math. 126, Springer-Verlag, New York, 1991.

[3] R. W. Brockett and C. I. Byrnes, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, 26 (1981), pp. 271–284.

[4] C. I. Byrnes, *Pole assignment by output feedback*, in Three Decades of Mathematical System Theory, Lecture Notes in Control and Inform. Sci. 135, H. Nijmeijer and J. M. Schumacher, eds., Springer-Verlag, New York, 1989, pp. 31–78.

[5] C. I. Byrnes and T. Duncan, *On certain topological invariants arising in systems theory*, in New Directions in Applied Mathematics, P. Hilton and G. Young, eds., Springer-Verlag, New York, 1982, pp. 29–71.
[6] C. I. Byrnes and T. E. Duncan, *A note on the topology of spaces of Hamiltonian transfer functions*, in Algebraic and Geometric Methods in Linear Systems Theory, AMS, Providence, RI, 1980, pp. 7–26.
[7] E. J. Davison and S. H. Wang, *On pole assignment in linear multivariable systems using output feedback*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 516–518.
[8] A. Eremenko and A. Gabrielov, *Pole placement by static output feedback for generic linear systems*, SIAM J. Control Optim., 41 (2002), pp. 303–312.
[9] P. A. Fuhrmann, *On symmetric rational transfer functions*, Linear Algebra Appl., 50 (1983), pp. 167–250.
[10] P. A. Fuhrmann and U. Helmke, *On complex parameterizations of real rational functions*, Int. J. Electron. Commun., 49 (1995), pp. 293–306.
[11] W. Fulton, *Intersection Theory*, Ergeb. Math. Grenzgeb. (3), Springer-Verlag, Berlin, Heidelberg, New York, 1984.
[12] W. Fulton, *Eigenvalues, invariant factors, highest weights, and Schubert calculus*, Bull. Amer. Math. Soc. (N.S.), 37 (2000), pp. 209–249.
[13] R. Hartshorne, *Algebraic Geometry*, Springer-Verlag, Berlin, 1977.
[14] U. Helmke and P. A. Fuhrmann, *Bezoutians*, Linear Algebra Appl., 122/123/124 (1989), pp. 1039–1097.
[15] W. Helton, J. Rosenthal, and X. Wang, *Matrix extensions and eigenvalue completions, the generic case*, Trans. Amer. Math. Soc., 349 (1997), pp. 3401–3408.
[16] R. Hermann and C. F. Martin, *Applications of algebraic geometry to systems theory.* I. IEEE Trans. Automat. Control, AC-22 (1977), pp. 19–25.
[17] H. Hiller, *Combinatorics and intersections of Schubert varieties*, Comment. Math. Helv., 57 (1982), pp. 41–59.
[18] B. Huber, F. Sottile, and B. Sturmfels, *Numerical Schubert calculus*, J. Symbolic Comput., 26 (1998), pp. 767–788.
[19] M. Kim, J. Rosenthal, and X. A. Wang, *Pole placement and matrix extension problems: A common point of view*, SIAM J. Control Optim., 42 (2004), pp. 2078–2093.
[20] H. Kimura, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 509–516.
[21] J. Leventides and N. Karcanias, *Global asymptotic linearisation of the pole placement map: A closed form solution for the constant output feedback problem*, Automatica J. IFAC, 31 (1995), pp. 1303–1309.
[22] I. G. Macdonald, *Symmetric Functions and Hall Polynomials*, 2nd ed., The Clarendon Press, Oxford University Press, New York, 1995.
[23] R. E. Mahony and U. Helmke, *System assignment and pole placement for symmetric realisations*, J. Math. Systems Estim. Control, 8 (1998), pp. 321–352.
[24] C. F. Martin and R. Hermann, *Applications of algebraic geometry to systems theory.* II. *Feedback and pole placement for linear Hamiltonian systems*, Proc. IEEE, 65 (1977), pp. 841–848.
[25] C. Martin and R. Hermann, *Applications of algebraic geometry to system theory: The McMillan degree and Kronecker indices of transfer functions as topological and holomorphic system invariants*, SIAM J. Control Optim., 16 (1978), pp. 743–755.
[26] D. Mumford, *Algebraic Geometry* I: *Complex Projective Varieties*, Springer-Verlag, Berlin, New York, 1976.
[27] M. S. Ravi, J. Rosenthal, and X. Wang, *Dynamic pole assignment and Schubert calculus*, SIAM J. Control Optim., 34 (1996), pp. 813–832.
[28] J. Rosenthal, *On dynamic feedback compensation and compactification of systems*, SIAM J. Control Optim., 32 (1994), pp. 279–296.
[29] J. Rosenthal, J. M. Schumacher, and J. C. Willems, *Generic eigenvalue assignment by memoryless real output feedback*, Systems Control Lett., 26 (1995), pp. 253–260.
[30] H. Schubert, *Kalkühl der abzählenden Geometrie*, Teubner, Leipzig, Germany, 1879.
[31] I. Schur, *Über die Darstellung der symmetrischen und der alternierenden Gruppen durch gebrochene lineare Substitutionen*, J. Reine Angew. Math., 139 (1911), pp. 155–250.
[32] I. R. Shafarevich, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, New York, 1977.
[33] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, Notices Amer. Math. Soc., 50 (2003), pp. 912–915; also available online from http://www.research.att.com/~njas/sequences/.
[34] F. Sottile, *Real rational curves in Grassmannians*, J. Amer. Math. Soc., 13 (2000), pp. 333–341.

[35] F. Sottile, *Rational curves on Grassmannians: Systems theory, reality, and transversality*, in Advances in Algebraic Geometry Motivated by Physics (Lowell, MA, 2000), Contemp. Math. 276, AMS, Providence, RI, 2001, pp. 9–42.

[36] B. Totaro, *Towards a Schubert calculus for complex reflection groups*, Math. Proc. Cambridge Philos. Soc., 134 (2003), pp. 83–93.

[37] X. Wang, *Pole placement by static output feedback*, J. Math. Systems Estim. Control, 2 (1992), pp. 205–218.

[38] X. Wang, *Grassmannian, central projection and output feedback pole assignment of linear systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 786–794.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 5, pp. 1915–1930

STABILITY AND STABILIZATION OF CONTINUOUS-TIME SWITCHED LINEAR SYSTEMS*

JOSÉ C. GEROMEL† AND PATRIZIO COLANERI‡

Abstract. This paper addresses two strategies for the stabilization of continuous-time, switched linear systems. The first one is of open loop nature (trajectory independent) and is based on the determination of a minimum dwell time by means of a family of quadratic Lyapunov functions. The relevant point on dwell time calculation is that the proposed stability condition does not require the Lyapunov function to be uniformly decreasing at every switching time. The second one is of closed loop nature (trajectory dependent) and is designed from the solution of what we call Lyapunov–Metzler inequalities from which the stability condition (including chattering) is expressed. Being nonconvex, a more conservative but simpler-to-solve version of the Lyapunov–Metzler inequalities is provided. The theoretical results are illustrated by means of examples.

Key words. switched systems, continuous-time systems, Lyapunov–Metzler inequalities, linear matrix inequality

AMS subject classifications. 34D20, 93D05

DOI. 10.1137/050646366

1. Introduction.

This paper aims to provide new results on stability analysis and stabilizing control synthesis for a continuous-time, switched linear system of the general form

$$\dot{x}(t) = A_{\sigma(t)}x(t)\ ,\ x(0) = x_0, \tag{1}$$

defined for all $t \geq 0$, where $x(t) \in \mathbb{R}^n$ is the state, $\sigma(t)$ is the switching rule, and x_0 is the initial condition. Considering a set of matrices $A_i \in \mathbb{R}^{n\times n}$, $i = 1, \ldots, N$, be given, the switching rule $\sigma(t)$, for each $t \geq 0$, is such that

$$A_{\sigma(t)} \in \{A_1, \ldots, A_N\}, \tag{2}$$

where it is clear that this model naturally imposes a discontinuity on $A_{\sigma(t)}$ since this matrix must jump instantaneously from A_i to A_j for some $i \neq j = 1, \ldots, N$ once switching occurs. In other words, $A_{\sigma(t)}$ is constrained to jump among the N vertices of the matrix polytope $\{A_1, \ldots, A_N\}$.

Stability of continuous-time switched linear systems has been addressed by several authors [3], [5], [11], [12], [14], [16], [17], [18], [20] and [23]. While the survey papers [5] and [17] give a complete and detailed description on the problems arising in this area, the recent paper [11], dealing with extensions of LaSalle's invariance principle, provides an interesting discussion on a collection of results on uniform stability of switched systems. The themes dealt with in the present paper have their roots in the conference paper [9].

*Received by the editors November 30, 2005; accepted for publication (in revised form) August 9, 2006; published electronically December 26, 2006. This research was supported by the Brazilian National Research Council (CNPq), by the Italian National Research Council (CNR), and by MIUR project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems."

http://www.siam.org/journals/sicon/45-5/64636.html

†School of Electrical and Computer Engineering, UNICAMP, CP 6101, 13083 - 970, Campinas, SP, Brazil (geromel@dsce.fee.unicamp.br).

‡Dipartimento di Elettronica e Informazione, Politecnico di Milano, 20133 Milano, Italy (colaneri@elet.polimi.it).

Generally speaking, when $\sigma(\cdot)$ is state independent, that is, when it is an a priori piecewise constant signal, the reported stability conditions are obtained using a family of symmetric and positive definite matrices $\{P_1, \ldots, P_N\}$, each one associated to the correspondent matrix of the set $\{A_1, \ldots, A_N\}$ such that a Lyapunov function $v(x(t))$ is nonincreasing for all $t \geq 0$; see [11]. In this paper, for minimum dwell time design preserving global stability it is assumed that each matrix of the set $\{A_1, \ldots, A_N\}$ is asymptotically stable but the above mentioned nonincreasing condition on the Lyapunov function is relaxed. It is replaced by the weaker condition that at every switching time t_k the sequence $v(x(t_k))$, for $k = 0, \ldots, \infty$, converges uniformly to zero. In some instances, our design procedure for the determination of the minimum dwell time, based on a quadratic guaranteed cost, is related to the results of [22] assuming further that the switching rule is not a priori given but can be taken arbitrarily, among the feasible ones; see [8]. For comparison purposes, a simple second-order example is solved, and it is shown that the estimation of the minimum dwell time provided in this paper is sensibly better than the one obtained from the classical result of [18] and the well-known *average dwell time stability* condition provided in [16], [10], and the references therein. The results obtained in this context have some resemblance with those achieved in [21], where the characterization of the exponential growth rate of switched systems is provided. However, much work is needed to establish the possible links between these two approaches.

For switched systems with $\sigma(\cdot)$ being state dependent, the stability condition is expressed in terms of a set of inequalities that we call *Lyapunov–Metzler inequalities* because the variables involved are a set of symmetric and positive definite matrices $\{P_1, \ldots, P_N\}$ and a Metzler matrix Π. The point to be noticed is that our asymptotic stability condition does not require any stability property associated to each individual matrix of the set $\{A_1, \ldots, A_N\}$ and it contains as a special case the quadratic stability condition. An important point of our main result is that it includes the stability of possible sliding modes, a fact that in the particular case $N = 2$ was observed in [16]. It is also important to stress that in [20] we can find some stability results related to the same problem (without the analysis of sliding modes) but restricted to the special case $N = 2$, which does not require the formalism based on the Lyapunov–Metzler inequalities introduced here. In our general case, the price to be paid, however, is the nonconvex nature of the the Lyapunov–Metzler inequalities being thus difficult to solve numerically. From this previous result, a more conservative but easier-to-solve asymptotic stability condition is proposed.

As a final remark, notice that the dwell time calculation provided in the first part of the paper also suggests a way to solve the state-feedback stabilization problem for a input-driven switched system characterized by the pairs (A_i, B_i). Under mild assumptions it is possible to design matrices K_i so as to stabilize the closed-loop systems $A_i + B_i K_i$. Hence, one can compute an upper bound of the dwell time to establish the time duration of the associated piecewise-constant control law. Notice that the general problem of minimization of the dwell time as a function of the design local control laws K_i is still open.

The paper is organized as follows. In section 2, time-switching control is analyzed. The switching rule $\sigma(t)$ is considered piecewise constant and a minimum dwell time preserving stability is determined. Section 3 is entirely devoted to state-switching control where the goal is to design a function $u(\cdot)$ such that the system (1) is globally asymptotically stable with the switching rule $\sigma(t) = u(x(t))$. Section 4 concludes the paper.

The notation used throughout is standard. Capital letters denote matrices, small letters denote vectors, and small Greek letters denote scalars. For matrices or vectors, $(')$ indicates transpose. For symmetric matrices, $X > 0$ (≥ 0) indicates that X is positive definite (positive semidefinite), and $\lambda_{min}(X)$, $\lambda_{max}(X)$ denote its minimum and maximum eigenvalue, respectively. The sets of real and natural numbers are denoted by $\mathbb{R}$ and $\mathbb{N}$, respectively. The $\mathcal{L}_2$ squared norm of $x(t) \in \mathbb{R}^n$ defined for all $t \geq 0$ equals $\|x(t)\|_2^2 = \int_0^\infty x(t)'x(t)dt$; see [4].

2. Time-switching control. This section is entirely dedicated to the design of a time-switching control law for the switched linear system defined by the model (1) and (2) where it is assumed that each matrix of the set $\{A_1, \ldots, A_N\}$ is asymptotically stable. The problem under consideration can be stated as follows. Determine a minimum dwell time $T_* > 0$ such that the equilibrium point $x = 0$ of the system (1) is globally asymptotically stable with the time switching control

$$\sigma(t) = i \in \{1, \ldots, N\}, \qquad t \in [t_k, t_{k+1}), \tag{3}$$

where t_k and t_{k+1} are successive switching times satisfying $t_{k+1} - t_k \geq T_*$ for all $k \in \mathbb{N}$ and the index $i \in \{1, \ldots, N\}$ selected at each instant of time $t \geq 0$ is arbitrary. Hence, asymptotic stability is preserved whenever $\sigma(t)$ remains unchanged for a period of time greater or equal to the minimum dwell time T_*. The next theorem provides the theoretical basis toward a possible solution of this problem by characterizing an upper bound for T_*. It uses the concept of multiple Lyapunov function with the innovation that the classical nonincreasing assumption at switching times is no longer needed.

THEOREM 1. *Assume that for some $T > 0$, there exists a collection of positive definite matrices $\{P_1, \ldots, P_N\}$ of compatible dimensions such that*

$$A_i'P_i + P_iA_i < 0 \quad \forall\ i = 1, \ldots, N \tag{4}$$

and

$$e^{A_i'T}P_je^{A_iT} - P_i < 0 \quad \forall\ i \neq j = 1, \ldots, N. \tag{5}$$

The time-switching control (3) *with $t_{k+1} - t_k \geq T$ makes the equilibrium solution $x = 0$ of* (1) *globally asymptotically stable.*

Proof. Consider, in accordance to (3), that $\sigma(t) = i \in \{1, \ldots, N\}$ for all $t \in [t_k, t_{k+1})$, where $t_{k+1} = t_k + T_k$ with $T_k \geq T > 0$, and that at $t = t_{k+1}$ the time-switching control jumps to $\sigma(t) = j \in \{1, \ldots, N\}$; otherwise the result trivially follows. From (4), it is seen that for all $t \in [t_k, t_{k+1})$, the time derivative of the Lyapunov function $v(x(t)) = x(t)'P_{\sigma(t)}x(t)$ along an arbitrary trajectory of (1) satisfies

$$\begin{aligned} \dot{v}(x(t)) &= x(t)'(A_i'P_i + P_iA_i)x(t) \\ &< 0, \end{aligned} \tag{6}$$

which enables us to conclude that there exist scalars $\alpha > 0$ and $\beta > 0$ such that

$$\|x(t)\|^2 \leq \beta e^{-\alpha(t-t_k)}v(x(t_k)) \quad \forall t \in [t_k, t_{k+1}). \tag{7}$$

On the other hand, using the inequalities (5) we have

$$\begin{aligned} v(x(t_{k+1})) &= x(t_{k+1})' P_j x(t_{k+1}) \\ &= x(t_k)' e^{A_i' T_k} P_j e^{A_i T_k} x(t_k) \\ &< x(t_k)' e^{A_i'(T_k - T)} P_i e^{A_i (T_k - T)} x(t_k) \\ &< x(t_k)' P_i x(t_k) \\ &< v(x(t_k)), \end{aligned} \tag{8}$$

where the second inequality holds from the fact that for every $\tau = T_k - T \geq 0$ it is true that $e^{A_i'\tau} P_i e^{A_i \tau} \leq P_i$. The consequence is that there exists $\mu \in (0,1)$ such that

$$v(x(t_k)) \leq \mu^k v(x_0) \quad \forall k \in \mathbb{N}, \tag{9}$$

which together with (7) implies that the equilibrium solution $x = 0$ of (1) is globally asymptotically stable. □

This result deserves some comments. First, it is simple to determine the scalars α, β and μ such that (7) and (9) hold. Indeed, assuming that $\{P_1, \ldots P_N\}$ satisfy the conditions of Theorem 1, then from (4) there exists $\epsilon > 0$ such that $A_i' P_i + P_i A_i' \leq -\epsilon I$ for all $i = 1, \ldots, N$ yielding $\alpha = \epsilon / \max_i \lambda_{max}(P_i) > 0$ and $\beta = 1/\min_i \lambda_{min}(P_i) > 0$. Furthermore, from (5) there exists $0 < \mu < 1$ such that $e^{A_i' T} P_j e^{A_i T} \leq \mu P_i$ for all $i \neq j = 1, \ldots, N$ leading to $v(x(t_{k+1})) \leq \mu v(x(t_k))$ and consequently (9). Second, since all matrices of the set $\{A_1, \ldots, A_N\}$ are supposed to be asymptotically stable, the constraints (4) are always feasible and the constraints (5) are satisfied when $T > 0$ is taken large enough. Third, assuming that matrices $A_1, \ldots, A_N$ are quadratically stable, which is the same as saying that they share a positive definite matrix P such that

$$A_i' P + P A_i < 0 \quad \forall i = 1, \ldots, N, \tag{10}$$

then the inequality (5) is satisfied for $P_1 = \cdots = P_N = P$ for any $T > 0$, meaning that even in the case the switching policy (3) jumps from i to j arbitrarily fast, the asymptotic stability is preserved. Hence, Theorem 1 contains, as a particular case, the quadratic stability condition. Finally, with $T > 0$ fixed it is always possible to define a time-switching control strategy (3) such that $A_{\sigma(t)}$ is periodic. As a consequence, a necessary condition for the feasibility of constraints (4) and (5) is

$$\theta(T) := \max_{q=1,\ldots,n} \left| \lambda_q \left(\prod_{p=1}^{N} e^{B_p T} \right) \right| < 1, \tag{11}$$

where $\lambda_q(\cdot)$ denotes a generic eigenvalue of $(\cdot)$ and $\{B_1, \ldots, B_N\}$ are matrices corresponding to any permutation among those of the set $\{A_1, \ldots, A_N\}$. However, since (3) may produce nonperiodic policies as well, the necessary condition (11) for the existence of a feasible solution to inequalities (4)–(5) generally does not meet sufficiency. In what follows, this aspect will be illustrated by means of an example.

Remark 1 (multiple Lyapunov functions). The important concept of multiple Lyapunov functions introduced in [3] is largely used in switching systems analysis; see also [11, 16]. In the present context, consider the family of Lyapunov functions $V_i(x(t)) = x(t)' P_i x(t)$ valid in each time interval such that $\sigma(t) = i$ for all $i = 1, \ldots, N$. From Theorem 3.1 of [16], the switched system (1) is globally asymptotically

stable whenever each Lyapunov function satisfies $V_i(x(t_q)) < V_i(x(t_p))$, where $\sigma(t_q) = \sigma(t_p) = i$ for every switching times $t_p < t_q$. For dwell time calculation, the use of this stability condition is not simple because it is difficult to impose the decreasing property for each function $V_i(\cdot)$ at nonsuccessive switching times. In Theorem 1 this constraint is replaced by $V_j(x(t_{k+1})) < V_i(x(t_k))$, where t_k and t_{k+1} are successive switching times such that $\sigma(t_k) = i$ and $\sigma(t_{k+1}) = j$. See [3] for a general discussion of these conditions applied to nonlinear switching systems. In [21], a similar stability condition is obtained based on the characterization of the exponential growth rate of switched systems.

From the previous result, an upper bound for the minimum dwell time T_* can be computed by taking the minimum value of T satisfying the conditions of Theorem 1. Hence, it can be calculated with no great difficulty from the optimal solution of the optimization problem[1]

$$\min_{T>0, P_1>0, \ldots, P_N>0} \{T \ : \ (4) \text{ and } (5)\}, \tag{12}$$

which, for each $T > 0$ fixed, reduces to a convex programming problem with linear matrix inequality constraints that can be handled by any LMI solver available in the literature to date; see [2] for an important study on systems and LMIs. A line search procedure is then used to deal with the scalar variable $T > 0$.

Finally, it is possible to generalize the result of Theorem 1 to define a guaranteed cost-to-go from an arbitrary initial point to the origin, associated to the stabilizing time-switching rule (3) with $t_{k+1} - t_k \geq T$ for any fixed $T > 0$. To this end, we make the assumption that $\mathcal{T} > 0$ is known such that $t_{k+1} - t_k \leq \mathcal{T}$ for all $k \in \mathbb{N}$. Clearly, these quantities are related through $\mathcal{T} \geq T \geq T_*$, where the second inequality assures global stability.

THEOREM 2. *Let $Q \geq 0 \in \mathbb{R}^{n\times n}$ and $\mathcal{T} \geq T > 0$ be given. Define the set of symmetric, nonnegative definite matrices*

$$R_i := \int_0^{\mathcal{T}} e^{A_i' t} Q e^{A_i t} dt, \qquad i = 1, \ldots, N. \tag{13}$$

Assume that there exists a collection of positive definite matrices $\{P_1, \ldots, P_N\}$ of compatible dimensions such that

$$A_i' P_i + P_i A_i + Q < 0 \ \ \forall\ i = 1, \ldots, N \tag{14}$$

and

$$e^{A_i' T} P_j e^{A_i T} - P_i + R_i < 0 \ \ \forall\ i \neq j = 1, \ldots, N. \tag{15}$$

The time-switching control (3) with $\mathcal{T} \geq t_{k+1} - t_k \geq T$ makes the equilibrium solution $x = 0$ of (1) globally asymptotically stable and

$$\int_0^{\infty} x(t)' Q x(t) dt < x_0' P_{\sigma(0)} x_0. \tag{16}$$

Proof. Since for $Q \geq 0$ and $\mathcal{T} \geq T > 0$ given, each matrix R_i defined in (13) is positive semidefinite and inequalities (14)–(15) are satisfied, then inequalities (4)–(5)

[1]This problem should be stated with inf instead of min. All feasible sets of problems expressed in terms of LMIs must be considered closed from the interior within a precision defined by the user.

are also satisfied. As a consequence, asymptotic stability follows from Theorem 1. On the other hand, using (13) together with the inequalities (14) and (15) we have that $P_i > R_i$ and

$$\begin{aligned} A_i'(P_i - R_i) + (P_i - R_i)A_i &< -Q - A_i'R_i - R_iA_i \\ &< -Q - \int_0^{\mathcal{T}} \frac{d}{dt} e^{A_i' t} Q e^{A_i t} dt \\ &< -e^{A_i'\mathcal{T}} Q e^{A_i \mathcal{T}} \\ &< 0 \end{aligned} \tag{17}$$

for all $i = 1, \ldots, N$. The important consequence of this calculation is that for each $i = 1, \ldots, N$ the inequality $e^{A_i'\tau}(P_i - R_i)e^{A_i\tau} \leq (P_i - R_i)$ holds for any $\tau \geq 0$. Using this property, taking into account the switching strategy (3) with $t_{k+1} - t_k = T_k \geq T$ and the inequalities (15) one obtains

$$\begin{aligned} v(x(t_{k+1})) &= x(t_{k+1})' P_j x(t_{k+1}) \\ &< x(t_k)' e^{A_i'(T_k - T)}(P_i - R_i) e^{A_i(T_k - T)} x(t_k) \\ &< x(t_k)'(P_i - R_i)x(t_k) \\ &< v(x(t_k)) - x(t_k)' R_{\sigma(t_k)} x(t_k), \end{aligned} \tag{18}$$

which summing up for all $k \in \mathbb{N}$ and taking into account that $\mathcal{T} \geq t_{k+1} - t_k$ allows us to write

$$\begin{aligned} \int_0^\infty x(t)' Q x(t) dt &= \sum_{k=0}^\infty \int_{t_k}^{t_{k+1}} x(t_k)' e^{A_i'(t - t_k)} Q e^{A_i(t - t_k)} x(t_k) dt \\ &\leq \sum_{k=0}^\infty x(t_k)' R_{\sigma(t_k)} x(t_k) \\ &< v(x_0), \end{aligned} \tag{19}$$

which proves the proposed theorem. □

It is interesting to observe that the conditions of Theorem 2 are feasible if and only if $\mathcal{T} \geq T \geq T_*$, and from (16) it is seen that a more accurate guaranteed cost is obtained whenever the value of $\mathcal{T}$ is chosen as small as possible. In addition, the choice $\mathcal{T} = +\infty$ enables us to conclude that the proposed time-switching rule (3) with $t_{k+1} - t_k \geq T_*$, makes the trajectory $y(t) = Q^{1/2}x(t), t \geq 0$, quadratically integrable. Theorem 2 admits the extreme situation $\mathcal{T} = T = +\infty$ for which no jump occurs and inequalities (14)–(15) are verified for

$$P_i = \int_0^\infty e^{A_i' t}(Q + \varepsilon I) e^{A_i t} dt > R_i \geq 0 \ , \ i = 1, \ldots, N \tag{20}$$

with $\varepsilon > 0$ arbitrary. When $\varepsilon > 0$ goes to zero, P_i goes to R_i and (16) becomes a well-known result. On the other hand, for $T > 0$ arbitrarily small and any $\mathcal{T} \geq T$, feasibility holds whenever the set of matrices $\{A_1, \ldots, A_N\}$ admits a common Lyapunov function.

Example 1. For illustration purposes of the theoretical results obtained so far, let us consider the example with $N = 2$ and matrices

$$A_1 = \begin{bmatrix} 0 & 1 \\ -10 & -1 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 \\ -0.1 & -0.5 \end{bmatrix}, \tag{21}$$

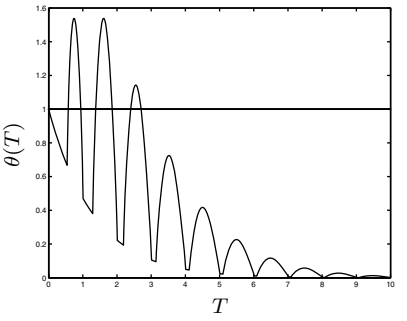


FIG. 1. *The function* $\theta(T)$.

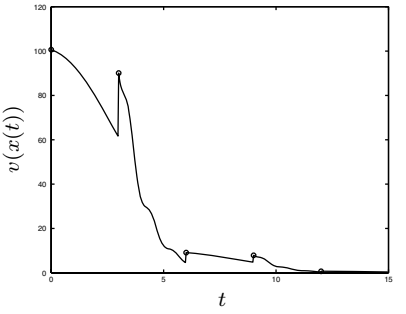


FIG. 2. *The Lyapunov function.*

which are not quadratically stable. First, from problem (12), we have calculated an upper bound for the minimum dwell time as being $T_* \leq 2.76$. To give an idea of its conservativeness we have calculated from the plot of Figure 1 the value $T_{per} = 2.71$ corresponding to the necessary condition for stability (11), arising from linear periodic systems. Both being very close indicates, for this simple example, a good precision on the determination of T_*. On the other hand, for comparison we have applied the classical result of [18] for the determination of an alternative upper bound for the minimum dwell time T_* given by $T_* \leq \max_{i=1,\ldots,N}\{T_i\}$, where

$$T_i = \inf_{\alpha>0,\beta>0} \left\{ \frac{\alpha}{\beta} \ : \ \|e^{A_i t}\| < e^{(\alpha-\beta t)} \quad \forall t \geq 0 \right\}. \tag{22}$$

For the matrices in (21), we have numerically determined $T_1 = 2.33$ and $T_2 = 6.66$, yielding an estimation for the minimum dwell time as being $T_* \leq 6.66$. Hence, in this particular example, the result provided by the solution of problem (12) is much more precise but at the expense of a more expressive computational effort. Moreover, using the method in [10] we have obtained the upper bound of the so-called average dwell time as being 16.5554, which is obviously greater than that obtained by our method.

Figure 2 has been constructed by simulation of system (1) with the time-switching rule (3), $t_{k+1} - t_k = 3.0$, initial conditions $x_0 = [1\ 1]'$, $\sigma(0) = 2$, and $Q = I$. The family of Lyapunov functions has been calculated from the optimal solution of the convex programming problem

$$\min_{P_1>0,\ldots,P_N>0} \max_{i=1,\ldots,N} \{x_0' P_i x_0 \ : \ (14) \text{ and } (15)\}, \tag{23}$$

which puts in evidence that a guaranteed cost can be determined for the worst case as far as the initial condition $\sigma(0)$ appearing in (16) is concerned. For $\mathcal{T} = T = 3.0$, we have obtained the minimum guaranteed cost equal to $\delta^* = 100.61$, valid for both initial conditions. As noted earlier, the Lyapunov function $v(x(t)) = x(t)'P_{\sigma(t)}x(t)$ goes to zero as t goes to infinity; however, it is not uniformly decreasing with respect to time. In Figure 2, due to the stability conditions of Theorem 2, the discontinuity points, marked with circles, defines a globally convergent sequence $v(x(t_k))$ for all $k \in \mathbb{N}$. Solving again problem (23) but for $\mathcal{T} = +\infty$ and $T = 3.0$, the minimum guaranteed cost increases to $\delta^* = 147.94$ as a consequence of allowing a more flexible switching rule (3) with $t_{k+1} - t_k \geq 3.0$.

The example above shows that there is a clear improvement on stability conditions, dwell time, and guaranteed cost calculations when compared to the results available in the literature to date; see [11], [18].

3. State-switching control. In this section, we consider once again the system (1) where the switching rule satisfies (2). The main difference from the previous section is that, presently, it is assumed that the state vector $x(t)$ is available for feedback for all $t \geq 0$. That is, our goal is to determine the function $u(\cdot) : \mathbb{R}^n \to \{1, \ldots, N\}$ such that

$$\sigma(t) = u(x(t)) \tag{24}$$

makes the equilibrium point $x = 0$ of (1) asymptotically stable. In this case, we do not assume that each matrix of the set $\{A_1, \ldots, A_N\}$ is asymptotically stable. To this end, let us define the simplex

$$\Lambda := \left\{ \lambda \in \mathbb{R}^N \ : \ \sum_{i=1}^{N} \lambda_i = 1 \ , \ \lambda_i \geq 0 \right\}, \tag{25}$$

which together with the set of positive definite matrices $\{P_1, \ldots, P_N\}$ enables us to introduce the following piecewise quadratic Lyapunov function:

$$v(x) := \min_{i=1,\ldots,N} x'P_i x = \min_{\lambda \in \Lambda} \left(\sum_{i=1}^{N} \lambda_i x'P_i x \right). \tag{26}$$

As it will be clear in what follows, this Lyapunov function is crucial to our purposes; see [1] and the references therein. However, it presents some difficulties to be handled, including the fact that it is not differentiable everywhere. To analyze this aspect, the set $I(x) = \{i : v(x) = x'P_i x\}$ plays a central role since $v(x)$ fails to be differentiable on $x \in \mathbb{R}^n$ such that $I(x)$ is composed of more than one element or, in other words, when the result of the minimization indicated in (26) is not unique, [19].

Before proceeding, let us recall the class of Metzler matrices denoted by $\mathcal{M}$ and consisting of all matrices $\Pi \in \mathbb{R}^{N \times N}$ with elements π_{ij}, such that

$$\pi_{ij} \geq 0 \ \forall i \neq j, \quad \sum_{i=1}^{N} \pi_{ij} = 0 \ \forall j. \tag{27}$$

It is clear that any $\Pi \in \mathcal{M}$ presents an eigenvalue at the origin of the complex plane since $c'\Pi = 0$, where $c' = [1 \ \cdots \ 1]$. In addition, it is well known from the Frobenius–Perron theorem that the eigenvector associated to the null eigenvalue of

Π is nonnegative, yielding the conclusion that there always exists $\lambda_\infty \in \Lambda$ such that $\Pi\lambda_\infty = 0$. The next theorem summarizes the main result of this section.

THEOREM 3. *Assume that there exist a set* $\{P_1, \ldots, P_N\}$ *of positive definite matrices and* $\Pi \in \mathcal{M}$ *satisfying the Lyapunov–Metzler inequalities*

$$A_i' P_i + P_i A_i + \sum_{j=1}^{N} \pi_{ji} P_j < 0, \qquad i = 1, \ldots, N. \tag{28}$$

The state-switching control (24) *with*

$$u(x(t)) = \arg \min_{i=1,\ldots,N} x(t)' P_i x(t) \tag{29}$$

makes the equilibrium solution $x = 0$ *of* (1) *globally asymptotically stable.*

Proof. It follows from the Lyapunov function (26), which, as we have said before, is not differentiable for all $t \geq 0$. For this reason we need to deal with the Dini derivative (see [7]):

$$D^+ v(x(t)) = \lim_{h \to 0^+} \sup \frac{v(x(t+h)) - v(x(t))}{h}. \tag{30}$$

Assume, in accordance to (29), that at an arbitrary $t \geq 0$, the state-switching control is given by $\sigma(t) = u(x(t)) = i$ for some $i \in I(x(t))$. Hence, from (30) and the system dynamic equation (1), applying the result of Theorem 1 of [15, p. 420] we have

$$\begin{aligned} D^+ v(x(t)) &= \lim_{h \to 0^+} \sup \frac{v(x(t) + hA_i x(t)) - v(x(t))}{h} \\ &= \min_{l \in I(x(t))} x(t)'(A_i' P_l + P_l A_i) x(t) \\ &\leq x(t)'(A_i' P_i + P_i A_i) x(t), \end{aligned} \tag{31}$$

where the inequality holds from the fact that $i \in I(x(t))$. Finally, remembering that (27) is valid for $\Pi \in \mathcal{M}$ and that $x(t)' P_j x(t) \geq x(t)' P_i x(t)$ for all $j \neq i = 1, \ldots, N$ once again due to the fact that $i \in I(x(t))$, using the Lyapunov–Metzler inequalities (28) one obtains

$$\begin{aligned} D^+ v(x(t)) &< -x(t)' \left(\sum_{j=1}^{N} \pi_{ji} P_j \right) x(t) \\ &< - \left(\sum_{j=1}^{N} \pi_{ji} \right) x(t)' P_i x(t) \\ &< 0, \end{aligned} \tag{32}$$

which proves the proposed theorem since the Lyapunov function $v(x(t))$ defined in (26) is radially unbounded. ∎

It is important to observe that Theorem 3 does not require the set $\{A_1, \ldots, A_N\}$ be exclusively composed of asymptotically stable matrices. Indeed, with $\Pi \in \mathcal{M}$, a necessary condition for the Lyapunov–Metzler inequalities to be feasible with respect to $\{P_1, \ldots, P_N\}$ is matrices $A_i + (\pi_{ii}/2)I$ for all $i = 1, \ldots, N$ be asymptotically stable. Since $\pi_{ii} \leq 0$ this condition does not imply the asymptotic stability of A_i. However,

an interesting case occurs when all matrices $\{A_1, \ldots, A_N\}$ are asymptotically stable, for which the choice $\Pi = 0$ is possible and the state-switching strategy proposed preserves stability. Furthermore, if the set $\{A_1, \ldots, A_N\}$ is quadratically stable, then the Lyapunov–Metzler inequalities admit a solution $P_1 = \cdots = P_N = P$ and $I(x(t)) = \{1, \ldots, N\}$ for all $t \geq 0$. In this classical but particular case, at any $t \geq 0$, the control law $u(x(t)) = i \in \{1, \ldots, N\}$ can be chosen arbitrarily and asymptotic stability is guaranteed. Hence, Theorem 3 contains as a particular case (since the Lyapunov–Metzler inequalities do not depend on Π anymore) the quadratic stability condition.

Remark 2 (chattering). Another important feature of Theorem 3 is that chattering in the switching, when it occurs, is always stable. Indeed, assume that $x \in \mathbb{R}^n$ belongs to a certain region $\mathcal{R}$ of the state space where the cardinality of $I(x)$ is greater than one. From the Lyapunov function (26) and the time derivative (31) a switching from any $i \in I(x)$ to some $j \in I(x)$ is possible only if $x'(A_i'P_j + P_jA_i)x \leq x'(A_i'P_i + P_iA_i)x < 0$, where the second inequality follows from (28). Hence, with scalars $\alpha_i \geq 0\ \forall i \in I(x)$ satisfying $\sum_{i \in I(x)} \alpha_i = 1$, we conclude that whenever $x \in \mathcal{R}$ the time derivative of the positive definite function $\nu(x) = x'P_jx$ is strictly negative along the trajectories of $\dot{x} = (\sum_{i \in I(x)} \alpha_i A_i)x$. From Filippov's definition this implies that the switched system is asymptotically stable. In the particular case characterized by $N = 2$, this aspect has already been treated in [16, pp. 70]. In [20] it is noted that a Lyapunov function like (26) but with min replaced by max does not exhibit this property, in which instance the chattering must be ruled out. In this sense, the numerical procedure proposed in [13] for the determination of a switching-state dependent control has to be further qualified to prevent chattering since when it occurs instability may be observed.

In the literature, the Lyapunov–Metzler inequalities with $\Pi \in \mathcal{M}$ fixed have been introduced to study the mean-square (MS) stability of Markov jump linear systems (MJLS). In that context, the Metzler matrix $\Pi = \Pi_0 \in \mathcal{M}$ is given and Π_0' represents the infinitesimal transition matrix of a Markov chain $\sigma(t)$ governing the dynamical system (1). In this respect, each component of the vector $\lambda(t) \in \Lambda$ is the probability of the Markov chain to be in the ith logical state and obeys the differential equation

$$\dot{\lambda}(t) = \Pi_0 \lambda(t)\ ,\ \lambda(0) = \lambda_0 \in \Lambda, \tag{33}$$

where the eigenvector $\lambda_\infty \in \Lambda$ associated to the null eigenvalue of Π_0 represents the stationary probability vector. Hence, using the fact that the stochastic system under consideration is said to be MS-stable if

$$\lim_{t \to +\infty} E(\|x(t)\|^2) = 0 \tag{34}$$

for any initial state $x(0)$ and any initial probability pattern $\lambda_0 \in \Lambda$, it has been shown (see, e.g., [6]) that the system is MS-stable if and only if there exists a set of positive definite matrices $\{P_1, \ldots, P_N\}$ satisfying the Lyapunov–Metzler inequalities (28) for $\Pi = \Pi_0$. Numerically speaking, this is a simple case, since (28) reduces to a set of linear matrix inequalities.

A relevant point to be discussed now concerns the existence of a solution of the Lyapunov–Metzler inequalities (28) with respect to the variables $\Pi \in \mathcal{M}$ and $\{P_1, \ldots, P_N\}$. Standard Kronecker calculus shows that for $\Pi \in \mathcal{M}$ fixed, a solution with respect to the remaining variables exists if and only if the Nn^2-dimensional

square matrix $\mathcal{J} := \mathcal{A} + \mathcal{BC}$ is asymptotically stable, where

$$\mathcal{A} = \begin{bmatrix} A_1' \oplus A_1' & 0 & \cdots & 0 \\ 0 & A_2' \oplus A_2' & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & A_N' \oplus A_N' \end{bmatrix} \tag{35}$$

and

$$\mathcal{B} = \Pi' \begin{bmatrix} 0_{N-1} \\ I_{N-1} \end{bmatrix} \otimes I_{n^2}, \qquad \mathcal{C} = \begin{bmatrix} -1_{N-1} & I_{N-1} \end{bmatrix} \otimes I_{n^2}, \tag{36}$$

with the symbols $\oplus$ and $\otimes$ indicating the Kronecker sum and Kronecker product respectively,[2] 0_{N-1} denoting a row vector of $N-1$ zeros components, and 1_{N-1} denoting a column vector of $N-1$ ones components. Hence, the existence of a solution to (28) reduces to the existence of $\Pi \in \mathcal{M}$ rendering matrix $\mathcal{J}$ asymptotically stable. A possible approach to verify the existence of such a matrix is based on the observation that any $\alpha \geq 0$ and $\Pi \in \mathcal{M}$ implies $\alpha\Pi \in \mathcal{M}$, which from the introduction of this new degree of freedom makes possible to verify the existence of $\alpha \geq 0$ such that $\mathcal{J}(\alpha) := \mathcal{A} + \alpha\mathcal{BC}$ is asymptotically stable. Putting aside the situation in which all matrices $\{A_1, \ldots, A_N\}$ are asymptotically stable (making it possible to set $\alpha = 0$), let us consider the other extreme situation corresponding to $\alpha \to +\infty$. Simple determinant manipulations show that a certain number of eigenvalues goes to $-\infty$ while the other ones that remain finite, coincide with the *invariant zeros* of the triple $(\mathcal{A}, \mathcal{B}, \mathcal{C})$.

Fortunately, these invariant zeros can be determined with no great difficulty from the definition

$$\begin{bmatrix} \mu I - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} = 0 \tag{37}$$

with the key observation that matrix $\mathcal{C}$, being constant, is independent of α and Π which imposes to the solution of $\mathcal{C}\xi = 0$ a vector of compatible dimension with the particular structure $\xi' = [x' \ \cdots \ x'], x \in \mathbb{R}^{n^2}$. In addition, taking $\lambda_\infty \in \Lambda$ such that $\Pi\lambda_\infty = 0$, multiplying each subequation above by $\lambda_{\infty i}$ and summing up, it follows that

$$\left(\mu I - \sum_{i=1}^{N} \lambda_{\infty i} A_i' \oplus A_i' \right) x = 0, \tag{38}$$

which can be rewritten as

$$\left(\mu I - A_{\lambda_\infty}' \oplus A_{\lambda_\infty}' \right) x = 0, \tag{39}$$

where $A_{\lambda_\infty} = \sum_{i=1}^{N} \lambda_{\infty i} A_i$. Therefore, as α goes to infinity, the eigenvalues of $\mathcal{J}(\alpha)$ that remain finite tend to the eigenvalues of $A_{\lambda_\infty}' \oplus A_{\lambda_\infty}'$ which are in the left-hand plane if and only if the eigenvalues of A_{λ_∞} are. This means that if there exists $\lambda_\infty \in \Lambda$

[2]While the Kronecker product is more or less standard, the sum requires a formal definition. In this respect we define the Kronecker sum of two matrices D and E as $D \oplus E = D \otimes I + I \otimes E$. It is important to recall that the eigenvalues of the Kronecker sum $D \oplus E$ are given by all sums of all eigenvalues of D and E.

such that A_{λ_∞} is asymptotically stable, then any $\Pi_0 \in \mathcal{M}$ satisfying $\Pi_0 \lambda_\infty = 0$ and α a sufficiently large positive number provide $\Pi = \alpha\Pi_0 \in \mathcal{M}$ such that the Lyapunov–Metzler inequalities are feasible with respect to the remaining variables $\{P_1, \ldots, P_N\}$.

Example 2. To illustrate the above point, let us consider a simple example with $N = 2$, the pair of matrices

$$A_1 = \begin{bmatrix} 0 & 1 \\ 2 & -9 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 \\ -2 & 8 \end{bmatrix}, \tag{40}$$

and

$$\Pi_0 = \begin{bmatrix} -0.51 & 0.49 \\ 0.51 & -0.49 \end{bmatrix} \in \mathcal{M}. \tag{41}$$

The eigenvector associated to the null eigenvalue of Π_0 is $\lambda_\infty' = [0.49 \;\; 0.51]$. We have determined numerically that the Lyapunov–Metzler inequalities (28) have a solution of the form $\Pi = \alpha\Pi_0$ for all $\alpha \geq 615.7374$, in accordance to the fact that the invariant zeros of the triple $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ are $-0.33, -0.33, -0.33 \pm j0.226$, which, as discussed before, can alternatively be obtained from the eigenvalues of the asymptotically stable matrix $A_{\lambda_\infty} = 0.49A_1 + 0.51A_2$, taking all sums.

The existence of this particular solution to the inequalities (28) meets exactly the already classical stability condition provided in [16] and more recently in [24] as a particular case of switched nonlinear systems. In our present context, let us assume that there exits $\lambda_\infty \in \Lambda$ such that A_{λ_∞} is asymptotically stable, making possible the determination of $P > 0$ satisfying the Lyapunov inequality $A_{\lambda_\infty}' P + P A_{\lambda_\infty} < 0$. Hence, the switching rule (24) with

$$u(x(t)) = \arg \min_{i=1,\ldots,N} x(t)' \left(A_i' P + P A_i\right) x(t) \tag{42}$$

makes the equilibrium point $x = 0$ of the switched system (1) globally asymptotically stable. Indeed, considering the Lyapunov function $v(x(t)) = x(t)'Px(t)$ we have

$$\begin{aligned} \dot{v}(x(t)) &= x(t)' \left(A_{\sigma(t)}' P + P A_{\sigma(t)}\right) x(t) \\ &= \min_{i=1,\ldots,N} x(t)' \left(A_i' P + P A_i\right) x(t) \\ &= \min_{\lambda \in \Lambda} x(t)' \left(A_\lambda' P + P A_\lambda\right) x(t) \\ &\leq x(t)' \left(A_{\lambda_\infty}' P + P A_{\lambda_\infty}\right) x(t) \\ &< 0. \end{aligned} \tag{43}$$

It is important to keep in mind that if the set of matrices $\{A_1, \ldots, A_N\}$ does not admit A_λ asymptotically stable for some $\lambda \in \Lambda$, then the above stabilizing switching rule cannot be determined. In addition, even if it is known that there exists $\lambda \in \Lambda$ such that A_λ is asymptotically stable, the numerical determination of $\lambda \in \Lambda$ and $P > 0$ such that $A_\lambda' P + P A_\lambda < 0$ is not a simple task due to the nonlinear nature of this inequality. The Lyapunov–Metzler inequalities introduced in Theorem 3 suffer the same difficulty, but fortunately a simple numerical procedure based on line search can be settled to determine its solution. This aspect will be considered next. First, let us introduce a guaranteed quadratic cost associated to the proposed state switching control law (29).

LEMMA 1. *Let $Q \geq 0$ be given. Assume that there exist a set of positive definite matrices $\{P_1, \ldots, P_N\}$ and $\Pi \in \mathcal{M}$ satisfying the Lyapunov–Metzler inequalities*

$$A_i' P_i + P_i A_i + \sum_{j=1}^{N} \pi_{ji} P_j + Q < 0, \qquad i = 1, \ldots, N. \tag{44}$$

The state-switching control (24) *with $u(x(t))$ given by* (29) *makes the equilibrium solution $x = 0$ of* (1) *globally asymptotically stable and*

$$\int_0^\infty x(t)' Q x(t) dt < \min_{i=1,\ldots,N} x_0' P_i x_0. \tag{45}$$

Proof. The proof has the same pattern of the proof of Theorem 3. The Lyapunov function (26) and the Lyapunov–Metzler inequalities (44) yield

$$D^+ v(x(t)) < -x(t)' Q x(t), \tag{46}$$

which after integration gives

$$\begin{aligned} v(x(t)) - v(x(0)) &= \int_0^t D^+ v(x(\tau)) d\tau \\ &< -\int_0^t x(\tau)' Q x(\tau) d\tau \quad \forall t \geq 0, \end{aligned} \tag{47}$$

thus proving the proposed lemma since, due to the asymptotic stability, $v(x(t))$ goes to zero as t goes to infinity. □

The numerical determination, if any, of a solution of the Lyapunov–Metzler inequalities with respect to the variables $(\Pi, \{P_1, \ldots, P_N\})$ is not a simple task and certainly deserves additional attention. The main source of difficulty stems from its nonconvex nature due to the products of variables and so LMI solvers do not apply. Perhaps a point to be further investigated is that its particular structure with π_{ji} being scalars may help with the design of an interactive method based on relaxation.

In this paper we pursue an alternative route. The main idea is to obtain a simpler, although certainly more conservative, stability condition that can be expressed by means of LMIs and thus solvable by the machinery available in the literature to date. The next theorem shows that working with a subclass of Metzler matrices, characterized by having the same diagonal elements, this goal is accomplished.

THEOREM 4. *Let $Q \geq 0$ be given. Assume that there exist a set of positive definite matrices $\{P_1, \ldots, P_N\}$ and a scalar $\gamma > 0$ satisfying the modified Lyapunov–Metzler inequalities*

$$A_i' P_i + P_i A_i + \gamma(P_j - P_i) + Q < 0, \qquad j \neq i = 1, \ldots, N. \tag{48}$$

The state-switching control (24) *with $u(x(t))$ given by* (29) *makes the equilibrium solution $x = 0$ of* (1) *globally asymptotically stable and*

$$\int_0^\infty x(t)' Q x(t) dt < \sum_{i=1}^{N} x_0' P_i x_0. \tag{49}$$

Proof. The proof follows from the choice of $\Pi \in \mathcal{M}$ such that $\pi_{ii} = -\gamma$ and the remaining elements satisfying

$$\gamma^{-1} \sum_{j \neq i = 1}^{N} \pi_{ji} = 1 \tag{50}$$

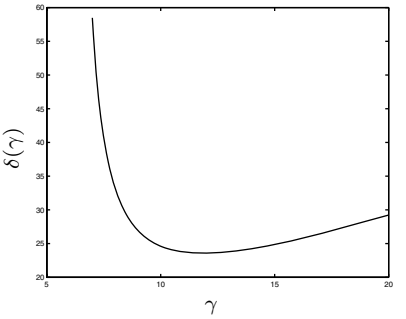


FIG. 3. *Guaranteed cost as a function of* γ.

for all $i = 1, \ldots, N$. Taking into account that $\pi_{ji} \geq 0$ for all $j \neq i = 1, \ldots, N$, multiplying (48) by π_{ji}, summing up for all $j \neq i = 1, \ldots, N$, and finally multiplying the result by $\gamma^{-1} > 0$ we obtain

$$\begin{aligned} A_i' P_i + P_i A_i + Q &< -\sum_{j \neq i = 1}^{N} \pi_{ji}(P_j - P_i) \\ &< -\sum_{j=1}^{N} \pi_{ji} P_j \end{aligned} \tag{51}$$

for all $i = 1, \ldots, N$, which are the Lyapunov–Metzler inequalities (44). From Lemma 1, the upper bound (45) holds, which trivially implies that (49) is verified. The proposed theorem is thus proved. □

The basic theoretical features of Theorem 3 and Lemma 1 are still present in Theorem 4. The most important is that the asymptotic stability of the set of matrices $\{A_1, \ldots, A_N\}$ is still not required. In addition, notice that the guaranteed cost (49) is clearly worse than the one provided by Lemma 1, but the former being convex makes it possible to solve the problem

$$\min_{\gamma > 0, P_1 > 0, \ldots, P_N > 0} \left\{ \sum_{i=1}^{N} x_0' P_i x_0 \ : \ (48) \right\} \tag{52}$$

by using LMI solvers and line search. The next example illustrates some aspects of the theoretical results obtained so far.

Example 3. Consider the system (1) with $N = 2$ and matrices $\{A_1, A_2\}$ given by

$$A_1 = \begin{bmatrix} 0 & 1 \\ 2 & -9 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 \\ -2 & 2 \end{bmatrix}, \tag{53}$$

which, as it can be easily verified by inspection, are both unstable. Considering $Q = I$ and the initial condition $x_0 = [1 \ 1]'$, problem (52) has been solved by line search fixing γ and minimizing its objective function, denoted by $\delta(\gamma)$, with respect to the remaining variables. Figure 3 shows the behavior of the function $\delta(\gamma)$ which enables us to determine its minimum value $\delta^* = 23.56$, corresponding to $\gamma^* = 11.80$. It is important to stress that, in this particular example, the function $\delta(\gamma)$ has a unique minimum. However, we do not have any evidence that this is a generic property valid

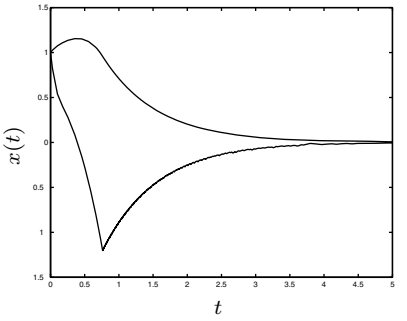


FIG. 4. *Time simulation of the state switching control.*

in all cases. Figure 4 shows the trajectories of the state variable $x(t) \in \mathbb{R}^2$ versus time for the system controlled by the state switching rule $\sigma(t) = u(x(t))$ given by (29) with the positive definite matrices

$$P_1 = \begin{bmatrix} 6.7196 & 1.6293 \\ 1.6293 & 1.0222 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 6.0825 & 2.1293 \\ 2.1293 & 2.2206 \end{bmatrix} \tag{54}$$

obtained from the optimal solution of problem (52). As it can be seen, the proposed control strategy is very effective in stabilizing the system under consideration.

4. Conclusion. In this paper we have introduced stability conditions for switched linear systems. They have been used for control synthesis of state-independent (open loop) and state-dependent (closed loop) switching rules. In both cases, the determination of a guaranteed cost associated to each control strategy has been addressed. Special attention has been devoted to the numerical solvability of the design problems by means of methods based on linear matrix inequalities.

Two issues deserve more attention. The first is related to the development of numerical algorithms for the solution of the Lyapunov–Metzler inequalities introduced in section 3. The second is the possible generalization of the stability conditions to cope with linear control design and guaranteed quadratic cost. Taking into account the nonlinear nature of the involved stability conditions, this point constitutes a real theoretical challenge.

Acknowledgment. The authors would like to thank the anonymous reviewers for many helpful comments and suggestions that certainly contributed to improve this paper.

REFERENCES

[1] F. BLANCHINI AND S. MIANI, *A new class of universal Lyapunov functions for the control of uncertain linear systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 641–647.

[2] S. P. BOYD, L. EL GHAOUI, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.

[3] M. S. BRANICKY, *Multiple Lyapunov functions and other analysis tools for switched and hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 475–482.

[4] P. COLANERI, J. C. GEROMEL, AND A. LOCATELLI, *Control Theory and Design—An RH_2 and RH_∞ viewpoint*, Academic Press, New York, 1997.

[5] R. A. DECARLO, M. S. BRANICKY, S. PETTERSSON, AND B. LENNARTSON, *Perspectives and results on the stability and stabilizability of hybrid systems*, Proc. IEEE, 88 (2000), pp. 1069–1082.

[6] X. Feng, K. A. Loparo, Y. Ji, and H. J. Chizeck, *Stochastic stability properties of jump linear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 38–53.
[7] K. M. Garg, *Theory of Differentiation: A Unified Theory of Differentiation Via New Derivate Theorems and New Derivatives*, Wiley-Interscience, New York, 1998.
[8] J. C. Geromel, M. C. de Oliveira, and L. Hsu, *LMI characterization of structural and robust stability*, Linear Algebra Appl., 285 (1998), pp. 69–80.
[9] J. C. Geromel and P. Colaneri, *Stabilization of Continuous-Time Switched Systems*, IFAC World Congress, Prague, 2005.
[10] J. P. Hespanha and A. S. Morse, *Stability of switched systems with average dwell-time*, in Proceedings of the 38th IEEE Conference on Decision and Control, 1999, pp. 2655–2660.
[11] J. P. Hespanha, *Uniform stability of switched linear systems: Extensions of LaSalle's principle*, IEEE Trans. Automat. Control, 49 (2004), pp. 470–482.
[12] J. Hockerman-Frommer, S. R. Kulkarni, and P. J. Ramadge, *Controller switching based on output predictions errors*, IEEE Trans. Automat. Control, 43 (1998), pp. 596–607.
[13] H. Ishii, T. Basar, and R. Tempo, *Randomized algorithms for synthesis of switching rules of multimodal systems*, IEEE Trans. Automat. Control, 50 (2005), pp. 754–767.
[14] M. Johansson and A. Rantzer, *Computation of piecewise quadratic Lyapunov functions for hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 555–559.
[15] L. S. Lasdon, *Optimization Theory for Large Scale Systems*, Series of Operation Research, MacMillan, New York, 1970.
[16] D. Liberzon, *Switching in Systems and Control*, Birkhäuser, Boston, 2003.
[17] D. Liberzon and A. S. Morse, *Basic problems in stability and design of switched systems*, IEEE Control Systems Magazine, 19 (1999), pp. 59–70.
[18] A. S. Morse, *Supervisory control of families of linear set-point controllers—Part* 1: *Exact matching*, IEEE Trans. Automat. Control, 41 (1996), pp. 1413–1431.
[19] R. Rockafellar, *Convex Analysis*, Princeton Press, Princeton, NJ, 1970.
[20] M. Wicks and R. A. DeCarlo, *Solution of coupled Lyapunov equations for the stabilization of multimodal linear systems*, in Proceedings of the American Control Conference, 1997, pp. 1709–1713.
[21] F. Wirth, *A converse Lyapunov theorem for linear parameter varying and linear switching systems*, SIAM J. Control Optim., 44 (2005), pp. 210–239.
[22] X. Xu and P. J. Antsaklis, *Optimal control of switched systems based on parameterization of the switching instants*, IEEE Trans. Automat. Control, 49 (2004), pp. 2–16.
[23] H. Ye, A. N. Michel, and L. Hou, *Stability theory for hybrid dynamical systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 461–474.
[24] J. Zhao and G. M. Dimirovski, *Quadratic stability of a class of switched nonlinear systems*, IEEE Trans. Automat. Control, 49 (2004), pp. 574–578.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 1931–1964

GLOBAL SMOOTH SOLUTIONS AND EXPONENTIAL STABILITY FOR A NONLINEAR BEAM*

PENG-FEI YAO† AND GEORGE WEISS‡

Abstract. In this paper we consider a dynamical system with boundary input and output describing the bending vibrations of a quasi-linear beam, where the nonlinearity comes from Hooke's law. First we derive an existence result for short-time solutions of the system of equations. Then we show that the structure of the boundary input and output forces the system to admit global solutions at least when the initial data and the boundary input are small in a certain sense. In particular, we prove that the norm of the state of the system decays exponentially if the input becomes zero after a finite time (the input being zero can be understood as a boundary feedback).

Key words. quasi-linear beam, exponential stabilization, energy-preserving system, boundary input and output

AMS subject classifications. 93C20, 35B35, 35B37, 35L65, 35L70

DOI. 10.1137/S0363012903435357

1. Introduction and main results.

We consider a system modeling the bending vibrations of a nonlinear beam of length $L > 0$. Let ϕ denote the deflection of the beam and let M be the bending torque. Assuming unit mass density, from Segel and Handelman [12] we have

$$\phi_{tt} - \kappa\phi''_{tt} + M'' = 0, \quad x \in [0, L], \quad t > 0,$$

where $\kappa > 0$, $\sqrt{\kappa}$ is the radius of gyration, and the superscript "$'$" and the subscript t denote differentiation with respect to the variables x (space) and t (time), respectively. We assume that the constitutive law (Hooke's law) is nonlinear, determined by a smooth function σ, such that $M = \sigma(\phi'')$. The beam equation then becomes

$$\phi_{tt} - \kappa\phi''_{tt} + \sigma(\phi'')'' = 0, \quad x \in [0, L], \quad t > 0. \tag{1.1}$$

If $\sigma(x) = ax$, with $a > 0$, then the model is linear, called the *Rayleigh beam* in the literature. Recent papers dealing with the boundary stabilization of Rayleigh beams are Ammari, Liu, and Tucsnak [2], Guo, Wang, and Yung [7], and Weiss and Curtain [17].

Remark 1.1. We can derive (1.1) also from the calculus of variations. Let $T > 0$ be fixed. We define the *kinetic energy* of the beam at time $t \in [0, T]$ by $\mathbf{K}(\phi_t(\cdot, t))$, where the quadratic functional $\mathbf{K}$ is defined by

$$\mathbf{K}(\psi(\cdot, t)) = \frac{1}{2}\int_0^L \left[\psi^2(x, t) + \kappa\psi'^2(x, t)\right] \mathrm{d}x.$$

*Received by the editors September 26, 2003; accepted for publication (in revised form) April 30, 2006; published electronically January 8, 2007. This work was supported by the EPSRC grant "Portfolio partnership: control and power" (GR/S61256/01), by the NSF of China, grants 60225003, 60221301, and 60334040, as well as by the Royal Society under a China–UK joint project grant.
http://www.siam.org/journals/sicon/45-6/43535.html

†Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (pfyao@iss.ac.cn).

‡Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London SW7 2AZ, UK (g.weiss@imperial.ac.uk).

The *potential energy* of the beam at time t is defined as

$$\mathbf{P}(\phi(\cdot,t)) = \int_0^L F(\phi''(x,t))\,\mathrm{d}x,$$

where F is a smooth function such that $\frac{\mathrm{d}F(z)}{\mathrm{d}z} = \sigma(z)$. Then it is easy to check that the partial differential equation (1.1) is equivalent to the stationary point condition derived from Hamilton's principle

$$\frac{\mathrm{d}}{\mathrm{d}s}\int_0^T \mathbf{L}(\phi + sv, \phi_t + sv_t)\,\mathrm{d}t\,\Big|_{s=0} = 0 \quad \forall v \in C_0^\infty(Q_T),$$

where $\mathbf{L}(\phi,\psi) = \mathbf{K}(\psi) - \mathbf{P}(\phi)$ and $Q_T = (0,L)\times(0,T)$. Indeed, this follows from

$$\left\langle \frac{\partial \mathbf{L}}{\partial \phi}(\phi,\psi), v \right\rangle = -\langle \sigma(\phi''), v''\rangle, \quad \left\langle \frac{\partial \mathbf{L}}{\partial \psi}(\phi,\psi), v_t \right\rangle = \langle \psi, v_t\rangle + \kappa\langle \psi', v_t'\rangle,$$

where all the inner products are taken in $L^2(Q_T)$.

Suppose that the initial data are given by

$$\phi(x,0) = \phi^0(x), \quad \phi_t(x,0) = \phi^1(x), \quad x \in [0,L]. \tag{1.2}$$

We suppose that the right end of the beam is hinged, i.e.,

$$\phi(L,t) = \phi''(L,t) = 0, \tag{1.3}$$

and at the left end we have actuators and sensors, linked in a damping feedback mechanism, satisfying the equations

$$\begin{cases} 2b_1u_1(t) = \sigma(\phi''(0,t)) - b_1^2\phi_t'(0,t), \quad t \ge 0, \\ 2b_2u_2(t) = \left(\sigma(\phi'')\right)'(0,t) - \kappa\phi_{tt}'(0,t) + b_2^2\phi_t(0,t), \quad t \ge 0, \end{cases} \tag{1.4}$$

$$\begin{cases} 2b_1y_1(t) = \sigma(\phi''(0,t)) + b_1^2\phi_t'(0,t), \quad t \ge 0, \\ 2b_2y_2(t) = \left(\sigma(\phi'')\right)'(0,t) - \kappa\phi_{tt}'(0,t) - b_2^2\phi_t(0,t), \quad t \ge 0, \end{cases} \tag{1.5}$$

where $b_i > 0$ are constants for $i = 1, 2$, $u = (u_1, u_2)$ is the input function, and $y = (y_1, y_2)$ is the output function. The term $2b_1u_1(t)$ is an external torque acting at the left end of the beam, while $2b_2u_2(t)$ is an external transversal force acting at the same end. The term $\sigma(\phi''(0,t))$ is the total torque acting at $x = 0$, while $(\sigma(\phi''))'(0,t) - \kappa\phi_{tt}'(0,t)$ is the total transversal force at $x = 0$. The terms $-b_1^2\phi_t'(0,t)$ and $b_2^2\phi_t(0,t)$ represent the boundary feedback torque and force, respectively.

To guarantee that the system (1.1)–(1.5) is hyperbolic, we further assume that

$$\sigma(0) = 0, \quad \sigma'(z) > 0, \quad z \in (-\infty, \infty). \tag{1.6}$$

If we define the *total energy* of the system (1.1)–(1.5) by

$$\mathbf{E}(t) = \mathbf{K}(\phi_t(\cdot,t)) + \mathbf{P}(\phi(\cdot,t)),$$

where $\mathbf{K}$ and $\mathbf{P}$ are defined as in Remark 1.1, we obtain by a simple computation

$$\begin{aligned} \frac{\mathrm{d}\mathbf{E}(t)}{\mathrm{d}t} &= -\sigma(\phi''(0,t))\phi_t'(0,t) + [\sigma(\phi'')'(0,t) - \kappa\phi_{tt}'(0,t)]\phi_t(0,t) \\ &= |u(t)|^2 - |y(t)|^2. \end{aligned} \tag{1.7}$$

This formula expresses that the rate of change of the energy is equal to the power supplied to the system by the input minus the power taken out by the output. The balance equation (1.7) means that the system (1.1)–(1.5) is (scattering) *energy-preserving*. We refer to Staffans [14] and to Malinen, Staffans, and Weiss [11] for the theory of energy-preserving linear systems.

We define the positive smooth function $\mathcal{N} : (\sigma_-, \sigma_+) \to (0, \infty)$ by

$$\mathcal{N}(z) = \sigma'(\sigma^{-1}(z)), \tag{1.8}$$

where σ^{-1} is the inverse function of σ and

$$\sigma_\pm = \lim_{z \to \pm\infty} \sigma(z). \tag{1.9}$$

By the assumption (1.6), the above limits exist (they may be infinite) and

$$0 \in (\sigma_-, \sigma_+), \quad \lim_{z \to \sigma_\pm} \sigma^{-1}(z) = \pm\infty. \tag{1.10}$$

In the linear case ($\sigma(z) = az$) we have $\mathcal{N}(z) = a$, of course.

We introduce an equivalent system by the change of unknowns

$$\varphi = \sigma(\phi''), \quad \psi = \phi_t.$$

Then the system (1.1)–(1.5) becomes

$$\begin{cases} \varphi_t = \mathcal{N}(\varphi)\psi'', \\ \psi_t = \kappa\psi_t'' - \varphi'', \\ \varphi(x,0) = \varphi_0(x), \quad \psi(x,0) = \psi_0(x), \\ \varphi(L,t) = \psi(L,t) = \psi''(L,t) = 0, \end{cases} \tag{1.11}$$

with the input signal $u = (u_1, u_2)$ acting through

$$\begin{cases} 2b_1 u_1(t) = \varphi(0,t) - b_1^2\psi'(0,t), \\ 2b_2 u_2(t) = \varphi'(0,t) - \kappa\psi_t'(0,t) + b_2^2\psi(0,t), \end{cases} \tag{1.12}$$

and with the output signal $y = (y_1, y_2)$ defined by

$$\begin{cases} 2b_1 y_1(t) = \varphi(0,t) + b_1^2\psi'(0,t), \\ 2b_2 y_2(t) = \varphi'(0,t) - \kappa\psi_t'(0,t) - b_2^2\psi(0,t). \end{cases} \tag{1.13}$$

The initial functions of this new system are obtained from those in (1.2) by

$$\varphi_0 = \sigma(\phi^{0\prime\prime}), \qquad \psi_0 = \phi^1.$$

It is well known that smooth solutions of quasi-linear hyperbolic systems usually develop singularities after some time; see, for example, Klainerman and Majda [10]. However, since the structure of the boundary conditions (1.12) causes energy dissipation (as seen in (1.7)), we expect that the introduction of the boundary input (1.12) prevents the formation of singularities and assures the existence of a global smooth solution. In fact, such problems have been considered for the nonlinear string in Greenberg and Li [6], Slemrod [13], Alber and Cooper [1], and Yao [18].

The main aim of this paper is to study the existence of smooth short-time solutions and smooth global solutions of the nonlinear beam system (1.11)–(1.12) if the initial data φ_0, ψ_0 and the boundary input u are given. For short-time solutions, we use the results of Kato [8] to establish the existence of solutions of a certain linear system and then use an argument with contraction mappings from Dafermos and Hrusa [4] to obtain the existence. Due to the special structure of (1.12), some modifications are necessary in the energy estimates which are used in the fixed-point argument. For global solutions, we establish an estimate for the boundary trace by invoking some ideas from Alber and Cooper [1] and from Cirina [3]. Such an estimate is necessary for the global existence. In the linear case (where $\sigma(x) = ax$), boundary trace estimates are usually derived by the multiplier technique.

We mention that in the linear case our system fits into the framework of Weiss and Tucsnak [16] and [15], which guarantees that the system is well-posed and conservative (this property is stronger than "energy-preserving").

If $b_2 = 0$, the system (1.11)–(1.13) can be transformed into a quasi-linear string with an input-output structure at the boundary that has been studied in [18] (see Remark 4.2). Here we study only the case $b_2 > 0$.

Motivated by the structure of the function $\mathcal{N}$, we introduce the following concept. We say that a continuous function φ on $[0, L]$ satisfies the *σ-condition* if

$$\sigma_- < \inf_{x\in[0,L]} \varphi(x) \le \sup_{x\in[0,L]} \varphi(x) < \sigma_+, \tag{1.14}$$

where σ_-, σ_+ are as in (1.9). If $\sigma_\pm = \pm\infty$, then this condition is trivially true.

For short-time solutions, we have the following existence result.

THEOREM 1.1. *Suppose that $u = (u_1, u_2) \in C^3[0,\infty) \times C^2[0,\infty)$ and $(\varphi_0, \psi_0) \in H^2(0, L) \times H^3(0, L)$ satisfies the boundary conditions*

$$\varphi_0(L) = \psi_0(L) = \psi_0''(L) = 0 \tag{1.15}$$

and the compatibility conditions

$$\begin{cases} 2b_1u_1(0) = \varphi_0(0) - b_1^2\psi_0'(0), \\ 2b_2u_2(0) = 2b_1^{-1}\kappa u_{1t}(0) + \varphi_0'(0) - \kappa b_1^{-2}\mathcal{N}(\varphi_0(0))\psi_0''(0) + b_2^2\psi_0(0). \end{cases} \tag{1.16}$$

Furthermore, we assume that φ_0 satisfies the σ-condition (1.14).

Then there is a $T_0 > 0$ such that the problem (1.11)–(1.12) *admits a unique solution (φ, ψ) on $Q_{T_0} = (0, L) \times (0, T_0)$ satisfying*

$$\varphi \in \bigcap_{i=0}^{2} C^i\left([0, T_0]; H^{2-i}(0, L)\right), \quad \psi \in \bigcap_{i=0}^{3} C^i\left([0, T_0)]; H^{3-i}(0, L)\right), \tag{1.17}$$

where $H^0(0, L)$ stands for $L^2(0, L)$.

The proof of this theorem is given in section 3.

When the input u and the initial data (φ_0, ψ_0) are appropriately small, the solutions are global. To describe this, we introduce an energy-like function $\mathcal{E}$ as follows. Let (φ, ψ) be a solution of the system (1.11)–(1.12). For any multi-index $\alpha = (\alpha_1, \alpha_2) \in \{0, 1, 2, \dots\}^2$, we denote $|\alpha| = \alpha_1 + \alpha_2$. Define

$$\mathcal{E}(t) = [\varphi'(0, t)]^2 + \|\psi_{tt}'(\cdot, t)\|^2 + \sum_{|\alpha|\le 2} (\|\partial^\alpha\varphi(\cdot, t)\|^2 + \|\partial^\alpha\psi(\cdot, t)\|^2), \tag{1.18}$$

where $\partial^{(\alpha_1,\alpha_2)} = \left(\frac{\partial}{\partial x}\right)^{\alpha_1}\left(\frac{\partial}{\partial t}\right)^{\alpha_2}$ and $\|\cdot\|$ is the usual norm on $L^2(0, L)$.

For $u = (u_1, u_2) \in [H^2(0,T)]^2$, denote

$$\|u\|^2_{H^2(0,T)} = \int_0^T |u|_2^2(t)\,\mathrm{d}t, \qquad |u|_2^2(t) = |u(t)|^2 + |u_t(t)|^2 + |u_{tt}(t)|^2. \tag{1.19}$$

THEOREM 1.2. *Suppose that $u = (u_1, u_2) \in H^2(0,\infty) \times H^2(0,\infty)$ and the initial data $(\varphi_0, \psi_0) \in H^2(0,L) \times H^3(0,L)$ are such that the boundary conditions* (1.15) *and the compatibility conditions* (1.16) *hold. Then there are $\epsilon > 0$, $c_1 > 0$, $c_2 > 0$, $c_3 > 0$, independent of (φ_0, ψ_0) and u, such that if*

$$\mathcal{E}(0) + \|u\|^2_{H^2(0,\infty)} \le \epsilon, \tag{1.20}$$

then the problem (1.11)–(1.12) *has a unique solution (φ, ψ) on $Q_\infty = (0,L) \times (0,\infty)$ and*

$$\mathcal{E}(t) + c_1 \int_0^t \mathcal{E}(\tau)\mathrm{d}\tau \le c_2\mathcal{E}(0) + c_3\|u\|^2_{H^2(0,\infty)} \quad \forall\ t \ge 0. \tag{1.21}$$

Note that from (φ_0, ψ_0) we can compute $\partial^\alpha\varphi(\cdot,0)$ and $\partial^\alpha\psi(\cdot,0)$ by solving (1.11) at $t = 0$, so that also $\mathcal{E}(0)$ can be computed.

The above theorem will be proved in section 5.

COROLLARY 1.1. *Under the assumptions of Theorem* 1.2, *if* (1.20) *holds, then the solution (φ, ψ) of* (1.11)–(1.12) *and the output $y = (y_1, y_2)$ from* (1.13) *satisfy*

$$\varphi \in H^2(Q_\infty), \quad \psi \in H^2(Q_\infty) \cap H^{1,2}(Q_\infty), \quad y \in H^2(0,\infty) \times H^2(0,\infty), \tag{1.22}$$

where $H^{1,2}(Q_\infty)$ is the Sobolev space with the norm

$$\|\psi\|^2_{H^{1,2}(Q_\infty)} = \sum_{0\le i\le 1;\, 0\le j\le 2} \int_0^\infty \|\partial^{(i,j)}\psi(\cdot,t)\|^2\mathrm{d}t.$$

The following stability result will also be proved in section 5.

THEOREM 1.3. *Under the assumptions of Theorem* 1.2, *suppose that (φ_0, ψ_0) with u satisfy* (1.20). *If there is a $T_0 > 0$ such that*

$$u(t) = 0 \quad \forall\ t \ge T_0, \tag{1.23}$$

then there are $C > 0$ and $\epsilon_1 > 0$ such that

$$\mathcal{E}(t) \le Ce^{-\epsilon_1 t} \quad \forall t \ge 0. \tag{1.24}$$

This theorem shows that the energy-like function $\mathcal{E}(t)$ for the boundary feedback system corresponding to $u = 0$ decays exponentially.

Remark 1.2. We do not assume that φ_0 satisfies the σ-condition in Theorems 1.2, 1.3 and in Corollary 1.1, because for suitable ε, the small initial data assumption (1.20) implies this. Indeed,

$$\sup_{x\in[0,L]} |\varphi_0(x)| \le L^{\frac{1}{2}}\mathcal{E}^{\frac{1}{2}}(0) \le L^{\frac{1}{2}}\varepsilon^{\frac{1}{2}}.$$

Then the σ-condition is true if $0 < \varepsilon < L^{-1}\min\{\sigma_-^2, \sigma_+^2\}$.

2. Linear time-varying equations. Let $T > 0$ and let N be a function on $Q_T = (0, L) \times (0, T)$ such that

$$N \in \bigcap_{i=0}^{2} C^i\left([0, T]; H^{2-i}(0, L)\right)$$

and there are $N_0 > 0$, $N_1 > 0$ such that

$$N_0 \le N(x, t) \le N_1, \quad (x, t) \in Q_T. \tag{2.1}$$

Motivated by (1.11)–(1.12), we consider the linear dynamical system described by

$$\begin{cases} \varphi_t = N\psi'', \\ \psi_t = \kappa\psi_t'' - \varphi'', \\ \varphi(x, 0) = \varphi_0(x), \qquad \psi(x, 0) = \psi_0(x), \\ \varphi(L, t) = \psi(L, t) = \psi''(L, t) = 0, \\ 2b_1u_1(t) = \varphi(0, t) - b_1^2\psi'(0, t), \\ 2b_2u_2(t) = \varphi'(0, t) - \kappa\psi_t'(0, t) + b_2^2\psi(0, t). \end{cases} \tag{2.2}$$

We make a change of unknowns by

$$v = \varphi, \quad w = \psi + g \quad \text{on} \quad Q_T, \tag{2.3}$$

where g is a smooth function on Q_T satisfying the boundary conditions

$$g(L, t) = 0, \quad b_1g'(0, t) = 2u_1(t), \quad t \in [0, T]. \tag{2.4}$$

We shall specify later how to choose g in order to satisfy additional requirements. By (2.3)–(2.4), the system (2.2) becomes, for $(x, t) \in Q_T$,

$$\begin{cases} v_t - Nw'' = -Ng'', \\ w_t - \kappa w_t'' + v'' = g_t - \kappa g_t'', \\ v(x, 0) = \varphi_0(x), \quad w(x, 0) = \psi_0(x) + g(x, 0), \\ v(L, t) = w(L, t), \qquad w''(L, t) = g''(L, t), \\ v(0, t) - b_1^2w'(0, t) = 0, \\ v'(0, t) - \kappa w_t'(0, t) + b_2^2w(0, t) = 2b_2u_2(t) + b_2^2g(0, t) - \kappa g_t'(0, t). \end{cases} \tag{2.5}$$

Let us formulate the problem (2.5) in an appropriate Hilbert space. First we introduce the following space:

$$H = \{f \mid f \in H^1(0, L),\ f(L) = 0\}, \quad \langle f, p\rangle_H = \langle f, p\rangle + \kappa\langle f', p'\rangle,$$

where $\langle\cdot, \cdot\rangle$ is the usual inner product of $L^2(0, L)$. To rewrite the second equation in (2.5), we introduce an operator B_0: $L^2(0, L) \to H$ by

$$\langle p, f'\rangle = \langle B_0p, f\rangle_H \quad \forall\ p \in L^2(0, L),\ f \in H. \tag{2.6}$$

LEMMA 2.1. *The operator B_0 from* (2.6) *is an isomorphism between $L^2(0, L)$ and H. In addition, for every $p \in L^2(0, L)$,*

$$(B_0p)(x) = \langle p, q_0\rangle \cosh\sqrt{a}x + a\int_0^x p(s)\cosh\sqrt{a}(x - s)\mathrm{d}\,s, \tag{2.7}$$

where

$$a = \frac{1}{\kappa}, \quad q_0(x) = -\frac{a\cosh\sqrt{a}(L-x)}{\cosh\sqrt{a}L}. \tag{2.8}$$

In particular, $\frac{\partial}{\partial x}B_0$ *is bounded from* $L^2(0,L)$ *to itself, and for* $p \in L^2(0,L)$

$$\left(\frac{\partial}{\partial x}B_0 p\right)(x) = \sqrt{a}\langle p, q_0\rangle \sinh\sqrt{a}x + a^{\frac{3}{2}}\int_0^x p(s)\sinh\sqrt{a}(x-s)\mathrm{d}s + ap(x). \tag{2.9}$$

Proof. First, we assume that $p \in H^1(0,L)$. Denote $y = B_0 p$. From (2.6), integration by parts gives the problem

$$y - \kappa y'' = -p', \quad y'(0) = ap(0), \quad y(L) = 0.$$

Solving this problem, we obtain (2.7) with (2.8). Since $H^1(0,L)$ is dense in $L^2(0,L)$, (2.7) is actually true for all $p \in L^2(0,L)$. Differentiating (2.7) we obtain (2.9).

It is easy to check that B_0 is invertible and that B_0^{-1}: $H \to L^2(0,L)$ is given by

$$\left(B_0^{-1}f\right)(x) = \kappa f'(x) - \int_0^x f(s)\mathrm{d}s \quad \forall\ f \in H. \qquad \square$$

Let (v, w) be a smooth solution of the system (2.5). We multiply the second equation in (2.5) by a function $f \in H$ and integrate over $[0,L]$. Using integration by parts and (2.6) we obtain that for all $f \in H$,

$$\langle w_t, f\rangle_H = \langle B_0 v' + g_t, f\rangle_H + \left[v'(0,t)(t) - \kappa w_t'(0,t) + \kappa g_t'(0,t)\right] f(0).$$

Using the last equation in (2.5), this becomes

$$\langle w_t, f\rangle_H = \langle B_0 v' + g_t, f\rangle_H + \left[2b_2u_2(t) - b_2^2 w(0,t) + b_2^2 g(0,t)\right] f(0). \tag{2.10}$$

Let $q_1 \in H$ be such that

$$f(0) = \langle q_1, f\rangle_H \quad \forall\ f \in H. \tag{2.11}$$

It follows from (2.10) and (2.11) that

$$w_t = B_0 v' - b_2^2 w(0,t)q_1 + g_t + [2b_2u_2(t) + b_2^2 g(0,t)]q_1 \quad \text{in} \quad H.$$

Now the problem (2.5) becomes a problem on $L^2(0,L)\times H$:

$$\begin{cases} v_t = Nw'' - Ng'', \\ w_t = B_0 v' - b_2^2 w(0,t)q_1 + g_t + [2b_2u_2(t) + b_2^2 g(0,t)]q_1, \\ v(x,0) = \varphi_0(x), \quad w(x,0) = \psi_0(x) + g(x,0), \\ v(L,t) = w(L,t) = 0, \quad v(0,t) = b_1^2 w'(0,t). \end{cases} \tag{2.12}$$

We wish to solve (2.12) by applying the results of Kato [8, 9] in the space $\mathcal{H} = L^2(0,L)\times H$, with the inner product

$$\left\langle \begin{bmatrix} p \\ f \end{bmatrix}, \begin{bmatrix} \hat{p} \\ \hat{f} \end{bmatrix} \right\rangle_{\mathcal{H}} = \langle p, \hat{p}\rangle + \langle f, \hat{f}\rangle_H \quad \forall\ \begin{bmatrix} p \\ f \end{bmatrix}, \begin{bmatrix} \hat{p} \\ \hat{f} \end{bmatrix} \in \mathcal{H}.$$

To this end, we introduce a family of closed operators on $\mathcal{H}$ by

$$\mathcal{A}_0(t) = \begin{bmatrix} 0 & N\partial^2 \\ B_0\partial & -b_2^2 q_1 \Gamma_0 \end{bmatrix}, \quad 0 \le t \le T, \tag{2.13}$$

where $\partial = \frac{\partial}{\partial x}$, the linear functional Γ_0 is given by

$$\Gamma_0 f = f(0) \quad \forall\, f \in H,$$

and the domain of $\mathcal{A}_0(t)$ is

$$D(\mathcal{A}_0(t)) = \left\{ \begin{bmatrix} p \\ f \end{bmatrix} \in H^1(0,L) \times H^2(0,L) \;\middle|\; \begin{matrix} p(L) = f(L) = 0, \\ p(0) = b_1^2 f'(0) \end{matrix} \right\}. \tag{2.14}$$

Then the problem (2.12) can be written as an evolution equation on $\mathcal{H}$:

$$X_t(t) = \mathcal{A}_0(t)X(t) + \mathcal{F}_0(t), \quad X(0) = X_0, \tag{2.15}$$

with

$$X = \begin{bmatrix} v \\ w \end{bmatrix}, \quad \mathcal{F}_0(t) = \begin{bmatrix} -Ng'' \\ g_t + [2b_2u_2(t) + b_2^2 g(0,t)]q_1 \end{bmatrix}. \tag{2.16}$$

More precisely, the connection between (2.12) and (2.15) is as follows: Let

$$\mathcal{H}_1 = D(\mathcal{A}_0(t)), \tag{2.17}$$

which is a Hilbert space with the graph norm. This space is independent of t and for different t we get on it equivalent norms.

LEMMA 2.2. *If $X(\cdot) \in C^1([0,l];\mathcal{H}_1)$ is a solution of* (2.15)*, then its components (v,w) are a solution of* (2.12).

The proof of this lemma is by simple verification.

Note that $X(\cdot) \in C^1([0,T];\mathcal{H}_1)$ implies $X(t) \in H^1(0,L) \times H^3(0,L)$ for all $t \in [0,T]$. For less regular solutions of (2.15), some of the equations in (2.12) become meaningless. For example, if we only have $X(\cdot) \in C([0,T];\mathcal{H}_1)$, which is still rather smooth, then the last equation in (2.12) becomes meaningless.

Since the domain $D(\mathcal{A}_0(t))$ is independent of t (see (2.14)), the key of applying the work of Kato to our problem (2.15) is to show that $\{\,\mathcal{A}_0(t)\,|\,0 \le t \le T\,\}$ is a stable family on $\mathcal{H}$, in the sense of [8]. We recall that $\{\,\mathcal{A}_0(t)\,|\,0 \le t \le T\,\}$ is called a *stable family* if the following conditions hold:

(1) For each $t \in [0,T]$, $\mathcal{A}_0(t)$ is the generator of a C_0-semigroup.

(2) There are constants M, β (called the constants of stability) such that

$$\left\| \prod_{j=1}^{k} (\lambda - \mathcal{A}_0(t_j))^{-1} \right\| \le M(\lambda - \beta)^{-k}, \quad \lambda > \beta,$$

for any finite family $\{t_j\}$ with $0 \le t_1 \le \cdots \le t_k \le T$, $k \in \mathbb{N}$. Here, the product Π is time-ordered; i.e., a factor with larger t_j stands to the left of one with smaller t_j.

LEMMA 2.3. *Let $\mathcal{A}_0(t)$ be defined in* (2.13) *and* (2.14)*. For each t, $0 \le t \le T$, $\mathcal{A}_0(t)$ generates a C_0-semigroup on $\mathcal{H}$.*

Proof. For $t \in [0,T]$ we introduce a family of new inner products $\langle \cdot, \cdot \rangle_t$ on $\mathcal{H}$ which are equivalent to the old one, to show that $\mathcal{A}_0(t)$ is dissipative on $\mathcal{H}$:

$$\left\langle \begin{bmatrix} p \\ f \end{bmatrix}, \begin{bmatrix} \hat{p} \\ \hat{f} \end{bmatrix} \right\rangle_t = \langle N^{-1}(\cdot,t)p, \hat{p} \rangle + \langle f, \hat{f} \rangle_H \quad \forall \begin{bmatrix} p \\ f \end{bmatrix}, \begin{bmatrix} \hat{p} \\ \hat{f} \end{bmatrix} \in \mathcal{H}. \tag{2.18}$$

Since N satisfies (2.1), these new norms are equivalent to the old one.

Let $\left[\begin{smallmatrix} p \\ f \end{smallmatrix}\right] \in D(\mathcal{A}_0(t))$. It follows from (2.13), (2.14), and (2.18) that

$$\begin{aligned} \left\langle \mathcal{A}_0(t) \begin{bmatrix} p \\ f \end{bmatrix}, \begin{bmatrix} p \\ f \end{bmatrix} \right\rangle_t &= \left\langle \begin{bmatrix} Nf'' \\ B_0 p' - b_2^2 f(0) q_1 \end{bmatrix}, \begin{bmatrix} p \\ f \end{bmatrix} \right\rangle_t \\ &= \langle f'', p \rangle + \langle p', f' \rangle - b_2^2 f^2(0) = p f'|_0^L - b_2^2 f^2(0) \\ &= -b_1^2 f'^2(0) - b_2^2 f^2(0) \le 0. \end{aligned}$$

We want to show that $\mathcal{A}_0(t)$ is m-dissipative. For this, it is enough to show that also $\mathcal{A}_0^*(t)$ is dissipative, where $\mathcal{A}_0^*(t)$ is the adjoint operator of $\mathcal{A}_0(t)$ with respect to the inner product $\langle \cdot, \cdot \rangle_t$. A simple computation shows that

$$\mathcal{A}_0^*(t) = -\begin{bmatrix} 0 & N\partial^2 \\ B_0 \partial & b_2^2 q_1 \Gamma_0 \end{bmatrix},$$

$$D(\mathcal{A}_0^*(t)) = \left\{ \begin{bmatrix} p \\ f \end{bmatrix} \in H^1(0,L) \times H^2(0,L) \;\middle|\; \begin{array}{l} p(L) = f(L) = 0, \\ p(0) = -b_1^2 f'(0) \end{array} \right\}$$

and for all $\left[\begin{smallmatrix} p \\ f \end{smallmatrix}\right] \in D(\mathcal{A}_0^*(t))$,

$$\left\langle \mathcal{A}_0^*(t) \begin{bmatrix} p \\ f \end{bmatrix}, \begin{bmatrix} p \\ f \end{bmatrix} \right\rangle_t = -b_1^2 f'^2(0) - b_2^2 f^2(0) \le 0.$$

Thus, $\mathcal{A}_0(t)$ is m-dissipative. According to the Lumer–Phillips theorem, $\mathcal{A}_0(t)$ is the generator of a contraction semigroup on $\mathcal{H}$ (with respect to $\langle \cdot, \cdot \rangle_t$). ∎

LEMMA 2.4. *For any $t \in [0,T]$, $\{ \mathcal{A}_0(\tau) \,|\, 0 \le \tau \le T \}$ is a stable family with respect to the norm $\| \cdot \|_t$ (which corresponds to the inner product* (2.18)*) on $\mathcal{H}$.*

Proof. In view of Proposition 3.4 in [8] and Lemma 2.3 above, it suffices to verify that there exists $c > 0$ such that the following inequality holds:

$$\|\mathcal{X}\|_t / \|\mathcal{X}\|_s \le e^{c|t-s|} \qquad \forall\, \mathcal{X} = \begin{bmatrix} p \\ f \end{bmatrix} \in \mathcal{H},\ s,t \in [0,T]. \tag{2.19}$$

To this end, we take

$$c \ge \sup_{(x,t) \in Q_T} |N_t| N^{-1}$$

and set

$$h(t) = \left\| \begin{bmatrix} p \\ f \end{bmatrix} \right\|_t^2 = \langle N^{-1}(\cdot,t)p, p \rangle + \|f\|_H^2, \quad t \in [0,T].$$

Then

$$|h_t(t)| \le 2ch(t) \qquad \forall \begin{bmatrix} p \\ f \end{bmatrix} \in \mathcal{H}, \quad t \in [0,T].$$

This inequality implies that the function $h(t)e^{-2ct}$ is decreasing while $h(t)e^{2ct}$ is increasing for $t \in [0,T]$. Therefore, (2.19) is true. ∎

In order to have a solution $X(\cdot) \in C([0,T];\mathcal{H}_1)$ to the inhomogeneous system (2.15), it will suffice to choose $X_0 \in \mathcal{H}_1$ and a function g such that the inhomogeneous term $\mathcal{F}_0(\cdot) \in C([0,T];\mathcal{H}_1)$. For this, we consider the 3×3 matrix $\mathcal{B}$ defined by

$$\mathcal{B} = (e_2, e_3, e_1),$$

where (e_1, e_2, e_3) is the standard basis of the Euclidian space $\mathbb{R}^3$, and we take a function $\rho \in C^\infty[0,L]$ such that

$$\rho(x) = 1 \ \text{ for } \ x \le L/3, \qquad \rho(x) = 0 \ \text{ for } \ x \ge 2L/3. \tag{2.20}$$

LEMMA 2.5. *Assume that $u = (u_1, u_2) \in C^2[0,T] \times C^1[0,T]$. Using the notation $\mathcal{B}$ and ρ introduced above, there exists a function g of the form*

$$g(x,t) = \rho(x)\beta(t)\cdot e^{\mathcal{B}x}e_1, \quad (x,t) \in Q_T, \tag{2.21}$$

where $\beta(t) = [\beta_1(t)\ \beta_2(t)\ \beta_3(t)]$, such that g satisfies (2.4) *and $\mathcal{F}_0(\cdot) \in C([0,T];\mathcal{H}_1)$, where $\mathcal{F}_0(t)$ is given in* (2.16).

Proof. It follows from (2.20)–(2.21) that

$$\begin{cases} g(L,t) = g''(L,t) = 0, \\ g(0,t) = \beta_1(t), \quad g'(0,t) = \beta_2(t), \quad g''(0,t) = \beta_3(t). \end{cases}$$

On the other hand, from (2.14) and (2.16) we see that the conditions (2.4) and $\mathcal{F}_0(t) \in D(\mathcal{A}_0(t))$ for $t \in [0,T]$ will be satisfied if

$$\begin{cases} g(L,t) = g''(L,t) = 0, \qquad b_1 g'(0,t) = 2u_1(t), \\ N(0,t)g''(0,t) + q_1'(0)b_1^2 b_2^2 g(0,t) = -2b_1 u_{1t}(t) - 2q_1'(0)b_1^2 b_2 u_2(t). \end{cases}$$

This can be achieved if we set

$$\begin{cases} \beta_1(t) = 0, \qquad \beta_2(t) = 2b_1^{-1}u_1(t), \\ \beta_3(t) = -2b_1 N^{-1}(0,t)[u_{1t}(t) + q_1'(0)b_1 b_2 u_2(t)], \end{cases}$$

and moreover we then have $\mathcal{F}_0(\cdot) \in C([0,T];\mathcal{H}_1)$. ∎

Since $D(\mathcal{A}_0(t))$ is constant, by Theorem 7.1 of [8] (see also Remark 6.2 of the same paper), for any $X_0 \in \mathcal{H}_1$ and for $\mathcal{F}_0$ as in the last lemma, the problem (2.15) admits a unique solution $X(\cdot) \in C([0,T];\mathcal{H}_1)$. However, from here we cannot go back to the problem (2.2) through the change of unknowns (2.3), because the functions do not have enough somoothness to make sense of the boundary conditions.

For our purposes in this paper, we need the problem (2.2) to have solutions (φ, ψ) such that $\varphi \in C\big([0,T];H^2(0,L)\big)$ and $\psi \in C\big([0,T];H^3(0,L)\big)$. For this, we consider an evolutionary problem, like (2.15), on the space $(\mathcal{H}_1, \|\cdot\|_1)$ where $\mathcal{A}_0(t) : D(\mathcal{A}_0^2(t)) \to \mathcal{H}_1$ and $\mathcal{H}_1 = D(\mathcal{A}_0(t))$ with the graph norm. Using Lemma 2.1, a simple computation shows that $D(\mathcal{A}_0^2(t))$ consists of all the pairs $(v,w) \in H^2(0,L)\times H^3(0,L)$ satisfying the boundary conditions

$$v(L) = w(L) = w''(L) = 0, \quad v(0) = b_1^2 w'(0),$$

$$N(0,t)w''(0) = ab_1^2 v'(0) + ab_1^2 b_2^2 w(0).$$

The last boundary condition shows that $D(\mathcal{A}_0^2(t))$ depends on the time t. Thus, we would meet difficulty in verifying the assumptions of [8], [9], although the work of Kato does not assume that the domain is independent of time t.

We overcome the above difficulty as follows. Suppose that we have a solution X of the problem (2.15) such that

$$X(\cdot) \in C\left([0,T]; D(\mathcal{A}_0^2(t))\right). \tag{2.22}$$

After making a change of unknowns by

$$Y(t) = \mathcal{C}(t)X(t), \quad \mathcal{C}(t) = \lambda I - \mathcal{A}_0(t),$$

where λ is large enough, we observe that $Y(\cdot) \in C\left([0,T]; \mathcal{H}_1\right)$ satisfies

$$Y_t(t) = [\mathcal{A}_0(t) + \mathcal{C}_t(t)\mathcal{C}^{-1}(t)]Y(t) + \mathcal{C}(t)\mathcal{F}_0(t), \quad Y(0) = \mathcal{C}(0)X_0 \in \mathcal{H}_1. \tag{2.23}$$

Since the above argument is reversible, in order to obtain a solution to the problem (2.15) to satisfy (2.22), we need only solve the problem (2.23) in the space $\mathcal{H}$. This is achieved by the following two lemmas.

LEMMA 2.6. *For $t \in [0,T]$, $\mathcal{C}_t(t)\mathcal{C}^{-1}(t)$ is bounded from $\mathcal{H}$ to $\mathcal{H}$. Furthermore, for any $X \in \mathcal{H}$, $\mathcal{C}_t(\cdot)\mathcal{C}^{-1}(\cdot)X \in C([0,T];\mathcal{H})$.*

Proof. From the formula (2.13), we have

$$\mathcal{C}_t(t) = \begin{bmatrix} 0 & -N_t\partial^2 \\ 0 & 0 \end{bmatrix}. \tag{2.24}$$

Thus, if $X = \left[\begin{smallmatrix} v \\ w \end{smallmatrix}\right] \in D(\mathcal{A}_0(t))$, then (recalling the operators B_0 and Γ_0 introduced earlier in this section) we have that for all $t \in [0,T]$,

$$\begin{aligned} \|C_t(t)X\|_{\mathcal{H}}^2 &= \|N_t w''\|^2 \le c\|Nw''\|^2 \\ &\le c\left(\|Nw''\|^2 + \|B_0 v' - b_2^2\Gamma_0(w)q_1\|_H^2\right) = c\|\mathcal{A}_0(t)X\|_{\mathcal{H}}^2. \end{aligned} \tag{2.25}$$

Here, $c = \sup_{(x,t)\in Q_T} N_t^2 N^{-2}$, and the norm $\|\cdot\|_{\mathcal{H}}$ corresponds to the inner product defined before (2.13). Lemma 2.6 follows from (2.24) and (2.25). ∎

An argument similar to that for Lemma 2.5 yields the following lemma.

LEMMA 2.7. *Assume that $u = (u_1, u_2) \in C^2[0,T] \times C^1[0,T]$. There is a function g on Q_T that satisfies the conditions* (2.4) *such that*

$$\mathcal{C}(\cdot)\mathcal{F}_0(\cdot) \in C([0,T], \mathcal{H}_1),$$

where $\mathcal{F}_0(t)$ is defined in (2.16).

After we apply [8] to the problem (2.23) and go back all the way to the problem (2.2), we actually obtain the following theorem.

THEOREM 2.1. *Suppose that $T > 0$ and $u = (u_1, u_2) \in C^3[0,T] \times C^2[0,T]$. Assume that $(\varphi_0, \psi_0) \in H^2(0,L) \times H^3(0,L)$ satisfies the boundary conditions*

$$\varphi_0(L) = \psi_0(L) = \psi_0''(L) = 0$$

and the compatibility conditions

$$\begin{cases} 2b_1 u_1(0) = \varphi_0(0) - b_1^2\psi_0'(0), \\ 2b_2 u_2(0) = 2\kappa b_1^{-1} u_{1t}(0) + \varphi_0'(0) - \kappa b_1^{-2} N(0,0)\psi_0''(0) + b_2^2\psi_0(0). \end{cases} \tag{2.26}$$

Then the problem (2.2) *admits a unique solution* (φ, ψ) *satisfying*

$$\varphi \in \bigcap_{i=0}^{2} C^i\left([0,T]; H^{2-i}(0,L)\right), \quad \psi \in \bigcap_{i=0}^{3} C^i\left([0,T]; H^{3-i}(0,L)\right).$$

Remark 2.1. For an integer $k \geq 2$ the method for Theorem 2.1 can be used to find solutions to the problem (2.2) such that

$$\varphi \in \bigcap_{i=0}^{k} C^i\left([0,T]; H^{k-i}(0,L)\right), \quad \psi \in \bigcap_{i=0}^{k+1} C^i\left([0,T]; H^{k-i}(0,L)\right),$$

where, of course, more boundary conditions and more compatibility conditions for the initial data and the input function are needed.

3. Quasi-linear equations. In this section we use Theorem 2.1 and a fixed-point argument to establish the existence of solutions of the quasi-linear problem (1.11)–(1.12). To this end, we need to establish some energy estimates of the solutions to the linear problem (2.2), which is also useful when we derive the global solutions later.

Let $T > 0$ and $u = (u_1, u_2) \in C^3(0,T) \times C^2(0,T)$. Suppose that $(\varphi_0, \psi_0) \in H^2(0,L) \times H^3(0,L)$ satisfies the boundary conditions (1.15) and the compatibility conditions (1.16), and φ_0 satisfies the σ-condition (1.14). We fix constants $\xi_\pm \in \mathbb{R}$ with $\xi_- < 0 < \xi_+$ such that

$$\sigma_- < \xi_- < \inf_{x\in[0,L]} \varphi_0(x) \leq \sup_{x\in[0,L]} \varphi_0(x) < \xi_+ < \sigma_+, \tag{3.1}$$

where $\sigma_\pm$ are given by (1.9).

For $r, T > 0$, let $M(r, T, \xi_-, \xi_+)$ denote the set of all functions v defined on Q_T which satisfy

$$v \in \bigcap_{i=0}^{2} C^i\left([0,T]; H^{2-i}(0,L)\right),$$

$$\xi_- \leq v(x,t) \leq \xi_+, \quad (x,t) \in Q_T, \tag{3.2}$$

$$v(x,0) = \varphi_0(x), \quad v_t(x,0) = \sigma'(\sigma^{-1}(\varphi_0(x)))\psi_0''(x), \quad x \in [0,L], \tag{3.3}$$

$$v(L,t) = 0, \quad t \in [0,T],$$

$$\operatorname*{ess\,sup}_{t\in[0,T]} \mathbf{h}(v,t) \leq r^2, \tag{3.4}$$

where

$$\mathbf{h}(v,t) = v'^2(0,t) + \sum_{|\alpha|\leq 2} \|\partial^\alpha v\|^2. \tag{3.5}$$

In the above formula $\|\cdot\|$ denotes, as usual, the norm in $L^2(0,L)$. It follows from the trace theorem that $M(r, T, \xi_-, \xi_+)$ is nonempty if r is sufficiently large and T small. For example, if we let

$$v(x,t) = \varphi_0(x)\cos t + \sigma'(\sigma^{-1}(\varphi_0(x)))\psi_0''(x)\sin t,$$

then $v \in M(r, T, \xi_-, \xi_+)$ for r suitably large and T small. In the present section, we always make this assumption.

For any $v \in M(r, T, \xi_-, \xi_+)$, we consider the linear problem

$$\begin{cases} \eta(v)\varphi_t - \psi'' = 0, \\ \psi_t - \kappa\psi_t'' + \varphi'' = 0, \\ \varphi(x,0) = \varphi_0(x), \qquad \psi(x,0) = \psi_0(x), \\ \varphi(L,t) = \psi(L,t) = \psi''(L,t) = 0, \\ 2b_1u_1(t) = \varphi(0,t) - b_1^2\psi'(0,t), \\ 2b_2u_2(t) = \varphi'(0,t) - \kappa\psi_t'(0,t) + b_2^2\psi(0,t), \end{cases} \tag{3.6}$$

with the output

$$\begin{cases} 2b_1y_1(t) = \varphi(0,t) + b_1^2\psi'(0,t), \\ 2b_2y_2(t) = \varphi'(0,t) - \kappa\psi_t'(0,t) - b_2^2\psi(0,t), \end{cases}$$

where

$$\eta(z) = 1/\mathcal{N}(z), \quad z \in (\sigma_-, \sigma_+) \tag{3.7}$$

and $\mathcal{N}(\cdot)$ is defined in (1.8). Note that the assumptions (3.1) and (3.2) imply

$$0 < \inf_{\sigma^{-1}(\xi_-) \le \xi \le \sigma^{-1}(\xi_+)} \sigma'(\xi) \le \mathcal{N}(v(x,t)) \le \sup_{\sigma^{-1}(\xi_-) \le \xi \le \sigma^{-1}(\xi_+)} \sigma'(\xi) < \infty$$

for all $(x,t) \in Q_T$, so that there are $\eta_1 > 0$, $\eta_2 > 0$ such that

$$\eta_1 \le \eta(v) \le \eta_2 \quad \forall\, v \in M(r, T, \xi_-, \xi_+).$$

It is easy to check that for $v \in M(r, T, \xi_-, \xi_+)$ and $N = \mathcal{N}(v)$ the compatibility conditions (2.26) hold. By Theorem 2.1, for any $v \in M(r, T, \xi_-, \xi_+)$, the problem (3.6) has a unique solution (φ, ψ) with the regularity $\varphi \in \cap_{i=0}^{2} C([0,T]; H^{2-i}(0,L))$ and $\psi \in \cap_{i=0}^{3} C([0,T]; H^{3-i}(0,L))$. We denote by $\mathcal{K}$ the map which carries $v \in M(r, T, \xi_-, \xi_+)$ into φ, i.e.,

$$\mathcal{K}v = \varphi, \tag{3.8}$$

where (φ, ψ) is the solution of the problem (3.6). We establish an a priori estimate to show that $\mathcal{K}$ maps $M(r, T, \xi_-, \xi_+)$ into itself if r is sufficiently large and T is sufficiently small relative to r. We then equip $M(r, T, \xi_-, \xi_+)$ with the complete metric ρ defined by

$$\rho(v, \hat{v}) = \sup_{t \in [0,T]} \|v(\cdot, t) - \hat{v}(\cdot, t)\|$$

and show that $\mathcal{K}$ is a strict contraction if T is sufficiently small. Then the Banach fixed-point theorem implies that $\mathcal{K}$ has a fixed point which is obviously a solution to

the problem (1.11)–(1.12). We define, omitting to write $(\cdot, t)$ in many places,

$$\Gamma(t) = \frac{1}{2}\varphi'^2(0,t),$$

$$E_{00}(t) = \frac{1}{2}\left[\|\psi\|^2 + \kappa\|\psi'\|^2 + \langle\eta(v)\varphi,\, \varphi\rangle\right],$$

$$E_{10}(t) = \frac{1}{2}\left[\|\psi'\|^2 + \kappa\|\psi''\|^2 + \langle\eta(v)\varphi',\, \varphi'\rangle\right],$$

$$E_{01}(t) = \frac{1}{2}\left[\|\psi_t\|^2 + \kappa\|\psi_t'\|^2 + \langle\eta(v)\varphi_t,\, \varphi_t\rangle\right],$$

$$E_{20}(t) = \frac{1}{2}\left[\|\psi''\|^2 + \kappa\|\left(\eta(v)\varphi_t\right)'\|^2 + \langle\eta(v)\varphi'',\, \varphi''\rangle\right],$$

$$E_{02}(t) = \frac{1}{2}\left[\|\psi_{tt}\|^2 + \kappa\|\psi_{tt}'\|^2 + \langle\eta(v)\varphi_{tt},\, \varphi_{tt}\rangle\right].$$

Moreover, we set

$$\wp = b_1^2 + b_1^{-2} + b_2^2 + b_2^{-2},$$

$$\aleph(v,t) = 8\wp[(\kappa^2+1)(\eta(v(0,t))+1)^2 + 1] + 1,$$

$$\Pi(t) = \Gamma(t) + \sum_{i=1}^{2} E_{i0}(t) + \aleph(v,t)\sum_{i=0}^{2} E_{0i}(t). \tag{3.9}$$

Remark 3.1. It is clear that the function $\aleph(v,t)$ depends on the function $v(0,t)$. However, there are constants $c_1(\xi_-,\xi_+) > 0$, $c_2(\xi_-,\xi_+) > 0$, which depend only on ξ_-, ξ_+, such that for any function v satisfying the condition (3.2),

$$c_1(\xi_1,\xi_+) \le \aleph(v,t) \le c_2(\xi_-,\xi_+).$$

Similar estimates hold for $\Pi(t)$. Further estimates on $\aleph$ will be given in (3.27)–(3.29).

THEOREM 3.1. *Let ξ_-, ξ_+ be such that*

$$\sigma_- < \xi_- < 0 < \xi_+ < \sigma_+, \tag{3.10}$$

where $\sigma_\pm$ are given by (1.9). *Let $v \in \cap_{i=0}^{2} C^i\left([0,T]; H^{2-i}(0,L)\right)$ satisfy*

$$\xi_- \le v(x,t) \le \xi_+, \quad (x,t) \in Q_T, \tag{3.11}$$

and the boundary condition

$$v(L,t) = 0, \quad 0 \le t \le T.$$

Suppose that (φ,ψ) is the solution of the problem (3.6) *on $Q_T = (0,L)\times(0,T)$.*

Then there is a number $C(\xi_-,\xi_+) > 0$, which may depend on the constants ξ_-, ξ_+ but is independent of the initial data (φ_0,ψ_0), such that for $0 \le t \le T$

$$\frac{\mathrm{d}\Pi(t)}{\mathrm{d}t} \le C(\xi_-,\xi_+)|u|_2^2(t) - |y|_2^2(t) + C(\xi_-,\xi_+)\left[\mathbf{h}^{\frac{1}{2}}(v,t) + \mathbf{h}(v,t)\right]\Pi(t), \tag{3.12}$$

where $|u|_2(t)$ is defined by (1.19) *and $\mathbf{h}(v,t)$ is defined by* (3.5).

Remark 3.2. Here we use a special combination of $E_{ij}(t)$, as in the formula (3.9), to control the boundary terms to have the term $-|y|_2^2$ on the right-hand side of the inequality (3.12). We need the estimate (3.12) for the existence of local smooth solutions of the quasi-linear system, and it also plays a key role in the proof of the existence of global smooth solutions.

The proof of Theorem 3.1 is divided into several lemmas. We introduce the notation ∇ for the gradient of a function φ defined on Q_T, so that $\nabla\varphi = [\varphi'\ \varphi_t]$. We denote

$$|\nabla\varphi(\cdot,t)|_\infty = \sup_{x\in[0,L]}\{|\varphi'(x,t)|,\ |\varphi_t(x,t)|\}.$$

LEMMA 3.1. *Let* $v \in \cap_{i=0}^{2} C^i\left([0,T]; H^{2-i}(0,L)\right)$ *satisfy the condition* (3.11). *Suppose that* (φ,ψ) *is the solution to the problem* (3.6). *Then*

$$|\nabla\varphi(\cdot,t)|_\infty^2 \le 2(L+1)\cdot \mathbf{h}(\varphi,t), \quad t\in[0,T], \tag{3.13}$$

$$|\nabla v(\cdot,t)|_\infty^2 \le 2(L+1)L\cdot \mathbf{h}(v,t), \quad t\in[0,T]. \tag{3.14}$$

This follows from the boundary conditions. By the boundary structure of the system (3.6), a simple computation shows that the following lemma holds.

LEMMA 3.2. *With the notation from the previous lemma,*

$$\begin{cases} b_2\psi(0) = u_2 - y_2, \\ b_1\psi'(0) = y_1 - u_1, \\ \varphi(0) = b_1(y_1+u_1), \\ \varphi'(0) - \kappa\psi_t'(0) = b_2(y_2+u_2), \\ \varphi'(0) = \kappa b_1^{-1}(y_{1t}-u_{1t}) + b_2(u_2+y_2). \end{cases} \tag{3.15}$$

In the formulas (3.15) we have suppressed the variable t for the sake of simplicity; for example, $\psi(0) = \psi(0,t)$, $\varphi(0) = \varphi(0,t)$, and so on. We shall always make this omission if there is no danger of confusion.

Denote

$$\check{c}(\xi_-,\xi_+) = \inf_{\xi_-\le x\le\xi_+}\eta(x), \quad \hat{c}(\xi_-,\xi_+) = \sup_{\xi_-\le x\le\xi_+}\eta(x), \tag{3.16}$$

$$c_1(\xi_-,\xi_+) = \sup_{\xi_-\le x\le\xi}|\eta'(x)|, \quad c_2(\xi_-,\xi_+) = \sup_{\xi_-\le x\le\xi_+}|\eta''(x)|. \tag{3.17}$$

Since $[\xi_-,\xi_+]\subset(\sigma_-,\sigma_+)$, the formulas (1.8), (1.9), (1.10), and (3.7) imply that

$$0 < \check{c}(\xi_-,\xi_+) \le \hat{c}(\xi_-,\xi_+) < \infty, \quad c_1(\xi_-,\xi_+) < \infty, \quad c_2(\xi_-,\xi_+) < \infty.$$

LEMMA 3.3. *Let the assumptions in Theorem* 3.1 *hold. Then there exists a number* $C(\xi_-,\xi_+) > 0$, *which may depend on* ξ_- *and* ξ_+, *such that, for* $0\le t\le T$,

$$\frac{\mathrm{d}E_{00}(t)}{\mathrm{d}t} \le |u(t)|^2 - |y(t)|^2 + C(\xi_-,\xi_+)\mathbf{h}^{\frac{1}{2}}(v,t)E_{00}(t).$$

Proof. Using the systems (3.6), the formulas (3.15), (3.16), and (3.17), and integration by parts, we find (suppressing the notation for the dependence on t)

$$\begin{aligned}\frac{\mathrm{d}E_{00}}{\mathrm{d}t} &= \langle\psi,\psi_t\rangle + \kappa\langle\psi',\psi_t'\rangle + \langle\varphi,\psi''\rangle + \frac{1}{2}\langle\eta'(v)v_t\varphi,\varphi\rangle\\ &\le (u_2+y_2)b_2\psi(0) - (u_1+y_1)b_1\psi'(0) + c_1(\xi_-,\xi_+)|\nabla v|_\infty\|\varphi\|^2\\ &= |u|^2 - |y|^2 + c_1(\xi_-,\xi_+)|\nabla v|_\infty\|\varphi\|^2\\ &\le |u|^2 - |y|^2 + 2\check{c}^{-1}(\xi_-,\xi_+)c_1(\xi_-,\xi_+)[2(L+1)]^{1/2}\mathbf{h}^{\frac{1}{2}}(v)E_{00},\end{aligned}$$

where the inequality (3.14) was used. ∎

LEMMA 3.4. *Let the assumptions in Theorem* 3.1 *hold. Then there exists a number $C(\xi_-,\xi_+)>0$, which may depend on ξ_-, ξ_+, such that, for $0\le t\le T$,*

$$\frac{\mathrm{d}E_{01}(t)}{\mathrm{d}t} \le |u_t(t)|^2 - |y_t(t)|^2 + C(\xi_-,\xi_+)\mathbf{h}^{\frac{1}{2}}(v,t)E_{01}(t). \tag{3.18}$$

Proof. By differentiating the first and second equations in (3.6) with respect to the time t, we obtain

$$\begin{cases}\eta(v)\varphi_{tt} = \psi_t'' - \eta'(v)v_t\varphi_t,\\ \psi_{tt} - \kappa\psi_{tt}'' + \varphi_t'' = 0.\end{cases}$$

This together with (3.15) yields

$$\begin{aligned}\frac{\mathrm{d}E_{01}}{\mathrm{d}t} &= \langle\psi_t,\psi_{tt}\rangle + \kappa\langle\psi_t',\psi_{tt}'\rangle + \langle\eta(v)\varphi_t,\varphi_{tt}\rangle + \frac{1}{2}\langle\eta'(v)v_t\varphi_t,\varphi_t\rangle\\ &= \kappa\psi_t\psi_{tt}'|_0^L + \varphi_t\psi_t'|_0^L - \varphi_t'\psi_t|_0^L - \frac{1}{2}\langle\eta'(v)v_t\varphi_t,\varphi_t\rangle\\ &= |u_t|^2 - |y_t|^2 - \frac{1}{2}\langle\eta'(v)v_t\varphi_t,\varphi_t\rangle.\end{aligned}$$

The inequality (3.18) follows from here. ∎

LEMMA 3.5. *Let the assumptions of Theorem* 3.1 *hold. Then there exists a number $C(\xi_-,\xi_+)>0$, which depends on ξ_-, ξ_+, such that, for $0\le t\le T$,*

$$\frac{\mathrm{d}E_{02}(t)}{\mathrm{d}t} \le |u_{tt}(t)|^2 - |y_{tt}(t)|^2 + C(\xi_-,\xi_+)\left[\mathbf{h}^{\frac{1}{2}}(v,t) + \mathbf{h}(v,t)\right]\mathbf{h}(\varphi,t).$$

Proof. After we differentiate the first and second equations in (3.6) twice with respect to t, we obtain (suppressing the dependence on t in the notation)

$$\begin{cases}\eta(v)\varphi_{ttt} = \psi_{tt}'' - 2\eta'(v)v_t\varphi_{tt} - \eta'(v)v_{tt}\varphi_t - \eta''(v)v_t^2\varphi_t,\\ \psi_{ttt} - \kappa\psi_{ttt}'' + \varphi_{tt}'' = 0.\end{cases} \tag{3.19}$$

It follows from (3.19), (3.13), and (3.14) that

$$\left\langle\eta(v)\varphi_{ttt},\varphi_{tt}\right\rangle = \left\langle\psi_{tt}'',\varphi_{tt}\right\rangle + I_1, \tag{3.20}$$

where

$$\begin{aligned}I_1 &= \left\langle -2\eta'(v)v_t\varphi_{tt} - \eta'(v)v_{tt}\varphi_t - \eta''(v)v_t^2\varphi_t,\ \varphi_{tt}\right\rangle\\ &\le C(\xi_-,\xi_+)\left[|\nabla v|_\infty\|\varphi_{tt}\|^2 + \|v_{tt}\||\nabla\varphi|_\infty\|\varphi_{tt}\| + |\nabla v|_\infty^2\|\varphi_t\|\|\varphi_{tt}\|\right]\\ &\le C(\xi_-,\xi_+)\left[\mathbf{h}^{\frac{1}{2}}(v) + \mathbf{h}(v)\right]\mathbf{h}(\varphi).\end{aligned}$$

Using this and (3.19), (3.20), (3.15), we find

$$\begin{aligned}\frac{\mathrm{d}E_{02}}{\mathrm{d}t} &= \langle\psi_{ttt},\psi_{tt}\rangle + \kappa\langle\psi'_{ttt},\psi'_{tt}\rangle + \langle\eta(v)\varphi_{ttt},\varphi_{tt}\rangle + \frac{1}{2}\langle\eta'(v)v_t\varphi_{tt},\varphi_{tt}\rangle\\ &\le \kappa\psi_{tt}\psi'_{ttt}|_0^L + \varphi_{tt}\psi'_{tt}|_0^L - \varphi'_{tt}\psi_{tt}|_0^L + I_1 + C(\xi_-,\xi_+)|\nabla v|_\infty\mathbf{h}(\varphi)\\ &= [\varphi'(0)-\kappa\psi'_t(0)]_{tt}\,\psi_{tt}(0) - \varphi_{tt}(0)\psi'_{tt}(0) + I_1 + C(\xi_-,\xi_+)|\nabla v|_\infty\mathbf{h}(\varphi)\\ &= (u_{2tt}+y_{2tt})b_2\psi_{tt}(0) - (u_{1tt}+y_{1tt})b_1\psi'_{tt}(0) + I_1 + C(\xi_-,\xi_+)|\nabla v|_\infty\mathbf{h}(\varphi)\\ &\le |u_{tt}|^2 - |y_{tt}|^2 + C(\xi_-,\xi_+)\left[\mathbf{h}^{\frac12}(v)+\mathbf{h}(v)\right]\mathbf{h}(\varphi). \qquad \square\end{aligned}$$

For $u=(u_1,u_2)\in H^1(0,T)^2$, we define $|u|_1(t)$ by

$$|u|_1^2(t) = |u(t)|^2 + |u_t(t)|^2.$$

This is similar to the notation $|u|_2(t)$ introduced in (1.19).

LEMMA 3.6. *There exists a number $C(\xi_-,\xi_+)>0$, which depends only on ξ_-, ξ_+, such that, for $0\le t\le T$,*

$$\frac{\mathrm{d}E_{10}(t)}{\mathrm{d}t} \le \left[2\eta(v(0,t))(\kappa^2+1)+1\right]\wp\left[|u|_1^2(t)+|y|_1^2(t)\right] + C(\xi_-,\xi_+)\mathbf{h}^{\frac12}(v,t)E_{10}(t).$$

Proof. From the boundary conditions and the second equation of the system (3.6), it follows that

$$\varphi''(L)=0, \qquad \varphi''(0)-\kappa\psi''_t(0) = -\psi_t(0). \tag{3.21}$$

Using the first equation of the system (3.6), we obtain

$$\eta(v)\varphi'_t = [\eta(v)\varphi_t]' - \eta'(v)v'\varphi_t = \psi^{(3)} - \eta'(v)\eta^{-1}(v)v'\psi''. \tag{3.22}$$

Using (3.22), the second equation in (3.6), (3.21), and the first equation in (3.6),

$$\begin{aligned}\frac{\mathrm{d}E_{10}(t)}{\mathrm{d}t} &= \langle\psi'_t,\psi'\rangle + \kappa\langle\psi''_t,\psi''\rangle + \langle\eta(v)\varphi'_t,\varphi'\rangle + \frac12\langle\eta'(v)v_t\varphi',\varphi'\rangle\\ &= \kappa\psi'\psi''_t|_0^L + \varphi'\psi''|_0^L - \varphi''\psi'|_0^L + \left\langle\psi'_t - \kappa\psi_t^{(3)} + \varphi^{(3)},\,\psi'\right\rangle\\ &\quad - \left\langle\eta'(v)\eta^{-1}(v)v'\psi'',\,\varphi'\right\rangle + \frac12\langle\eta'(v)v_t\varphi',\varphi'\rangle\\ &\le -\psi_t(0)\psi'(0) - \eta(v(0))\varphi_t(0)\varphi'(0) + C(\xi_-,\xi_+)|\nabla v|_\infty E_{10}(t)\\ &\le \frac12\left[\psi_t^2(0)+\psi'^2(0)\right] + \frac12\eta(v(0))\left[\varphi_t^2(0)+\varphi'^2(0)\right] + C(\xi_-,\xi_+)|\nabla v|_\infty E_{10}(t).\end{aligned}$$

Moreover, a simple computation using (3.15) shows that

$$\begin{aligned}&\frac12\left[\psi_t^2(0)+\psi'^2(0)\right] + \frac12\eta(v(0))\left[\varphi_t^2(0)+\varphi'^2(0)\right]\\ &\qquad\le \left[2\eta(v(0))(\kappa^2+1)+1\right]\wp(|u|^2+|u_t|^2+|y|^2+|y_t|^2). \qquad \square\end{aligned}$$

LEMMA 3.7. *There exists a number $C(\xi_-,\xi_+)>0$, which depends only on ξ_-, ξ_+ such that, for $0\le t\le T$,*

$$\begin{aligned}\frac{\mathrm{d}E_{20}(t)}{\mathrm{d}t} \le{}& \left\{4\wp\left[(2\kappa^2+1)\eta^2(v(0,t))+1\right]+\eta(v(0,t))\right\}\left[|u|_2^2(t)+|y|_2^2(t)\right]\\ &+ C(\xi_-,\xi_+)\left[\mathbf{h}(v,t)+\mathbf{h}^{\frac12}(v,t)\right]\mathbf{h}(\varphi,t).\end{aligned}$$

Proof. Differentiating the first equation of (3.6) twice with respect to x gives

$$\eta(v)\varphi_t'' = \psi^{(4)} - 2\eta'(v)v'\varphi_t' - \eta'(v)v''\varphi_t - \eta''(v)v'^2\varphi_t.$$

This together with Lemma 3.1 then yields

$$\langle \eta(v)\varphi_t'', \varphi''\rangle = \left\langle \psi^{(4)}, \varphi''\right\rangle - I_0,$$

where

$$\begin{aligned} I_0 &= \left\langle 2\eta'(v)v'\varphi_t' + \eta'(v)v''\varphi_t + \eta''(v)v'^2\varphi_t,\ \varphi''\right\rangle \\ &\le C(\xi_-,\xi_+)\left[\mathbf{h}^{\frac{1}{2}}(v) + \mathbf{h}(v)\right]\mathbf{h}(\varphi). \end{aligned} \tag{3.23}$$

Next, noting that $(\eta(v)\varphi_t)' = \psi^{(3)}$, we obtain via integration by parts and the estimate (3.23) that

$$\begin{aligned} \frac{\mathrm{d}E_{20}}{\mathrm{d}t} &= \langle \psi'', \psi_t''\rangle + \kappa\left\langle \psi^{(3)}, \psi_t^{(3)}\right\rangle + \langle \eta(v)\varphi'', \varphi_t''\rangle + \frac{1}{2}\langle \eta'(v)v_t\varphi'', \varphi''\rangle \\ &\le \kappa\psi''\psi_t^{(3)}|_0^L + \varphi''\psi^{(3)}|_0^L - \varphi^{(3)}\psi''|_0^L + \left\langle \psi'',\ \psi_t'' - \kappa\psi_t^{(4)} + \varphi^{(4)}\right\rangle - I_0 \\ &= \left[\varphi^{(3)}(0) - \kappa\psi_t^{(3)}(0)\right]\psi''(0) - \varphi''(0)\psi^{(3)}(0) - I_0 \\ &= -\psi_t'(0)\psi''(0) - \varphi''(0)\psi^{(3)}(0) - I_0. \end{aligned} \tag{3.24}$$

From the formulas in (3.15) we see that

$$\begin{aligned} -\psi_t'(0)\psi''(0) &= -\psi_t'(0)\eta(v(0))\varphi_t(0) = -\eta(v(0))b_1\psi_t'(0)(y_{1t} + u_{1t}) \\ &= \eta(v(0))(u_{1t}^2 - y_{1t}^2). \end{aligned} \tag{3.25}$$

Let us estimate the term $-\varphi''(0)\psi^{(3)}(0)$. From the equations in (3.6) and the formulas in (3.15) we have

$$\begin{aligned} \varphi''(0) &= \kappa\psi_t''(0) - \psi_t(0) = \kappa(\eta(v)\varphi_t)_t(0) - \psi_t(0) \\ &= \kappa\eta'(v(0))v_t(0)\varphi_t(0) + \kappa b_1\eta(v(0))(u_{1tt} + y_{1tt}) - b_2^{-1}(u_{2t} - y_{2t}). \end{aligned}$$

It follows from here that

$$\varphi''^2(0) \le C(\xi_-,\xi_+)|\nabla\varphi|_\infty^2|\nabla v|_\infty^2 + 8[\kappa^2\eta^2(v(0)) + 1]\wp\left[u_{2t}^2 + y_{2t}^2 + u_{1tt}^2 + y_{1tt}^2\right]. \tag{3.26}$$

In addition, a similar computation gives the estimate

$$[\psi^{(3)}(0)]^2 \le C(\xi_-,\xi_+)|\nabla\varphi|_\infty^2|\nabla v|_\infty^2 + 8\eta^2(v(0))(\kappa^2+1)\wp\left[u_{1tt}^2 + y_{1tt}^2 + u_{2t}^2 + y_{2t}^2\right].$$

Combining this with (3.26), (3.25), (3.24), and (3.23) completes the proof. ∎

A similar computation proves the following lemma.

LEMMA 3.8.

$$\frac{\mathrm{d}\Gamma(t)}{\mathrm{d}t} \le 4(\kappa^2+1)\wp\left[|u|_2^2 + |y|_2^2\right].$$

Proof of Theorem 3.1. Under the assumption (3.11), there are $c(\xi_-,\xi_+) > 0$, $C(\xi_-,\xi_+) > 0$, which may depend on ξ_-,ξ_+ but are independent of the initial data and the input, such that, if v satisfies the condition (3.2), then

$$1 \le \aleph(v,t) \le C(\xi_-,\xi_+), \tag{3.27}$$

$$\left|\frac{\mathrm{d}\aleph(v,t)}{\mathrm{d}t}\right| \le C(\xi_-,\xi_+)\mathbf{h}^{\frac{1}{2}}(v,t), \tag{3.28}$$

$$\mathbf{h}(\varphi,t) \le C(\xi_-,\xi_+)\Pi(t) \tag{3.29}$$

for all $t \in [0,T]$. By summing up the differential inequalities from Lemmas 3.3–3.8 and using (3.27)–(3.29), we compute that for $t \in [0,T]$,

$$\begin{aligned}
\frac{\mathrm{d}\Pi(t)}{\mathrm{d}t} &= \frac{\mathrm{d}\Gamma(t)}{\mathrm{d}t} + \sum_{i=1}^{2}\frac{\mathrm{d}E_{i0}(t)}{\mathrm{d}t} + \aleph(v,t)\sum_{i=0}^{2}\frac{\mathrm{d}E_{0i}(t)}{\mathrm{d}t} + \frac{\mathrm{d}\aleph(v,t)}{\mathrm{d}t}\sum_{i=0}^{2}E_{0i}(t)\\
&\le [\aleph(v,t)-1]\left[|u|_2^2(t)+|y|_2^2(t)\right] + \aleph(v,t)\left[|u|_2^2(t)-|y|_2^2(t)\right]\\
&\quad + C(\xi_-,\xi_+)[\mathbf{h}(v,t)+\mathbf{h}^{\frac{1}{2}}(v,t)]\Pi(t)\\
&= [2\aleph(v,t)-1]|u|_2^2(t) - |y|_2^2(t) + C(\xi_-,\xi_+)\left[\mathbf{h}(v,t)+\mathbf{h}^{\frac{1}{2}}(v,t)\right]\Pi(t)\\
&\le C(\xi_-,\xi_+)|u|_2^2(t) - |y|_2^2(t) + C(\xi_-,\xi_+)\left[\mathbf{h}(v,t)+\mathbf{h}^{\frac{1}{2}}(v,t)\right]\Pi(t). \qquad \square
\end{aligned}$$

Proof of Theorem 1.1. Since φ_0 satisfies the σ-condition (1.14), we can fix constants ξ_-, ξ_+ such that the relations (3.1) and (3.10) are true.

Recall the notation $M(r,T,\xi_-,\xi_+)$ from the beginning of this section. Fix r large and T small such that $M(r,T,\xi_-,\xi_+)$ is nonempty. For any $v \in M(r,T,\xi_-,\xi_+)$, by Theorem 2.1, the condition (3.2) implies that the problem (3.6) has a unique solution (φ,ψ). We need to verify that $\varphi \in M(r,T,\xi_-,\xi_+)$ and the map $\mathcal{K}$, given by (3.8), is a strict contraction if r is sufficiently large and T is small.

By Theorem 3.1, there exists $C(\xi_-,\xi_+) > 0$ such that

$$\Pi(t) \le \Pi(0) + C(\xi_-,\xi_+)\|u\|^2_{H^2[0,T]} + C(\xi_-,\xi_+)\int_0^t \left[\mathbf{h}^{\frac{1}{2}}(v,\tau)+\mathbf{h}(v,\tau)\right]\Pi(\tau)\mathrm{d}\tau$$

for any $t \in [0,T]$. It follows from here, (3.29), and Gronwall's inequality that

$$\begin{aligned}
\mathbf{h}(\varphi,t) &\le C(\xi_-,\xi_+)\Pi(t)\\
&\le C(\xi_-,\xi_+)[\Pi(0)+C(\xi_-,\xi_+)\|u\|^2_{H^2[0,T]}]e^{C(\xi_-,\xi_+)\int_0^t[\mathbf{h}^{\frac{1}{2}}(v,\tau)+\mathbf{h}(v,\tau)]\mathrm{d}\tau}
\end{aligned} \tag{3.30}$$

for any $t \in [0,T]$. From the initial data (3.3) and the system (3.6), we can check that the constant $\Pi(0)$ is independent of the choice of $v \in M(r,T,\xi_-,\xi_+)$.

We now take $r_0 > r$ sufficiently large such that

$$r_0^2 \ge 2C(\xi_-,\xi_+)\left[\Pi(0)+C(\xi_-,\xi_+)\|u\|^2_{H^2[0,T]}\right].$$

It follows from (3.6), (3.4), (3.30), and the above estimate that

$$\mathbf{h}(\varphi,t) \le \frac{r_0^2}{2}e^{C(\xi_-,\xi_+)r_0(1+r_0)t}, \quad t \in [0,T],$$

for any $v \in M(r_0, T, \xi_-, \xi_+)$. Using this, we select $T_1 \in (0, T]$ such that for $t \in (0, T_1]$ we have

$$\operatorname*{ess\,sup}_{s \in [0,t]} \mathbf{h}(\varphi, s) \le r_0^2. \tag{3.31}$$

To find a $T_2 \in (0, T_1]$ such that $\varphi \in M(r_0, T_2, \xi_-, \xi_+)$, the only thing left to do is to verify that the condition (3.2) is satisfied by φ, since the conditions (3.3) are implied by the equations (3.6). In fact, by (3.31), for all $(x,t) \in Q_{T_1}$,

$$|\varphi(x,t) - \varphi_0(x)| = \left| \int_0^t \varphi_t(x,t) \mathrm{d}t \right| \le L^{1/2} \int_0^t \mathbf{h}^{\frac{1}{2}}(\varphi, s) \mathrm{d}s \le t L^{1/2} r_0.$$

Thus (3.2) is true for φ for small $T_2 \in (0, T_1]$, because of the assumption (3.1). With the above choice of r_0 and T_2, $\mathcal{K}$ maps $M(r_0, t, \xi_-, \xi_+)$ into itself for every $t \in (0, T_2]$.

Now we shall find a $T_3 \in (0, T_2]$ such that $\mathcal{K}$ is contractive from $M(r_0, T_3, \xi_-, \xi_+)$ to itself. Let $v_1, v_2 \in M(r_0, T_2, \xi_-, \xi_+)$ and set $\varphi_1 = \mathcal{K} v_1$, $\varphi_2 = \mathcal{K} v_2$, $v = v_1 - v_2$, $\varphi = \varphi_1 - \varphi_2$, $\psi = \psi_1 - \psi_2$. A simple computation shows that

$$\begin{cases} \eta(v_1)\varphi_t = \psi'' - [\eta(v_1) - \eta(v_2)]\varphi_{2t}, \\ \psi_t - \kappa \psi_t'' + \varphi'' = 0, \\ \varphi(x,0) = 0, \quad \psi(x,0) = 0, \\ \varphi(L,t) = \psi(L,t) = \psi''(L,t) = 0, \\ \varphi(0,t) - b_1^2 \psi'(0,t) = 0, \\ \varphi'(0,t) - \kappa \psi_t'(0,t) + b_2^2 \psi(0,t) = 0. \end{cases} \tag{3.32}$$

Denoting

$$\hat{E}_{00}(t) = \frac{1}{2} \left[\|\psi\|^2 + \kappa \|\psi'\|^2 + \langle \eta(v_1)\varphi, \varphi \rangle \right] \tag{3.33}$$

(where the argument $(\cdot, t)$ has been omitted in several places), from (3.32) we obtain

$$\begin{aligned} \frac{\mathrm{d}\hat{E}_{00}}{\mathrm{d}t} &= \kappa \psi \psi_t' |_0^L + \varphi \psi' |_0^L - \varphi' \psi |_0^L + I_2 \\ &= -b_2^2 \psi^2(0,t) - b_1^2 \psi'^2(0,t) + I_2 \le I_2, \end{aligned} \tag{3.34}$$

where

$$I_2 = \langle [\eta(v_2) - \eta(v_1)]\varphi_{2t}, \varphi \rangle + \frac{1}{2} \langle \eta'(v_1) v_{1t} \varphi, \varphi \rangle .$$

It follows from here and Lemma 3.1 that there exists a number $C(\xi_-, \xi_+) > 0$, which may depend on ξ_-, ξ_+, such that

$$\begin{aligned} |I_2| &\le C(\xi_-, \xi_+) \left[|\nabla \varphi_2|_\infty \|\varphi\| \|v_1 - v_2\| + |\nabla v_1|_\infty \|\varphi\|^2 \right] \\ &\le C(\xi_-, \xi_+) \left[\mathbf{h}^{\frac{1}{2}}(\varphi_2) \rho(\mathcal{K} v_1, \mathcal{K} v_2) \rho(v_1, v_2) + \mathbf{h}^{\frac{1}{2}}(v_1) \|\varphi\|^2 \right] \\ &\le C(\xi_-, \xi_+) \left[r_0 \rho(\mathcal{K} v_1, \mathcal{K} v_2) \rho(v_1, v_2) + \mathbf{h}^{\frac{1}{2}}(v_1) \|\varphi\|^2 \right], \end{aligned} \tag{3.35}$$

since $\varphi_2 = \mathcal{K} v_2 \in M(r_0, T_2, \xi_-, \xi_+)$. By virtue of (3.32), (3.33), (3.34), and (3.35),

for $T_3 \in (0, T_2]$ and for $t \in (0, T_3]$,

$$\|\varphi(\cdot,t)\|^2 \le C(\xi_-,\xi_+)\hat{E}_{00}(t) \le C(\xi_-,\xi_+)\int_0^t I_2(s)\mathrm{d}s$$

$$\le T_3 r_0 C(\xi_-,\xi_+)\rho(\mathcal{K}v_1,\mathcal{K}v_2)\rho(v_1,v_2) + C(\xi_-,\xi_+)\int_0^t \mathbf{h}^{\frac{1}{2}}(v_1,s)\|\varphi(\cdot,s)\|^2\mathrm{d}s.$$

Applying Gronwall's inequality to the last estimate yields

$$\|\varphi(\cdot,t)\|^2 \le T_3 r_0 C(\xi_-,\xi_+)\rho(\mathcal{K}v_1,\mathcal{K}v_2)\rho(v_1,v_2)e^{C(\xi_-,\xi_+)r_0T_3}$$

for all $t \in (0, T_3]$, where $T_3 \in (0, T_2]$. Since $\varphi = \mathcal{K}v_1 - \mathcal{K}v_2$,

$$\rho(\mathcal{K}v_1,\mathcal{K}v_2) \le T_3 r_0 C(\xi_-,\xi_+)\rho(v_1,v_2)e^{C(\xi_-,\xi_+)r_0T_3}, \quad T_3 \in (0,T_2].$$

The last inequality shows that $\mathcal{K}$ is strictly contractive if T_3 is sufficiently small. By the Banach fixed-point theorem, for $T_3 > 0$ sufficiently small $\mathcal{K}$ has a unique fixed point $\varphi \in M(r_0, T_3, \xi_-, \xi_+)$. Then (φ, ψ) is the desired solution of the system (1.11)–(1.12). □

Let (φ, ψ) be a solution of (1.11)–(1.12) on Q_T such that (1.17) holds. By the argument above, (φ, ψ) is the solution of the problem (3.6) on Q_T where $v = \varphi$. Thus, by the inequality (3.29), Theorem 3.1 can be rewritten as follows.

THEOREM 3.1*. *Let $T > 0$ and $\sigma_- < \xi_- < 0 < \xi_+ < \sigma_+$. There exists $C(\xi_-, \xi_+) > 0$, which depends on ξ_-, ξ_+ such that, if the solution (φ, ψ) of the problem (1.11)–(1.12) on Q_T satisfies $\xi_- \le \varphi \le \xi_+$, then for $0 \le t \le T$,*

$$\frac{\mathrm{d}\Pi(t)}{\mathrm{d}t} + |y|_2^2(t) \le C(\xi_-,\xi_+)|u|_2^2(t) + C(\xi_-,\xi_+)^2\left[\Pi^2(t) + \Pi^{3/2}(t)\right], \tag{3.36}$$

where $\Pi(t)$ is given by the formula (3.9) with $\aleph(v,t) = \aleph(\varphi,t)$.

4. Estimates of the boundary trace. Let $T > 0$ and $u = (u_1, u_2) \in H^2(0,T) \times H^2(0,T)$. Suppose that $(\varphi_0, \psi_0) \in H^2(0,L) \times H^3(0,L)$ satisfies the boundary conditions (1.15) and the compatibility conditions (1.16), where φ_0 satisfies the σ-condition (1.14). We assume that (φ, ψ) solves the problem (1.11)–(1.12) on $Q_T = (0,L)\times(0,T)$ such that the regularity (1.17) holds. Let $\xi_-, \xi_+ \in \mathbb{R}$ be such that

$$\sigma_- < \xi_- < 0 < \xi_+ < \sigma_+.$$

We define

$$T_{\xi_-,\xi_+} = L\kappa^{\frac{1}{2}} \sup_{\xi_- \le z \le \xi_+} \eta^{\frac{1}{2}}(z), \tag{4.1}$$

where (as usual) $\eta = 1/\mathcal{N}$ and $\mathcal{N}$ is defined in (1.8).

The purpose of this section is to prove the following theorem.

THEOREM 4.1. *Suppose that $\varsigma > 0$, $T > 3T_{\xi_-,\xi_+}$ and (φ, ψ) is a solution of the problem (1.11)–(1.12) on Q_T. Then there exists a number $C(\xi_-, \xi_+, \varsigma) > 0$, which depends on ξ_-, ξ_+ and ς such that, if (φ, ψ) satisfies*

$$\xi_- \le \varphi \le \xi_+, \quad \check{\mathcal{E}}_T \le \varsigma,$$

where

$$\check{\mathcal{E}}_T = \sup_{0\le t\le T} \mathcal{E}(t) \tag{4.2}$$

and $\mathcal{E}(t)$ *is defined by* (1.18)*, then for* $0 \le s \le t \le T$ *with* $t - s \ge 3T_{\xi_-,\xi_+}$,

$$\int_s^t \mathcal{E}(\tau)\mathrm{d}\tau \le C(\xi_-,\xi_+,\varsigma)\left(\|u\|^2_{H^2(s,t)} + \|y\|^2_{H^2(s,t)} + \int_s^t [\mathcal{E}^2(\tau) + \mathcal{E}^{\frac{3}{2}}(\tau)]\,\mathrm{d}\tau\right). \tag{4.3}$$

Remark 4.1. In fact, for any $\varepsilon > 0$, we can prove that the inequality (4.3) holds for $t - s \ge 2T_{\xi_-,\xi_+} + \varepsilon$, where the constant $C(\xi_-,\xi_+,\varsigma)$ may also depend on ε. In general, the conclusion of Theorem 4.1 is not true when $t - s < 2T_{\xi_-,\xi_+}$.

The proof of Theorem 4.1 is based on several lemmas below. Set

$$\vartheta = \begin{bmatrix}\varphi\\ \psi'\end{bmatrix}, \qquad A(\varphi) = \begin{bmatrix}0 & -\kappa\\ -\eta(\varphi) & 0\end{bmatrix}, \tag{4.4}$$

where $\eta(z) = 1/\mathcal{N}(z) = 1/\sigma'(\sigma^{-1}(z))$. We diagonalize $A(\varphi)$ as

$$SA(\varphi)S^{-1} = \begin{bmatrix}\kappa^{\frac{1}{2}}\eta^{\frac{1}{2}}(\varphi) & 0\\ 0 & -\kappa^{\frac{1}{2}}\eta^{\frac{1}{2}}(\varphi)\end{bmatrix},$$

where

$$S = \frac{1}{2}\begin{bmatrix}1 & -\kappa^{\frac{1}{2}}\eta^{-\frac{1}{2}}(\varphi)\\ 1 & \kappa^{\frac{1}{2}}\eta^{-\frac{1}{2}}(\varphi)\end{bmatrix}, \quad S^{-1} = \begin{bmatrix}1 & 1\\ -\kappa^{-\frac{1}{2}}\eta^{\frac{1}{2}}(\varphi) & \kappa^{-\frac{1}{2}}\eta^{\frac{1}{2}}(\varphi)\end{bmatrix}. \tag{4.5}$$

LEMMA 4.1. *Let* $\xi_- \le \varphi \le \xi_+$ *hold on* Q_T. *Let* $\vartheta = [\begin{smallmatrix}\varphi\\ \psi'\end{smallmatrix}]$, *where* (φ,ψ) *is the solution of the problem* (1.11)–(1.12). *Then*

$$A(\varphi)\vartheta_t + \vartheta' = \begin{bmatrix}\Lambda\varphi - \kappa b_2[u_2(t) + y_2(t)]q_1'\\ 0\end{bmatrix}, \quad (x,t) \in Q_T, \tag{4.6}$$

where

$$(\Lambda\varphi)(x) = a^{-1/2}\langle\varphi, q_0'\rangle \sinh a^{\frac{1}{2}}x - a\int_0^x \varphi(s)\cosh a^{\frac{1}{2}}(x-s)\mathrm{d}s. \tag{4.7}$$

Proof. From the second equation in (1.11), a computation as in (2.10) yields

$$\psi_t = B_0\varphi' - b_2^2\psi(0,t)q_1 + 2b_2u_2(t)q_1,$$

where B_0 is defined in Lemma 2.1 and q_1 by (2.11). We then employ (2.9), (2.8), and $b_2\psi(0,t) = u_2(t) - y_2(t)$ (see (3.15) with $\psi = w$) to obtain

$$\begin{aligned}\kappa\psi_t' &= \kappa\frac{\partial}{\partial x}B_0\varphi' - \kappa b_2^2\psi(0,t)q_1' + 2\kappa b_2u_2(t)q_1'\\ &= \varphi' + a^{-\frac{1}{2}}\langle\varphi', q_0\rangle\sinh a^{\frac{1}{2}}x + a^{\frac{1}{2}}\int_0^x \varphi'(s)\sinh a^{\frac{1}{2}}(x-s)\mathrm{d}s\\ &\quad - \kappa b_2^2\psi(0,t)q_1' + 2\kappa b_2u_2(t)q_1'\\ &= \varphi' - \Lambda\varphi + \kappa b_2[u_2(t) + y_2(t)]q_1'. \end{aligned} \tag{4.8}$$

The formula (4.6) follows from (4.8) and the first equation in (1.11). ∎

Remark 4.2. Assume for a moment that $b_2 = 0$. Using (4.8) we observe that, if we make the change of unknowns

$$\mu = \varphi, \qquad \nu = \psi',$$

then the problem (1.11)–(1.12) becomes

$$\begin{cases} \mu_t = \eta^{-1}(\mu)\nu', \\ \kappa\nu_t = \mu' - \Lambda(\mu), \\ \mu(x,0) = \varphi_0(x), \quad \nu(x,0) = \psi_0(x), \\ \mu(L,t) = 0, \end{cases} \tag{4.9}$$

with the input-output at $x = 0$:

$$\begin{cases} 2b_1 u_1(t) = \mu(0,t) - b_1^2\nu(0,t), \\ 2b_1 y_1(t) = \mu(0,t) + b_1^2\nu(0,t). \end{cases} \tag{4.10}$$

It is clear that the problem (4.9)–(4.10) represents a quasi-linear string with the input-output structure at the boundary. The existence of global solutions of the above problem has been studied in [18].

If $b_2 \neq 0$, then the term $\kappa[\mu'(0,t) - \kappa\nu_t'(0,t)]q_1'$ appears on the right-hand side of the second equation in (4.9), and this makes things quite different.

For $(x,t) \in Q_T$ let $\tau_i(s) = \tau_i(s;x,t)$ be the solutions of

$$\frac{\mathrm{d}\tau_i}{\mathrm{d}s} = \lambda_i(s,\tau_i), \qquad \tau_i(x;x,t) = t, \qquad i = 1,2, \tag{4.11}$$

where

$$\lambda_1(x,t) = \kappa^{\frac{1}{2}}\eta^{\frac{1}{2}}(\varphi(x,t)), \qquad \lambda_2(x,t) = -\kappa^{\frac{1}{2}}\eta^{\frac{1}{2}}(\varphi(x,t)).$$

Then $(s, \tau_i(s;x,t))$ are the forward and backward characteristics through (x,t), respectively, for $i = 1,2$. An argument similar to that in [1] implies the following lemma.

LEMMA 4.2. *For $i = 1,2$,*

$$\tau_i'(s;x,t) = -\lambda_i(x,t)\tau_{it}(s;x,t).$$

Suppose that $\xi_- \le \varphi \le \xi_+$ holds on Q_T. There exist numbers $C(\xi_-,\xi_+) > 0$ and $c(\xi_-,\xi_+) > 0$, which depend only on ξ_-, ξ_+ such that for $0 \le s \le x \le L$, $0 \le t \le T$,

$$e^{-c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \le \tau_{it}(s;x,t) \le e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \tag{4.12}$$

and

$$C^{-1}(\xi_-,\xi_+)e^{-c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \le |\tau_i'(s;x,t)| \le C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}},$$

where $\check{\mathcal{E}}$ is defined in (4.2).

LEMMA 4.3. *Assume that $\xi_- \le \varphi \le \xi_+$ holds on Q_T and let T_{ξ_-,ξ_+} be given by* (4.1). *Let $\omega \in H^1(Q_T) \times H^1(Q_T)$ such that*

$$A(\varphi)\omega_t + \omega' = \Psi(\varphi), \quad (x,t) \in Q_T, \tag{4.13}$$

where $A(\varphi)$ is defined in (4.4) *and $\Psi(\varphi)$ satisfies the estimate*

(4.14)
$$|\Psi(\varphi)| \le C(\xi_-,\xi_+)\left[|\Psi_0(\varphi)| + |\nabla\varphi|_\infty|\Psi_1(\varphi)| + |\nabla\varphi|_\infty^3 + |\nabla\varphi|_\infty^2 + |\nabla\varphi|_\infty|\omega|\right]$$

for $(x,t) \in Q_T$, where $\Psi_0(\varphi)$ and $\Psi_1(\varphi)$ (the components of $\Psi(\varphi)$) are functions determined by φ. Set

$$\Upsilon = S\omega, \tag{4.15}$$

where S is given in (4.5)*. Then there exist numbers $C(\xi_-,\xi_+) > 0$ and $c(\xi_-,\xi_+) > 0$, which depend only on ξ_-, ξ_+, such that for $2T_{\xi_-,\xi_+} \le t \le T$*

(4.16)
$$\int_0^t \|\Upsilon(\cdot,\tau)\|^2 \mathrm{d}\tau \le C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac12}} \int_0^t \left[|\Upsilon(0,\varsigma)|^2 + |\Upsilon(L,\varsigma)|^2 + \Xi(\Upsilon)(\varsigma)\right]\mathrm{d}\varsigma,$$

where $\check{\mathcal{E}}_T$ is given in (4.2) *and*

$$\Xi(\Upsilon)(t) = \|\Psi_0(\varphi)(\cdot,t)\|^2 + \|\Psi_1(\varphi)(\cdot,t)\|^2$$

$$+ \left[|\nabla\varphi(\cdot,t)|_\infty^2 + |\nabla\varphi(\cdot,t)|_\infty\right]\|\Upsilon(\cdot,t)\|^2 + |\nabla\varphi(\cdot,t)|_\infty^4 + |\nabla\varphi(\cdot,t)|_\infty^3. \tag{4.17}$$

Furthermore, denoting the components of Υ by Υ_1, Υ_2, we have for $2T_{\xi_-,\xi_+} \le t \le T$,

$$\int_{T_\zeta}^t \Upsilon_1^2(L,\tau)\mathrm{d}\tau \le C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac12}} \int_0^t \left[\Upsilon_1^2(0,\varsigma) + \Xi(\Upsilon)(\varsigma)\right]\mathrm{d}\varsigma, \tag{4.18}$$

$$\int_0^{T_\zeta} \Upsilon_2^2(L,\tau)\mathrm{d}\tau \le C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac12}} \int_0^t \left[\Upsilon_2^2(0,\varsigma) + \Xi(\Upsilon)(\varsigma)\right]\mathrm{d}\varsigma. \tag{4.19}$$

Proof. From the formulas (4.13) and (4.15), a simple computation yields

$$SAS^{-1}\Upsilon_t + \Upsilon' = \Theta, \tag{4.20}$$

where

$$\Theta = \left(SAS^{-1}S_t + S'\right)S^{-1}\Upsilon + S\Psi(\varphi). \tag{4.21}$$

Thus we have from (4.20) that for $(x,t) \in Q_T$,

$$\kappa^{\frac12}\eta^{\frac12}(\varphi)\Upsilon_{1t} + \Upsilon_1' = \Theta\cdot e_1, \qquad -\kappa^{\frac12}\eta^{\frac12}(\varphi)\Upsilon_{2t} + \Upsilon_2' = \Theta\cdot e_2, \tag{4.22}$$

where $e_1 = (1,0)$ and $e_2 = (0,1)$.

Let $(x,t) \in Q_T$ be given. Using (4.22), we integrate Υ_1^2 along the forward characteristic $(s,\tau_1(s))$ from 0 to x to obtain

$$\begin{aligned}\Upsilon_1^2(x,t) &= \Upsilon_1^2(0,\tau_1(0;x,t)) + 2\int_0^x \Upsilon_1\frac{\mathrm{d}}{\mathrm{d}s}\Upsilon_1\,\mathrm{d}s \\ &= \Upsilon_1^2(0,\tau_1(0;x,t)) + 2\int_0^x (\Upsilon_1\Theta\cdot e_1)(s,\tau_1(s))\,\mathrm{d}s.\end{aligned} \tag{4.23}$$

On the other hand, from (4.21) and (4.14) we can derive that for $(x,t)\in Q_T$,

$$|\Theta|\le C(\xi_-,\xi_+)\left[|\Psi_0(\varphi)|+|\nabla\varphi|_\infty|\Psi_1(\varphi)|+|\nabla\varphi|_\infty^3+|\nabla\varphi|_\infty^2+|\nabla\varphi|_\infty|\Upsilon|\right].$$

By combining the above estimate with (4.23), we obtain

$$\Upsilon_1^2(x,t)\le\Upsilon_1^2(0,\tau_1(0;x,t))+C(\xi_-,\xi_+)\int_0^x f(s,\tau_1(s;x,t))\mathrm{d}\,s, \tag{4.24}$$

where

$$\begin{aligned}f(x,t)=&|\Upsilon_1(x,t)|\cdot|\Psi_0(\varphi)(x,t)|+|\Psi_1(\varphi)(x,t)|\cdot|\nabla\varphi(\cdot,t)|_\infty|\Upsilon_1(x,t)|\\&+\left[|\nabla\varphi(\cdot,t)|_\infty^2+|\nabla\varphi(\cdot,t)|_\infty\right]|\Upsilon(x,t)|^2+|\nabla\varphi(\cdot,t)|_\infty^3+|\nabla\varphi(\cdot,t)|_\infty^4.\end{aligned}$$

We now integrate both sides of (4.24) from T_{ξ_-,ξ_+} to t with respect to the time (denoted ν) and arrive at

$$\begin{aligned}\int_{T_{\xi_-,\xi_+}}^t\Upsilon_1^2(x,\nu)\,\mathrm{d}\,\nu\le&\int_{T_{\xi_-,\xi_+}}^t\Upsilon_1^2(0,\,\tau_1(0;x,\nu))\,\mathrm{d}\,\nu\\&+C(\xi_-,\xi_+)\int_0^x\int_{T_{\xi_-,\xi_+}}^t f(s,\tau_1(s;x,\nu))\mathrm{d}\nu\mathrm{d}\,s.\end{aligned}\tag{4.25}$$

Now we estimate the two terms on the right-hand side of (4.25) by a change of the integration variable. By the choice of the constant T_{ξ_-,ξ_+}, it is easy to see that

$$0\le\tau_1(s;x,t)\le t,\quad s\in[0,x],\ x\in[0,L],\ t\in[T_{\xi_-,\xi_+},T]. \tag{4.26}$$

For each fixed $s\in[0,x]$, we make a change of the integration variable by

$$\varsigma=\tau_1(s;x,t)$$

and use the inequalities (4.12) and (4.26) to obtain

$$\int_{T_{\xi_-,\xi_+}}^t f(s,\tau_1(s;x,\nu))\mathrm{d}\nu=\int_{\tau_1(s;x,T_{\xi_-,\xi_+})}^{\tau_1(s;x,t)}\frac{f(s,\varsigma)}{\tau_{1t}(s;x,\tau_1^{-1}(\varsigma))}\mathrm{d}\varsigma\le e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac12}}\int_0^t f(s,\nu)\mathrm{d}\nu.$$

Similarly, we have

$$\int_{T_{\xi_-,\xi_+}}^t\Upsilon_1^2(0,\,\tau_1(0;x,\nu))\,\mathrm{d}\,\nu\le e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac12}}\int_0^t\Upsilon_1^2(0,\nu)\mathrm{d}\nu.$$

We now substitute the last two inequalities into (4.25) and then integrate both sides from 0 to L with respect to the variable x to obtain

$$\begin{aligned}\int_{T_{\xi_-,\xi_+}}^t\|\Upsilon_1(\cdot,\nu)\|^2\mathrm{d}\nu\le&\,\varepsilon\int_0^t\|\Upsilon_1(\cdot,\nu)\|^2\mathrm{d}\nu\\&+C(\xi_-,\xi_+,\varepsilon)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac12}}\int_0^t\left[\Upsilon_1^2(0,\varsigma)+\Xi(\Upsilon)(\varsigma)\right]\mathrm{d}\,\varsigma,\end{aligned}\tag{4.27}$$

where we have used the inequality $|\Upsilon_1||\Psi_0(\varphi)|\le\epsilon|\Upsilon_1|^2+|\Psi_0(\varphi)|^2/4\epsilon$ for ε small.

Clearly, in order to obtain an estimate for $\int_0^t \|\Upsilon_1(\cdot,\nu)\|^2 \mathrm{d}\nu$, we have to estimate $\int_0^{T_{\xi_-,\xi_+}} \|\Upsilon_1(\cdot,\nu)\|^2 \mathrm{d}\nu$. To this end, we repeat the above process by using the forward characteristic $(s,\tau_1(s))$ again but this time on the segment $x \le s \le L$. We integrate $\Upsilon_1^2(s,\tau_1(s))$ from x to L and then we have

$$\Upsilon_1^2(x,t) \le \Upsilon_1^2(L,\tau_1(L;x,t)) + C(\xi_-,\xi_+)\int_0^L f(s,\tau_1(s;x,t))\mathrm{d}s. \tag{4.28}$$

From the choice of the T_{ξ_-,ξ_+} it is easy to see that

$$t \le \tau_1(s;x,t) \le \tau_1(L;x,t) \le T_{\xi_-,\xi_+} + t, \quad s \in [x,L],\ t \in [0,T].$$

We integrate both sides of (4.28) from 0 to T_{ξ_-,ξ_+} with respect to t. Noting that

$$\begin{aligned}\int_0^{T_{\xi_-,\xi_+}} \Upsilon_1^2(L,\tau_1(L;x,t))\mathrm{d}t &= \int_{\tau_1(L;x,0)}^{\tau_1(L;x,T_{\xi_-,\xi_+})} \Upsilon_1^2(L,\varsigma)\tau_{1t}^{-1}\mathrm{d}\varsigma \\ &\le e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^{2T_{\xi_-,\xi_+}} \Upsilon_1^2(L,t)\mathrm{d}t,\end{aligned}$$

a computation similar to that for the inequality (4.27) yields

$$\begin{aligned}\int_0^{T_{\xi_-,\xi_+}} \|\Upsilon_1(\cdot,t)\|^2\mathrm{d}t \le{}& \varepsilon\int_0^{2T_{\xi_-,\xi_+}} \|\Upsilon_1(\cdot,t)\|^2\mathrm{d}t \\ &+ C(\xi_-,\xi_+,\varepsilon)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^{2T_{\xi_-,\xi_+}} [\Upsilon_1^2(L,\varsigma) + \Xi(\Upsilon)(\varsigma)]\mathrm{d}\varsigma.\end{aligned}$$

It follows from the above inequality and (4.27) that for $2T_{\xi_-,\xi_+} \le t \le T$,

$$\int_0^t \|\Upsilon_1(\cdot,\nu)\|^2\mathrm{d}t \le C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^t \left[\Upsilon_1^2(0,\varsigma) + \Upsilon_1^2(L,\varsigma) + \Xi(\Upsilon)(\varsigma)\right]\mathrm{d}\varsigma.$$

Now, we do the same thing along the backward characteristic $(s,\tau_2(s))$ and then we obtain an inequality which is the same as the last one if Υ_1 is changed into Υ_2. Therefore, addition of these two inequalities actually yields (4.16).

Next, we prove the inequalities (4.18) and (4.19). We go back to the inequality (4.24) and set $x = L$ in (4.24). Then we integrate both sides of (4.24) with respect to time from T_{ξ_-,ξ_+} to t to give the inequality (4.18). As to the inequality (4.19), we do the same thing on the backward characteristic $(s,\tau_2(s))$ $(0 \le s \le L)$. Noting the inequality

$$0 \le \tau_2(0;L,0) \le \tau_2(0;L,T_{\xi_-,\xi_+}) \le 2T_{\xi_-,\xi_+} \le t,$$

we obtain the inequality (4.19). □

LEMMA 4.4. *Assume that $\xi_- \le \varphi \le \xi_+$ holds on Q_T. If $T > 2T_{\xi_-,\xi_+}$, then there exist numbers $C(\xi_-,\xi_+) > 0$ and $c(\xi_-,\xi_+) > 0$, which depend only on ξ_-,ξ_+, such that for $2T_{\xi_-,\xi_+} \le t \le T$,*

$$\begin{aligned}\int_0^t \mathcal{E}(\tau)\mathrm{d}\tau \le{}& C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \left(\|u\|_{H^2(0,t)}^2 + \|y\|_{H^2(0,t)}^2\right) \\ &+ C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^t \left[\mathbf{h}(\Lambda\varphi,\tau) + \mathcal{E}^2(\tau) + \mathcal{E}^{\frac{3}{2}}(\tau)\right]\mathrm{d}\tau,\end{aligned} \tag{4.29}$$

where $\Lambda\varphi$ and $\mathbf{h}(\Lambda\varphi,\tau)$ are given by the formulas (4.7) *and* (3.5)*, respectively.*

Proof. Let (φ, ψ) be a solution of the problem (1.11)–(1.12). Take $\omega = \left[\begin{smallmatrix} \varphi \\ \psi' \end{smallmatrix}\right]$. The formula (4.6) yields

$$A(\varphi)\omega_t + \omega' = \Psi(\varphi),$$

where

$$\Psi(\varphi) = \begin{bmatrix} \Lambda\varphi - \kappa b_2[u_2(t) + y_2(t)]q_1' \\ 0 \end{bmatrix}. \tag{4.30}$$

Set $\Upsilon = S\omega$. Then the estimate (4.14) is trivially true for $\Psi(\varphi)$ with $\Psi_1(\varphi) = 0$.

From the boundary condition $\varphi(L,t) = 0$ in (1.11) and the formula for S in (4.5), we have

$$\Upsilon_1(L,t) + \Upsilon_2(L,t) = \varphi(L,t) = 0, \quad t \geq 0. \tag{4.31}$$

Applying the inequalities (4.18) and (4.19) to $\Upsilon = (\Upsilon_1, \Upsilon_2)$ and using (4.31) gives

$$\begin{aligned} \int_0^t |\Upsilon(L,\nu)|^2 \mathrm{d}\nu &= 2\int_0^{T_{\xi_-,\xi_+}} \Upsilon_2^2(L,\nu)\mathrm{d}\nu + 2\int_{T_{\xi_-,\xi_+}}^t \Upsilon_1^2(L,\nu)\mathrm{d}\nu \\ &\leq C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^t [|\Upsilon|^2(0,\nu) + \Xi(\Upsilon)(\nu)]\mathrm{d}\nu. \end{aligned}$$

Combining the above inequality and (4.16), we get for $2T_{\xi_-,\xi_+} \leq t \leq T$,

$$\int_0^t \|\Upsilon(\cdot,\nu)\|^2 \mathrm{d}\nu \leq C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^t \left[|\Upsilon(0,\nu)|^2 + \Xi(\Upsilon)(\nu)\right] \mathrm{d}\nu. \tag{4.32}$$

On the other hand, from (3.13), (4.17), and (4.30), we obtain

$$\Xi(\Upsilon)(t) \leq C(\xi_-,\xi_+)\left[u_2^2(t) + y_2^2(t) + \|\Lambda\varphi(\cdot,t)\|^2 + \mathcal{E}^2(t) + \mathcal{E}^{\frac{3}{2}}(t)\right] \tag{4.33}$$

for all $t \in [0,T]$. In addition, it is easy to check that

$$C_1(\xi_-,\xi_+)[\varphi^2(x,t) + \psi'^2(x,t)] \leq |\Upsilon(x,t)|^2 \leq C_2(\xi_-,\xi_+)[\varphi^2(x,t) + \psi'^2(x,t)]$$

for all $(x,t) \in Q_T$. Using the inequalities (4.33), (4.32) and the boundary values in (3.15), where $v = \varphi$, $w = \psi$, we then obtain

(4.34)

$$\begin{aligned} &\int_0^t \left[\|\varphi(\cdot,\nu)\|^2 + \|\psi'(\cdot,\nu)\|^2\right] \mathrm{d}\nu \\ &\quad \leq C(\xi_-,\xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}} \int_0^t \left[|u(\nu)|^2 + |y(\nu)|^2 + \|\Lambda\varphi(\cdot,\nu)\|^2 + \mathcal{E}^2(\nu) + \mathcal{E}^{\frac{3}{2}}(\nu)\right] \mathrm{d}\nu. \end{aligned}$$

Next, we take $\omega = \left[\begin{smallmatrix} \varphi' \\ \psi'' \end{smallmatrix}\right]$. We differentiate (4.6) with respect to x, obtaining

$$A(\varphi)\omega_t + \omega' = \Psi(\varphi),$$

where, this time,

$$\Psi(\varphi) = \Psi_0(\varphi) - [A(\varphi)]'\vartheta_t,$$

$$\Psi_0(\varphi) = \begin{bmatrix} (\Lambda\varphi)' - \kappa b_2 \left[u_2(t) + y_2(t)\right] q_1'' \\ 0 \end{bmatrix}.$$

In order to apply Lemma 4.3 to $\Upsilon = S(\varphi', \psi'')$ this time, we have to find $\Psi_1(\varphi)$ such that the estimate (4.14) is true, where $\Psi(\varphi)$ is as above. In fact, using the identity (4.8) and $|\Lambda\varphi| \le C(\xi_-, \xi_+)|\nabla\varphi|_\infty$, we obtain (suppressing the variable t)

$$\begin{aligned}|\Psi(\varphi)| &\le C(\xi_-, \xi_+)\left[|\Psi_0(\varphi)| + |\nabla\varphi|_\infty(|\varphi_t| + |\psi_t'|)\right]\\ &\le C(\xi_-, \xi_+)\left[|\Psi_0(\varphi)| + |\nabla\varphi|_\infty|\Psi_1(\varphi)| + |\nabla\varphi|_\infty^2\right],\end{aligned}$$

where $\Psi_1(\varphi) = [u_2(t) + y_2(t)]q_1'$. Then a computation similar to that for (4.34) but this time with $\Upsilon = S(\varphi', \psi'')$ yields

$$\int_0^t \left[\|\varphi'(\cdot,\nu)\|^2 + \|\psi''(\cdot,\nu)\|^2\right]\mathrm{d}\nu$$

$$\le C(\xi_-, \xi_+)e^{c(\xi_-,\xi_+)\check{\mathcal{E}}_T^{\frac{1}{2}}}\int_0^t \left[\|(\Lambda\varphi)'(\cdot,\nu)\|^2 + |u|_1^2(\nu) + |y|_1^2(\nu) + \check{\mathcal{E}}^2(\nu) + \check{\mathcal{E}}^{\frac{3}{2}}(\nu)\right]\mathrm{d}\nu.$$

Similarly, when we apply Lemma 4.3 to $\Upsilon = S[\begin{smallmatrix}\varphi''\\ \psi'''\end{smallmatrix}]$, $\Upsilon = S[\begin{smallmatrix}\varphi_t\\ \psi_t'\end{smallmatrix}]$, and $\Upsilon = [\begin{smallmatrix}\varphi_{tt}\\ \psi_{tt}'\end{smallmatrix}]$, respectively, we get the corresponding estimates for $\int_0^t(\|\varphi''\|^2 + \|\varphi_t'\|^2)\mathrm{d}\nu$, $\int_0^t(\|\varphi_t\|^2 + \|\psi_t'\|^2)\mathrm{d}\nu$, and $\int_0^t(\|\varphi_{tt}\|^2 + \|\psi_{tt}'\|^2)\mathrm{d}\nu$, respectively. After we add all those inequalities together, we obtain (4.29). We omit the details here. □

Remark 4.3. To prove Theorem 4.1, our next step is to absorb the lower order term $\int_0^t \mathbf{h}(\Lambda\varphi, \tau)\mathrm{d}\tau$ in the inequality (4.29). Since the operator Λ: $L^2(Q_T) \to L^2(Q_T)$ (given by (4.7)) is compact, $\int_0^t \mathbf{h}(\Lambda\varphi, \tau)\mathrm{d}\tau$ is a lower order term with respect to the energy level $\int_0^t \mathcal{E}\mathrm{d}\tau$. This is shown by a compactness/uniqueness argument in the following lemma.

LEMMA 4.5. *Let $\xi_- \le \varphi \le \xi_+$ hold on $Q_{3T_{\xi_-,\xi_+}}$ and let $\varsigma > 0$ be given. Let T_{ξ_-,ξ_+} and $\check{\mathcal{E}}$ be defined by* (4.1) *and* (4.2)*, respectively. Suppose that (φ_0, ψ_0) and $u \in [H^2(0, 3T_{\xi_-,\xi_+})]^2$ are such that the solution of the system* (1.11)–(1.12) *satisfies*

$$\check{\mathcal{E}}_{3T_{\xi_-,\xi_+}} \le \varsigma.$$

Then there exists $C(\xi_-, \xi_+, \varsigma) > 0$, which may depend on ξ_-, ξ_+, and ς but is independent of the initial data (φ_0, ψ_0), such that

$$\begin{aligned}\int_0^{3T_{\xi_-,\xi_+}} \mathbf{h}(\Lambda\varphi, \nu)\mathrm{d}\nu \le\ & C(\xi_-, \xi_+, \varsigma)\left(\|u\|^2_{H^2(0,3T_{\xi_-,\xi_+})} + \|y\|^2_{H^2(0,3T_{\xi_-,\xi_+})}\right)\\ &+ C(\xi_-, \xi_+, \varsigma)\int_0^{3T_{\xi_-,\xi_+}} [\mathcal{E}^2(\nu) + \mathcal{E}^{\frac{3}{2}}(\nu)]\mathrm{d}\nu,\end{aligned}$$

where $\mathbf{h}(\Lambda\varphi, \nu)$ is given by the formula (3.5).

Proof. By contradiction, assume that the conclusion of the lemma does not hold. Then, there exist initial data $(\varphi_{0,n}, \psi_{0,n})$, an input u_n, and the corresponding solutions (φ_n, ψ_n) of (1.11)–(1.12) over $(0, 3T_{\xi_-,\xi_+})$ such that, denoting $\tau = 3T_{\xi_-,\xi_+}$,

$$\xi_- \le \varphi_n \le \xi_+, \quad \check{\mathcal{E}}_{n,\tau} \le \varsigma,$$

$$\frac{1}{n}\int_0^\tau \mathbf{h}(\Lambda\varphi_n, \nu)\mathrm{d}\nu \ge \|u_n\|^2_{H^2(0,\tau)} + \|y_n\|^2_{H^2(0,\tau)} + \int_0^\tau \left[\mathcal{E}_n^2(\nu) + \mathcal{E}_n^{\frac{3}{2}}(\nu)\right]\mathrm{d}\nu. \tag{4.35}$$

Since $\mathbf{h}(\Lambda\varphi_n,\nu)\le C(\xi_-,\xi_+)\check{\mathcal{E}}_{n,\tau}$,

$$\|u_n\|^2_{H^2(0,\tau)}+\|y_n\|^2_{H^2(0,\tau)}+\int_0^\tau\left[\mathcal{E}_n^2(\nu)+\mathcal{E}_n^{\frac{3}{2}}(\nu)\right]\mathrm{d}\,\nu\to 0,$$

which further implies

$$\|(\varphi_n,\psi_n)\|^2_{H^2(Q_\tau)}+\int_0^\tau[{\varphi_n'}^2(0,\nu)+\|\psi_{tt}'(\cdot,\nu)\|^2]\mathrm{d}\,\nu=\int_0^\tau\mathcal{E}_n(\nu)\mathrm{d}\,\nu\to 0, \tag{4.36}$$

$$\sup_{(x,t)\in Q_\tau}|\varphi_n(x,t)|\to 0. \tag{4.37}$$

Set

$$c_n^2=\int_0^\tau\mathbf{h}(\Lambda\varphi_n,\nu)\mathrm{d}\,\nu,\quad \rho_n=\frac{\varphi_n}{c_n},\quad \varrho_n=\frac{\psi_n}{c_n},\quad \mu_n=\frac{u_n}{c_n},\quad \nu_n=\frac{y_n}{c_n},$$

$$\mathcal{E}(\rho_n,\varrho_n)(t)={\rho'}_n^2(0,t)+\|\varrho_{ntt}'(\cdot,t)\|^2+\sum_{|\alpha|\le 2}\left[\|\partial^\alpha\rho_n(\cdot,t)\|^2+\|\partial^\alpha\varrho_n(\cdot,t)\|^2\right].$$

Then

$$\|\Lambda\rho_n\|^2_{H^2(Q_\tau)}+\int_0^\tau(\Lambda\rho_n')^2(0,t)\mathrm{d}\,t=\int_0^\tau\mathbf{h}(\Lambda\rho_n,t)\mathrm{d}\,t=1, \tag{4.38}$$

and by (4.35)

$$\|\mu_n\|^2_{H^2(0,\tau)}+\|\nu_n\|^2_{H^2(0,\tau)}+\frac{1}{c_n^2}\int_0^\tau\left[\mathcal{E}_n^2(t)+\mathcal{E}_n^{\frac{3}{2}}(t)\right]\mathrm{d}\,t\to 0. \tag{4.39}$$

We divide both sides of (4.29) by c_n^2 and use the relation (4.39) to see that the expressions below form a bounded sequence (indexed by n):

$$\|(\rho_n,\varrho_n)\|^2_{H^2(Q_\tau)}+\int_0^\tau\left[{\rho'}_n^2(0,t)+\|\varrho_{ntt}'(\cdot,t)\|^2\right]\mathrm{d}\,t=\int_0^\tau\mathcal{E}(\rho_n,\varrho_n)(t)\mathrm{d}t.$$

Since the operator Λ is compact, there are $\rho\in H^2(Q_\tau)$ and $\varrho\in H^{1,2}(Q_\tau)\cap H^2(Q_\tau)$ such that (by taking a subsequence)

$$\rho_n\to\rho\quad\text{weakly in } H^2(Q_\tau), \tag{4.40}$$

$$\Lambda\rho_n\to\Lambda\rho\quad\text{strongly in } H^2(Q_\tau), \tag{4.41}$$

$$\varrho_n\to\varrho\quad\text{weakly in } H^{1,2}(Q_\tau). \tag{4.42}$$

Moreover, it is easy to check that (ρ_n,ϱ_n) and φ_n satisfy for $(x,t)\in Q_\tau$

$$\begin{cases}\rho_{nt}=\mathcal{N}(\varphi_n)\varrho_n'',\\ \varrho_{nt}=\kappa\varrho_{nt}''-\rho_n'',\\ \rho_n(L,t)=\varrho_n(L,t)=\varrho_n''(L,t)=0,\\ \rho_n(0,t)=b_1[\mu_{n1}(t)+\nu_{1n}(t)],\\ \varrho_n'(0,t)=b_1^{-1}[\nu_{1n}(t)-\mu_{1n}(t)],\\ \rho_n'(0,t)-\kappa\varrho_{nt}'(0,t)=b_2[\mu_{2n}(t)+\nu_{2n}(t)],\\ \varrho_n(0,t)=b_2^{-1}[\mu_{n2}(t)-\nu_{2n}(t)].\end{cases} \tag{4.43}$$

Therefore, the relations (4.36), (4.37), (4.39), and (4.40)–(4.43) imply that

$$(\rho, \varrho) \in H^2(Q_\tau) \times [H^{1,2}(Q_\tau) \cap H^2(Q_\tau)]$$

solves the following linear boundary-value problem for $(x,t) \in Q_\tau$:

$$\begin{cases} \rho_t = \mathcal{N}(0)\varrho'', \\ \varrho_t = \kappa\varrho_t'' - \rho'', \\ \rho(L,t) = \varrho(L,t) = \varrho''(L,t) = 0, \\ \rho(0,t) = \varrho(0,t) = 0, \\ \rho'(0,t) = \varrho'(0,t) = 0. \end{cases} \tag{4.44}$$

However, since $\tau > 2T_{\xi_-,\xi_+}$ it is easy to check that the linear problem (4.44) has only the zero solution, i.e., $(\rho, \varrho) = 0$ on Q_τ. This is a contradiction, since (4.38) and (4.41) imply $\int_0^\tau \mathbf{h}(\Lambda\rho, t)\mathrm{d}t = 1$. □

Remark 4.4. To guarantee that the problem (4.44) has only the zero solution, it is enough to assume that (4.44) holds on Q_{T_0}, where $T_0 > 2T_{\xi_-,\xi_+}$.

Proof of Theorem 4.1. It suffices to prove the inequality (4.3) for $s = 0$. We use a trick from Feng and Feng [5]. From Lemma 4.4 we see that (4.3) is true for $s = 0$ if we can find $C(\xi_-, \xi_+, \varsigma) > 0$ such that for $3T_{\xi_-,\xi_+} \le t \le T$

$$\int_0^t \mathbf{h}(\Lambda\varphi, \nu)\mathrm{d}\nu \le C(\xi_-, \xi_+, \varsigma)\left[\|u\|^2_{H^2(0,t)} + \|y\|^2_{H^2(0,t)} + \int_0^t \left[\mathcal{E}^2(\nu) + \mathcal{E}^{\frac{3}{2}}(\nu)\right]\mathrm{d}\nu\right]. \tag{4.45}$$

In fact, for any $h > 0$, from Lemma 4.5 we have, denoting again $\tau = 3T_{\xi_-,\xi_+}$,

$$\begin{aligned} \int_h^{\tau+h} \mathbf{h}(\Lambda\varphi, t)\mathrm{d}t &= \int_0^\tau \mathbf{h}(\Lambda\varphi_h, t)\mathrm{d}t \\ &\le C(\xi_-, \xi_+, \varsigma)\left[\|u_h\|^2_{H^2(0,\tau)} + \|y_h\|^2_{H^2(0,\tau)} + \int_0^\tau \left[\mathcal{E}_h^2(t) + \mathcal{E}_h^{\frac{3}{2}}(t)\right]\mathrm{d}t\right] \\ &= C(\xi_-, \xi_+, \varsigma)\left[\|u\|^2_{H^2(h,\tau+h)} + \|y\|^2_{H^2(h,\tau+h)}\right] \\ &\quad + C(\xi_-, \xi_+, \varsigma)\int_h^{\tau+h} \left[\mathcal{E}^2(t) + \mathcal{E}^{\frac{3}{2}}(t)\right]\mathrm{d}t, \end{aligned} \tag{4.46}$$

where $\varphi_h(t) = \varphi(t+h)$, $u_h(t) = u(t+h)$, $y_h(t) = y(t+h)$, and $\mathcal{E}_h$ is defined by (1.18) with (φ_h, ψ_h) in place of (φ, ψ). Denote the right-hand side of (4.46) by $I(h, \tau+h)$. Then for $m\tau \le t \le (m+1)\tau$ with $m \ge 2$ an integer, we obtain

$$\begin{aligned} \int_0^t \mathbf{h}(\Lambda\varphi, \nu)\mathrm{d}\nu &\le \int_0^\tau \mathbf{h}(\Lambda\varphi, \nu)\mathrm{d}\nu + \cdots + \int_{(m-1)\tau}^{m\tau} \mathbf{h}(\Lambda\varphi, \nu)\mathrm{d}\nu + \int_{t-\tau}^t \mathbf{h}(\Lambda\varphi, \nu)\mathrm{d}\nu \\ &\le I(0,\tau) + \cdots + I((m-1)\tau,\, m\tau) + I(t-\tau,\ t) \le 2I(0,t), \end{aligned}$$

which is just the inequality (4.45). □

From the definitions (1.18) and (3.9), it is easy to check that if the solution (φ, ψ) of the system (1.11)–(1.12) satisfies $\xi_- \le \varphi \le \xi_+$, then there are $C_1(\xi_-, \xi_+) > 0$ and $C(\xi_-, \xi_+) > 0$ such that

$$C_1(\xi_-, \xi_+)\mathcal{E}(t) \le \Pi(t) \le C(\xi_-, \xi_+)\mathcal{E}(t), \quad 0 \le t \le T. \tag{4.47}$$

For convenience, we rewrite Theorem 4.1 as follows.

THEOREM 4.1*. *Suppose that* $\sigma_- < \xi_- < 0 < \xi_+ < \sigma_+$, $\varsigma > 0$, *and* $T > 3T_{\xi_-,\xi_+}$. *Then there exists a number* $C(\xi_-, \xi_+, \varsigma) > 0$, *which depends on* ξ_-, ξ_+ *and* ς, *such that if the solution* (φ, ψ) *of the system* (1.11)–(1.12) *satisfies the assumptions*

$$\xi_- \le \varphi \le \xi_+, \quad \sup_{0\le t\le T} \Pi(t) \le \varsigma,$$

then for $0 \le s \le t \le T$ *with* $t - s \ge 3T_{\xi_-,\xi_+}$,

$$\int_s^t \Pi(\nu)\mathrm{d}\nu \le C(\xi_-, \xi_+, \varsigma) \left[\|u\|^2_{H^2(s,t)} + \|y\|^2_{H^2(s,t)} + \int_s^t \left[\Pi^2(\nu) + \Pi^{\frac{3}{2}}(\nu)\right] \mathrm{d}\nu \right].$$

5. Global smooth solutions. In this section we establish the existence of the global smooth solutions and the exponential decay of the Sobolev norm for small initial data.

Let $u(t) = (u_1, u_2) \in H^2(0,\infty) \times H^2(0,\infty)$. Suppose that $(\varphi_0, \psi_0) \in H^2(0, L) \times H^3(0, L)$ satisfies the boundary conditions (1.15) and the compatibility conditions (1.16). We assume that (φ, ψ) solves the problem (1.11)–(1.12) on $Q_T = (0, L)\times(0, T)$ for some $T > 0$ such that the regularity (1.17) holds.

Throughout this section we fix ξ such that

$$\sigma_- < -\xi < 0 < \xi < \sigma_+.$$

We introduce some constants that we need later: Let $C(-\xi, \xi) > 0$ be given by Theorem 3.1*, (4.47), and (3.29) for $\xi_- = -\xi$ and $\xi_+ = \xi$. Set

$$\varsigma = \min\left\{1, \frac{1}{C(-\xi,\xi)}, \frac{\xi^2}{2LC(-\xi,\xi)}\right\}.$$

Let $C(-\xi, \xi, \varsigma)$ be given by Theorem 4.1* for $\xi_- = \xi$, $\xi_+ = \xi$, and the above ς. We then fix $\gamma > 0$ small such that

$$\varpi(\gamma) = C^{-1}(-\xi, \xi, \varsigma) - 2[2\gamma + \gamma^{\frac{1}{2}}] > 0, \tag{5.1}$$

$$c(\gamma) = \varpi - 2C(-\xi, \xi)[2\gamma + \gamma^{\frac{1}{2}}] > 0, \tag{5.2}$$

$$T_0(\gamma) = \frac{1}{2C(-\xi,\xi)\gamma^{\frac{1}{2}}} > 3T_{-\xi,\xi}, \tag{5.3}$$

where $T_{-\xi,\xi}$ is given by (4.11) for $\xi_- = \xi$ and $\xi_+ = \xi$.

Proof of Theorem 1.2. We fix

$$0 < \gamma < \varsigma/4 \tag{5.4}$$

such that the inequalities (5.1)–(5.3) hold.

We assume that the initial data (φ_0, ψ_0) and the input u are such that

$$\mathcal{E}(0) + \|u\|^2_{H^2(0,\infty)} < \min\left\{\frac{\xi^2}{2L}, \frac{\gamma}{C(-\xi,\xi)+1}\right\}, \tag{5.5}$$

where $\mathcal{E}(t)$ is given by (1.18).

Step I. We prove that, under the small initial condition (5.5), the problem (1.11)–(1.12) has global solutions. The condition (5.5) yields

$$\sup_{x\in[0,L]}|\varphi_0(x)| \le L^{\frac{1}{2}}\mathcal{E}^{\frac{1}{2}}(0) < \frac{\xi}{\sqrt{2}}, \tag{5.6}$$

which, by Theorem 1.1, implies that the short-time solution (φ,ψ) of the problem (1.11)–(1.12) exists. For $0\le s\le t$, set

$$h(s,t) = \Pi(t) + \|y\|^2_{H^2(s,t)},$$

$$g(s,t) = \Pi(s) + C(-\xi,\xi)\|u\|^2_{H^2(s,t)}. \tag{5.7}$$

By (4.47), (5.5), and (5.4), we have

$$g(0,\infty) \le C(-\xi,\xi)[\mathcal{E}(0) + \|u\|^2_{H^2(0,\infty)}] \le \gamma < \varsigma/4. \tag{5.8}$$

In particular, $\Pi(0)\le\varsigma/4$. This inequality together with (5.6) implies that the local solution must satisfy

$$-\xi\le\varphi\le\xi, \quad x\in[0,L], \quad \Pi(t)\le\varsigma, \tag{5.9}$$

for t in some interval $[0,\delta]$.

Let δ_0 be the largest number such that the estimates in (5.9) hold for $t\in[0,\delta_0)$. We shall prove that $\delta_0=\infty$ by contradiction. Then the problem (1.11)–(1.12) has global solutions under the small initial data condition (5.5).

Let $\delta_0<\infty$. Then (3.36) implies that for $t\in[0,\ \min\{\delta_0,T_0(\gamma)\}]$

$$\begin{aligned} h(0,t) &\le g(0,t) + C(-\xi,\xi)\int_0^t \Pi^{\frac{3}{2}}(\Pi^{\frac{1}{2}}+1)\mathrm{d}t \\ &\le g\left(0,\ \min\{\delta_0,T_0(\gamma)\}\right) + 2C(-\xi,\xi)\int_0^t h^{\frac{3}{2}}(0,t)\mathrm{d}t, \end{aligned} \tag{5.10}$$

since $g(0,\cdot)$ is increasing. By (5.8), we have

$$\min\{\delta_0,T_0(\gamma)\} \le T_0(\gamma) \le \frac{1}{2C(-\xi,\xi)g^{\frac{1}{2}}\left(0,\ \min\{\delta_0,T_0(\gamma)\}\right)}. \tag{5.11}$$

We obtain from (5.8), (5.10), and (5.11) that

$$\begin{aligned} \Pi(t) + \|y\|^2_{H^2(0,t)} &\le g\left(0,\ \min\{\delta_0,T_0(\gamma)\}\right)/[1 - tC(-\xi,\xi)g^{\frac{1}{2}}\left(0,\ \min\{\delta_0,T_0(\gamma)\}\right)]^2 \\ &\le 4g\left(0,\ \min\{\delta_0,T_0(\gamma)\}\right) \le 4\gamma < \varsigma \end{aligned} \tag{5.12}$$

for $0\le t\le\min\{\delta_0,T_0(\gamma)\}$.

We assume that $T_0(\gamma)\ge\delta_0$. Then the inequality (5.12) shows that $\Pi(t)\le 4\gamma<\varsigma$ holds for $0\le t\le\delta_0$, which gives, via (3.29),

$$\begin{aligned} \sup_{0\le x\le L}|\varphi(x,t)| &\le L^{\frac{1}{2}}\mathbf{h}^{\frac{1}{2}}(\varphi,t) \le [LC(-\xi,\xi)\Pi(t)]^{\frac{1}{2}} \\ &\le 2[LC(-\xi,\xi)]^{\frac{1}{2}}\gamma^{\frac{1}{2}} \le \frac{\xi}{\sqrt{2}}, \quad 0\le t\le\delta_0. \end{aligned}$$

Then the conditions (5.9) are true for $t \in [0, \delta_1)$ with $\delta_1 > \delta_0$. This is a contradiction. Thus, we obtain that $T_0(\gamma) < \delta_0$.

Now the inequality (5.12) implies

$$\Pi(t) \le 4\gamma, \quad 0 \le t \le T_0(\gamma). \tag{5.13}$$

We apply Theorem 4.1* to give for $T_0(\gamma) \ge t \ge 3T_{-\xi,\xi}$

$$\varpi \int_0^t \Pi(t)dt \le \|u\|^2_{H^2(0,t)} + \|y\|^2_{H^2(0,t)}. \tag{5.14}$$

We then add the inequalities (3.36) with $\xi_- = -\xi$, $\xi_+ = \xi$, and (5.14) to deduce from (5.13) that, for $3T_{-\xi,\xi} \le t \le T_0(\gamma)$,

$$\begin{aligned} &\Pi(t) + \varpi \int_0^t \Pi(\tau)\mathrm{d}\tau + \|y\|^2_{H^2(0,t)} \\ &\quad \le \Pi(0) + [C(-\xi,\xi)+1]\|u\|^2_{H^2(0,t)} + \|y\|^2_{H^2(0,t)} \\ &\quad\quad + 2C(-\xi,\xi)[2\gamma + \gamma^{\frac{1}{2}}] \int_0^t \Pi(\tau)\mathrm{d}\tau, \end{aligned} \tag{5.15}$$

that is,

$$\Pi(t) + c(\gamma) \int_0^t \Pi(\tau)d\tau \le \Pi(0) + [C(-\xi,\xi)+1]\,\|u\|^2_{H^2(0,t)}, \tag{5.16}$$

for $3T_{-\xi,\xi} \le t \le T_0(\gamma)$. From (5.7), (5.16), (4.47), and (5.5) we have

$$\begin{aligned} g\,(T_0(\gamma),\,\infty) &\le \Pi(0) + [C(-\xi,\xi)+1]\,\|u\|^2_{H^2(0,T_0(\gamma))} + C(-\xi,\xi)\|u\|^2_{H^2(T_0(\gamma),\infty)} \\ &\le \Pi(0) + [C(-\xi,\xi)+1]\,\|u\|^2_{H^2(0,\infty)} \\ &\le C(-\xi,\xi)\mathcal{E}(0) + [C(-\xi,\xi)+1]\,\|u\|^2_{H^2(0,\infty)} \le \gamma. \end{aligned}$$

Next, we restart the above procedure from $g\,(T_0(\gamma),\infty) < \gamma$ and using $T_0(\gamma) \le t \le 2T_0(\gamma)$ instead of $0 \le t \le T_0(\gamma)$ and we arrive at

$$2T_0(\gamma) < \delta_0.$$

If we continue, by induction we obtain that for any $k \in \mathbb{N}$,

$$kT_0(\gamma) < \delta_0.$$

This is also a contradiction.

Step II. By Step I, under the condition (5.5), the estimates (5.9) hold for all $t \in [0,\infty)$. In the interval $[0,\infty)$ we apply Theorems 3.1* and 4.1* and, by the extension from $[iT_0(\gamma),\ (i+1)T_0(\gamma)]$ to $[(i+1)T_0(\gamma),\ (i+2)T_0(\gamma)]$ for $i \ge 0$, obtain that the inequality (5.12) holds for all $t \in [0,\infty)$. Then, an argument similar to that for (5.16) yields that for $t \in [3T_{-\xi,\xi},\infty)$,

$$\Pi(t) + c(\gamma) \int_0^t \Pi(\tau)\mathrm{d}\tau \le \Pi(0) + [C(-\xi,\xi)+1]\,\|u\|^2_{H^2(0,t)}.$$

By (4.47), the proof is complete. ∎

Proof of Theorem 1.3. Let the input u satisfy

$$u(t) = 0 \quad \forall\, t \geq T_0.$$

We repeat the argument in the proof of Theorem 1.2 for $s \geq t \geq T_0$ to obtain that for $s \geq t \geq T_0$,

$$\Pi(s) + c(\gamma) \int_t^s \Pi(\tau) d\tau \leq \Pi(t) + [C(-\xi, \xi) + 1] \|u\|^2_{H^2(t,\infty)} = \Pi(t),$$

which yields

$$\Pi(t) \leq \Pi(T_0) e^{-c(\gamma)(t-T_0)} \qquad \forall\, t \geq T_0.$$

The decay rate estimate (1.24) is true because of the inequality (4.47). ∎

REFERENCES

[1] H. D. Alber and J. Cooper, *Quasilinear hyperbolic* 2×2 *systems with a free, damping boundary condition*, J. Reine Angew. Math., 406 (1990), pp. 10–43.

[2] K. Ammari, Z. Liu, and M. Tucsnak, *Decay rates for a beam with pointwise force and moment feedback*, Math. Control Signals Systems, 15 (2002), pp. 229–255.

[3] M. Cirina, *Nonlinear hyperbolic problems with solutions on preassigned sets*, Michigan Math. J., 17 (1970), pp. 193–209.

[4] C. M. Dafermos and W. J. Hrusa, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Rational Mech. Anal., 87 (1985), pp. 267–292.

[5] S. J. Feng and D. X. Feng, *Nonlinear internal damping of wave equations with variable coefficients*, Acta Math. Sin. (Engl. Ser.), 20 (2004), pp. 1057–1072.

[6] J. M. Greenberg and T. T. Li, *The effect of boundary damping for the quasilinear wave equation*, J. Differential Equations, 52 (1984), pp. 66–75.

[7] B.-Z. Guo, J.-M. Wang, and S.-P. Yung, *Boundary stabilization of a flexible manipulator with rotational inertia*, Differential Integral Equations, 18 (2005), pp. 1013–1038.

[8] T. Kato, *Linear evolution equations of "hyperbolic" type*, J. Fac. Sci. Univ. Tokyo Sect. I, 17 (1970), pp. 241–258.

[9] T. Kato, *Linear evolution equations of "hyperbolic" type*. II, J. Math. Soc. Japan, 25 (1973), pp. 648–666.

[10] S. Klainerman and A. Majda, *Formation of singularities for wave equations including the nonlinear vibrating string*, Comm. Pure Appl. Math., 33 (1980), pp. 241–263.

[11] J. Malinen, O. Staffans, and G. Weiss, *When is a linear system conservative?*, Quart. Appl. Math., 64 (2006), pp. 61–91.

[12] L. A. Segel and G. H. Handelman, *Mathematics Applied to Continuum Mechanics*, Macmillan, New York, 1977.

[13] M. Slemrod, *Boundary feedback stabilization for a quasilinear wave equation*, in Control Theory for Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 54, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer-Verlag, Berlin, 1983, pp. 211–237.

[14] O. Staffans, *Passive and conservative continuous-time impedance and scattering systems. Part* I: *Well-posed systems*, Math. Control Signals Systems, 15 (2002), pp. 291–315.

[15] M. Tucsnak and G. Weiss, *How to get a conservative well-posed linear system out of thin air. Part* II. *Controllability and stability*, SIAM J. Control Optim., 42 (2003), pp. 907–935.

[16] G. Weiss and M. Tucsnak, *How to get a conservative well-posed linear system out of thin air.* I. *Well-posedness and energy balance*, ESAIM-Control Optim. Calc. Var., 9 (2003), pp. 247–274.

[17] G. Weiss and R. Curtain, *Exponential stabilization of a Rayleigh beam using colocated control*, IEEE Trans. Automat. Control, to appear.

[18] P.-F. Yao, *An energy-preserving nonlinear system modeling a string with input and output on the boundary*, Nonlinear Anal., to appear.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 1965–1987

FROM DIFFERENTIAL CALCULUS TO 0-1 TOPOLOGICAL OPTIMIZATION*

MAATOUG HASSINE†, SOPHIE JAN‡, AND MOHAMED MASMOUDI‡

Abstract. The topological asymptotic expansion gives the variation of a cost function when a small hole is created in a domain. This approach leads to very powerful algorithms in topological optimization. Unfortunately, these asymptotic expansions are obtained with the use of complicated mathematical tools. The goal of this paper is to provide a straightforward way to derive a topological asymptotic expansion using a classical gradient. We will illustrate this general approach by some numerical experiments for the elasticity and the Stokes problems.

Key words. topological gradient, classical gradient, shape optimization, topological optimization, differential calculus

AMS subject classifications. 49Q10, 49Q12, 74P05, 74P10, 74P15

DOI. 10.1137/050631720

1. Introduction.

The topological sensitivity analysis consists of studying the variation of a cost function with respect to a modification of the topology of a domain, such as the inclusion of holes or insertion of obstacles. It also provides an asymptotic expansion of a cost function with respect to a finite perturbation of the physical properties in a small region of the domain.

In order to present the basic idea, let us consider a domain Ω of $\mathbb{R}^d$, $d = 1, 2, 3$, and a cost function $j(\Omega) = J(\Omega, u_\Omega)$ which is to be minimized. We denote by u_Ω the solution to a given PDE (partial differential equation) problem for $d = 2, 3$, or to a differential equation for $d = 1$, defined in Ω. For $\varepsilon > 0$, let $\Omega_\varepsilon = \Omega \backslash \overline{(x_0 + \varepsilon\omega)}$ be the domain obtained by removing a small part $\overline{(x_0 + \varepsilon\omega)}$ from Ω, where $x_0 \in \Omega$ and $\omega \subset \mathbb{R}^d$ is a fixed bounded domain containing the origin. Then, an asymptotic expansion of the function j is obtained in the following form:

$$j(\Omega_\varepsilon) = j(\Omega) + \rho(\varepsilon) g(x_0) + o(\rho(\varepsilon)),$$

where $\rho(\varepsilon)$ is a scalar positive function going to zero with ε. This expression is called the topological asymptotic expansion and g the topological gradient. The function g is very easy to compute. In order to minimize the cost function, the best location to insert a small hole in Ω is where g is negative. In fact if $g(x_0) < 0$, we have $j(\Omega_\varepsilon) < j(\Omega)$ for small ε. The function g can be used as a descent direction of the domain optimization process.

The optimality condition $g(x_0) \geq 0$ in the domain coincides with the one obtained by Buttazzo and Dal Maso [12] for the Laplace equation using the homogenization theory. The notion of topological asymptotic expansion allows us to build fast numerical algorithms that can solve a large class of problems in applications.

In structural mechanics the total energy variation with respect to the creation of a small hole is well known [25]. Schumacher [35] introduced the bubble method that

*Received by the editors May 17, 2005; accepted for publication (in revised form) July 3, 2006; published electronically January 8, 2007.

http://www.siam.org/journals/sicon/45-6/63172.html

†Campus Universitaire, ENIT-LAMSIN and FSM, Le Belvédaire B.P. 37, 1002 Tunis, Tunisia (maatoug.hassine@enit.rnu.tn).

‡Mathématiques pour l'Industrie et la Physique, UMR 5640, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 4, France (jan@mip.ups-tlse.fr, masmoudi@mip.ups-tlse.fr).

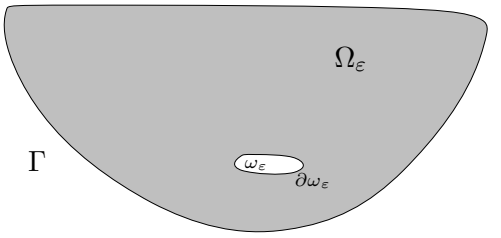


FIG. 1. *The perturbed domain* Ω_ε.

uses this energy variation for topological optimization. Then Sokołowski [36] extended this idea to more general cost functions using the adjoint approach but still with a Neumann boundary condition. A topological sensitivity framework allowing us to consider a Dirichlet condition on the boundary of the hole has been introduced in [29]. The key ideas in this work are to use a generalized Lagrangian method and a domain truncation technique to obtain a fixed functional space. The fundamental property of an adjoint method [13, 29] is to provide the variation of a function with respect to a parameter using a solution u_Ω and an adjoint state v_Ω which do not depend on the chosen parameters. From the numerical point of view, only two problems have to be solved to obtain $g(x)$ for all $x \in \Omega$. This observation leads to very efficient numerical algorithms. The work in [29] has been generalized to the elasticity equations in the case of arbitrarily shaped holes and a Neumann boundary condition [20].

Recently, similar results have been obtained using a Dirichlet boundary condition and noncircular holes in the Poisson equation [23], Stokes equations [24], quasi-Stokes equations [26], Helmholtz equations [8, 33], and Navier–Stokes equations [7]. One can remark that in all these papers the topological gradient has the same expression. Such a result can be interpreted by the fact that all these operators have the same principal part. It is proved that when J is an integral over the boundary, we have $g(x) = u_\Omega(x).v_\Omega(x)$. Others terms may appear if J involves ∇u_Ω at x.

The goal of this paper is to provide an easy way to obtain this type of topological sensitivity in the case of Dirichlet boundary conditions. We will show that it is possible, under weak assumptions, to derive topological asymptotic expansion using differential calculus. Also, when these hypotheses do not hold, the classical gradient [22, 28, 32, 34] provides a "topological descent direction." More precisely, let us consider the following elliptic problem:

$$-\mathrm{div}(\alpha\nabla u) + c\,u = F \text{ in } \Omega,$$

where α and c are two functions defined in Ω.

Let ω_ε be a small domain included in Ω (see Figure 1). We suppose that ω_ε has the form $\omega_\varepsilon = x_0 + \varepsilon\omega$, where $x_0 \in \Omega$, $\varepsilon > 0$, and where ω is a given fixed open and bounded domain of $\mathbb{R}^d$, containing the origin, whose boundary $\partial\omega$ is connected and of class $\mathcal{C}^1$. Our aim is to derive topological sensitivity information from the classical gradient with respect to α or c. If α goes to zero in ω_ε, for a given ε, the corresponding solution u_α tends to the solution u_ε of the following problem:

$$-\mathrm{div}(\alpha\nabla u_\varepsilon) + c\,u_\varepsilon = F \text{ in } \Omega_\varepsilon,$$

with a Neumann boundary condition $\partial_n u_\varepsilon = 0$ on $\partial\omega_\varepsilon$.

If c goes to infinity in ω_ε, for a given ε, the corresponding solution u_c tends to the solution u_ε of the following problem:

$$-\mathrm{div}(\alpha\nabla u_\varepsilon) + c\,u_\varepsilon = F \text{ in } \Omega_\varepsilon,$$

with a Dirichlet boundary condition $u_\varepsilon = 0$ on $\partial\omega_\varepsilon$. This reminds us of the well-known penalization technique used in the finite element method for the implementation of a Dirichlet condition.

We will show in subsection 2.1 that the gradient with respect to c is equal to the topological gradient in the case of a Dirichlet condition. To this end, we will consider a topological perturbation δc_ε of the form

$$\delta c_\varepsilon(x) = \begin{cases} \kappa \text{ if } x \in \omega_\varepsilon, \\ 0 \text{ if } x \in \Omega\backslash\overline{\omega_\varepsilon}, \end{cases}$$

where κ is a constant and $\varepsilon > 0$ is small enough.

This perturbation is small in $L^p(\Omega)$ for $1 \le p < +\infty$, and the variation $f(c + \delta c_\varepsilon) - f(c)$ of a given differentiable functional f defined on $L^p(\Omega)$ can be estimated by a classical gradient. But this result is not valid for the gradient with respect to α. We will give some counterexamples in subsections 2.2 and 2.3, where the classical gradient is not equal to the topological gradient but still provides a "topological descent direction." Some numerical results are given for the elasticity problem in section 3 and for the Stokes problem in section 4.

2. From classical gradient to topological sensitivity. Let $1 \le p < +\infty$. We denote by f a differentiable functional on $L^p(\Omega)$,

$$\begin{array}{ccc} f : L^p(\Omega) & \to & \mathbb{R}, \\ c & \mapsto & f(c), \end{array} \tag{2.1}$$

and we denote by g the Riesz representation of its differential $f'(c)$.

Then for all $\delta c \in L^p(\Omega)$, we have

$$f'(c)\delta c = \int_\Omega g(x)\,\delta c(x)dx.$$

Now, we wish to study the variation of the functional f with respect to a finite topological perturbation δc_ε of c.

We make the following assumptions.

HYPOTHESIS 2.1.

- *The function f is differentiable, and there exists a constant $\gamma_1 > 0$ such that*

$$\forall c \in L^p(\Omega), \quad \Big|f(c+\delta c) - f(c) - f'(c)\delta c\Big| \le \gamma_1 \|\delta c\|^2_{L^p(\Omega)} \quad \forall \delta c \in L^p(\Omega). \tag{2.2}$$

- *The function g is Lipschitz continuous, that is, there exists a constant $\gamma_2 > 0$ such that*

$$|g(x) - g(y)| \le \gamma_2 \|x - y\| \quad \forall\, x, y \in \Omega. \tag{2.3}$$

The following theorem gives the variation of f.

THEOREM 2.2. *If f is differentiable on $L^p(\Omega)$ with $1 \le p < 2$ and if Hypothesis 2.1 holds, then we have*

$$f(c + \delta c_\varepsilon) - f(c) = \kappa\rho(\varepsilon)g(x_0) + o\Big(\rho(\varepsilon)\Big), \tag{2.4}$$

where $\rho(\varepsilon) = meas(\omega_\varepsilon)$. The function g is called the topological gradient.

Proof. First, we have

$$|f(c+\delta c_\varepsilon) - f(c) - \kappa\rho(\varepsilon)g(x_0)| \leq |f(c+\delta c_\varepsilon) - f(c) - f'(c)\delta c_\varepsilon| + |f'(c)\delta c_\varepsilon - \kappa\rho(\varepsilon)g(x_0)|. \tag{2.5}$$

Thanks to (2.2), we derive

$$|f(c+\delta c_\varepsilon) - f(c) - f'(c)\delta c_\varepsilon| \leq \gamma_1 \|\delta c_\varepsilon\|^2_{L^p(\Omega)}.$$

Using the fact that $p < 2$, we obtain

$$\|\delta c_\varepsilon\|^2_{L^p(\Omega)} = \kappa^2(\rho(\varepsilon)^{1/p})^2 = \kappa^2\rho(\varepsilon)^{2/p} = o\Big(\rho(\varepsilon)\Big).$$

Therefore

$$|f(c+\delta c_\varepsilon) - f(c) - f'(c)\delta c_\varepsilon| = o\Big(\rho(\varepsilon)\Big). \tag{2.6}$$

For the second part in (2.5), we have

$$\left|\int_\Omega g(x)\,\delta c_\varepsilon(x)dx - \kappa\rho(\varepsilon)g(x_0)\right| = \left|\int_\Omega \Big(g(x) - g(x_0)\Big)\,\delta c_\varepsilon(x)dx\right|.$$

It follows from (2.3) that

$$\begin{aligned}\left|\int_\Omega g(x)\,\delta c_\varepsilon(x)dx - \kappa\rho(\varepsilon)g(x_0)\right| &\leq \gamma_2\int_\Omega \|x-x_0\|\,|\delta c_\varepsilon(x)|\,dx = \gamma_2\,|\kappa|\int_{\omega_\varepsilon}\|x-x_0\|\,dx \\ &\leq \gamma_2\,|\kappa|\,\varepsilon\rho(\varepsilon).\end{aligned} \tag{2.7}$$

Finally, using (2.5), (2.6), and (2.7), we deduce the desired result. ∎

2.1. The Dirichlet condition. Consider a Hilbert space $\mathcal{V} \subset H^1(\Omega)^d$. Let a be a bilinear, symmetric, continuous, and coercive form on $\mathcal{V}$, and let l be a linear and continuous form on $\mathcal{V}$; that is, there exist constants $M > 0$, $\gamma > 0$, and $L > 0$ such that

$$\begin{aligned} a(u\,,\,v) &\leq M\,\|u\|\,\|v\| && \forall u,\, v \in \mathcal{V},\\ a(u\,,\,u) &\geq \gamma\,\|u\|^2 && \forall u \in \mathcal{V},\\ l(v) &\leq L\,\|v\| && \forall v \in \mathcal{V}.\end{aligned}$$

For all $c \in L^p(\Omega)$, we denote by a_c the following bilinear form:

$$\begin{aligned} a_c:\quad \mathcal{V}\times\mathcal{V} &\longrightarrow \mathbb{R},\\ (u,v) &\longmapsto a_c(u,v) = a(u,v) + \int_\Omega c(x)u(x).v(x)dx.\end{aligned}$$

We have the following lemma.

LEMMA 2.3. *The map*

$$\begin{aligned} L^p(\Omega) &\longrightarrow \mathcal{L}_2(\mathcal{V}),\\ c &\longmapsto a_c\end{aligned} \tag{2.8}$$

is continuous for $p \geq 1$ if $d = 1$, for $p > 1$ if $d = 2$, and for $p \geq \frac{3}{2}$ if $d = 3$.

Proof. Thanks to Theorem IX.16 from Brezis [11], we know that $H^1(\Omega)^d \subset L^q(\Omega)^d$ for $1 \leq q \leq +\infty$ when $d = 1$, for $1 \leq q < +\infty$ when $d = 2$, and for $1 \leq q \leq 6$ when $d = 3$.

It is well known that if $u \in L^{p_1}(\Omega)$ and $v \in L^{p_2}(\Omega)$, the product $u.v \in L^p(\Omega)$ with $\frac{1}{p} = \frac{1}{p_1} + \frac{1}{p_2}$. So, for all u and v in $L^q(\Omega)^d$, the product $u.v$ belongs to $L^{q/2}(\Omega)$.

Since Ω is bounded, the map $c \mapsto a_c$ is continuous for

$$\frac{1}{p} + \frac{2}{q} \leq 1. \tag{2.9}$$

Then, combining (2.9) and the conditions on q for which we have $H^1(\Omega)^d \subset L^q(\Omega)^d$, we deduce that the map is continuous for $p \geq 1$ if $d = 1$, for $p > 1$ if $d = 2$, and for $p \geq \frac{3}{2}$ if $d = 3$. □

Next, we denote by $u_c \in \mathcal{V}$ the unique solution of the following variational problem:

$$a_c(u_c, v) = l(v) \quad \forall\, v \in \mathcal{V}, \tag{2.10}$$

and we consider the cost function f defined by

$$\begin{array}{rccc} f: & L^p(\Omega) & \longrightarrow & \mathbb{R}, \\ & c & \longmapsto & f(c) = J(u_c), \end{array}$$

where J is a given functional defined on the boundary of Ω.

If J is differentiable, using the Lagrangian method, one can prove that the function f is differentiable, and we have

$$f'(c)\delta c = \int_\Omega \delta c(x) u_c(x).v_c(x)dx \quad \forall \delta c \in L^p(\Omega), \tag{2.11}$$

where u_c is the solution to the direct problem (2.10) and $v_c \in \mathcal{V}$ is the solution to the associated adjoint problem

$$a_c(w, v_c) = -\frac{\partial J}{\partial u}(u_c)w \quad \forall w \in \mathcal{V}. \tag{2.12}$$

From relation (2.11), we deduce the Riesz representation g of the differential $f'(c)$:

$$g = u_c.v_c.$$

The regularity of g depends on that of the bilinear form a, the linear mapping l, and the functional J.

Let us now consider the following particular case:

$$c \equiv 0 \text{ and } \delta c = \delta c_\varepsilon.$$

Denoting $j(\varepsilon) = J(u_{\delta c_\varepsilon})$ and using Theorem 2.2, we deduce the following result.

PROPOSITION 2.4. *The cost function j has the following asymptotic expansion:*

$$j(\varepsilon) - j(0) = \kappa\, \rho(\varepsilon)\, u_0.v_0 + o(\rho(\varepsilon)). \tag{2.13}$$

When κ is large, the solution $u_{\delta c_\varepsilon}$ is close to zero in ω_ε. Using the penalization technique, this corresponds to a good approximation of the solution in $\Omega \backslash \overline{\omega_\varepsilon}$ with a Dirichlet condition on the boundary of the hole ω_ε. The topological gradient $g = u_0.v_0$ is exactly what we obtain by a much more complicated approach (see [7, 8, 19, 20, 23, 24, 26]).

2.2. Analytical examples. We consider here two one-dimensional examples for which the solution can be calculated explicitly. In the first, we give an example in which the classical gradient is equal to the topological gradient. In the second, we consider the case when the classical gradient and the topological one are different.

2.2.1. First example. Consider the following one-dimensional elliptic state equation:

$$\left\{\begin{array}{rcl} -u''(x)+c(x)u(x) &=& 0 \quad \text{for } x\in]0,1[,\\ u(0) &=& 0,\\ u'(1) &=& 1. \end{array}\right. \tag{2.14}$$

One can show that (2.14) have the following variational formulation:

$$\left\{\begin{array}{l} \text{find } u\in\mathcal{V} \text{ such that}\\ \displaystyle\int_0^1 u'(x)w'(x)dx+\int_0^1 c(x)u(x)w(x)dx=w(1) \quad \forall\, w\in\mathcal{V}, \end{array}\right. \tag{2.15}$$

where $\mathcal{V}=\{w\in H^1(\Omega): w(0)=0\}$.

If c is in $L^1([0,1])$, thanks to the Lax–Milgram theorem, problem (2.15) admits a unique solution. We denote this solution by u_c.

We now want to minimize the functional

$$\begin{array}{rccc} f: & L^1([0,1]) & \longrightarrow & \mathbb{R},\\ & c & \longmapsto & u_c(1). \end{array}$$

The associated Lagrangian is given by

$$L(c,u,w)=u(1)+\int_0^1 u'(x)w'(x)dx+\int_0^1 c(x)u(x)w(x)dx-w(1).$$

Using (2.15), we deduce

$$L(c,u_c,w)=f(c) \quad \forall w\in\mathcal{V}.$$

Therefore,

$$f'(c)\delta c=\frac{\partial L}{\partial c}(c,u_c,w)\delta c+\frac{\partial L}{\partial u}(c,u_c,w)\frac{\partial u_c}{\partial c}\delta c \quad \forall w\in\mathcal{V}.$$

We denote by v_c the solution to the adjoint problem

$$\frac{\partial L}{\partial u}(c,u_c,v_c)w=0 \quad \forall w\in\mathcal{V}.$$

Then, $v_c\in\mathcal{V}$ is a solution to

$$w(1)+\int_0^1 w'(x)v_c'(x)dx+\int_0^1 c(x)w(x)v_c(x)dx=0 \quad \forall w\in\mathcal{V}. \tag{2.16}$$

From (2.15) and (2.16), one can observe that $v_c=-u_c$.

Thus, we derive

$$f'(c)\delta c=\frac{\partial L}{\partial c}(c,u_c,v_c)\delta c=-\int_0^1 \delta c(x)\,u_c^2(x)dx.$$

Let δc_ε be the characteristic function of $[x_0, x_0+\varepsilon]$:

$$\delta c_\varepsilon = \begin{cases} 1 & \text{if } x \in [x_0, x_0+\varepsilon], \\ 0 & \text{elsewhere.} \end{cases}$$

Then, using Theorem 2.2, we derive

$$f(c+\delta c_\varepsilon) - f(c) = \varepsilon\Big(- u_c(x_0)^2\Big) + o(\varepsilon).$$

Recall that $j(\varepsilon) = J(u_{\delta c_\varepsilon}) = f(\delta c_\varepsilon)$.

In the particular case when $c \equiv 0$, the function j has the following asymptotic expansion:

$$j(\varepsilon) - j(0) = \varepsilon\Big(- u_0(x_0)^2\Big) + o(\varepsilon). \tag{2.17}$$

But, in this simple case, the solution of (2.14) can be calculated explicitly: we have $u_0(x) = x$, for all $x \in]0,1[$.

Equation (2.17) can thus be rewritten as

$$j(\varepsilon) - j(0) = \varepsilon(-x_0^2) + o(\varepsilon). \tag{2.18}$$

In the case when $c = \delta c_\varepsilon$, the solution $u_{\delta c_\varepsilon}$ is given by

$$u_{\delta c_\varepsilon}(x) = \begin{cases} \dfrac{2x}{(x_0+1)e^{\varepsilon} - (x_0-1)e^{-\varepsilon}} & \text{if } x \in [0, x_0], \\ \dfrac{(x_0+1)e^{x-x_0} + (x_0-1)e^{-(x-x_0)}}{(x_0+1)e^{\varepsilon} - (x_0-1)e^{-\varepsilon}} & \text{if } x \in [x_0, x_0+\varepsilon], \\ x - (x_0+\varepsilon) + \dfrac{(x_0+1)e^{\varepsilon} + (x_0-1)e^{-\varepsilon}}{(x_0+1)e^{\varepsilon} - (x_0-1)e^{-\varepsilon}} & \text{if } x \in [x_0+\varepsilon, 1]. \end{cases}$$

Therefore,

$$\begin{aligned} j(\varepsilon) - j(0) = u_{\delta c_\varepsilon}(1) - u_0(1) &= -(x_0+\varepsilon) + \frac{(x_0+1)e^{\varepsilon} + (x_0-1)e^{-\varepsilon}}{(x_0+1)e^{\varepsilon} - (x_0-1)e^{-\varepsilon}} \\ &= -x_0^2\varepsilon + (-1+x_0^2)x_0\varepsilon^2 + o(\varepsilon^2). \end{aligned}$$

We obtain the same result as in (2.18) which confirms that given by Theorem 2.2.

2.2.2. Second example. Let us consider the following problem:

$$\begin{cases} \Big(c(x)u'(x)\Big)' &= 0 \quad \text{for } x \in]0,1[, \\ u(0) &= 0, \\ c(1)u'(1) &= 1, \end{cases} \tag{2.19}$$

which admits the following variational formulation:

$$\text{find } u \in \mathcal{V} \text{ such that } \forall\, w \in \mathcal{V}, \quad \int_0^1 c(x)u'(x)w'(x)dx = w(1), \tag{2.20}$$

where $\mathcal{V} = \{w \in H^1(\Omega) : w(0) = 0\}$.

If c is in $L^\infty([0,1])$ and there exists $\gamma > 0$ such that $c \geq \gamma > 0$, thanks to the Lax–Milgram theorem, problem (2.20) admits a unique solution $u_c \in \mathcal{V}$.

Next, we want to minimize the following functional:

$$f: \begin{array}{ccc} L^\infty([0,1]) & \longrightarrow & \mathbb{R}, \\ c & \longmapsto & u_c(1). \end{array} \tag{2.21}$$

The associated Lagrangian is defined by

$$L(c,u,w) = u(1) + \int_0^1 c(x)u'(x)w'(x)dx - w(1).$$

Taking into account that u_c is the solution to (2.20), we have $L(c,u_c,w) = f(c)$ for all $w \in \mathcal{V}$. Therefore,

$$f'(c)\delta c = \frac{\partial L}{\partial c}(c,u_c,w)\delta c + \frac{\partial L}{\partial u}(c,u_c,w)\frac{\partial u_c}{\partial c}\delta c \quad \forall w \in \mathcal{V}.$$

Let $v_c \in \mathcal{V}$ be the solution to the associated adjoint problem; then v_c satisfies

$$\left\{ \begin{array}{l} v_c \in \mathcal{V}, \\ w(1) + \displaystyle\int_0^1 c(x)w'(x)v_c'(x)dx = 0 \quad \forall w \in \mathcal{V}. \end{array} \right. \tag{2.22}$$

As in the first example, from (2.20) and (2.22) we deduce that $v_c = -u_c$, which implies

$$f'(c)\delta c = \frac{\partial L}{\partial c}(c,u_c,v_c)\delta c = -\int_0^1 \delta c(x)\Big(u_c'(x)\Big)^2 dx.$$

As a consequence, the classical gradient is given by $g = -(u_c')^2$.

Let us consider the perturbation

$$\delta c_\varepsilon(x) = \left\{ \begin{array}{ll} \kappa & \text{if } x \in [x_0, x_0+\varepsilon], \\ 0 & \text{elsewhere,} \end{array} \right.$$

with $\kappa \in \mathbb{R}$ a fixed given constant.

As in the previous case, we can use analytic calculus.

Let c_0 be a constant in $\mathbb{R}$. Then if $c \equiv c_0$, the solution of (2.19) is given by $u_0(x) = \frac{x}{c_0}$.

If $c = c_0 + \delta c_\varepsilon$, the solution $u_{c_0+\delta c_\varepsilon}$ of (2.19) has the following expression

$$u_{c_0+\delta c_\varepsilon}(x) = \left\{ \begin{array}{ll} \dfrac{x}{c_0} & \text{if } x \in [0,x_0], \\ \dfrac{x-x_0}{c_0+\kappa} + \dfrac{x}{c_0} & \text{if } x \in [x_0, x_0+\varepsilon], \\ \dfrac{x-\varepsilon}{c_0} + \dfrac{\varepsilon}{c_0+\kappa} & \text{if } x \in [x_0+\varepsilon, 1]. \end{array} \right.$$

Therefore,

$$j(\varepsilon) - j(0) = f(c_0+\delta c_\varepsilon) - f(c_0) = \varepsilon\left(\frac{1}{c_0+\kappa} - \frac{1}{c_0}\right) = \frac{-\kappa\varepsilon}{c_0^2 + \kappa c_0}. \tag{2.23}$$

The assumptions of Theorem 2.2 are not satisfied in this case, since the map f (see (2.21)) is differentiable on $L^p([0,1])$ only for $p = +\infty$. But if they were, we should have

$$j(\varepsilon) - j(0) = -\frac{\kappa\varepsilon}{c_0^2} + o(\varepsilon). \tag{2.24}$$

Here is an example in which the classical gradient (2.24) is different from the topological gradient (2.23). These two expressions are close when κ is small, but this is no more the case when κ is close to $-c_0$. However, when $\kappa > -c_0$, both expressions have the same sign. Thus, the classical gradient could be seen as a "topological descent direction."

2.3. Another example: The Maxwell problem. We consider the Maxwell equations in a domain $\Omega \subset \mathbb{R}^3$ containing a small inhomogeneity ω_ε. We denote by H_ε the magnetic field. It is the solution to

$$\left\{ \begin{array}{ll} \nabla \times (\alpha_\varepsilon \nabla \times H_\varepsilon) + \beta_\varepsilon H_\varepsilon = 0 & \text{in } \Omega, \\ \alpha_\varepsilon (\nabla \times H_\varepsilon) \times \mathbf{n} = \phi & \text{on } \partial\Omega. \end{array} \right. \tag{2.25}$$

Here, $\mathbf{n}$ denotes the outward normal to $\partial\Omega$, and the functions α_ε and β_ε are given by

$$\alpha_\varepsilon = \left\{ \begin{array}{ll} \alpha_0 & \text{in } \Omega \backslash \overline{\omega_\varepsilon}, \\ \alpha_1 & \text{in } \omega_\varepsilon, \end{array} \right. \qquad \beta_\varepsilon = \left\{ \begin{array}{ll} \beta_0 & \text{in } \Omega \backslash \overline{\omega_\varepsilon}, \\ \beta_1 & \text{in } \omega_\varepsilon, \end{array} \right.$$

where α_0, α_1, β_0, and β_1 are positive constants.

A sensitivity analysis with respect to the presence of small inhomogeneities for the Maxwell equations is given in [4, 5, 6, 30]. A topological asymptotic expansion is derived in [30] when inserting a dielectric object and when creating a small hole. In the case of a spherical metallic object, it is proved that the topological gradient is given by

$$g(H_0, v_0) = \mathcal{R}e\Big(-2\nabla \times H_0 . \overline{\nabla \times v_0} + k^2 \, H_0 . \overline{v_0} \Big),$$

where $H_0 \in H(curl, \Omega) = \{u \in L^2(\Omega)^3;\ \nabla \times u \in L^2(\Omega)^3\}$ is the magnetic field in the absence of any inhomogeneities, v_0 is the solution to the associated adjoint problem, and k is the wave number. As we observe, the topological gradient $g(H_0, v_0)$ is not equal to the classical gradient $g_c(H_0, v_0) = \mathcal{R}e(-2\nabla \times H_0 . \overline{\nabla \times v_0})$. The reason for this is that $H(curl, \Omega)$ is a subset of $L^2(\Omega)^3$, but not a subset of $L^p(\Omega)^3$ for $p < 2$.

The numerical experiments that are performed in [30] show that it is possible to obtain satisfying results using the classical gradient instead of the topological one.

3. Application to the elasticity problem. The linear elasticity problem [17, 20] is described by the following system:

$$\left\{ \begin{array}{rcl} -\operatorname{div}\sigma(u) = & 0 & \text{in } \Omega, \\ \sigma(u).\mathbf{n} = & T_N & \text{in } \Gamma_N, \\ u = & 0 & \text{in } \Gamma_D, \end{array} \right. \tag{3.1}$$

where u is the displacement, $(\sigma_{ij})_{1 \le i,j \le d}$ $(d = 2, 3)$ is the stress tensor

$$\sigma_{ij}(u) = \lambda \operatorname{div} u \, \delta_{ij} + 2\mu \varepsilon_{ij}(u),$$

λ and μ are the Lamé coefficients, and $(\varepsilon_{ij})_{1 \le i,j \le d}$ is the strain tensor

$$\varepsilon_{ij}(u) = \frac{1}{2}\Big(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \Big), \quad 1 \le i, j \le d.$$

Here δ_{ij} is the Kronecker symbol, and $\mathbf{n}$ denotes the outward normal to the boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, where both Γ_D and Γ_N have a nonnegative Lebesgue

TABLE 1
Expressions of the topological gradient.

Boundary condition on $\partial\omega_\varepsilon$	Topological gradient g
2D Dirichlet	$-\dfrac{4\pi\mu(\mu+\eta)}{2\mu+\eta}\,u.v$
3D Dirichlet	$\dfrac{12\pi\mu(\lambda+2\mu)}{2\lambda+5\mu}\,u.v$
2D Neumann	$-\dfrac{\pi(\mu+\eta)}{2\eta\mu}\{4\mu\sigma(u):\varepsilon(v)+(\eta-2\mu)\ \mathrm{tr}\sigma(u)\ \mathrm{tr}\varepsilon(v)\}$
3D Neumann	$-\dfrac{\pi(\lambda+2\mu)}{\mu(9\lambda+14\mu)}\{20\mu\sigma(u):\varepsilon(v)+(3\lambda-2\mu)\ \mathrm{tr}\sigma(u)\ \mathrm{tr}\varepsilon(v)\}$

measure and $\Gamma_D \cap \Gamma_N = \emptyset$. Equations (3.1) describe the displacement u of an isotropic solid which is loaded by a surface traction T_N on Γ_N and which is clamped on Γ_D.

Structural optimization involves relaxed formulations, homogenization, and the level-set method (see, e.g., [1, 2, 3, 9, 16, 27, 31]). In the context of the topological sensitivity analysis, the asymptotic behavior of a design function with respect to the creation of a small hole ω_ε in Ω has been studied in [20]. In the case of a spherical hole, we present in Table 1 the different expressions of the topological sensitivity for the Dirichlet and Neumann boundary conditions in two and three dimensions, where v is the solution to the associated adjoint problem, $\mathrm{tr}\sigma = \sum_i \sigma_{ii}$ is the usual trace operator, and the parameter η is given by

$$\eta = \begin{cases} \dfrac{\mu(3\lambda+2\mu)}{\lambda+2\mu} & \text{plane stress,} \\ \lambda+\mu & \text{plane strain.} \end{cases}$$

One can observe that in both cases, the topological gradient $g(u,v)$ is not equal to the classical gradient, which is given by

$$g_c(u,v) = \sigma(u):\varepsilon(v).$$

However, we will show on a simple two-dimensional plane stress elasticity problem that it is possible to use the classical gradient as a "topological descent direction" in the optimization algorithm. For the sake of completeness, we recall briefly the optimization algorithm we use. For more details, one may consult [7, 8, 19, 20, 23, 24, 26, 33]). We apply an iterative process to build a sequence of geometries $(\Omega_k)_{k\geq 0}$ with $\Omega_0 = \Omega$. At the kth iteration, the topological gradient is denoted by g_k, and the new geometry Ω_{k+1} is defined by a level set curve of g_k.

THE ALGORITHM.

- Initialization: choose $\Omega_0 = \Omega$ and set $k = 0$.
- Repeat until $g_k \geq 0$ in Ω_k:
  - Solve (4.1) and its associated adjoint problem in Ω_k;
  - compute the topological sensitivity g_k;
  - set $\Omega_{k+1} = \{x \in \Omega_k, \quad g_k(x) \geq c_k\}$, where c_k is chosen in such a way that the cost function decreases as most as possible;
  - $k \longleftarrow k+1$.

This algorithm can be seen as a descent method, where the descent direction is determined by the topological sensitivity g_k and the step length is given by the volume variation $meas(\Omega_k \backslash \Omega_{k+1})$.

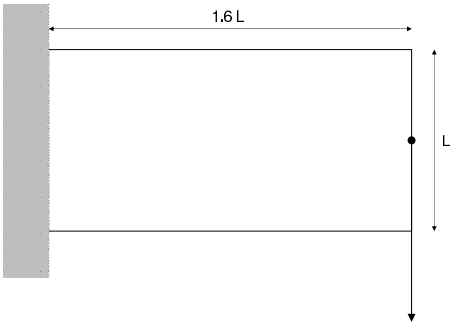


FIG. 2. *2D cantilever beam.*

The natural optimality condition is

$$g_k(x) \geq 0 \quad \forall x \in \Omega_k.$$

It coincides with the one obtained by Buttazzo and Dal Maso [12] for the Laplace equation using homogenization theory. In practice, some other stopping criteria can be successfully implemented, such as the material volume to be removed [14, 20, 23], the number of holes to be inserted [26], or the obstacles number to be detected [24].

We consider here the maximization of the compliance for a two-dimensional cantilever beam represented in Figure 2. The initial domain is a plain rectangle with one edge clamped, and a pointwise load is applied to the middle of the opposite edge. In the numerical results (Figures 3 and 4), 10% of the material is removed at each step, and the optimal domains are obtained when their volumes are less than 40% of the initial one; i.e., after nine iterations. One can observe that the results obtained with the classical gradient are not so different from those obtained with the topological expansion, and this is the case for various materials such as steel or concrete. The reason for this is that the difference between the topological gradient and the classical one is small.

Other numerical results on some structural shape optimization examples are presented in [20] that show also that it is possible to use the classical gradient as a "topological descent direction" in the optimization algorithm.

4. Application to the Stokes equations. In this section, we consider the Stokes equations case. Let Ω be a bounded domain of $\mathbb{R}^d$, $d = 2, 3$. We denote by Γ its boundary.

4.1. Formulation of the problem. The standard form of the Stokes equations describing the motion of an incompressible fluid in Ω is given by

$$\begin{aligned} -\nu\Delta u + \nabla p &= F \quad \text{in } \Omega, \\ \operatorname{div} u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.1}$$

where u is the velocity, p is the pressure, F is a given body force per unit of mass, and ν is the kinematic viscosity of the fluid.

One can show that the velocity u satisfies the following variational problem:

$$\begin{cases} \text{find } u \in \mathcal{V} \text{ such that} \\ \quad a(u, w) = l(w) \quad \forall w \in \mathcal{V}, \end{cases} \tag{4.2}$$

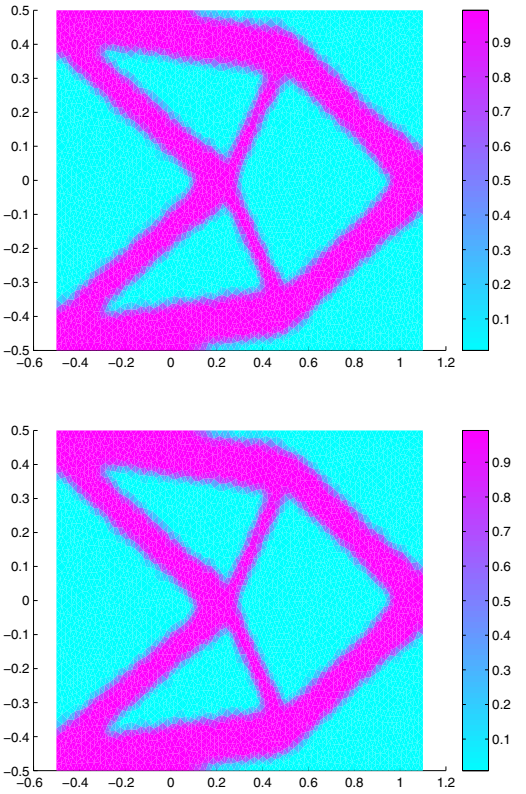


FIG. 3. *Comparison: density obtained by the topological algorithm using the classical gradient (top) and the topological one (bottom).*

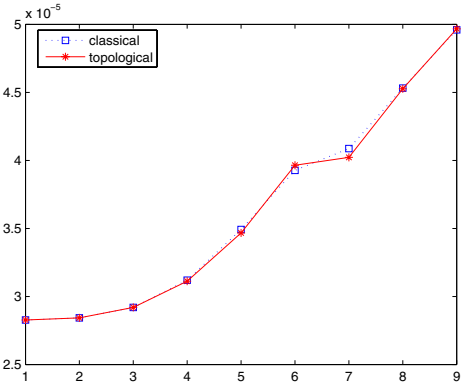


FIG. 4. *Evolution of the compliance with respect to the number of iterations using the classical (square) and the topological (star) gradients.*

with

$$\mathcal{V} = \Big\{ v \in H_0^1(\Omega)^d, \ \text{div } v = 0 \Big\},$$
$$a(v,w) = \nu \int_\Omega \nabla v \nabla w \ \text{dx} \quad \forall v, w \in \mathcal{V},$$
$$l(w) = \int_\Omega F\, w \ \text{dx} \quad \forall w \in \mathcal{V}.$$

It is easy to verify that $\mathcal{V} \subset H^1(\Omega)^d$ is a Hilbert space; a is a bilinear, symmetric, continuous, and coercive form on $\mathcal{V}$; and l is a linear and continuous form on $\mathcal{V}$.

In order to apply the tools presented in section 2, we consider the following penalized problem:

$$\begin{aligned} -\text{ div } (\nu \nabla u_\varepsilon) + \delta c_\varepsilon u_\varepsilon + \nabla p_\varepsilon &= F \quad \text{in } \Omega, \\ \text{div } u_\varepsilon &= 0 \quad \text{in } \Omega, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.3}$$

where δc_ε is a piecewise constant function defined in Ω by

$$\delta c_\varepsilon = \begin{cases} \kappa \text{ if } x \in \omega_\varepsilon, \\ 0 \text{ if } x \in \Omega_\varepsilon, \end{cases}$$

with κ a positive constant and $\Omega_\varepsilon = \Omega \backslash \overline{\omega_\varepsilon}$.

Then, if κ goes to infinity in ω_ε, for a given ε, the corresponding solution u_ϵ converges to the solution of the following Stokes system:

$$\begin{aligned} -\nu \Delta u_\varepsilon + \nabla p_\varepsilon &= F \quad \text{in } \Omega_\varepsilon, \\ \text{div } u_\varepsilon &= 0 \quad \text{in } \Omega_\varepsilon, \\ u_\varepsilon &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.4}$$

with a Dirichlet condition on the boundary of ω_ε:

$$u_\varepsilon = 0 \text{ on } \partial\omega_\varepsilon.$$

Consider now a cost function

$$j(\varepsilon) = J(u_\varepsilon) \in \mathbb{R},$$

where $u_\varepsilon \in \mathcal{V}$ is the solution to (4.3).

Using Proposition 2.4 (see (2.13)), we derive the asymptotic expansion for j. Then, if J is defined on the boundary of Ω, we have

$$j(\varepsilon) - j(0) = \kappa |\omega| \varepsilon^d u_0 . v_0 + o(\varepsilon^d),$$

where u_0 is the solution to (4.1) and v_0 is the solution to the associated adjoint problem.

Let us recall the topological asymptotic expansion obtained in [7, 24, 26]:

$$j(\varepsilon) - j(0) = m(\varepsilon) u_0 . v_0 + o(m(\varepsilon)).$$

When ω is the unit ball $B(x_0, 1)$, we have $\omega_\varepsilon = B(x_0, \varepsilon)$, and $m(\varepsilon)$ is given by

$$m(\varepsilon) = \begin{cases} \dfrac{-4\pi\nu}{\log(\varepsilon)} & \text{in two dimensions,} \\ \\ 6\pi\nu\varepsilon & \text{in three dimensions.} \end{cases}$$

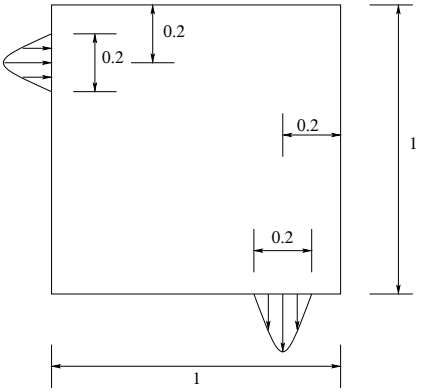


FIG. 5. *Design domain for the pipe bend example.*

We remark that the straightforward method gives the right topological gradient,

$$g(x) = u_0(x).v_0(x), \quad x \in \Omega.$$

From the numerical point of view, the behavior with respect to ε is not so important. In the following subsections, we give some numerical applications of this result.

4.2. Numerical results. Here, we limit ourselves to the two-dimensional case. As an application of the previous theoretical results, we consider in this subsection some engineering applications commonly found in the fluid mechanics literature. In the first part of this subsection we consider two numerical examples that were also considered in [10, 21]. In the second part, we deal with other interesting examples concerning the shape optimization of tubes in a cavity.

Our implementation is based on the topological optimization algorithm recalled in section 3. We proceed by iterations. At each step, we remove the zones in which the topological gradient g_k is nonpositive. In this test case, we stop the process when g_k is positive everywhere. The direct and the adjoint problems are discretized by a finite element method using a Reynolds number $Re = 100$. The computation of the approximate solution is achieved by the Uzawa algorithm.

4.2.1. A comparison. In order to test the advantage of our approach, we compare our results to those obtained in [10, 21]. We consider here two numerical examples. The first one is the pipe bend that is depicted in Figure 5. This test case is treated by Borrvall and Petersson in [10]. The second one is the double pipe shown in Figure 7. It was also considered by Borrvall and Petersson in [10] and recently by Guest and Prévost in [21]. The aim here is to obtain the optimal shape minimizing the dissipated power in the fluid. In both cases the inflow and the outflow conditions are of a prescribed parabolic flow profiles type.

Pipe bend example. We present in Figure 5 the design domain for the pipe bend example. Here the prescribed maximum flow velocity is equal to 1 at the inlet (left) and at the outlet (bottom). The inlet and outlet flows are shown in the figure, and velocity is prescribed to be zero elsewhere on the boundary of the domain. The results of this example are obtained in only three iterations. We present the geometries obtained during the optimization process in Figure 6. The optimal solution is nearly identical to those presented in [10]. However, we obtain this result in three iterations,

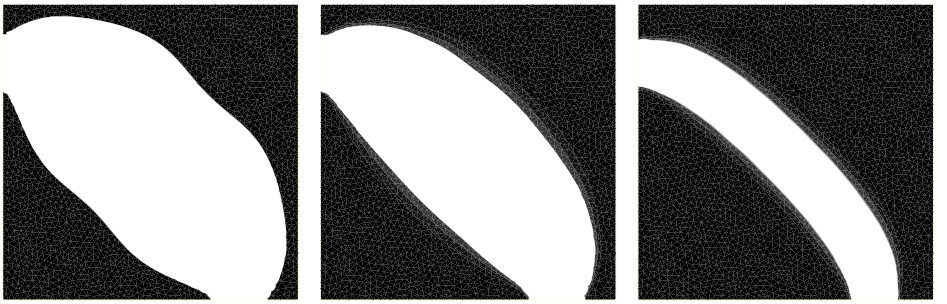


FIG. 6. *Geometries obtained during the optimization process.*

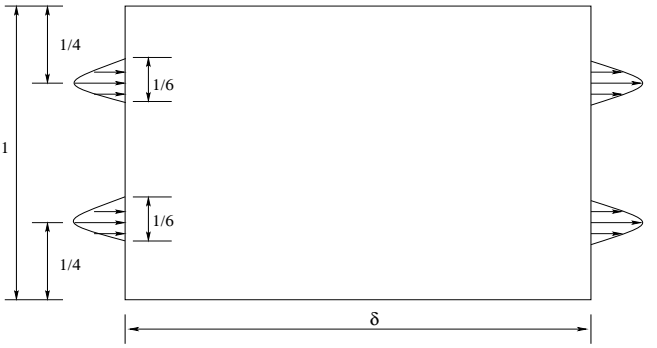


FIG. 7. *Design domain for the double pipe example.*

where Borrvall and Petersson needed more than sixty. As can be seen, we have a more torus-shaped pipe than in [10], like most pipe bends in fluid mechanics literature. It is stated in [21] that the solution in Borrvall and Petersson [10] contains regions of artificial material and does not sufficiently take into account the adherence condition.

Double pipe example. The design domain of this example is shown in Figure 7. The domain has two inlet and outlet flows, each with a maximum velocity equal to 1, and velocity is prescribed to be zero elsewhere on the boundary of the domain. The problem is solved using an 80×120 element mesh. We present here the solution obtained for the longer flow distances (e.g., $\delta = 1.5$). The optimal geometry is obtained using only five iterations (where Borrvall and Peterson needed more than sixty iterations). The geometries obtained during the optimization process are given in Figure 8. However, the two pipes join to form a single, wider pipe through the center of the domain. This design decreases the length of the fluid-solid interface, thereby decreasing the power lost. As can be seen, the optimal solution is identical to that obtained by Guest and Prévost [21], but it does not match that of Borrvall and Petersson [10]. As for the pipe bend example, the solution in [10] contains regions of artificial material and does not sufficiently take into account the adherence condition.

4.2.2. Shape optimization of tubes in a cavity. In this section we treat the shape optimization of tubes in a cavity. We consider an incompressible fluid in a cavity Ω having one inlet Γ_{in} and some outlets Γ_{out}^i, $i = 1, m$. (see Figure 9).

The aim here is to determine the optimal shape of the tubes that connect the inlet to the outlets of the cavity maximizing the outflow rate. It consists of inserting small obstacles in the cavity in order to maximize the outflow rate at Γ_{out}^i, $i = 1, m$.

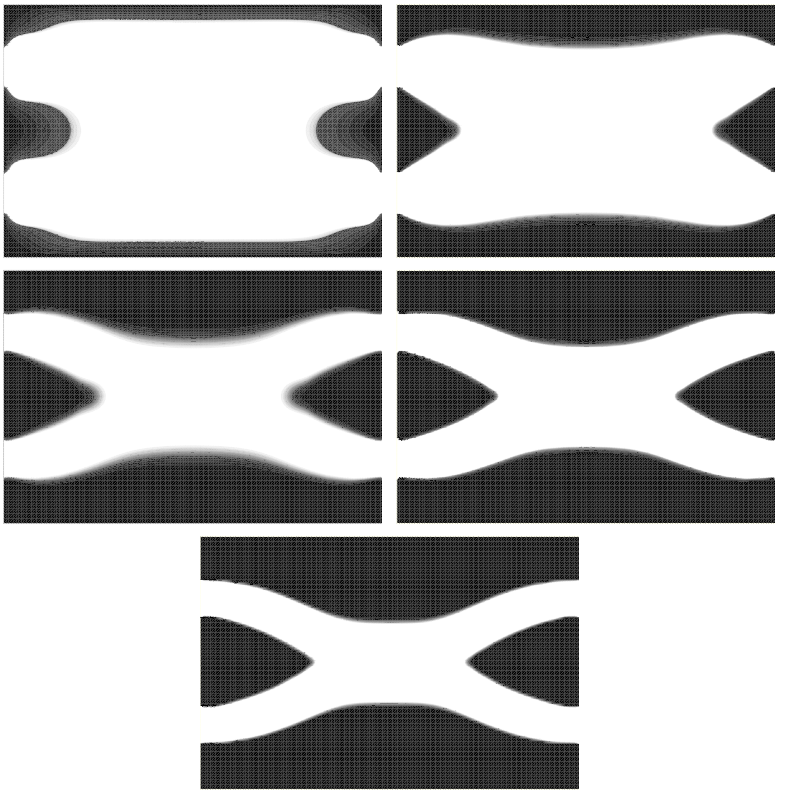


FIG. 8. *Geometries obtained during the optimization process (double pipe, $\delta = 1.5$).*

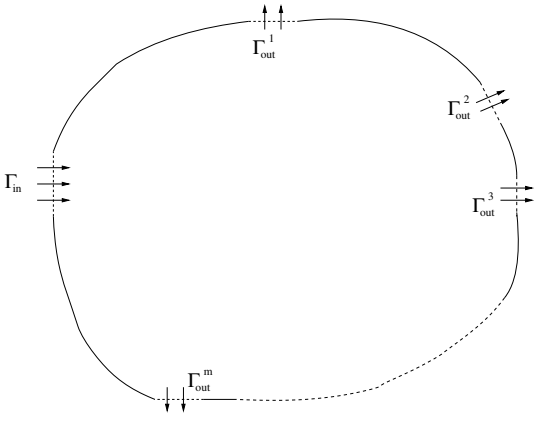


FIG. 9. *The cavity Ω.*

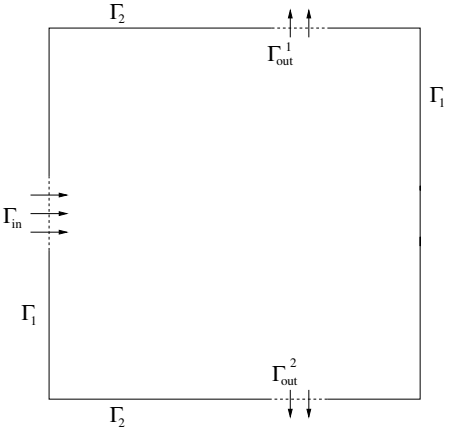


FIG. 10. *The initial geometry for the first test case.*

We assume that the flow satisfies the steady state Stokes equations

$$\begin{aligned} -\nu\Delta u + \nabla p &= 0 \quad \text{in } \Omega, \\ \text{div } u &= 0 \quad \text{in } \Omega. \end{aligned} \tag{4.5}$$

On the inlet Γ_{in} we define the normal component of the stress tensor,

$$\sigma(u).\mathbf{n} = \phi \text{ on } \Gamma_{in},$$

where $\sigma(u) = \nu(\nabla u + \nabla u^T) - pI$, I is the 2×2 identity matrix, and $\mathbf{n}$ denotes the outward normal to the boundary.

On the outlets, we use a free surface boundary condition

$$\sigma(u).\mathbf{n} = 0 \text{ on } \cup_{i=1}^{m} \Gamma_{out}^{i}.$$

The velocity is prescribed to be zero elsewhere on the boundary of the cavity.

The cost function measuring the outflow rate is given by

$$J(u) = \sum_{i=1}^{m} \int_{\Gamma_{out}^{i}} |u.\mathbf{n}| \text{ ds}.$$

4.2.3. First case: A cavity with one inlet and two outlets. We consider a cavity with one inlet and two outlets having the same section (see Figure 10).

The results of this case are given in Figures 11, 12, and 13. In Figure 11, we illustrate the variation of the outflow rate. The optimal design is presented in Figure 12. We present the topological gradient and the geometries obtained during the optimization process in Figure 13. The topological gradient corresponding to the initial geometry is shown in Figure 13(a). The topological gradient corresponding to the final geometry is presented in Figure 13(d). One can observe that the optimality condition is reached after only four iterations. Indeed, the topological gradient is almost positive everywhere (see Figure 13(d)) in the current domain ($g_4(x) \geq -8.5610^{-4}$ for all $x \in \Omega_4$).

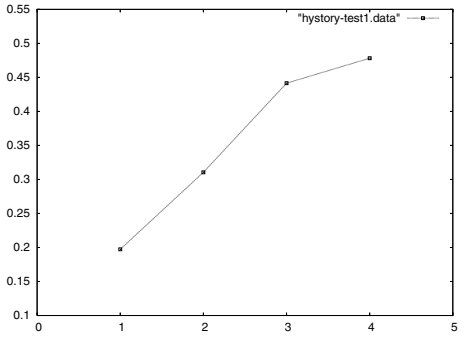


FIG. 11. *Variation of the outflow rate (first case).*

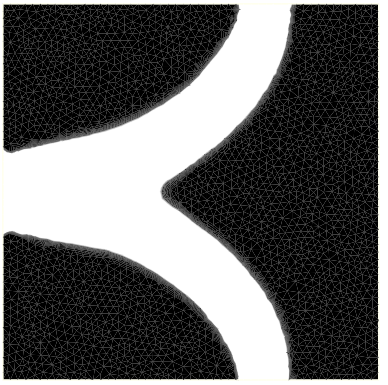


FIG. 12. *The optimal design (first case).*

4.2.4. Second case: A cavity with one inlet and three outlets. In the second case, we use a cavity with one inlet Γ_{in} and three outlets Γ_{out}^1, Γ_{out}^2, and Γ_{out}^3 (see Figure 14).

The results of this case are given in Figures 15, 16, and 17. As in the first case, we present the variation of the outflow rate in Figure 15 and the optimal design in Figure 16. The topological gradient and the geometries that we obtained during the optimization process are given in Figure 17. The topological gradient corresponding to the initial geometry is shown in Figure 17(a). The topological gradient corresponding to the final geometry is presented in Figure 17(d). As can be seen, the optimal solution is obtained using only four iterations.

4.2.5. Third case: A cavity with one inlet and three outlets (nonsymmetric case). We consider now a cavity with one inlet and three nonsymmetric outlets. The initial geometry is described in Figure 18. The obtained results are given in Figures 19, 20, and 21. As in the previous cases, we illustrate the variation of the outflow rate (see Figure 19), we present the optimal design (see Figure 20), and we give the topological gradient and the geometries that we obtained during the optimization process (see Figure 21). The initial geometry is given in Figure 21(a); the optimal one is given in Figure 21(d).

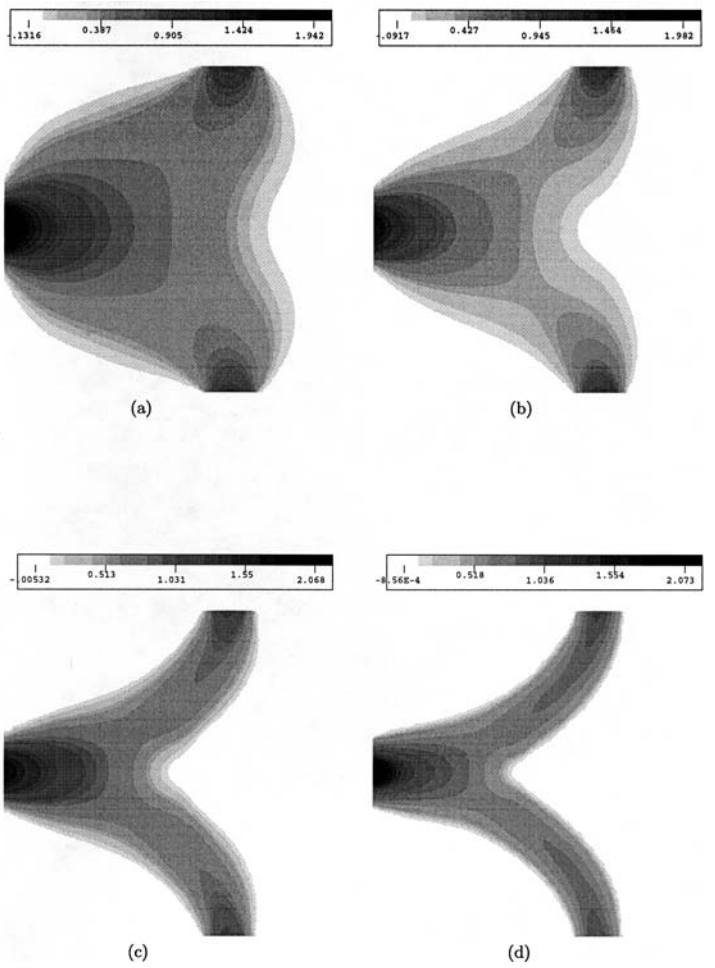


FIG. 13. *The topological gradient describing the geometry during the optimization process (first case).*

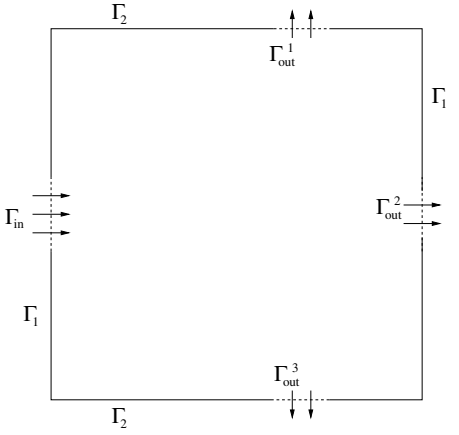


FIG. 14. *The initial geometry for the second test case.*

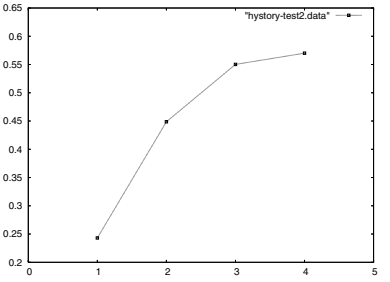


FIG. 15. *Variation of the outflow rate (second case).*

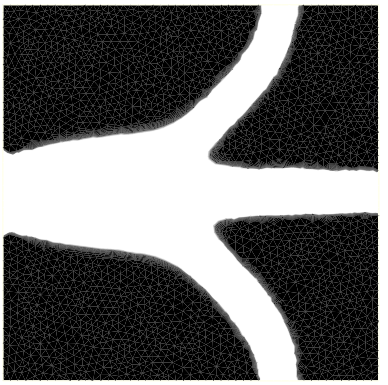


FIG. 16. *The optimal design (second case).*

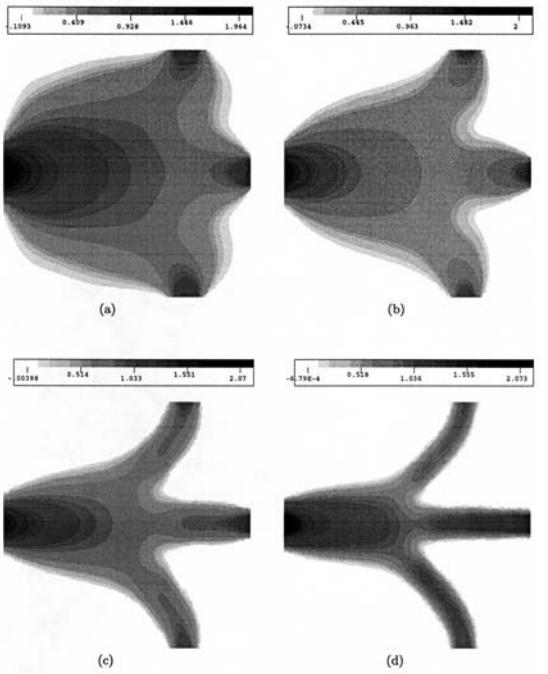


FIG. 17. *The topological gradient describing the geometry during the optimization process (second case).*

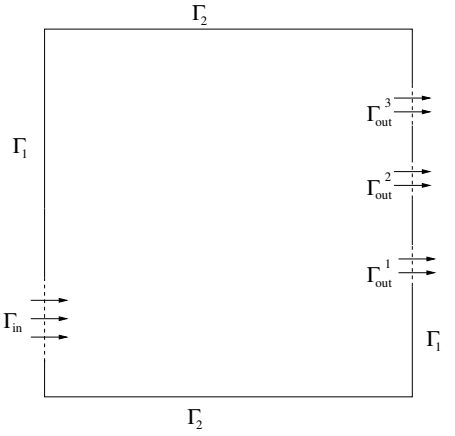


FIG. 18. *The initial geometry for the third test case.*

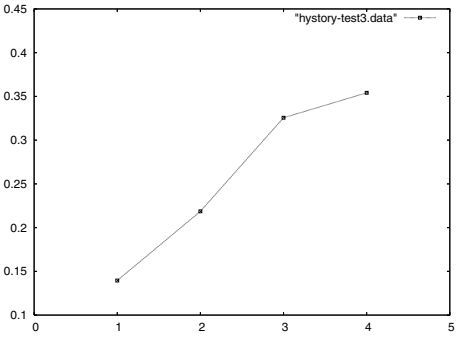


FIG. 19. *Variation of the outflow rate (third case).*

FIG. 20. *The optimal design (third case).*

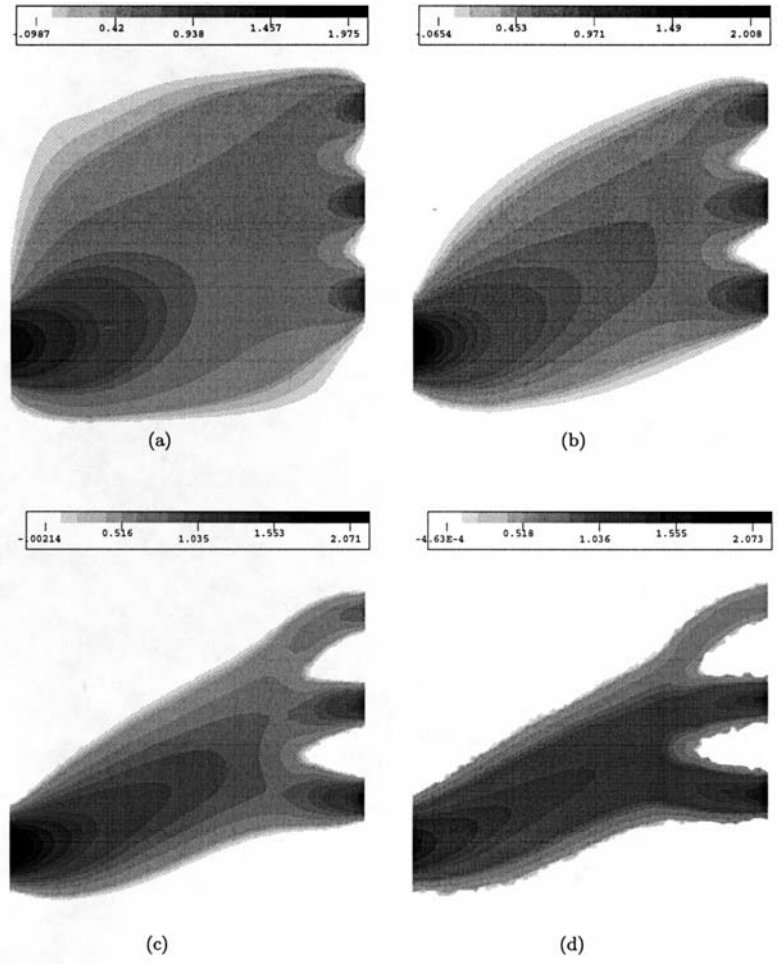


FIG. 21. *The topological gradient describing the geometry during the optimization process (third case).*

REFERENCES

[1] G. ALLAIRE, F. JOUVE, AND A.-M. TOADER, *Structural optimization using sensitivity analysis and a level-set method*, J. Comput. Phys., 194 (2004), pp. 363–393.

[2] G. ALLAIRE AND A. HENROT, *On some recent advances in shape optimization*, C. R. Acad. Sci. Paris, Ser. IIb, 329 (2001), pp. 383–396.

[3] G. ALLAIRE AND R. KOHN, *Optimal bounds on the effective behavior of a mixture of two well-orderd elastic materials*, Quart. Appl. Math., 51 (1993), pp. 643–674.

[4] H. AMMARI AND H. KANG, *Reconstruction of Small Inhomogeneties from Boundary Measurements*, Lecture Notes in Math. 1846, Springer-Verlag, Berlin, 2004.

[5] H. AMMARI, S. MOSKOW, AND M. VOGELIUS, *Boundary integral formulae for the reconstruction of electric and electromagnetic inhomogeneities of small volume*, ESAIM Control. Optim. Calc. Var., 9 (2003), pp. 49–66.

[6] H. AMMARI, M. VOGELIUS, AND D. VOLKOV, *Asymptotic formulas for perturbations in the electromagnetic fields due to presence of inhomogeneities of small diameter* II. *The full Maxwell's equations*, J. Math. Pures Appl., 80 (2001), pp. 769–814.

[7] S. AMSTUTZ, *The topological asymptotic for Navier–Stokes equations*, ESAIM Control. Optim. Calc. Var., 11 (2005), pp. 401–425.

[8] B. SAMET, S. AMSTUTZ, AND M. MASMOUDI, *The topological asymptotic for the Helmholtz equation*, SIAM J. Control. Optim., 42 (2003), pp. 1523–1544.

[9] M. Bendsoe, *Optimal Topology Design of Continuum Structure: An Introduction*, Technical report, Department of Mathematics, Technical University of Denmark, Lyngby, Denmark, 1996.

[10] T. Borrvall and J. Petersson, *Topology optimization of fluids in Stokes flow*, Internat. J. Numer. Methods Fluids, 41 (2003), pp. 77–107.

[11] H. Brezis, *Analyse fonctionnelle: Théorie et applications*, Masson, Paris, 1983.

[12] G. Buttazzo and G. Dal Maso, *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions*, Appl. Math. Optim., 23 (1991), pp. 17–49.

[13] J. Céa, *Conception optimale ou identification de formes: Calcul rapide de la dérivée directionnelle de la fonction coût*, RAIRO Modél. Math. Anal. Numér, 20 (1986), pp. 371–402.

[14] J. Céa, S. Garreau, Ph. Guillaume, and M. Masmoudi, *The shape and topological optimizations connection*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 713–726.

[15] D. J. Cedio-Fengya, S. Moskow, and M. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–595.

[16] M. Chipot and G. Dal Maso, *Relaxed shape optimization: The case of nonnegative data for the Dirichlet problems*, Adv. Math. Sci. Appl., 1 (1992), pp. 47–81.

[17] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[18] A. Friedman and M. Vogelius, *Identification of small inhomogeneities of extreme conductivity by boundary measurements: A theorem on continuous dependence*, Arch. Ration. Mech. Anal., 105 (1989), pp. 299–326.

[19] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological sensitivity for linear isotropic elasticity*, in Proceedings of the European Congress on Computational Mechanics (ECCM), Munich, Aug. 31–Sept. 3, 1999, W. Wunderlich, ed., CD-ROM.

[20] S. Garreau, Ph. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: The elasticity case*, SIAM J. Control Optim., 39 (2001), pp. 1756–1778.

[21] J. K. Guest and J. H. Prévost, *Topology optimization of creeping fluid flows using a Darcy–Stokes finite element*, Internat. J. Numer. Methods Fluids, 66 (2006), pp. 461–484.

[22] Ph. Guillaume and M. Masmoudi, *Computation of high order derivatives in optimal shape design*, Numer. Math., 67 (1994), pp. 231–250.

[23] Ph. Guillaume and K. Sid Idris, *The topological asymptotic expansion for the Dirichlet problem*, SIAM J. Control Optim., 41 (2002), pp. 1042–1072.

[24] Ph. Guillaume and K. Sid Idris, *Topological sensitivity and shape optimization for the Stokes equations*, SIAM J. Control Optim., 43 (2004), pp. 1–31.

[25] P. Germain and P. Muller, *Introduction à la mécanique des milieux continus*, Masson, Paris, 1994.

[26] M. Hassine and M. Masmoudi, *The topological asymptotic expansion for the quasi-Stokes problem*, ESAIM Control Optim. Calc. Var. J., 10 (2004), pp. 478–504.

[27] J. Jacobsen, N. Olhoff, and E. Ronholt, *Generalized Shape Optimization of Three-Dimensionnal Structures Using Materials with Optimum Microstructures*, Technical report, Institute of Mechanical Engineering, Aalborg University, Aalborg, Denmark, 1996.

[28] M. Masmoudi, *Outils pour la conception optimale de formes*, Thése d'état, Université de Nice, Nice, France, 1987.

[29] M. Masmoudi, *The topological asymptotic*, in Computational Methods for Control Applications, H. Kawarada and J. Periaux, eds., International Séries GAKUTO, 2002, CD-ROM.

[30] M. Masmoudi, J. Pommier, and B. Samet, *The topological asymptotic expansion for the Maxwell equations and some applications*, J. Inverse Problems, 21 (2005), pp. 547–564.

[31] F. Murat and L. Tartar, *Calcul des variations et homogénéisation*, in Les méthodes de l'homogénéisation: Théorie et Applications en Physique, Eyrolles, Paris, 1985, pp. 319–369.

[32] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer, Berlin, 1984.

[33] J. Pommier and B. Samet, *The topological asymptotic for the Helmholtz equation with Dirichlet condition on the boundary of an arbitrarily shaped hole*, SIAM J. Control Optim., 43 (2004), pp. 899–921.

[34] J. Simon, *Domain variation for drag in Stokes flow*, in Control Theory of Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 159, X. Li and J. Yang, eds., Springer, Berlin, 1991, pp. 28–42.

[35] A. Schumacher, *Topologieoptimierung von bauteilstrukturen unter verwendung von lopchpositionierungkrieterien*, thesis, Universität Gesamthochschule Siegen, Siegen, Germany, 1995.

[36] J. Sokołowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 1988–2010

COPRIME FACTORIZATION AND DYNAMIC STABILIZATION OF TRANSFER FUNCTIONS*

KALLE M. MIKKOLA†

Abstract. It is known that a matrix-valued transfer function P has a stabilizing dynamic controller Q (i.e., $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$) iff P has a right (or left) coprime factorization. We show that the same result is true in the operator-valued case. Thus, the standard Youla–Bongiorno parameterization applies to every dynamically stabilizable function. We then derive further equivalent conditions, one of them being that P has a stabilizing controller with internal loop; this and some others are new even in the scalar-valued case. We also establish certain related results. For example, we extend the classical results on coprime factorization and partial feedback (measurement-feedback) stabilization to nonrational transfer functions. All our results apply in both discrete- and continuous-time settings, except that in the latter it is not clear whether the controller Q can always be chosen so that it is "continuous-time proper" (holomorphic and bounded on a right half-plane) unless, e.g., $P(z) \to 0$ as $\operatorname{Re} z \to +\infty$.

Key words. dynamic stabilization, internal stabilization, right coprime factorization, measurement feedback, dynamic partial feedback, dynamically stabilizing controllers with internal loop, operator-valued transfer functions, infinite-dimensional systems

AMS subject classifications. 93D15, 93D25, 47A56

DOI. 10.1137/050641740

1. Introduction.

In this introductory section we present our main results for discrete-time transfer functions (those defined on a subset of the unit disc $\mathbb{D} := \{z \in \mathbb{C} \mid |z| < 1\}$). Corresponding results for continuous-time functions (those defined on the right half-plane) and others are given in section 7.

Let $\mathtt{U}$, $\mathtt{W}$, $\mathtt{Y}$, and $\mathtt{Z}$ be complex Hilbert spaces. By $\mathcal{B}(\mathtt{U},\mathtt{Y})$ we denote bounded linear operators $\mathtt{U} \to \mathtt{Y}$ and by $\mathrm{H}^\infty(\mathtt{U},\mathtt{Y})$ we denote bounded holomorphic functions $\mathbb{D} \to \mathcal{B}(\mathtt{U},\mathtt{Y})$ with supremum norm. We set $\mathcal{B}(\mathtt{U}) := \mathcal{B}(\mathtt{U},\mathtt{U})$, $\mathrm{H}^\infty(\mathtt{U}) := \mathrm{H}^\infty(\mathtt{U},\mathtt{U})$, $\mathcal{GB} := \{F \in \mathcal{B} \mid \text{there exists } F^{-1} \in \mathcal{B}\}$, and $\mathcal{G}\,\mathrm{H}^\infty := \{F \in \mathrm{H}^\infty \mid \text{there exists } F^{-1} \in \mathrm{H}^\infty\}$. By I or $I_{\mathtt{U}}$ we denote the identity operator $I \in \mathcal{B}(\mathtt{U})$ (or the corresponding constant function $I \in \mathrm{H}^\infty(\mathtt{U})$).

A holomorphic function P ("the plant") defined on a neighborhood of the origin is called *proper*. It is *strictly proper* if $P(0) = 0$. We identify a holomorphic function on a disc $r\mathbb{D} = \{z \in \mathbb{C} \mid |z| < r\}$ with its restriction to any open subset of $r\mathbb{D}$.

A proper $\mathcal{B}(\mathtt{Y},\mathtt{U})$-valued function Q is called a (dynamic feedback) *proper stabilizing controller* for a proper $\mathcal{B}(\mathtt{U},\mathtt{Y})$-valued function P if the "input-to-error" map $E : \left[\begin{smallmatrix} u_{\mathrm{in}} \\ y_{\mathrm{in}} \end{smallmatrix}\right] \mapsto \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$ in Figure 1.1 is in H^∞.[1] The map E is obviously given by

*Received by the editors October 3, 2005; accepted for publication (in revised form) July 12, 2006; published electronically January 8, 2007. This work was supported by the Academy of Finland under grant 203946 and by the Magnus Ehrnrooth Foundation.

http://www.siam.org/journals/sicon/45-6/64174.html

†Helsinki University of Technology, Institute of Mathematics, P.O. Box 1100, FIN-02015 HUT, Espoo, Finland (Kalle.Mikkola@iki.fi).

[1]This means that some $E \in \mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$ satisfies $E\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] = \left[\begin{smallmatrix} I & 0 \\ 0 & I \end{smallmatrix}\right] = \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] E$ on a neighborhood of 0. By a direct computation, (1.1) follows (on a neighborhood of 0). Recall that the inverse of a holomorphic operator-valued function is always holomorphic. (This kind of algebraic, function-theoretic and other well-known results used in this article can be found in our generality in [11, Appendices A and D].)

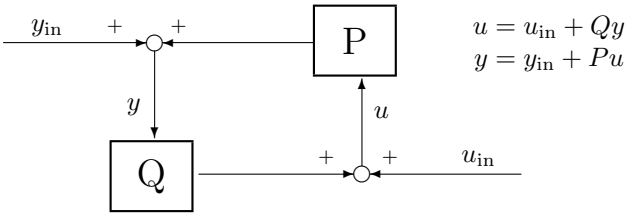


FIG. 1.1. *Controller Q for the transfer function P.*

$$E := \begin{bmatrix} I & -Q \\ -P & I \end{bmatrix}^{-1} = \begin{bmatrix} (I-QP)^{-1} & Q(I-PQ)^{-1} \\ P(I-QP)^{-1} & (I-PQ)^{-1} \end{bmatrix}. \tag{1.1}$$

(Observe that then P is also a proper stabilizing controller for Q.)

Two functions $M, N \in \mathrm{H}^\infty$ are called (Bézout) *r.c.* (right coprime) if $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ is left-invertible in H^∞, i.e., if there exist $\tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$ satisfying the "Bézout identity"

$$\tilde{X}M - \tilde{Y}N \equiv I. \tag{1.2}$$

We call the factorization $P = NM^{-1}$ a *r.c.f.* (right coprime factorization) of P if $N \in \mathrm{H}^\infty(\mathtt{U}, \mathtt{Y})$ and $M \in \mathrm{H}^\infty(\mathtt{U})$ are r.c., $M(0) \in \mathcal{GB}$, and $P = NM^{-1}$ (near 0).

The following is our main result.

THEOREM 1.1 (dynamic feedback stabilization). *The following are equivalent for any proper $\mathcal{B}(\mathtt{U}, \mathtt{Y})$-valued function P:*

(i) *P has a strictly proper stabilizing controller.*
(ii) *P has a proper stabilizing controller.*
(iii) *P has a stabilizing controller with internal loop.*[2]
(iv) *P has a r.c.f.*
(v) *$\left[\begin{smallmatrix} P & 0 \\ 0 & I_{\mathtt{Z}} \end{smallmatrix}\right]$ has a r.c.f. for some (hence any) Hilbert space $\mathtt{Z}$.*

Assume that P has a r.c.f. $P = NM^{-1}$. Then $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right] \in \mathrm{H}^\infty(\mathtt{U}, \mathtt{U} \times \mathtt{Y})$ can be extended to an invertible element of $\mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$, say $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$. Denote its inverse by $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] \in \mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$. Then all stabilizing controllers for P are given by the Youla(–Bongiorno) parameterization[3]

$$Q = (Y + MV)(X + NV)^{-1} \qquad (= (\tilde{X} + V\tilde{N})^{-1}(\tilde{Y} + V\tilde{M})), \tag{1.3}$$

where $V \in \mathrm{H}^\infty(\mathtt{Y}, \mathtt{U})$ is arbitrary (the controller is proper iff $(X + NV)^{-1}$ is proper, or equivalently, iff $(\tilde{X} + V\tilde{N})^{-1}$ is proper). The map $V \mapsto Q$ is one-to-one.

If P is strictly proper, then all these controllers are proper.

Usually one excludes the values of the parameter V that make the controller (1.3) nonproper. However, sometimes only such controllers possess the properties that one would like to obtain in practical applications [4]. To also include such controllers, the

[2]These will be defined in section 3. They may be nonproper.

[3]For some functions P and V, the inverse $(X + NV)^{-1}$ in (1.3) need not exist at the origin (or anywhere; e.g., $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, $V = 0$). Even so, the "nonproper" controller (1.3) can be interpreted as a "stabilizing controller with internal loop," as described in section 3, where also properness is explained in detail. Nevertheless, for each P (and $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$), some $V \in \mathrm{H}^\infty$ makes $(X + NV)(0)$ invertible in $\mathcal{B}(\mathtt{Y})$. The parameterization (1.3) covers all stabilizing controllers with internal loop in the sense described in section 3. Moreover, every proper stabilizing controller equals exactly one of these Q on a neighborhood of the origin.

theory of "controllers with internal loop" (which cover both the proper and nonproper controllers) was developed in [32] and [4]. Also nonproper controllers with internal loop can be physically realized. In section 3 we shall define them and explain their relation to proper controllers.

Further necessary and sufficient conditions for (i) will be presented later, particularly in Theorem 2.1 and Proposition 2.2. One such condition is the existence of a stabilizable and detectable realization. Conditions (iii) and Theorem 2.1(ii') are weaker forms of (ii). Their equivalence to (ii) means that if P is dynamically stabilizable in any reasonable sense, then it is dynamically stabilizable in the standard sense (possibly by a different, nonequivalent controller).

By combining the above results with [27], [30], and [7], we obtain the following result.

COROLLARY 1.2 (matrix-valued case). *Assume that* $\dim \mathtt{U} < \infty$ *and* $\dim \mathtt{Y} < \infty$. *Then also the following conditions are equivalent to* (i) *of Theorem* 1.1 *for a proper* $\mathcal{B}(\mathtt{U},\mathtt{Y})$*-valued function* P*:*

(vi) P *has a stable (hence proper) stabilizing controller* ($Q \in \mathrm{H}^\infty(\mathtt{Y},\mathtt{U})$).

(vii) $P = NM^{-1}$, *where* $N, M \in \mathrm{H}^\infty$, $N^*N + M^*M \geq \epsilon I$ *on* $\mathbb{D}$, $\epsilon > 0$, *and* $\det M \not\equiv 0$.

(The corona condition in (vii) is not sufficient for coprimeness in the operator-valued case [26]. It is not known whether (vi) is necessary in general.)

For rational transfer functions, (i)–(vii) always hold, and also the rest of Theorem 1.1 is well known [6]. The study of corresponding results for nonrational functions started in the 1970s and soon became intensive. An introduction to coprime factorization and dynamic stabilization of infinite-dimensional systems can be found in, e.g., [5] or [31]. Several sufficient conditions for some of the conditions (i)–(v) have been established earlier, but our proof of the equivalence would not have been possible without the results in [23], [32], [11], [14], [2], and [16].

In the matrix-valued case, the implication (ii)⇒(iv) was independently established in [9] and [22] and the enhanced converse (iv)⇒(vi) in [27] (in the scalar-valued case, which is equivalent to the matrix-valued case, by [30, Theorem 3]; the exact statement can be found in [19]). The "Carleson corona theorem" (iv)⇔(vii) was extended to the matrix-valued case in [7] (see [28] for the operator-valued case, where (vii)⇒(iv) is not true without additional assumptions).

In the general case, the implication (iv)⇒(iii) was established in [32] and [4], and (iv)⇒(v) and (i)⇒(ii)⇒(iii) are trivial. The existence of $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U}\times\mathtt{Y})$ is from [16] (based on [28] and [12]); the matrix-valued case is a well-known consequence of Tolokonnikov's lemma [25]. The Youla parameterization (including the "if" part of the properness of Q) is straightforward [4]. The "only if" part of properness, implications (iii)⇒(v)⇒(iv)⇒(i), and the strictly proper case are new (except that (iv)⇒(i) was already known in the matrix-valued case). The differences between continuous- and discrete-time results are otherwise insignificant, but properness and strict properness become more complicated in continuous time; see section 7 for details.

With certain other commutative unital rings in place of $\mathrm{H}^\infty(\mathbb{C})$, Theorem 1.1 becomes false. Related results for such settings are given by Quadrat [19], [18], [20] in the matrix-valued case.

In section 2 we present further conditions that are equivalent to (i), such as coprime factorization or stabilization with invertibility at some other $\alpha \in \mathbb{D}$ instead of 0. In section 3 we define controllers with internal loop, present corresponding details of Theorem 1.1, and develop related new results. The results in sections 2 and 3 are needed in the proof of Theorem 1.1, but they are also important by themselves.

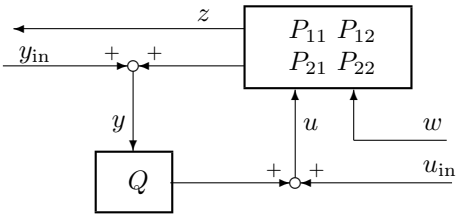


FIG. 1.2. *DPF-controller Q for P.*

In section 4 we present analogous results for "measurement feedback" or dynamic partial feedback, where the controller can use only a part of the output and can affect only a part of the input of $P = \left[\begin{smallmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{smallmatrix}\right]$, where $P(z) \in \mathcal{B}(\mathtt{U} \times \mathtt{W},\ \mathtt{Z} \times \mathtt{Y})$, as in Figure 1.2. We obtain direct generalizations of the classical results, such as those in [6] or [8]. In particular, we show that if P is stabilizable by dynamic partial feedback, then a $\mathcal{B}(\mathtt{Y}, \mathtt{U})$-valued controller Q stabilizes P by dynamic partial feedback iff it stabilizes P_{21} by dynamic feedback.

In section 5 we observe that practically all our results also hold for "power stabilization" (or "exponential stabilization" in the continuous-time setting of section 7), mutatis mutandis, where the "closed-loop" map (1.1) is required to be holomorphic on an open set that contains $\overline{\mathbb{D}}$. In section 6 we show that even if we allow the domain of Q to be an arbitrary region, we meet no ambiguity with holomorphic extensions and the identification of controllers.

In section 7 we establish our results in the continuous-time setting, where the properness notion is different. Proofs and some further results are given in the appendices.

In our generality, corresponding state-space results can be found in [33], [11], and [24] (and [32]), where many assumptions can be weakened, by our results. Robust stabilization with state-space results are given in [1]. Further state-space results will be presented in a subsequent article by the author.

2. Dynamic stabilization. In this section we show how any reasonable variants of the above conditions (i)–(v) are equivalent to (i). We also present realization-based conditions that are equivalent to (i).

In the matrix-valued case, as in (vii), one need not care where M or $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1}$ is invertible, since it is invertible a.e. anyway (if it is invertible somewhere). In Theorem 2.1 we show that invertibility at any reasonable point is sufficient also in the operator-valued case. The definitions below are used to formulate these facts.

Let $\alpha \in \mathbb{D}$. We call NM^{-1} an *α-r.c.f. of P* if $N, M \in \mathrm{H}^\infty$ are r.c., $M(\alpha) \in \mathcal{GB}(\mathtt{U})$, and $NM^{-1} = P$ on a neighborhood of α. We call $\tilde{M}^{-1}\tilde{N}$ an *α-l.c.f. of P* if $\tilde{N}, \tilde{M} \in \mathrm{H}^\infty$ are *l.c.* (i.e., $\tilde{M}X - \tilde{N}Y = I$ for some $X, Y \in \mathrm{H}^\infty$), $\tilde{M}(\alpha) \in \mathcal{GB}(\mathtt{U})$, and $\tilde{M}^{-1}\tilde{N} = P$ on a neighborhood of α. We call $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$ an *α-d.c.f. of P* if $M(\alpha) \in \mathcal{GB}(\mathtt{U})$ and $P = NM^{-1}$ on a neighborhood of α (it follows that $P = NM^{-1}$ is an α-r.c.f. and $P = \tilde{M}^{-1}\tilde{N}$ is an α-l.c.f., where $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] := \left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]^{-1}$ [11]; conversely, any α-r.c.f. and α-l.c.f. can be extended to an α-d.c.f., by Lemma 3.4). A 0-d.c.f. (resp., 0-l.c.f.) is called a *d.c.f.* (resp., *l.c.f.*).

Now we can present further equivalent conditions (see section 3 for (iii')).

THEOREM 2.1 (dynamic feedback stabilization). *Assume that $\Omega_P \subset \mathbb{D}$ is open and connected, $P : \Omega_P \to \mathcal{B}(\mathtt{U}, \mathtt{Y})$ is holomorphic, and $0, \alpha, \beta \in \Omega_P$. Then the follow-*

ing conditions are equivalent to (iv) *of Theorem* 1.1*:*

(iv') *P has an α-r.c.f.*

(iv") *P has an α-l.c.f.*

(iv"') *P has an α-d.c.f.*

(ii') *For some open and connected $\Omega_Q \subset \Omega_P$ there exists a holomorphic function $Q : \Omega_Q \to \mathcal{B}(\mathtt{Y},\mathtt{U})$ such that $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$.*

(ii") *For some neighborhood Ω_Q of α, condition* (ii') *holds with $Q(\alpha) = 0$.*

(iii') *P has a stabilizing canonical controller.*

Any α-r.c.f. of P is a β-r.c.f. of P. The same holds with "l.c.f." or "d.c.f." in place of "r.c.f."

Note that, by duality, we get "left results" from all "right results" of this article (because, e.g., $P = \tilde{M}^{-1}\tilde{N}$ is an α-l.c.f. iff $P^{\mathrm{d}} = \tilde{N}^{\mathrm{d}}(\tilde{M}^{\mathrm{d}})^{-1}$ is an $\bar{\alpha}$-r.c.f., where the *dual* P^{d} is defined by $P^{\mathrm{d}}(s) := P(\bar{s})^*$). See section 6 for further variants of (ii').

We recall the following from [16] (which contains the definitions of (viii)–(viii"')).

PROPOSITION 2.2 (realizations). *Assume that P is a proper $\mathcal{B}(\mathtt{U},\mathtt{Y})$-valued function. Then also the following are equivalent to* (i) *of Theorem* 1.1*:*

(viii) *P has a jointly stabilizable and detectable realization.*

(viii') *P has a stabilizable and detectable realization.*

(viii") *P has an output-stabilizable and input-detectable realization.*

(viii"') *P has a realization Σ such that Σ and its dual satisfy the finite cost condition.*

See [2] or [16] for an equivalent condition in terms of Riccati equations, which also yield a constructive formula for the r.c.f. The original proof of "(viii')⇒(iv)" is due to [2], and that of "(viii)⇔(iv"')" due to [23], both in continuous time.

By combining Proposition 2.2 and Theorem 1.1 with [11] one can obtain further equivalent conditions, such as having a dynamically stabilizable realization. For (continuous-time) exponential dynamic stabilization of realizations, the necessity of exponential stabilizability and detectability was shown in [33]; their sufficiency follows from [17] (or [33]), Remark 5.1, and Theorems 1.1 and 7.3.

Constructive formulae for doubly coprime factorizations in terms of realizations can be found in, e.g., [2], [3], and [11] under different assumptions; in [32], [11], and [1] formulae for stabilizing dynamic controllers are given. They also provide further historical remarks. For constructive formulae for mere r.c.f.'s, see also the end of section 7.

3. Controllers with internal loop. In this section we present certain results on controllers with internal loop and explain the rest of Theorem 1.1. As before, we work in the discrete-time setting, but we show in section 7 that practically everything below holds in the continuous-time setting too.

Controllers with internal loop were defined in [32] both to complete the theory of dynamic stabilization of nonrational transfer functions and to cover also the "short circuit control"-type applications. Their theory has been further developed in [4], [33], and [11]. As explained in [32] and [4], without them some aspects of the standard theory for finite-dimensional systems cannot be satisfactorily generalized to general infinite-dimensional systems. For example, the standard observer-based controller need not have a proper transfer function, but it can be identified with a proper 2×2-matrix-valued transfer function [32, Example 6.5], which has a well-posed realization. See also the rational SISO example at the end of this section.

We start this section from the definitions and then explain the correspondence to the proper controllers presented in the introduction.

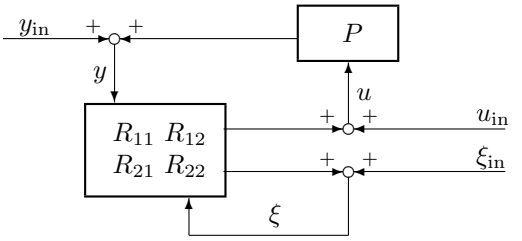


FIG. 3.1. *Controller R with internal loop for P.*

We say that R is a (possibly nonproper) *stabilizing controller with internal loop* for P if $R = \left[\begin{smallmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{smallmatrix}\right]$ is a proper $\mathcal{B}(\mathtt{Y} \times \Xi, \mathtt{U} \times \Xi)$-valued function for some Hilbert space Ξ and

$$(I - P^R)^{-1} \in \mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y} \times \Xi), \quad \text{where } P^R = \begin{bmatrix} 0 & R_{11} & R_{12} \\ P & 0 & 0 \\ 0 & R_{21} & R_{22} \end{bmatrix}. \tag{3.1}$$

Note that $(I - P^R)^{-1}$ maps $\left[\begin{smallmatrix} u_{\text{in}} \\ y_{\text{in}} \\ \xi_{\text{in}} \end{smallmatrix}\right] \mapsto \left[\begin{smallmatrix} u \\ y \\ \xi \end{smallmatrix}\right]$ in Figure 3.1. Thus, R is stabilizing iff the maps $\left[\begin{smallmatrix} u_{\text{in}} \\ y_{\text{in}} \\ \xi_{\text{in}} \end{smallmatrix}\right] \mapsto \left[\begin{smallmatrix} u \\ y \\ \xi \end{smallmatrix}\right]$ are "well-posed and stable."

If $R = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$, then R is completely equivalent to the stabilizing controller Q. Thus, the proper controllers presented in the introduction essentially form a subset of the controllers with internal loop.

In general, R corresponds to "$R_{11} + R_{12}(I - R_{22})^{-1}R_{21}$" (cf. Lemma 3.2); this "function" need not be proper (we may even have $R_{22} \equiv I$). In the nonproper case the ξ-loop in Figure 3.1 becomes ill-posed if R is disconnected from P; i.e., physically one must connect R to P before closing the internal loop.

A nonproper controller is a proper controller for an extended system.

LEMMA 3.1. *A proper $\mathcal{B}(\mathtt{Y} \times \Xi, \mathtt{U} \times \Xi)$-valued function R is a stabilizing controller with internal loop for P iff R is a stabilizing controller for $P_I := \left[\begin{smallmatrix} P & 0 \\ 0 & I_\Xi \end{smallmatrix}\right]$.*

(This follows because $\left[\begin{smallmatrix} I & -R \\ -P_I & I \end{smallmatrix}\right]^{-1}$ consists of $(I - P^R)^{-1}$ and of some copies of its elements, as observed in [11, Proposition 7.2.5(c)]. An alternative proof is to observe that the equations that determine the latter reduce to those that determine the former. In fact, this is rather obvious, since I_Ξ corresponds to the identity feedthrough of ξ in Figure 3.1.)

Two stabilizing controllers with internal loop are considered *equivalent* for P iff they lead to the same closed-loop map $\left[\begin{smallmatrix} u_{\text{in}} \\ y_{\text{in}} \end{smallmatrix}\right] \mapsto \left[\begin{smallmatrix} u \\ y \end{smallmatrix}\right]$ (i.e., if the (1–2, 1–2)-blocks of corresponding $(I - P^R)^{-1}$'s are equal), even if the maps from ξ_{in} and the maps to ξ (i.e., those describing the internal loop in the controller) would differ.

In the lemma below we show that R corresponds to a proper controller Q iff the internal loop of R can be closed (while R is disconnected from P).

LEMMA 3.2 (proper R). *Assume that $R = \left[\begin{smallmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{smallmatrix}\right]$ is a stabilizing controller with internal loop for P.*

Then R is equivalent to a stabilizing controller with internal loop of form $\tilde{R} = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$ iff $I - R_{22}(0) \in \mathcal{GB}$. If $I - R_{22}(0) \in \mathcal{GB}$, then the unique solution is given by $Q = R_{11} + R_{12}(I - R_{22})^{-1}R_{21}$.

(Note that then $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$, as in the introduction, and that if we close the internal loop of R, then its top-left block becomes $R_{11} + R_{12}(I - R_{22})^{-1}R_{21}$.)

Any such R is called a *proper* stabilizing controller (with internal loop) for P (and R is identified with Q). Since equivalence is an equivalence relation, R is equivalent to a proper stabilizing controller with internal loop iff R is proper, by Lemma 3.2.

If $Y \in \mathrm{H}^\infty(\mathtt{Y},\mathtt{U})$ and $X \in \mathrm{H}^\infty(\mathtt{U})$ are r.c., then $R := \left[\begin{smallmatrix} 0 & Y \\ I & I-X \end{smallmatrix}\right]$ is called a *canonical controller* (see [4] or [11]; in [11], the term controller with a coprime internal loop was used). Sometimes we denote it by YX^{-1}, as in the Youla parameterization (1.3) above. In particular, we say that YX^{-1} *stabilizes* P iff $\left[\begin{smallmatrix} 0 & Y \\ I & I-X \end{smallmatrix}\right]$ is a stabilizing controller with internal loop for P.

(If YX^{-1} is a stabilizing canonical controller for P, then it is equivalent to $\left[\begin{smallmatrix} 0 & I \\ \tilde{Y} & I-\tilde{X} \end{smallmatrix}\right]$ (or $\tilde{X}^{-1}\tilde{Y}$), where $\tilde{X}$ and $\tilde{Y}$ are obtained by $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] := \left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]^{-1}$ for any r.c.f. NM^{-1} of P, as one observes from Lemma 3.5 and its dual.)

Modulo equivalence, there are no other controllers than the canonical ones.

LEMMA 3.3 (equivalent canonical controller). *Let R be a stabilizing controller with internal loop for P. Then some stabilizing canonical controller $\tilde{X}^{-1}\tilde{Y}$ for P is equivalent to R (and so is one of form YX^{-1}).*

The Youla parameterization (1.3) gives all stabilizing canonical controllers for P. Here we have identified the canonical controllers that are equivalent (see above), i.e., the ones determined by $\left[\begin{smallmatrix} Y \\ X \end{smallmatrix}\right] V$ for a fixed r.c. pair $\left[\begin{smallmatrix} Y \\ X \end{smallmatrix}\right]$ and arbitrary $V \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U})$. Any stabilizing controller with internal loop is equivalent to exactly one stabilizing canonical controller, so actually all stabilizing controllers with internal loop are covered by (1.3) modulo equivalence. In particular, this parameterization contains all proper stabilizing controllers, by Lemma 3.2. (Indeed, any proper stabilizing controller is equivalent to one of form $\tilde{R} := \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and to a canonical controller $R := \left[\begin{smallmatrix} 0 & Y \\ I & I-X \end{smallmatrix}\right]$. By Lemma 3.2, $X(0) = I - R_{22}(0) \in \mathcal{GB}$ and $Q = YX^{-1}$.)

Since any r.c.f. and l.c.f. can be extended to a d.c.f. (Lemma 3.4), we can apply (1.3) when we have either.

A "Bézout identity" $\tilde{X}M - \tilde{Y}N = I$ (or $\tilde{M}X - \tilde{N}Y = I$) can always be extended to a d.c.f.

LEMMA 3.4 (r.c.f.→d.c.f.). *Let $M, N, \tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$ be such that NM^{-1} is a $\mathcal{B}(\mathtt{U},\mathtt{Y})$-valued r.c.f. and $\tilde{X}M - \tilde{Y}N = I$. Then, for any l.c.f. $\tilde{M}^{-1}\tilde{N}$ of NM^{-1}, there exist $X, Y \in \mathrm{H}^\infty$ such that $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right]^{-1} \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U}\times\mathtt{Y})$.*

(This follows from the proof of [23, Lemma 4.3(iii)]; observe from Theorem 2.1 that the l.c.f. necessarily exists.)

A canonical controller YX^{-1} stabilizes P iff $\left[\begin{smallmatrix} Y \\ X \end{smallmatrix}\right]$ can be extended to a d.c.f. of P. We state the dual result here.

LEMMA 3.5. *Let P be a proper $\mathcal{B}(\mathtt{U},\mathtt{Y})$-valued function and $R' = \left[\begin{smallmatrix} 0 & I \\ \tilde{Y} & I-\tilde{X} \end{smallmatrix}\right] \in \mathrm{H}^\infty(\mathtt{Y}\times\mathtt{U},\, \mathtt{U}\times\mathtt{U})$.*

Then R' is a stabilizing controller with internal loop for P iff for some (hence any) r.c.f. NM^{-1} of P we have $\tilde{X}M - \tilde{Y}N \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U})$, or equivalently, iff P has a r.c.f. NM^{-1} such that $\tilde{X}M - \tilde{Y}N = I$.

Assume that R' is a stabilizing controller with internal loop for P. Then there exists a d.c.f. $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty(\mathtt{U}\times\mathtt{Y})$ of P with these particular $\tilde{X}$ and $\tilde{Y}$. Moreover, for any such d.c.f., we have

$$\begin{bmatrix} I & -R_{11} & -R_{12} \\ -P & I & 0 \\ 0 & -R_{21} & I-R_{22} \end{bmatrix}^{-1} = \begin{bmatrix} Y\tilde{N}+I & M\tilde{Y} & * \\ X\tilde{N} & N\tilde{Y}+I & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} M\tilde{X} & Y\tilde{M} & * \\ N\tilde{X} & X\tilde{M} & * \\ * & * & * \end{bmatrix}, \tag{3.2}$$

where R is any stabilizing controller with internal loop that is equivalent to $\tilde{X}^{-1}\tilde{Y}$. If $R = \left[\begin{smallmatrix} 0 & I \\ \tilde{Y} & I-\tilde{X} \end{smallmatrix}\right]$, then

$$\begin{bmatrix} I & -R_{11} & -R_{12} \\ -P & I & 0 \\ 0 & -R_{21} & I-R_{22} \end{bmatrix}^{-1} = \begin{bmatrix} M\tilde{X} & M\tilde{Y} & M \\ N\tilde{X} & N\tilde{Y}+I & N \\ M\tilde{X}-I & M\tilde{Y} & M \end{bmatrix}. \tag{3.3}$$

(Thus, any such stabilizing R' is actually a canonical controller. However, also some noncanonical functions $R' = \left[\begin{smallmatrix} 0 & I \\ \tilde{Y} & I-\tilde{X} \end{smallmatrix}\right] \in \mathrm{H}^\infty$ do lead to (3.2), but the $*$'s (which denote unimportant entries) do not become stable unless $\tilde{X}$ and $\tilde{Y}$ are l.c.)

We record here an obvious consequence of the dual of Lemma 3.5.

COROLLARY 3.6. *Let P be a proper $\mathcal{B}(\mathtt{U},\mathtt{Y})$-valued function. If a canonical controller YX^{-1} stabilizes P, then P has a r.c.f. and any r.c.f. NM^{-1} of P satisfies $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty$. Conversely, if there exists $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty$ such that $M(0) \in \mathcal{GB}(\mathtt{U})$ and $P = NM^{-1}$, then YX^{-1} stabilizes P.*

If P is strictly proper, then the X in Corollary 3.6 is necessarily invertible at 0.

LEMMA 3.7. *If P is strictly proper, then any stabilizing controller with internal loop for P is proper.*

(This follows from Lemmata 3.2, 3.3, and 3.5, because $P(0) = 0 \;\Rightarrow\; N(0) = 0 \;\Rightarrow\; \tilde{X}(0)M(0) = I \;\Rightarrow\; \tilde{X}(0) = M(0)^{-1} \in \mathcal{GB}(\mathtt{U})$.)

Example. The function $P = NM^{-1} = (1-z)^{-1}$, where $N = 1$, $M = 1-z$, has the stabilizing controller $R := \left[\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right]$ with internal loop (the canonical controller $\tilde{\mathscr{X}} = 0$, $\tilde{\mathscr{Y}} = -1$). Since $1 - R_{22}$ is nowhere invertible, this controller is not equivalent to any proper controller, by Lemma 3.2. However, by Theorem 1.1, there are also proper stabilizing controllers for P (e.g., $Q(z) = -z$). The continuous-time equivalent of this example was presented in [32, p. 6], where the nonproper controller was shown to be the natural engineering solution (short circuit tracking) for the problem.

Notes for section 3. The "if" part of Lemma 3.2 is from [32]. With the additional assumption that P has a d.c.f., Lemma 3.3 is contained in [4]. However, the proof in [4] is seven pages long, so we present a short, self-contained proof in Appendix A. Also most of Lemma 3.5 can be found in [4]. For further similar results, see [11]; for practical examples, see [32] and [4]. Lemma 3.7 becomes less obvious and even more important in the continuous-time setting of Theorem 7.3.

4. Partial feedback. In this section we treat *dynamic partial feedback (DPF)*, where the controller Q sees only a part of the output and can affect only a part of the input.

Throughout this section we assume that P is a proper $\mathcal{B}(\mathtt{U}\times\mathtt{W},\ \mathtt{Z}\times\mathtt{Y})$-valued function. A proper $\mathcal{B}(\mathtt{Y},\mathtt{U})$-valued function Q is called a *stabilizing DPF-controller* for P if $\left[\begin{smallmatrix} 0 & Q \\ 0 & 0 \end{smallmatrix}\right]$ is a stabilizing controller for P. This obviously corresponds to Figure 1.2. Analogously, a $\mathcal{B}(\mathtt{Y}\times\Xi,\ \mathtt{U}\times\Xi)$-valued proper function R is called a *stabilizing DPF-controller with internal loop* for P if

$$R_{\mathrm{DPF}} := \begin{bmatrix} 0 & R_{11} & R_{12} \\ 0 & 0 & 0 \\ 0 & R_{21} & R_{22} \end{bmatrix} \quad \text{(whose values lie in } (\mathtt{Z}\times\mathtt{Y}\times\Xi,\ \mathtt{U}\times\mathtt{W}\times\Xi)) \tag{4.1}$$

is a stabilizing controller with internal loop for P; see Figure 4.1. If such an R exists, then we call P *DPF-stabilizable* (with internal loop). (Further details are given in [11, section 7.3]. Observe that the [DPF-]controller $R = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$ with internal loop functions exactly as the [DPF-]controller Q; we identify the two.)

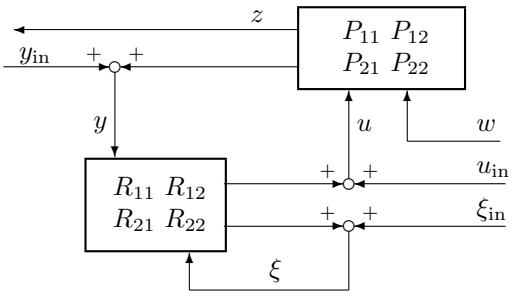


FIG. 4.1. *DPF-controller R with internal loop for P.*

We call two stabilizing controllers with internal loop for P (say, R and R') *equivalent* if they lead to same maps $u_{\text{in}}, y_{\text{in}} \mapsto u, y$ (or equivalently, to same maps $u_{\text{in}}, w, y_{\text{in}} \mapsto u, y, z$, or equivalently, if R_{DF} and R'_{DF} are equivalent for P, or equivalently, if R and R' are equivalent for P_{21}; see [11, Lemma 7.3.8] for this equivalence).

DPF is the standard setting in the general (four-block) H^∞ regulator problem (see [11, Chapter 12] for this general case with internal loops).

Note that the second input (column) of P is the exogenous input (or disturbance) and the first input (column) is the one connected to the controller output, not vice versa (both variants can be found in the literature; the other choice would move the I's out of the diagonal in Theorem 4.2 below).

With the aid the Theorem 1.1, we can derive the following two theorems, which are direct generalizations of well-known results for rational functions. The first theorem reduces DPF-stabilization problems to ordinary dynamic stabilization problems.

THEOREM 4.1 (P iff P_{21}). *Assume that P is DPF-stabilizable. Then a proper $\mathcal{B}(\mathtt{Y} \times \Xi, \mathtt{U} \times \Xi)$-valued function R is a stabilizing DPF-controller with internal loop for P iff R is a stabilizing controller with internal loop for P_{21}.*

In particular, a proper $\mathcal{B}(\mathtt{Y}, \mathtt{U})$-valued function Q is a stabilizing DPF-controller for P iff Q is a stabilizing controller for P_{21}. It also follows that every stabilizing DPF-controller with internal loop for P is equivalent to one of the canonical controllers (for P_{21}) given by the Youla parameterization.

Observe that $P_{21} : u \mapsto y - y_{\text{in}}$ is the control-to-measurement part of P.

A function is DPF-stabilizable iff it has a coprime factorization "through P_{21}."

THEOREM 4.2 (DPF). *The following are equivalent:*

(i) *P has a strictly proper stabilizing DPF-controller.*
(ii) *P has a proper stabilizing DPF-controller.*
(iii) *P has a stabilizing DPF-controller with internal loop.*
(iv) *P has a r.c.f. of the form $P = \left[\begin{smallmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{smallmatrix}\right] \left[\begin{smallmatrix} M_{11} & M_{12} \\ 0 & I \end{smallmatrix}\right]^{-1}$ such that N_{21} and M_{11} are r.c.*
(v) *P has a l.c.f. of the form $P = \left[\begin{smallmatrix} I & \tilde{M}_{12} \\ 0 & \tilde{M}_{22} \end{smallmatrix}\right]^{-1} \left[\begin{smallmatrix} \tilde{N}_{11} & \tilde{N}_{12} \\ \tilde{N}_{21} & \tilde{N}_{22} \end{smallmatrix}\right]$ such that $\tilde{N}_{21}$ and $\tilde{M}_{22}$ are l.c.*

Assume (iv) *and* (v). *Then $N_{21}M_{11}^{-1}$ is a r.c.f. of P_{21} and $\tilde{M}_{22}^{-1}\tilde{N}_{21}$ is a l.c.f. of P_{21}.*

(Thus, the above r.c.f. of P contains a r.c.f. of P_{21} that can be used for the Youla parameterization of all stabilizing DPF-controllers with internal loop for P, by Theorem 4.1.)

One can also derive sufficient conditions for DPF-stabilizability in terms of realizations. One sufficient condition is the power stabilizability and detectability of the subsystem corresponding to P_{21}; see [11, Lemma 7.3.6(c)] for details.

Notes for section 4. For rational matrix-valued functions, the above two theorems can be found in, e.g., [6] or [8]; most of them were extended to the Callier–Desoer class in [5].

If (iv) holds, then the stabilizing DPF-controllers for P are, modulo equivalence, exactly the canonical controllers $\tilde{X}^{-1}\tilde{Y}$ for any $\tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$ such that $\tilde{X}M_{11} - \tilde{Y}N_{21} = I$, by Lemma 3.5. An equivalent characterization is those YX^{-1} for which $\left[\begin{smallmatrix} M_{11} & M_{12} & Y \\ M_{21} & M_{22} & 0 \\ N_{21} & N_{22} & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U} \times \mathtt{W} \times \mathtt{Y})$ for some (hence any) r.c.f. NM^{-1} of P [11, Lemma 7.3.22]. A third equivalent characterization is the Youla parameterization (for P_{21}). By Theorem 4.1, also all proper stabilizing DPF-controllers for P are contained in any of these (as in Theorem 1.1).

The coprimeness condition in (iv) cannot be weakened: the r.c.f. $P = NM^{-1} := \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} (z+1)/(z+2) & 0 \\ 0 & 1 \end{smallmatrix}\right]^{-1}$ is of the form $\left[\begin{smallmatrix} * & * \\ * & * \end{smallmatrix}\right] \left[\begin{smallmatrix} * & * \\ 0 & I \end{smallmatrix}\right]^{-1}$ (hence P and P_{21} both have a d.c.f. and thus have stabilizing controllers), but yet P is not DPF-stabilizable with internal loop, since $P_{11} = (z+2)/(z+1)$ is unstable ($\notin \mathrm{H}^\infty$) and unaffected by any DPF-controller, because $P_{21} = 0$. However, if some P is DPF-stabilizable, then any r.c.f. of that P of the form $P = \left[\begin{smallmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{smallmatrix}\right] \left[\begin{smallmatrix} M_{11} & M_{12} \\ 0 & I \end{smallmatrix}\right]^{-1}$ has N_{21}, M_{11} r.c. [11, Corollary 7.3.17]. A corresponding continuous-time example is given by $P(s) = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} s/(s+1) & 0 \\ 0 & 1 \end{smallmatrix}\right]^{-1}$.

A rational right factorization is r.c. iff M has no other zeros than the poles of P. The coprimeness condition on N_{21} and M_{11} says that as we multiply the zeros of P away by M_{11}, we do not introduce to $N_{21} = P_{21}M_{11}$ any new zeros (in addition to those of P_{21}), i.e., that the poles of P are also poles of P_{21}. In other words, this says that the poles of P are visible through P_{21}; other kind of poles of P could not be stabilized by partial feedback having access to P_{21} only, as is the case in the above example $P = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] \left[\begin{smallmatrix} (z+1)/(z+2) & 0 \\ 0 & 1 \end{smallmatrix}\right]^{-1}$.

Using the above two theorems and the other results in this article, [16], and [14], one could generalize to nonrational functions also the other classical results, as presented in, e.g., [6] or [8]. Part of this can be found in [11], whose Hypothesis 7.3.15 holds iff P is DPF-stabilizable with internal loop, by Theorem 4.2. This simplifies section 7.3 of [11] significantly; similarly, Theorem 1.1 simplifies sections 7.1 and 7.2. Partially the same applies to state-space results.

5. Power stabilization. One sometimes wants to power-stabilize systems or transfer functions (or stabilize exponentially in the continuous-time setting). In this section we observe that the "power-variants" of our results hold and follow easily. (However, from the "power-variants" one cannot obtain the original results. Moreover, in the power stabilization of systems, there are some results whose nonpower-stabilization variants are false.)

We write $N \in \mathrm{H}^\infty_{\mathrm{power}}$ if $N(r\cdot) \in \mathrm{H}^\infty$ for some $r > 1$ (i.e., $N \in \mathrm{H}^\infty$ has a holomorphic extension to an open disc that contains $\overline{\mathbb{D}}$). We define *power*-variants of the following definitions by replacing H^∞ by $\mathrm{H}^\infty_{\mathrm{power}}$: r.c., l.c., r.c.f., l.c.f., d.c.f., α-r.c.f., α-l.c.f., α-d.c.f., stabilizing [DPF-]controller, stabilizing [DPF-]controller with internal loop, canonical controller, DPF-stabilizable. Thus, e.g., a d.c.f. $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$ is a *power-d.c.f.* iff $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right], \left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty_{\mathrm{power}}$, and Q is a *power-stabilizing* controller for P iff $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty_{\mathrm{power}}$.

REMARK 5.1 (power stabilization). *With the above "power-"concepts in place of the original ones and* $\mathrm{H}^\infty_{\mathrm{power}}$ *in place of* H^∞*, Theorem* 2.1*, Lemmata* A.5*,* A.10*, and* A.11*, and the results of sections* 1*,* 3*, and* 4 *hold.*

(Also the power form of Proposition 2.2 holds in the sense explained in [16].)

Proof. This follows easily from the original results. For example, if P has a power-stable r.c.f. NM^{-1}, then $P(r\cdot) = N(r\cdot)M(r\cdot)^{-1}$ is a r.c.f. for some $r > 1$; hence then $P(r\cdot)$ has a proper stabilizing controller $\tilde{Q}$, and hence $Q := \tilde{Q}(r^{-1}\cdot)$ is power-stabilizing for P (because $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} I & -\tilde{Q} \\ -P(r\cdot) & I \end{smallmatrix}\right]^{-1}(r^{-1}\cdot) \in \mathrm{H}^\infty_{\mathrm{power}}$), by Theorem 1.1. ∎

Observe that a rational function is in H^∞ iff it does not have a pole in $\overline{\mathbb{D}}$, or equivalently, iff it is in $\mathrm{H}^\infty_{\mathrm{power}}$. Similarly, also finite-dimensional state-space stability coincides with state-space power stability. In the infinite-dimensional setting, both forms of stability are very popular.

6. Nonproper controller functions. In this section we study "stabilizing controllers" of the form of a possibly nonproper function. We also show that all such controllers are canonical controllers and explain how they relate to each other.

In the matrix-valued case, a factorization NM^{-1} is well defined everywhere on $\mathbb{D}$ except possibly for some isolated points (assuming that $M, N \in \mathrm{H}^\infty$, $\det M \not\equiv 0$). In the operator-valued case, one may easily end up with functions having disconnected domains. Moreover, in dynamic stabilization one often meets the question whether two functions can be identified when they coincide on the intersection of their domains.

We show that if a function Q stabilizes P in a reasonable sense, then $Q = YX^{-1}$, where YX^{-1} is a stabilizing canonical controller for P in the standard sense, and then YX^{-1} (on the subset of $\mathbb{D}$ where X^{-1} exists) is the maximal holomorphic extension of Q (within $\mathbb{D}$). A similar claim holds for P. It follows that any such function element of Q stabilizes any such function element of P (on the intersection of their domains) in the same sense.

LEMMA 6.1. *Let* $\Omega \subset \mathbb{D}$ *be open and let* $\alpha, \beta \in \Omega$*. Let* $P : \Omega \to \mathcal{B}(\mathtt{U}, \mathtt{Y})$ *and* $Q : \Omega \to \mathcal{B}(\mathtt{Y}, \mathtt{U})$*.*

Then some $E \in \mathrm{H}^\infty(\mathtt{Y} \times \mathtt{U})$ *satisfies* $E \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] = I = \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] E$ *on* Ω *iff there exists* $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U}\times\mathtt{Y})$ *such that* $M(z), X(z) \in \mathcal{GB}$ *for all* $z \in \Omega$ *and* $P = NM^{-1}$ *and* $Q = YX^{-1}$ *on* Ω*.*

If such a quadruple $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$ *exists, and* $P = N_1M_1^{-1}$ *and* $Q = Y_1X_1^{-1}$ *are* α*-r.c.f.'s, then they are* β*-r.c.f.'s and* $\left[\begin{smallmatrix} M_1 & Y_1 \\ N_1 & X_1 \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$*. Moreover, then any holomorphic extension (to a connected open subset of* $\mathbb{D}$*) of any restriction of* NM^{-1} *(to an open set) is a restriction of* NM^{-1} *(with domain* $\{z \in \mathbb{D} \mid M(z)^{-1}$ *exists*$\}$*).*

(Obviously, then $\left[\begin{smallmatrix} X & N \\ Y & M \end{smallmatrix}\right]$ is a β-d.c.f. of Q.)

We observed above that even if the domain of P and Q is not connected, a single d.c.f. applies at each component of the domain (if Q stabilizes P "at each component with the same inverse E"; otherwise the different components of P could be arbitrary). Moreover, there is no problem of extending P or Q holomorphically within the unit disc (the values of the functions at a certain point do not depend of the domain). This is an alternative proof of the fact that the function $P = \log$ (or any other function with different branches) is not dynamically stabilizable.

Next we define (possibly nonproper) stabilizing controller functions. Assume, for awhile, that $\Omega_P \subset \mathbb{D}$ is open and connected. Let $P : \Omega_P \to \mathcal{B}(\mathtt{U}, \mathtt{Y})$ be holomorphic. If $\Omega \subset \Omega_P$ is open and $Q : \Omega \to \mathcal{B}(\mathtt{Y}, \mathtt{U})$ is holomorphic, then we call Q a *stabilizing*

controller function for P if $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$.[4] We define *stabilizing DPF-controller functions* analogously (i.e., we do not require them to be proper).

By Lemma 6.1, the above definitions are in a complete accordance with the old ones and any stabilizing controller function is a stabilizing canonical controller. In particular, the existence of a stabilizing controller function (for a proper function P) is equivalent to Theorem 1.1(i). However, not all canonical controllers are functions, as one observes from the example at the end of section 3.

7. Continuous-time results. In this section we shall show that almost all *DT* (discrete-time) results of the other sections also hold in their *CT* (continuous-time) forms, even if we use the standard CT properness (defined later). But first we record the following obvious consequence of the Riemann mapping theorem.

REMARK 7.1 (Cayley). *The results in sections* 1–4, *section* 6, *and Appendix* A *except Proposition* 2.2 *hold true even if we replace* $\mathbb{D}$ *by any simply connected open* $\mathbb{D}' \subsetneq \mathbb{C}$ *and the origin* 0 *by any* $\zeta \in \mathbb{D}'$.

(We shall often use this implicitly when referring to those results. In fact, many of these results would hold even if $\mathbb{D}'$ was not simply connected (some results would hold with essentially the same proof; some others could be reduced to the simply connected case if, e.g., $\mathbb{D}'$ is a finite union of simply connected open sets containing ζ).)

In the most important special case, where $\mathbb{D}' := \mathbb{C}^+$ and $\zeta \in \mathbb{C}^+$, we can use the "Cayley" mapping $f : s \mapsto \frac{\zeta - s}{\zeta + s}$ to map $\mathbb{C}^+ \to \mathbb{D}$ conformally with $\zeta \mapsto 0$. Then we can apply the earlier results to $P \circ f^{-1}$, $M \circ f^{-1}$, $N \circ f^{-1}$, etc. in place of P, M, N, etc.

In CT, the right half-plane $\mathbb{C}^+$ takes the role of $\mathbb{D}$. Therefore, for the rest of this section, we redefine some concepts (cf. Theorem 7.3).

DEFINITION 7.2 (CT forms). *Given* $\omega \in \mathbb{R}$ *we set* $\mathbb{C}^+_\omega := \{z \in \mathbb{C} \mid \operatorname{Re} z > \omega\}$, *and by* $\mathrm{H}^\infty_\omega(\mathtt{U}, \mathtt{Y})$ *we denote the Banach space of bounded holomorphic functions* $\mathbb{C}^+_\omega \to \mathcal{B}(\mathtt{U}, \mathtt{Y})$ *with the supremum norm.*

We call P proper *if* $P \in \mathrm{H}^\infty_\infty := \cup_{\omega \in \mathbb{R}} \mathrm{H}^\infty_\omega$, *i.e., if* P *is a bounded holomorphic function on some right half-plane. (We identify a holomorphic function on a right half-plane* $\mathbb{C}^+_\omega$ *with its restriction to any open subset of* $\mathbb{C}^+_\omega$.*) It is* strictly proper *if, in addition,* $P(z) \to 0$ *as* $\operatorname{Re} z \to +\infty$.[5] *We set* $\mathbb{C}^+ := \mathbb{C}^+_0$, $\mathrm{H}^\infty := \mathrm{H}^\infty_0$. *Moreover, in all definitions and results in the other sections, we replace* $\mathbb{D}$ *by* $\mathbb{C}^+$ *and invertibility at* 0 *by the existence of a proper inverse.*

(The main motivation for the above properness concept is that a function is proper iff it is the transfer function of a well-posed linear system [21].)

Thus, e.g., if $N, M \in \mathrm{H}^\infty$ are r.c., $M^{-1} \in \mathrm{H}^\infty_\infty(\mathtt{U})$ and $P = NM^{-1}$ (on a right half-plane), then we call $P = NM^{-1}$ an *r.c.f. of* P; similarly, if P and Q are proper and $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$, then Q is a *proper stabilizing controller* for P. Recall that $(I - P^R)^{-1} \in \mathrm{H}^\infty$ in (3.1) means that $(I - P^R)^{-1}$ is the restriction of some element of H^∞, or equivalently, that some $E \in \mathrm{H}^\infty$ satisfies $E(I - P^R) = I = (I - P^R)E$ on some right half-plane (since P and R were assumed to be proper in (3.1)). If $I - R_{22}$ has a proper inverse, then we again (see below Lemma 3.2) identify R with the proper controller $R_{11} + R_{12}(I - R_{22})^{-1}R_{21}$.

---

[4] Naturally, this means that there exists $E \in \mathrm{H}^\infty$ such that $E\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] \equiv I \equiv \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] E$ on Ω. Note that if(f) $0 \in \Omega$, then this is equivalent to the definition of a proper stabilizing controller. Thus, "proper stabilizing controller function" means the same as "proper stabilizing controller."

[5] This means that for each $\epsilon > 0$ there exists $\omega_\epsilon \in \mathbb{R}$ such that $\|P(z)\| < \epsilon$ for all $z \in \mathbb{C}$ such that $\operatorname{Re} z > \omega_\epsilon$.

When N and M are r.c., $\alpha \in \mathbb{C}^+$, $M(\alpha) \in \mathcal{GB}(\mathtt{U})$, and $P = NM^{-1}$ on a neighborhood of α, we call $P = NM^{-1}$ an *α-r.c.f. of P*.

Next we define (possibly nonproper) stabilizing controller functions. Let, for awhile, $\Omega_P \subset \mathbb{C}^+$ be open and connected. Let $P : \Omega_P \to \mathcal{B}(\mathtt{U},\mathtt{Y})$ be holomorphic. If $\Omega \subset \Omega_P$ is open and $Q : \Omega \to \mathcal{B}(\mathtt{Y},\mathtt{U})$ is holomorphic, then we call Q a *stabilizing controller function for P* if $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$.[6] We define *stabilizing DPF-controller functions* analogously (i.e., we do not require them to be proper, whereas we still require that a *stabilizing [DPF-]controller with internal loop* is determined by a proper function R, as above).

By the arguments of Remark 7.1, the corresponding DT comments (below Lemma 6.1) apply here too (with Theorem 7.4 in place of Theorem 1.1).

THEOREM 7.3 (CT forms). *Propositions* A.1 *and* A.2*, Lemmata* A.3*,* A.4*,* A.9*,* A.10*, and* A.11*, and the results in sections* 1–3 *and* 6 *hold in their CT forms too if we replace Theorem* 1.1 *(resp., Theorem* 2.1*, Lemma* 3.2*) by Theorem* 7.4 *(resp., Theorem* 7.5*, Lemma* 7.7*).*

See Theorem 7.8 (resp., Theorem 7.9, Remark 7.10) for Theorem 4.1 (resp., Theorem 4.2, Remark 5.1). See [16] for the CT definitions for Proposition 2.2.

The main theorem, Theorem 1.1, holds in its CT form too once we remove "[strictly] proper" from (i) and (ii). We write this explicitly below with a new condition (i).

THEOREM 7.4 (CT: Dynamic feedback stabilization). *Let P be a proper $\mathcal{B}(\mathtt{U},\mathtt{Y})$-valued function, i.e., $P \in \mathrm{H}^\infty_\omega(\mathtt{U},\mathtt{Y})$ for some $\omega \geq 0$. Let $\zeta \in \mathbb{C}^+_\omega$. Then the following are equivalent:*

(i) *There exists a holomorphic $\mathcal{B}(\mathtt{Y},\mathtt{U})$-valued function Q on a neighborhood of ζ such that $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$.*
(ii) *P has a stabilizing controller function.*
(iii) *P has a stabilizing controller with internal loop.*
(iv) *P has a r.c.f.*
(v) *$\left[\begin{smallmatrix} P & 0 \\ 0 & I_{\mathtt{Z}} \end{smallmatrix}\right]$ has a r.c.f. for some (hence any) Hilbert space $\mathtt{Z}$.*

Assume that P has a r.c.f. $P = NM^{-1}$. Then $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right] \in \mathrm{H}^\infty(\mathtt{U},\mathtt{U}\times\mathtt{Y})$ can be extended to an invertible element of $\mathrm{H}^\infty(\mathtt{U}\times\mathtt{Y})$, say $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$. Denote its inverse by $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] \in \mathrm{H}^\infty(\mathtt{U}\times\mathtt{Y})$. Then all stabilizing controllers for P are given by the Youla(–Bongiorno) parameterization

$$Q = (Y + MV)(X + NV)^{-1} \qquad (= (\tilde{X} + V\tilde{N})^{-1}(\tilde{Y} + V\tilde{M})), \tag{7.1}$$

where $V \in \mathrm{H}^\infty(\mathtt{Y},\mathtt{U})$ is arbitrary (the controller is proper iff $(X + NV)^{-1}$ is proper, or equivalently, iff $(\tilde{X} + V\tilde{N})^{-1}$ is proper). The map $V \mapsto Q$ is one-to-one.

If P is strictly proper, then all these controllers are proper.

(Note that in PDE systems, the transfer function is usually strictly proper. That is also the case for well-posed systems having a bounded input or output operator and no feedthrough [13, Theorem 1.2].)

If, in Theorem 7.4, we set $\Omega_P := \mathbb{C}^+_\omega$ and fix some $\alpha, \beta \in \Omega_P$, then also the six conditions listed below become equivalent to (i).

THEOREM 7.5. *Assume that $\Omega_P \subset \mathbb{C}^+$ is open and connected and contains a right half-plane, $P : \Omega_P \to \mathcal{B}(\mathtt{U},\mathtt{Y})$ is holomorphic and proper, and $\alpha, \beta \in \Omega_P$.*

---

[6] Naturally, this means that there exists $E \in \mathrm{H}^\infty$ such that $E\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] \equiv I \equiv \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right] E$ on Ω. Note that if(f) Ω contains a right half-plane and P and Q are proper, then this is equivalent to the definition of a proper stabilizing controller.

Then the conditions (ii'), (ii"), (iii'), (iv'), (iv"), *and* (iv"') *of Theorem* 2.1 *are equivalent to* (iv) *of Theorem* 7.4. *Moreover, then any* α*-r.c.f. of* P *is a* β*-r.c.f. of* P. *The same holds with "l.c.f." or "d.c.f." in place of "r.c.f."*

(See Proposition 2.2 and Corollary 1.2 for further equivalent conditions.)

Condition (ii") says that for any point α in the domain Ω_P of P, there exists a stabilizing controller function whose domain includes α. We do not know whether a proper stabilizing controller always exists even if we assume that P is proper. Naturally, a similar comment applies to Theorem 7.9. In the matrix-valued case, a proper stabilizing controller $Q \in \mathrm{H}^\infty$ exists, by Corollary 1.2(vi) (through Theorem 7.3). If P is strictly proper, then any stabilizing controller (with or without internal loop) for P is proper, by Lemma 3.7. Moreover, whenever P has a sufficiently regular right factorization, a strictly proper Q exists.

THEOREM 7.6 (CT: strictly proper Q). *Assume that* P *has a r.c.f. and that* $P = NM^{-1}$, *where* $N \in \mathrm{H}^\infty(\mathtt{U},\mathtt{Y})$, $M \in \mathrm{H}^\infty(\mathtt{U})$, *and* M^{-1} *is proper. If* $M(+\infty) := \lim_{\mathrm{Re}\, s \to +\infty} M(s)$ *exists, then there exists a strictly proper* Q *such that* $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty(\mathtt{Y} \times \mathtt{U})$.

(Note that this NM^{-1} need not be a r.c.f.; the existence of a r.c.f. is needed only for guaranteeing the existence of a stabilizing controller.)

A r.c.f. of P (if any exists) can be determined from the LQR Riccati equation for an output-stabilizable realization as the resulting closed-loop transfer function $\left[\begin{smallmatrix} N \\ M \end{smallmatrix}\right]$; see [2] or [14]. For sufficient regularity (for Theorem 7.6) of this particular factorization, many different assumptions can be found in the literature, such as the analytic semigroup setting of [10] or certain assumptions on the unboundedness of the control and/or observation operators [29], [11].

Next we rewrite Lemma 3.2, which says that a controller with internal loop is proper iff $(I - R_{22})^{-1}$ is proper.

LEMMA 7.7 (CT: proper R). *Assume that* $R = \left[\begin{smallmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{smallmatrix}\right]$ *is a stabilizing controller with internal loop for* P.

Then R *is equivalent to a stabilizing controller with internal loop of form* $\tilde{R} = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$ *iff* $(I - R_{22})^{-1} \in \mathrm{H}_\infty^\infty$. *If* $(I - R_{22})^{-1} \in \mathrm{H}_\infty^\infty$, *then the unique solution is given by* $Q = R_{11} + R_{12}(I - R_{22})^{-1}R_{21}$.

(This holds by the original proof (with H_∞^∞-invertibility in place of invertibility at the origin).)

As in Theorem 4.1, we can reduce DPF-stabilization problems to ordinary dynamic stabilization problems.

THEOREM 7.8 (CT: P iff P_{21}). *Assume that* P *is a proper* $\mathcal{B}(\mathtt{U} \times \mathtt{W},\ \mathtt{Z} \times \mathtt{Y})$*-valued DPF-stabilizable function. Then a proper* $\mathcal{B}(\mathtt{Y} \times \Xi, \mathtt{U} \times \Xi)$*-valued function* R *is a stabilizing DPF-controller with internal loop for* P *iff* R *is a stabilizing controller function with internal loop for* P_{21}.

Moreover, a $\mathcal{B}(\mathtt{Y},\mathtt{U})$*-valued function* Q *is a stabilizing DPF-controller function for* P *iff* Q *is a stabilizing controller fucntion for* P_{21}. *It also follows that every stabilizing DPF-controller with internal loop for* P *is equivalent to one of the canonical controllers (for* P_{21}*) given by the Youla parameterization.*

(Also Theorem 4.1 holds in this CT terminology (and vice versa); the only difference is that here we do not require Q to be proper.)

As in Theorem 4.2, a function P is DPF-stabilizable iff it has a coprime factorization "through P_{21}."

THEOREM 7.9 (CT: DPF). *The following are equivalent for a proper* $\mathcal{B}(\mathtt{U} \times \mathtt{W},\ \mathtt{Z} \times \mathtt{Y})$*-valued function* P*:*

(ii) *P has a stabilizing DPF-controller function.*

(iii) *P has a stabilizing DPF-controller with internal loop.*

(iv) *P has a r.c.f. of the form $P = \left[\begin{smallmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{smallmatrix}\right] \left[\begin{smallmatrix} M_{11} & M_{12} \\ 0 & I \end{smallmatrix}\right]^{-1}$ such that N_{21} and M_{11} are r.c.*

(v) *P has a l.c.f. of the form $P = \left[\begin{smallmatrix} I & \tilde{M}_{12} \\ 0 & \tilde{M}_{22} \end{smallmatrix}\right]^{-1} \left[\begin{smallmatrix} \tilde{N}_{11} & \tilde{N}_{12} \\ \tilde{N}_{21} & \tilde{N}_{22} \end{smallmatrix}\right]$ such that $\tilde{N}_{21}$ and $\tilde{M}_{22}$ are l.c.*

Assume (iv) *and* (v). *Then $N_{21}M_{11}^{-1}$ is a r.c.f. of P_{21} and $\tilde{M}_{22}^{-1}\tilde{N}_{21}$ is a l.c.f. of P_{21}.*

(The changes above are essentially the same as those in Theorem 7.4. Naturally, by Theorem 7.8, we could add a condition resembling Theorem 7.4(i).)

Also Remark 5.1 holds in CT, mutatis mutandis.

REMARK 7.10 (exponential stabilization). *Define the power concepts above Remark* 5.1 *with $\mathrm{H}^\infty_{\mathrm{exp}} := \cup_{\omega<0} \mathrm{H}^\infty_\omega$ in place of $\mathrm{H}^\infty_{\mathrm{power}}$.*

With such "power-"concepts in place of the original ones and $\mathrm{H}^\infty_{\mathrm{exp}}$ in place of H^∞, Theorems 7.4–7.6, *Lemma* 7.7, *Theorems* 7.8 *and* 7.9, *and the CT forms (see Definition* 7.2, *Theorem* 7.3, *and the text between them) of Corollary* 1.2, *Theorem* 2.1, *Lemmata* A.10 *and* A.11, *and the results of section* 3 *hold if we rewrite the CT form of Corollary* 1.2(vii) *as follows:*

(vii) *$P = NM^{-1}$, where $N, M \in \mathrm{H}^\infty_{\mathrm{exp}}$, $N^*N + M^*M \geq \epsilon I$ on $\mathbb{C}^+_{-\epsilon}$, $\epsilon > 0$, and $\det M \not\equiv 0$.*

(Also the "power form" of Proposition 2.2 holds in the sense explained in [15]. Thus, all results in sections 1–4 are covered with some slight modifications.)

In the CT terminology for the "power concepts" of section 5, one usually replaces the component "power-" by the word "exponential[ly]" (see, e.g., [24] or [11] for details).

Despite the "different properness and different power stability" in CT, the "same" results hold as in DT, with the exception that we do not guarantee the existence of a proper stabilizing [DPF-]controller in general (just in the three special cases mentioned below Theorem 7.5) and we made the slight "change" ($\mathbb{C}^+_{-\epsilon}$ instead of $\mathbb{C}^+$) in (vii) at the end of Remark 7.10.

Appendix A. Discrete-time proofs. In this appendix we shall prove all our nontrivial results except those of section 7. We start by showing that every dynamically stabilizable function has a r.c.f. For that purpose we need to recall part of [16], particularly the fact that any $\mathrm{H}^\infty / \mathrm{H}^\infty$ fraction can be written as a fraction of so-called weakly r.c. functions. This requires the following definitions.

If $N \in \mathrm{H}^\infty(\mathtt{U}, \mathtt{Y})$, $M \in \mathrm{H}^\infty(\mathtt{U})$, and $M(0) \in \mathcal{GB}(\mathtt{U})$, then we call NM^{-1} a *right factorization (of P*, if $P = NM^{-1}$ near 0). We call such a factorization a *weakly right coprime factorization (w.r.c.f.)* if, in addition,

$$\begin{bmatrix} N \\ M \end{bmatrix} f \in \mathrm{H}^2 \;\Rightarrow\; f \in \mathrm{H}^2 \tag{A.1}$$

for every proper $\mathtt{U}$-valued function f, i.e., if a holomorphic $\mathtt{U}$-valued function f defined on a neighborhood of 0 is a restriction of an element of $\mathrm{H}^2(\mathtt{U})$ whenever $\left[\begin{smallmatrix} N \\ M \end{smallmatrix}\right] f$ is a restriction of an element of $\mathrm{H}^2(\mathtt{Y} \times \mathtt{U})$.

We recall the following two propositions from [16].

PROPOSITION A.1 (w.r.c.f.). *A $\mathcal{B}(\mathtt{U}, \mathtt{Y})$-valued function P has a right factorization iff it has a w.r.c.f.*

Moreover, if $P = NM^{-1}$ is a w.r.c.f., then all right factorizations of P are parameterized by $P = (NV)(MV)^{-1}$, where $V \in \mathrm{H}^\infty(\mathtt{U})$ and V^{-1} is proper. The

w.r.c.f.'s are those for which $V^{-1} \in \mathrm{H}^\infty$ *too. In particular, if a function* P *has a r.c.f., then every w.r.c.f. of* P *is a r.c.f.*

PROPOSITION A.2. *If* NM^{-1} *is a w.r.c.f. with* $M \in \mathrm{H}^\infty(\mathtt{U})$, R *is a proper* $\mathcal{B}(\mathtt{Z},\mathtt{U})$*-valued function, and* $NR, MR \in \mathrm{H}^\infty$, *then* $R \in \mathrm{H}^\infty$.

LEMMA A.3. *If* $P = NM^{-1}$ *is a w.r.c.f. and* $(I-P)^{-1} \in \mathrm{H}^\infty$, *then* $(M-N)^{-1} \in \mathrm{H}^\infty$.

Proof. Since $I - P = (M - N)M^{-1}$, we have $M(M - N)^{-1} = (I - P)^{-1} \in \mathrm{H}^\infty$ and

$$\text{(A.2)}\quad N(M-N)^{-1} = NM^{-1}M(M-N)^{-1} = P(I-P)^{-1} = (I-P)^{-1} - I \in \mathrm{H}^\infty.$$

By Proposition A.2, it follows that $(M - N)^{-1} \in \mathrm{H}^\infty$. □

Now we are ready to prove the implication (ii)⇒(iv) in Theorem 1.1.

LEMMA A.4. *If* Q *is a proper stabilizing controller to a proper* $\mathcal{B}(\mathtt{U},\mathtt{Y})$*-valued function* P, *then* P *has a r.c.f.*

In the proof we take the factorizations (of P and Q) determined by (1.1), replace them by weakly coprime factorizations ($P = NM^{-1}$ and $Q = YX^{-1}$), and use Lemma A.3 to prove the invertibility of $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$ in H^∞.

Proof of Lemma A.4. By (1.1), $P = N_1 M_1^{-1}$, where $M_1 := (I - QP)^{-1} \in \mathrm{H}^\infty$ and $N_1 := PM_1 \in \mathrm{H}^\infty$. By Proposition A.1, P has a w.r.c.f. $P = NM^{-1}$. Similarly, Q has a w.r.c.f. $Q = YX^{-1}$. It obviously follows that

$$\text{(A.3)}\qquad \begin{bmatrix} 0 & Q \\ P & 0 \end{bmatrix} = \begin{bmatrix} 0 & Y \\ N & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & X \end{bmatrix}^{-1}$$

is a w.r.c.f. Since $\left(I - \left[\begin{smallmatrix} 0 & Q \\ P & 0 \end{smallmatrix}\right]\right)^{-1} \in \mathrm{H}^\infty$ (see (1.1)), it follows from Lemma A.3 that

$$\text{(A.4)}\qquad \left(\begin{bmatrix} M & 0 \\ 0 & X \end{bmatrix} - \begin{bmatrix} 0 & Y \\ N & 0 \end{bmatrix}\right)^{-1} = \begin{bmatrix} M & -Y \\ -N & X \end{bmatrix}^{-1} \in \mathrm{H}^\infty.$$

Therefore, $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]^{-1} = \left(\left[\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right]\left[\begin{smallmatrix} M & -Y \\ -N & X \end{smallmatrix}\right]\left[\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right]\right)^{-1} \in \mathrm{H}^\infty$. □

(The main part of the above is essentially the proof of [11, Lemmata 7.1.5 and 6.6.6] with w.r.c.f.'s in place of r.c.f.'s.)

Next we observe that a r.c.f. can be extended to a d.c.f. that has $Y(0) = 0$ and $X(0) \in \mathcal{GB}$.

LEMMA A.5. *Let* NM^{-1} *be a* $\mathcal{B}(\mathtt{U},\mathtt{Y})$*-valued r.c.f. Then* NM^{-1} *has a d.c.f.* $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$ *such that* $X(0), \tilde{X}(0), \tilde{M}(0) \in \mathcal{GB}$ *and* $Y(0) = 0$. *Moreover, the strictly proper function* $Q := YX^{-1} = \tilde{X}^{-1}\tilde{Y}$ *stabilizes* $P := NM^{-1}$ *in the sense that* $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty$. *Furthermore, we can have above* $X(0) = I = \tilde{M}(0)$ *(and* $\tilde{X}(0) = I = M(0)$ *if we replace* N *by* NV *and* M *by* MV, *where* $V := M(0) \in \mathcal{GB}(\mathtt{U})$*).*

Proof. By [16], we can extend the function $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ to a d.c.f. $\left[\begin{smallmatrix} M & Y_0 \\ N & X_0 \end{smallmatrix}\right]$. Set $\left[\begin{smallmatrix} M & Y_1 \\ N & X_1 \end{smallmatrix}\right] := \left[\begin{smallmatrix} M & Y_0 \\ N & X_0 \end{smallmatrix}\right]\left[\begin{smallmatrix} I & -M(0)^{-1}Y_0 \\ 0 & I \end{smallmatrix}\right]$ to have $Y_1(0) = 0$. It follows that $X_1(0)$ must be invertible. Set $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] := \left[\begin{smallmatrix} M & Y_1 \\ N & X_1 \end{smallmatrix}\right]\left[\begin{smallmatrix} I & 0 \\ 0 & X_1(0)^{-1} \end{smallmatrix}\right]$ to have $X(0) = I$; hence $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right](0) = \left[\begin{smallmatrix} M(0)^{-1} & 0 \\ P(0) & I \end{smallmatrix}\right]$, where $P := NM^{-1}$ (and $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] := \left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]^{-1}$). The final claim obviously follows. One easily verifies that $\left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} M\tilde{X} & Y\tilde{M} \\ N\tilde{X} & X\tilde{M} \end{smallmatrix}\right] = I$. □

LEMMA A.6. *Conditions* (i)–(v) *of Theorem* 1.1 *are equivalent.*

Proof. 1° (i), (ii), (iv), *and* (v): By Lemmata A.4 and A.5, (i), (ii), and (iv) are equivalent. If $P = NM^{-1}$ is a r.c.f., then obviously so is $\left[\begin{smallmatrix} P & 0 \\ 0 & I \end{smallmatrix}\right] = \left[\begin{smallmatrix} N & 0 \\ 0 & I \end{smallmatrix}\right] \left[\begin{smallmatrix} M & 0 \\ 0 & I \end{smallmatrix}\right]^{-1}$; thus, (iv) implies (v). Assume then (v). By Proposition 2.2, $\left[\begin{smallmatrix} P & 0 \\ 0 & I \end{smallmatrix}\right]$ has a realization $\tilde{\Sigma}$ that satisfies the finite cost condition and its dual. Remove the bottom row of $\tilde{\Sigma}$ to obtain a realization Σ of P that satisfies the finite cost condition and its dual (this follows very easily). By Proposition 2.2, this shows that P has a r.c.f.

2° (iii): Trivially, (i) implies (iii) (take $R := \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$). If (iii) holds, then so does (v), by Lemmata 3.1 and A.4 (with $\left[\begin{smallmatrix} P & 0 \\ 0 & I \end{smallmatrix}\right]$ in place of P). ∎

The following is well known [24].

PROPOSITION A.7 (Schur decomposition). (a) *If* $A = B^{-1} \in \mathcal{GB}(\mathtt{X}_1 \times \mathtt{X}_2, \mathtt{Y}_1 \times \mathtt{Y}_2)$, *then* $A_{22} \in \mathcal{GB}(\mathtt{X}_2, \mathtt{Y}_2) \Leftrightarrow B_{11} \in \mathcal{GB}(\mathtt{Y}_1, \mathtt{X}_1) \Rightarrow B_{11}^{-1} = A_{11} - A_{12}A_{22}^{-1}A_{21}$.

(b) *If* $A \in \mathcal{B}(\mathtt{X}_1 \times \mathtt{X}_2,\ \mathtt{Y}_1 \times \mathtt{Y}_2)$ *and* $A_{22} \in \mathcal{GB}(\mathtt{X}_2, \mathtt{Y}_2)$, *then* $A \in \mathcal{GB} \Leftrightarrow A_{11} - A_{12}A_{22}^{-1}A_{21} \in \mathcal{GB}(\mathtt{X}_1, \mathtt{Y}_1)$.

Now we can prove the results of section 3.

Proof of Lemma 3.2. 1° Set

$$A := I - P^R = \left[\begin{array}{cc|c} I & -R_{11} & -R_{12} \\ -P & I & 0 \\ \hline 0 & -R_{21} & I - R_{22} \end{array}\right], \tag{A.5}$$

$B := A^{-1}$. Assume that $A_{22}(0) := I - R_{22}(0) \in \mathcal{GB}$. By Proposition A.7(a), $B_{11}^{-1} = A_{11} - A_{12}A_{22}^{-1}A_{21} = \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]$, where $Q := R_{11} - R_{12}(I - R_{22})^{-1}R_{21}$; hence $B_{11} = \tilde{B}_{11}$, where $\tilde{B} := (I - P^{\tilde{R}})^{-1}$, $\tilde{R} := \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$. Thus, R and $\tilde{R}$ are "equivalent." But $\tilde{B}_{21} = 0$, $\tilde{B}_{12} = 0$, and $\tilde{B}_{22} = I$; hence $\tilde{B} \in \mathrm{H}^\infty$, so also $\tilde{R}$ is stabilizing.

2° Assume that R and $\tilde{R}$ are equivalent. Then we again have $B_{11} = \tilde{B}_{11}$. But $\tilde{B}_{11}^{-1} = \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]$ is proper; hence $B_{11}(0) \in \mathcal{GB}$. By Proposition A.7(a), this implies that $\mathcal{GB} \ni A_{22}(0) = I - R_{22}(0)$. ∎

Proof of Lemma 3.3. By Lemma A.6, P has a r.c.f. NM^{-1}. Set

$$T := \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} := (I - P^R)^{-1}, \qquad \tilde{X} := M^{-1}T_{11}, \qquad \tilde{Y} := M^{-1}T_{12}. \tag{A.6}$$

Since $T(I - P^R) = I = (I - P^R)T$, we have (see (A.5)) $T_{11}I - T_{12}P = I$, $-PT_{11} + T_{21} = 0$, and $-PT_{12} + T_{22} = I$; hence $\tilde{X} - \tilde{Y}P = M^{-1}$, i.e., $\tilde{X}M - \tilde{Y}N = I$, and also $-N\tilde{X} + T_{21} = 0$, $-N\tilde{Y} + T_{22} = 0$. Thus, $N\tilde{X}, N\tilde{Y} \in \mathrm{H}^\infty$. But $M\tilde{X} = T_{11} \in \mathrm{H}^\infty$ and $M\tilde{Y} = T_{12} \in \mathrm{H}^\infty$; hence $\tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$, by Proposition A.2, and hence $\tilde{X}$ and $\tilde{Y}$ are l.c. By duality (because R^{d} is stabilizing for P^{d}), we get X and Y. ∎

The following is a direct consequence of Proposition A.7.

LEMMA A.8. *Let* P *be* $\mathcal{B}(\mathtt{U}, \mathtt{Y})$*-valued and proper and let* $R = \left[\begin{smallmatrix} 0 & I \\ \tilde{Y} & I - \tilde{X} \end{smallmatrix}\right]$ *be* $\mathcal{B}(\mathtt{Y} \times \mathtt{U},\ \mathtt{Y} \times \mathtt{Y})$*-valued and proper. Then* $R(0) \in \mathcal{GB}$ *iff* $(\tilde{X} - \tilde{Y}P)(0) \in \mathcal{GB}$. *Moreover, if* $R(0) \in \mathcal{GB}$, *then (set* $M := (\tilde{X} - \tilde{Y}P)^{-1}$*)*

$$(I - P^R)^{-1} = \begin{bmatrix} I + M\tilde{Y}P & M\tilde{Y} & M \\ P(I + M\tilde{Y}P) & I + PM\tilde{Y} & PM \\ M\tilde{Y}P & M\tilde{Y} & M \end{bmatrix} = \begin{bmatrix} M\tilde{X} & M\tilde{Y} & M \\ PM\tilde{X} & I + PM\tilde{Y} & PM \\ M\tilde{X} - I & M\tilde{Y} & M \end{bmatrix}. \tag{A.7}$$

Proof of Lemma 3.5. 1° One easily verifies (3.3). By the definition of equivalence (and the equation $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] = I$), (3.2) follows.

2° If $P = NM^{-1}$ is a r.c.f., $\tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$, and $\Delta := \tilde{X}M - \tilde{Y}N \in \mathcal{G}\mathrm{H}^\infty$ (by Proposition A.1, the same then holds for all r.c.f.'s of P), then $N_1 := N\Delta^{-1}$, $M_1 := M\Delta^{-1}$ form a r.c.f. of P and satisfy $\tilde{X}M_1 - \tilde{Y}N_1 = I$. By Lemma 3.4 and (3.3), only the "only if" from the first "iff" remains to be proved.

3° Assume that R' is stabilizing. By the proof of Lemma 3.3, R' is equivalent to $\tilde{X}_1^{-1}\tilde{Y}_1$, where $\tilde{X}_1, \tilde{Y}_1 \in \mathrm{H}^\infty$ and $\tilde{X}_1 M_1 - \tilde{Y}_1 N_1 = I$ for some r.c.f. $P = N_1 M_1^{-1}$. By (A.7) and the equivalence of R' to $\tilde{X}_1^{-1}\tilde{Y}_1$, we have $\begin{bmatrix} M_1\tilde{X}_1 & M_1\tilde{Y}_1 \end{bmatrix} = \begin{bmatrix} M\tilde{X} & M\tilde{Y} \end{bmatrix}$; hence $\tilde{X} = \Delta\tilde{X}_1$ and $\tilde{Y} = \Delta\tilde{Y}_1$, where

$$\Delta := M^{-1}M_1 = \tilde{X}M_1 - \tilde{Y}N_1 \in \mathrm{H}^\infty(\mathtt{U}) \tag{A.8}$$

(because $M^{-1} = \tilde{X} - \tilde{Y}P$). But $\Delta^{-1} = M_1^{-1}M$; hence $M_1\Delta^{-1} = M$ and $N_1\Delta^{-1} = PM$. Since $M, PM \in \mathrm{H}^\infty$, by (A.7), we have $\Delta^{-1} \in \mathrm{H}^\infty$, by Proposition A.2. Thus, $\tilde{X}M_1 - \tilde{Y}N_1 = \Delta \in \mathcal{G}\,\mathrm{H}^\infty$. □

Now we are ready to complete the proof of our main result.

Proof of Theorem 1.1. Lemma A.6 contains the equivalence. Assume then that $P = NM^{-1}$ is a r.c.f. By [16], $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right]$ can be extended to a d.c.f. It obviously follows that $\tilde{M}(0) \in \mathcal{GB}$ and $P = \tilde{M}^{-1}\tilde{N}$ (see [11] or [24]). The Youla parameterization is essentially from [4, Theorem 5.5] or [11, Theorem 7.2.14] (and can easily be obtained from Lemmata 3.3 and 3.5). The claim on properness is from Lemma 3.2 and that on strict properness is from Lemma 3.7. □

Proof of Theorem 2.1. 1° If $P = NM^{-1}$ is an α-r.c.f. and $\tilde{X}M - \tilde{Y}N = I$, $\tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$, then $\tilde{X} - \tilde{Y}P = M^{-1}$ near α, hence on Ω_P; hence then $P = NM^{-1}$ is a β-r.c.f. too. The rest of the last paragraph follows analogously.

2° (iii)–(iv"'): By 1°, (iv) and (iv') are equivalent. The implication (iv')⇒(iv"') is from Theorem 1.1 (and Remark 7.1) and the converse is trivial. By duality, we get (iv")⇔(iv"'). The equivalence of (iii) and (iii') follows from Lemma 3.3.

3° As in Remark 7.1, we observe that Theorem 1.1 holds with α in place of 0, so the equivalence of (ii") and (iv') follows from that of (i) and (iv). Similarly, if (ii') holds and $z \in \Omega_Q$, then P has a z-r.c.f.; hence then (iv) holds, by 1°. Trivially, (ii") implies (ii). □

Proof of Corollary 1.2. The implications (vi)⇒(ii) and (iv)⇒(vii) are trivial (take $\epsilon := 1/\|[\tilde{X} - \tilde{Y}]\|^2$). Implication (iv)⇒(vi) is from [19, Corollary 6.6] and (vii)⇒(iv') (see Theorem 2.1) holds for some suitable $\alpha \in \mathbb{D}$ by [7]. □

(In fact, (vii) is equivalent to (i)–(v) even if $\dim \mathtt{Y} = \infty$; it suffices that $\dim \mathtt{U} < \infty$.)

Next we need the following generalization of a classical result.

LEMMA A.9 (R stabilizes P_{21}). *If R is stabilizing DPF-controller with internal loop for P, then R is a stabilizing controller with internal loop for P_{21}.*

(This was shown rigorously in [11, Lemma 7.3.5], but this can be observed from Figure 4.1: For $w = 0$ the closed-loop equations obviously define the map $(I - (P_{21})^R)^{-1}$ (see (3.1)) if we ignore the equation for z; thus, $(I - (P_{21})^R)^{-1}$ is contained in the "DPF closed-loop map" $(I - P^{R_{\mathrm{DPF}}})^{-1} \in \mathrm{H}^\infty$ for P and R.)

Proof of Theorems 4.1 *and* 4.2. 1° We first note that if R DPF-stabilizes P with internal loop, then it stabilizes P_{21}, by Lemma A.9; hence then P_{21} has a d.c.f. and R is equivalent to a canonical controller, by Theorem 1.1.

2° (iii)⇒(iv): Let $\tilde{X}^{-1}\tilde{Y}$ be a canonical controller that DPF-stabilizes P. Then $\tilde{X}^{-1}\tilde{Y}$ stabilizes P_{21}. One can verify that $\left[\begin{smallmatrix}\tilde{X} & 0\\ 0 & I\end{smallmatrix}\right]^{-1}\left[\begin{smallmatrix}0 & \tilde{Y}\\ 0 & 0\end{smallmatrix}\right]$ is a canonical controller that stabilizes P (see [11, Lemma 7.3.10] for details). By Lemma 3.5 it follows that there exists a r.c.f. $P = NM^{-1}$ such that

$$\text{(A.9)}\qquad \begin{bmatrix} I & 0\\ 0 & I\end{bmatrix} = \begin{bmatrix}\tilde{X} & 0\\ 0 & I\end{bmatrix} M - \begin{bmatrix}0 & \tilde{Y}\\ 0 & 0\end{bmatrix} N = \begin{bmatrix}\tilde{X}M_{11} - \tilde{Y}N_{21} & \tilde{X}M_{12} - \tilde{Y}N_{22}\\ M_{21} & M_{22}\end{bmatrix};$$

in particular, $\tilde{X}M_{11} - \tilde{Y}N_{21} = I$, $M_{21} = 0$, and $M_{22} = I$, and hence (iv) holds.

3° (iv)⇒(iii): Assume (iv). By Lemma 3.5 and (A.9), $\tilde{X}^{-1}\tilde{Y}$ DPF-stabilizes P iff we have

$$\text{(A.10)}\qquad \begin{bmatrix}\tilde{X}M_{11} - \tilde{Y}N_{21} & \tilde{X}M_{12} - \tilde{Y}N_{22}\\ 0 & I\end{bmatrix} \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U}\times\mathtt{W}),$$

or equivalently, $\tilde{X}M_{11} - \tilde{Y}N_{21} \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U})$, or equivalently, iff the canonical controller $\tilde{X}^{-1}\tilde{Y}$ stabilizes $N_{21}M_{11}^{-1} = P_{21}$. By Theorem 1.1 (applied to P_{21}), such $\tilde{X}, \tilde{Y} \in \mathrm{H}^\infty$ do exist.

4° By duality, we get "(iii)⇔(v)"; hence (iii)–(v) are equivalent. Moreover, in 3° we observed that if (iv) holds, then a canonical controller DPF-stabilizes P iff it stabilizes P_{21}. By Theorem 1.1, all such canonical controllers are given by the Youla parameterization (1.3) and at least one of them is strictly proper; hence (i) and (ii) are equivalent to (iii). Moreover, by Lemmata A.9 and 3.3, any DPF-stabilizing controller with internal loop for P is equivalent to some canonical controller. Thus, both theorems have been established. □

In a r.c.f. $P = NM^{-1}$, the inverse M^{-1} has the same maximal domain as P.

LEMMA A.10. *If $P : \Omega \to \mathcal{B}(\mathtt{U},\mathtt{Y})$ is holomorphic, where $\Omega \subset \mathbb{D}$ is connected, and $P = NM^{-1}$ is a α-r.c.f. for some $\alpha \in \Omega$, then $P = NM^{-1}$ is a z-r.c.f. for every $z \in \Omega$.*

(Indeed, if $\tilde{X}M - \tilde{Y}N = I$, then $I = (\tilde{X} - \tilde{Y}P)M = M(\tilde{X} - \tilde{Y}P)$ near α, hence on Ω.)

Thus, then any connected holomorphic extension of P (within $\mathbb{D}$) is a restriction of NM^{-1}.

Next we conclude the same with a "possibly nonconnected Ω." For example, if NM^{-1} has an α_1-r.c.f. and an α_2-r.c.f., then these are the same (modulo a unit).

LEMMA A.11. *Let $\alpha_1, \alpha_2 \in \mathbb{D}$, $N \in \mathrm{H}^\infty(\mathtt{U},\mathtt{Y})$, $M \in \mathrm{H}^\infty(\mathtt{U})$, $M(\alpha_1), M(\alpha_2) \in \mathcal{GB}(\mathtt{U})$. Let $N_kM_k^{-1}$ be an α_k-r.c.f. of NM^{-1} ($k = 1, 2$). Then $\left[\begin{smallmatrix}N_2\\ M_2\end{smallmatrix}\right] = \left[\begin{smallmatrix}N_1\\ M_1\end{smallmatrix}\right] V$ for some $V \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U})$; in particular, $N_1M_1^{-1}$ is also an α_2-r.c.f. of NM^{-1}.*

Proof. Set $F := \left[\begin{smallmatrix}N\\ M\end{smallmatrix}\right]$, $F_k := \left[\begin{smallmatrix}N_k\\ M_k\end{smallmatrix}\right]$, and pick $G_k \in \mathrm{H}^\infty$ such that $G_kF_k = I$ ($k = 1, 2$). By Proposition A.1 (and Remark 7.1), we have $F = F_kV_k$ for some $V_k \in \mathrm{H}^\infty(\mathtt{U})$ such that $V_k(\alpha_k) \in \mathcal{GB}(\mathtt{U})$ ($k = 1, 2$). Therefore $V_1 = G_1F_2V_2$ and $V_2 = G_2F_1V_1 = G_2F_1G_1F_2V_2$; hence $G_2F_1G_1F_2 = I$ near α_2, hence on $\mathbb{D}$. Similarly, $TS = I$, where $T := G_1F_2$, $S := G_2F_1$. But $F_2V_2 = F_1V_1 = F_1TV_2$; hence $F_2 = F_1T$ near α_2, hence on $\mathbb{D}$. Set $V := T$. □

Proof of Lemma 6.1. 1° *"If"*: Set $\left[\begin{smallmatrix}\tilde{X} & -\tilde{Y}\\ -\tilde{N} & \tilde{M}\end{smallmatrix}\right] := \left[\begin{smallmatrix}M & Y\\ N & X\end{smallmatrix}\right]^{-1}$. By (3.2) (and Remark 7.1 and Lemma 3.2), $\left[\begin{smallmatrix}I & -Q\\ -P & I\end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix}Y\tilde{N}+I & Y\tilde{M}\\ X\tilde{N} & X\tilde{M}\end{smallmatrix}\right] =: E$ on Ω.

2° *"Only if"*: By (1.1), $N_0, M_0 \in \mathrm{H}^\infty$, where $N_0 := P(I - QP)^{-1}$ and $M_0 = (I - QP)^{-1}$. Since $P = N_0M_0^{-1}$ on Ω, P has an α-r.c.f. and a β-r.c.f., by Lemma A.4 (cf. Remark 7.1). By Lemma A.11, any α-r.c.f. of P is a β-r.c.f. of P.

By Lemmata 3.3, 3.5, and 3.2, there exists an α-d.c.f. $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right]^{-1}$ of P such that $X(\alpha) \in \mathcal{GB}(\mathtt{Y})$ and $Q = YX^{-1}$ on a neighborhood of α. By the above, $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right]$ is also a β-d.c.f. of P. Exchange the roles of P and Q to observe that $Q = YX^{-1}$ is also a β-r.c.f. (in particular, $X(\beta) \in \mathcal{GB}(\mathtt{Y})$).

3° By Proposition A.1 (cf. Remark 7.1), $\left[\begin{smallmatrix} M_1 & Y_1 \\ N_1 & X_1 \end{smallmatrix}\right] = \left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] \left[\begin{smallmatrix} V & 0 \\ 0 & V_Q \end{smallmatrix}\right]$ for some $V \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{U})$, $V_Q \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{Y})$; hence the claim on β-r.c.f.'s follows. The last claim now follows from Lemma A.10. ∎

Appendix B. Continuous-time proofs (for section 7). In this appendix we prove the results of section 7.

Proof of Remark 7.1. This is straightforward, but we give some examples below. For clarity, in this proof we add the prefix "$(\mathbb{D}', \zeta)$-" when we refer to the redefined terminology of Remark 7.1; otherwise we refer to the original DT terminology of sections 1–6.

By the Riemann mapping theorem, there exists a holomorphic function $f : \mathbb{D}' \to \mathbb{D}$ that has a holomorphic inverse and satisfies $f(\zeta) = 0$. Obviously, a function P' is $(\mathbb{D}', \zeta)$-proper (resp., $(\mathbb{D}', \zeta)$-H^∞) iff $P := P' \circ f^{-1}$ is proper (resp., H^∞). Moreover, a function Q is a proper stabilizing controller for P iff $Q' := Q \circ f$ is a $(\mathbb{D}', \zeta)$-proper stabilizing controller for $P \circ f = P'$, i.e., iff Q' is $(\mathbb{D}', \zeta)$-proper and $\left[\begin{smallmatrix} I & -Q' \\ -P' & I \end{smallmatrix}\right]^{-1}$ is a bounded holomorphic function on $\mathbb{D}'$ (because $\left[\begin{smallmatrix} I & -Q' \\ -P' & I \end{smallmatrix}\right]^{-1} = \left[\begin{smallmatrix} I & -Q \\ -P & I \end{smallmatrix}\right]^{-1} \circ f$ wherever either exists).

Similarly, also all other required old terminology is mapped to corresponding new $(\mathbb{D}', \zeta)$-terminology; hence all results mentioned in Remark 7.1 hold also for this new $(\mathbb{D}', \zeta)$-terminology.

For example, if a ζ-proper function P' has a $(\mathbb{D}', \zeta)$-proper stabilizing controller, then $P := P' \circ f^{-1}$ has a r.c.f. $P = NM^{-1}$, by Theorem 1.1; hence then P' has the r.c.f. $N'(M')^{-1}$ with $N' := N \circ f$, $M' := M \circ f$. ∎

The original proofs of all DT results apply in their CT forms too except that part of the proof of Theorem 1.1 would require additional details. Also Remark 7.1 (with some additional work) could be used to prove almost all CT results, but in some cases it becomes more cumbersome than the use of the original proofs, so we shall in each proof below select the simplest method.

From now on we use the CT terminology defined in Definition 7.2 and below it unless we explicitly use the prefix *DT-* (when referring to the original DT terminology of sections 1–6)) or the prefix ζ- (when referring to the terminology defined in Remark 7.1 for fixed $\zeta \in \mathbb{C}^+$ and $\mathbb{D}' := \mathbb{C}^+$). Thus, e.g., proper means H^∞_∞, but *DT-proper* (resp., ζ-*proper*) means holomorphic on a neighborhood of 0 (resp., ζ). Observe that in, e.g., ζ-*canonical controller* or ζ-H^∞ the prefix is redundant (but *DT*-H^∞ means holomorphic and bounded on $\mathbb{D}$, not on $\mathbb{C}^+$).

However, when referring to some result, we refer to its original form unless we use the prefix "CT-" or "ζ-". When we write, e.g., *CT-Corollary* 3.6, we refer to the CT-form of Corollary 3.6 that is established in Theorem 7.3 (its proof is given later below). Similarly, ζ-*Corollary* 3.6 refers to the form established in Remark 7.1 (with $\mathbb{D}' = \mathbb{C}^+$).

We shall often need the fact that the (CT) r.c.f.'s of a proper function are the same as its α-r.c.f.'s:

Lemma B.1 (α-r.c.f.). *Let $\omega \geq 0$, $P \in \mathrm{H}^\infty_\omega(\mathtt{U}, \mathtt{Y})$, and $\alpha \in \mathbb{C}^+_\omega$. Then an α-r.c.f. of P is a r.c.f., and a r.c.f. of P is an α-r.c.f. A similar claim obviously holds for l.c.f.'s and d.c.f.'s.*

Proof. If $\tilde{X}, \tilde{Y}, M, N \in \mathrm{H}^\infty$, $I = \tilde{X}M - \tilde{Y}N$ (on $\mathbb{C}^+$), and $NM^{-1} = P$ on a neighborhood Ω of α, then $I = (\tilde{X} - \tilde{Y}P)M$ on $\mathbb{C}_\omega^+$ and $M(\alpha)^{-1}$ exists; hence then $I = M(\tilde{X} - \tilde{Y}P)$ on a neighborhood of α, hence on $\mathbb{C}_\omega^+$ too, so then $P = NM^{-1}$ is a (CT) r.c.f. The converse is analogous. ∎

Proof of Theorem 7.5. As in the proof of Lemma B.1, we observe that also here α-r.c.f.'s coincide with r.c.f.'s. The rest of Theorem 7.5 is already contained in ζ-Theorem 2.1 (for any fixed $\zeta \in \Omega_P$). ∎

Proof of Theorem 7.4 *except CT-Lemma* 3.7. Pick $\omega \geq 0$ such that $P \in \mathrm{H}_\omega^\infty$, and set $\Omega_P := \mathbb{C}_\omega^+$ (and $\mathbb{D}' := \mathbb{C}^+$). Observe that a (CT) "canonical controller" is such also in terms of Remark 7.1 (when, e.g., $\zeta \in \mathbb{C}_\omega^+$).

1° The equivalence of (i)–(v) follows from Theorem 7.5 as follows: The equivalence "(iv)⇔(iv')" is contained in Theorem 7.5. For $\alpha = \zeta \in \mathbb{C}_\omega^+$, we get "(iv)⇔(v)" from "(iv)⇔(iv')" (twice) and ζ-Theorem 1.1 (P has a r.c.f. ⇔ it has a ζ-r.c.f. ⇔ $\left[\begin{smallmatrix} P & 0 \\ 0 & I \end{smallmatrix}\right]$ has a ζ-r.c.f. ⇔ $\left[\begin{smallmatrix} P & 0 \\ 0 & I \end{smallmatrix}\right]$ has a r.c.f.). Each of conditions (i) and (ii) is obviously equivalent to (ii') and/or (ii") (of Theorem 7.5). (iii') implies (iii); if (iii) holds, then P has a ζ-r.c.f. for some $\zeta \in \mathbb{C}_\omega^+$ (pick one in the definition of the controller), by ζ-Theorem 1.1(iii) and (iv); i.e., then (iv') holds (take $\alpha = \zeta$).

2° For any $\zeta \in \mathbb{C}_\omega^+$, we get the d.c.f. from ζ-Theorem 1.1 and Lemma B.1. Obviously, a canonical controller $\left[\begin{smallmatrix} 0 & Y \\ I & I-X \end{smallmatrix}\right]$ stabilizes P iff it ζ-stabilizes P (for some $\zeta \in \mathbb{C}_\omega^+$). A stabilizing controller $R \in \mathrm{H}_\infty^\infty$ with internal loop for P is a ζ-stabilizing controller with internal loop for P (for some $\zeta \in \mathbb{C}_\omega^+$), hence equivalent to a [ζ-]canonical controller, by ζ-Lemma 3.3, so the Youla parameterization claim holds (recall Lemma 7.7). ∎

Proof of Theorem 7.3. *All propositions*: The CT forms are given in [15].

Lemmata A.3, A.4, *and* 3.3: Use the original proofs.

Lemmata 3.4 *and* 3.5 *and Corollaries* 3.6 *and* 1.2: Use the original proofs, mutatis mutandis (see Definition 7.2).

Lemmata A.9, A.10, A.11, *and* 3.1: These follow easily from Remark 7.1.

Lemma 3.7: Pick a d.c.f. $\left[\begin{smallmatrix} M & Y \\ N & X \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right]^{-1} \in \mathrm{H}^\infty(\mathtt{U} \times \mathtt{Y})$ of P. Pick $\omega' > 0$ such that $\|\tilde{X}(s) - M^{-1}(s)\| = \|\tilde{Y}(s)P(s)\| < 1/2\|M\|_{\mathrm{H}^\infty}$ for all $s \in \mathbb{C}_{\omega'}^+$ to have $\tilde{X}(s)^{-1}$ bounded on $\mathbb{C}_{\omega'}^+$. Thus, all canonical controllers are proper (see CT-Corollary 3.6). By CT-Lemmata 3.2 and 3.3, all stabilizing controllers with internal loop for P are proper.

Lemma 6.1: This follows easily from Remark 7.1 and Lemma B.1. ∎

Proof of Theorem 7.6. One easily verifies that $M(+\infty) \in \mathcal{GB}(\mathtt{U})$ [11, Proposition 6.3.1(c)]. Let $P = N_0 M_0^{-1}$ be a r.c.f. By CT-Proposition A.1, $\left[\begin{smallmatrix} N \\ M \end{smallmatrix}\right] = \left[\begin{smallmatrix} N_0 \\ M_0 \end{smallmatrix}\right] V$ for some $V \in \mathrm{H}^\infty$ such that V^{-1} is proper. Let $\left[\begin{smallmatrix} M_0 & Y \\ N_0 & X \end{smallmatrix}\right] \in \mathcal{G}\,\mathrm{H}^\infty(\mathtt{Y} \times \mathtt{U})$ (see Theorem 7.4) and set

(B.1)
$$\begin{bmatrix} M_0 & Y_0 \\ N_0 & X_0 \end{bmatrix} := \begin{bmatrix} M_0 & Y \\ N_0 & X \end{bmatrix} \begin{bmatrix} I & -VM(+\infty)^{-1}Y \\ 0 & I \end{bmatrix} = \begin{bmatrix} M_0 & (1 - MM(+\infty)^{-1})Y \\ N_0 & X - NM(+\infty)^{-1}Y \end{bmatrix}.$$

Set $\left[\begin{smallmatrix} \tilde{X} & -\tilde{Y} \\ -\tilde{N} & \tilde{M} \end{smallmatrix}\right] := \left[\begin{smallmatrix} M_0 & Y_0 \\ N_0 & X_0 \end{smallmatrix}\right]^{-1} \in \mathcal{G}\,\mathrm{H}^\infty$. Since $Y_0(+\infty) = 0$, the function $\tilde{M}X_0 = I - \tilde{N}Y_0$ is boundedly invertible on some right half-plane; hence so is X_0 (because so is $\tilde{M}$). Thus, $Q = Y_0 X_0^{-1}$ is strictly proper (and stabilizing, by Theorem 7.4). ∎

Proof of Theorem 7.9. The equivalence of (ii) and (iii) follows from Theorem 7.4 (and the definitions). The rest follows from ζ-Theorem 4.2 and Lemma B.1 for a suitable ζ. ∎

Proof of Theorem 7.8. From ζ-Theorem 4.1 (for some ζ in the domain of P and R) we observe that the first equivalence and the last claim hold.

If Q is a stabilizing DPF-controller function for P, then it is a [ζ-]stabilizing canonical controller for P_{21} given by the Youla parameterization (pick some ζ in the domain of Q), by ζ-Theorem 4.1. The same holds when Q is a stabilizing controller function for P_{21}, by Theorem 7.4. Therefore, the second equivalence in Theorem 7.8 follows from the first. □

Proof of Remark 7.10. In the proof of Remark 5.1, use $(\cdot - r)$ in place of $(r\cdot)$, for a suitable $r > 0$. □

REFERENCES

[1] R. F. Curtain, *Robustly stabilizing controllers with respect to left-coprime factor perturbations for infinite-dimensional linear systems*, Systems Control Lett., 55 (2006), pp. 509–517.

[2] R. F. Curtain and M. R. Opmeer, *Normalized doubly coprime factorizations for infinite-dimensional linear systems*, Math. Control Signals Systems, 18 (2006), pp. 1–31.

[3] R. F. Curtain, G. Weiss, and M. Weiss, *Coprime factorization for regular linear systems*, Automatica J. IFAC, 32 (1996), pp. 1519–1531.

[4] R. F. Curtain, G. Weiss, and M. Weiss, *Stabilization of irrational transfer functions by controllers with internal loop*, in Systems, Approximation, Singular Integral Operators, and Related Topics, Oper. Theory Adv. Appl. 129, Birkhäuser, Basel, 2001, pp. 179–207.

[5] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[6] B. A. Francis, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, Berlin, 1987.

[7] P. A. Fuhrmann, *On the corona theorem and its application to spectral problems in Hilbert space*, Trans. Amer. Math. Soc., 132 (1968), pp. 55–66.

[8] M. Green, *$\mathscr{H}_\infty$ controller synthesis by J-lossless coprime factorization*, SIAM J. Control Optim., 30 (1992), pp. 522–547.

[9] Y. Inouye, *Parametrization of Compensators for Linear Systems with Transfer Functions of Bounded Type*, Technical report 88-01, Osaka University, Osaka, Japan, 1988.

[10] I. Lasiecka and R. Triggiani, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories.* I. *Abstract Parabolic Systems*, Encyclopedia Math. Appl. 74, Cambridge University Press, Cambridge, UK, 2000.

[11] K. M. Mikkola, *Infinite-Dimensional Linear Systems, Optimal Control and Algebraic Riccati Equations*, Doctoral dissertation, Technical report A452, Helsinki University of Technology, Institute of Mathematics, Espoo, Finland, http://www.math.hut.fi/~kmikkola/research/ (2002).

[12] K. M. Mikkola, *Hankel and Toeplitz operators on nonseparable Hilbert spaces*, submitted.

[13] K. M. Mikkola, *Characterization of transfer functions of Pritchard-Salamon or other realizations with a bounded input or output operator*, Integral Equations Operator Theory, 54 (2006), pp. 427–440.

[14] K. M. Mikkola, *State-feedback stabilization of well-posed linear systems*, Integral Equations Operator Theory, 55 (2006), pp. 249–271.

[15] K. M. Mikkola, *Weakly coprime factorization and continuous-time systems*, submitted.

[16] K. M. Mikkola, *Weakly coprime factorization and state-feedback stabilization of discrete-time systems*, submitted.

[17] K. A. Morris, *Justification of input-output methods for systems with unbounded control and observation*, IEEE Trans. Automat. Control, 44 (1999), pp. 81–85.

[18] A. Quadrat, *The fractional representation approach to synthesis problems: An algebraic analysis viewpoint Part* II: *Internal stabilization*, SIAM J. Control Optim., 42 (2003), pp. 300–320.

[19] A. Quadrat, *On a general structure of the stabilizing controllers based on stable range*, SIAM J. Control Optim., 42 (2004), pp. 2264–2285.

[20] A. Quadrat, *An algebraic interpretation to the operator-theoretic approach to stabilizability.* I. *SISO systems*, Acta Appl. Math., 88 (2005), pp. 1–45.

[21] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[22] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[23] O. J. Staffans, *Coprime factorizations and well-posed linear systems*, SIAM J. Control Optim., 36 (1998), pp. 1268–1292.
[24] O. J. Staffans, *Well-Posed Linear Systems*, Encyclopedia Math. Appl. 103, Cambridge University Press, Cambridge, UK, 2005.
[25] V. A. Tolokonnikov, *Estimates in the Carleson corona theorem, ideals of the algebra H^∞, a problem of Sz.-Nagy*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 113 (1981), pp. 178–198, 267.
[26] S. R. Treil, *Angles between co-invariant subspaces, and the operator corona problem. The Szökefalvi-Nagy problem*, Soviet Math. Dokl., 38 (1989), pp. 394–399.
[27] S. R. Treil, *The stable rank of the algebra H^∞ equals* 1, J. Funct. Anal., 109 (1992), pp. 130–154.
[28] S. R. Treil, *An operator Corona theorem*, Indiana Univ. Math. J., 53 (2004), pp. 1763–1780.
[29] B. van Keulen, *H_∞-Control for Distributed Parameter Systems: A State Space Approach*, Birkhäuser Verlag, Basel, Boston, Berlin, 1993.
[30] L. N. Vaserštein, *Stable rank of rings and dimensionality of topological spaces*, Funct. Anal. Appl., 5 (1971), pp. 102–110.
[31] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press Ser. Signal Process. Optim. Control 7, MIT Press, Cambridge, MA, 1985.
[32] G. Weiss and R. F. Curtain, *Dynamic stabilization of regular linear systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 4–21.
[33] G. Weiss and R. Rebarber, *Optimizability and estimatability for infinite-dimensional linear systems*, SIAM J. Control Optim., 39 (2000), pp. 1204–1232.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2011–2034

SIMPLIFIED FORMULA FOR A CONTROLLER IN OPTIMAL CONTROL PROBLEMS*

JOVAN STEFANOVSKI†

Abstract. We derive a simplified frequency-domain formula for a controller transfer matrix for LQ optimal output feedback control of stochastic systems. For this purpose we apply a generalization of the Wiener–Hopf method. The generalization is characterized by the following three properties: (1) We generalize the usual operator $\{\cdot\}_+$ on rational matrices in the traditional Wiener–Hopf approach. The same algorithms as for computation of $\{\cdot\}_+$ are applicable to compute the new operator. (2) The matrices of the Youla–Kučera parametrization do not appear in the optimal controller transfer matrix $\mathbf{C}$ (they appear only in its derivation), even if the plant is unstable. (3) Unlike the traditional Wiener–Hopf method [D. C. Youla, H. A. Jabr, and J. J. Bongiorno, Jr., *IEEE Trans. Automat. Control*, AC-21 (1976), pp. 319–338], where the spectral factors in the two spectral factorizations are both stable and minimum phase, our spectral factors need to be minimum phase only. Finally, three state-space applications of the formula are presented.

Key words. parametrization of stabilizing controllers, Laplace transformation, Wiener–Hopf approach, partial fraction expansion, spectral factorization

AMS subject classifications. 93C35, 93C95, 93E20, 47A68

DOI. 10.1137/050624765

1. Introduction.

Consider the system of Figure 1. This system is classical in control theory and practice and it originates from the problem of optimal tracking with rejection of disturbances in the presence of measurement noise, which is one of the most important control problems. The variable w represents the plant disturbances, v is measurement noise, and r is reference input (to-be-tracked signal).

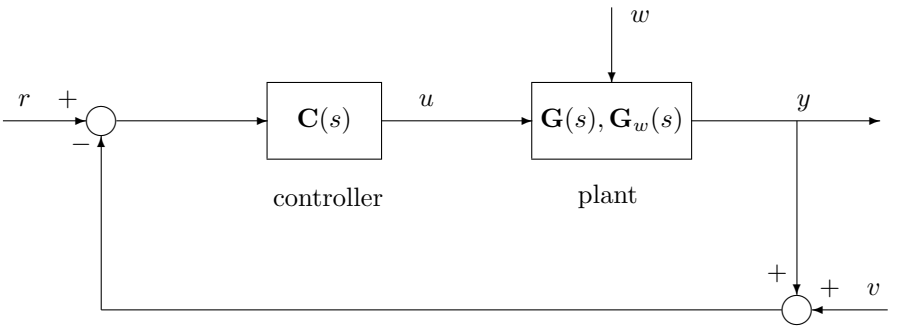


FIG. 1. *Control scheme for tracking and disturbance rejecting.*

The two approaches for linear quadratic (LQ) optimal control by output feedback are the time-domain approach, operating with state-space equations, and the frequency-domain approach, operating with polynomial and/or rational matrices.

The state-space approach (see [3], [13], and references therein), did not solve the optimization control problem for the system of Figure 1, where the spectral characteristics of the signals r, w, and v are arbitrary, because problems with colored

*Received by the editors February 19, 2005; accepted for publication (in revised form) July 6, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sicon/45-6/62476.html
†Control & Informatics Division, JP "Streževo," 7000 Bitola, Republic of Macedonia (jovanstef@mt.net.mk).

measurement noise, problems with a singular covariance matrix of a white measurement noise, problems with a colored reference signal, problems with shape-deterministic inputs, and problems with frequency-dependent weighting functions are not close to the philosophy of this approach. In particular, the methodology of [3] and its generalization in [13] cannot directly solve our state-space applications presented in sections 4.1 and 4.3, because they require frequency-dependent spectral characteristics of inputs, or shape-deterministic inputs. The application in section 4.2 have zero matrix D_{12}, where the matrix D_{12} is defined in [3], but the algorithm of [3] requires $D_{12} = \left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$.

The second approach, with polynomial and/or rational matrices (see [17], [8], [5], [2], [4], [6], [7], [11], and references therein), is based on solving a Wiener–Hopf problem in the frequency domain [17], and uses the parametrization of all stabilizing controllers, known in the control literature as Youla–Kučera parametrization [17], [5]. Unlike the state-space approach, this approach can handle colored input signals, frequency-dependent weighting matrices, shape-deterministic inputs, and any structure at infinity of the plant, so the general LQ optimization problem of the system of Figure 1 can be solved by the existing frequency-domain algorithms.

However, a disadvantage of the frequency-domain algorithms is that the matrices of the Youla–Kučera parametrization appear in the controller transfer matrix, so that it is quite complex, and unsuitable for large-scale systems because the pole-zero cancellation in the controller transfer matrix among its products cannot be done numerically in a reliable way. As a result, the controller order is greater than necessary, and, more important, if the cancellation among the products in $\mathbf{C}$ is not done correctly, the closed-loop system of Figure 1 can be unstable.[1]

On the other hand, given the optimal controller transfer matrix in the frequency domain, when state-space equations are employed, matrices of Youla–Kučera parametrization can be removed from the optimal controller transfer matrix. Two different approaches for this removal are presented in [10] and [14]. These results are satisfactory, from a viewpoint of practical controller construction, but the solved class of systems is limited. For example, in [10], the inputs are white noises (except the reference signal), having positive definite spectral densities, the weighting matrices in the minimizing criterion are also positive definite, the plant transfer matrices are strict proper, and the minimal realizations of $\mathbf{G}_w$ and $\mathbf{G}$ have the same poles.

In this work we solve the problem of obtaining a simplified controller transfer matrix (see (4.1)), which is free of the matrices of Youla–Kučera parametrization. This result allows us to solve, in respect to the cited results, more general optimal control problems.

Before we proceed with an explanation of our contribution, some remarks are in order.

Remarks on the notation and techniques used. All considered stochastic processes are stationary. By $\mathcal{E}\{\cdot\}$ we denote the mean value of a stochastic variable.

Vectors are denoted by lowercase symbols and matrices are denoted by uppercase symbols.

If $x(t)$ is a vector stochastic variable, then the matrix autocorrelation function is defined by $\Phi_{xx}(\tau) = \mathcal{E}\left\{x(t) \cdot x^{\mathrm{T}}(t-\tau)\right\}$.

By $\mathbf{\Phi}(s)$ we denote the two-sided Laplace transformation of a function $\Phi(t)$, i.e., $\mathbf{\Phi}(s) = \int_{-\infty}^{\infty} \Phi(t) e^{-st} \mathrm{d}t$, $\alpha_1 \leq \Re(s) \leq \alpha_2$, $\alpha_1 < 0 < \alpha_2$, where by $\Re(s)$ we denote the

[1] The matrices of the Youla–Kučera parametrization do not appear in the controller in Corollary 1 in [17], if the plant is stable.

real part of s, and the inverse Laplace transformation is $\Phi(t) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \mathbf{\Phi}(s)e^{st}\mathrm{d}s$. All considered functions of s are rational and with real coefficients. If not ambiguous, we shall omit the Laplace argument s in the functions.

Poles and zeros of a rational matrix are defined through its McMillan form.

The notation $\mathbf{\Phi}(s) \leq O(s^{\nu})$ means that no entry of $\mathbf{\Phi}$ grows faster than s^{ν} as $s \to \infty$. The rational matrix $\mathbf{\Phi}(s)$ is proper if $\mathbf{\Phi}(s) \leq O(1)$ and is strict proper if $\mathbf{\Phi}(s) \leq O(s^{-1})$. If $\mathbf{\Phi}_1 \leq O(s^{\nu_1})$ and $\mathbf{\Phi}_2 \leq O(s^{\nu_2})$, then $\mathbf{\Phi}_1\mathbf{\Phi}_2 \leq O(s^{\nu_1+\nu_2})$.

The transfer matrix $\mathbf{H}(s)$ is *stable* if $\mathbf{H}(s)$ is analytic in $\Re(s) \geq 0$. The transfer matrix $\mathbf{H}(s)$ is *minimum phase* if the zeros of $\mathbf{H}(s)$ are in $\Re(s) < 0$. Matrix A is stable if $(sI - A)^{-1}$ is stable.

If $\mathbf{\Phi}(s)$ is a rational matrix that is analytic on the imaginary axis, the integral $\int_{-i\infty}^{i\infty} \mathbf{\Phi}(s)\mathrm{d}s$ is finite if and only if $\mathbf{\Phi}(s) \leq O(s^{-2})$. If $\mathbf{\Phi}(s) \leq O(s^{-2})$ and $\mathbf{\Phi}$ is analytic in $\Re(s) \geq 0$ (or $\Re(s) \leq 0$), then $\int_{-i\infty}^{i\infty} \mathbf{\Phi}(s)\mathrm{d}s = 0$.

By $\mathrm{Tr}\{\cdot\}$ we denote the trace of a square matrix, and by the superscript T we denote the matrix transposition. By $\mathbf{H}^{\sim}$ we denote the matrix $\mathbf{H}^{\mathrm{T}}(-s)$. By $\mathbf{H}^{-\sim}$ we denote the matrix $(\mathbf{H}^{\sim})^{-1}$.

We introduce a generalization of the usual operators $\{\cdot\}_+$ and $\{\cdot\}_-$ [8], [17]. Let τ be a set of complex numbers and $\mathbf{\Phi}$ be a rational matrix. There exists a partial fraction expansion $\mathbf{\Phi} = \mathbf{\Phi}_\tau + \mathbf{\Phi}_\sigma$ such that the poles of $\mathbf{\Phi}_\tau$ are in τ, the poles of $\mathbf{\Phi}_\sigma$ are not in τ, and $\mathbf{\Phi}_\sigma$ is strict proper. Then we define the operators $\{\mathbf{\Phi}\}_\sigma = \mathbf{\Phi}_\sigma$ and $\{\mathbf{\Phi}\}_\tau = \mathbf{\Phi}_\tau$. If τ is the open right complex half-plane, then $\{\mathbf{\Phi}\}_+ = \{\mathbf{\Phi}\}_\sigma$.

The operators $\{\cdot\}_\sigma$ and $\{\cdot\}_\tau$ can be computed exactly as the operators $\{\cdot\}_+$ and $\{\cdot\}_-$, that is, by partial fraction expansion. □

To state the problem of the paper, consider the plant in Figure 1 with its proper transfer matrices $\mathbf{G}$ and $\mathbf{G}_w$, so that

$$\mathbf{y} = \mathbf{G}\mathbf{u} + \mathbf{G}_w\mathbf{w} \ .$$

Define the tracking error as $\mathbf{e} = \mathbf{r} - \mathbf{y}$. In respect to Figure 1, we have

$$\mathbf{e} = \mathbf{r} - \mathbf{y} = \mathbf{H}\mathbf{x} + (I - \mathbf{H})\mathbf{v} \ , \tag{1.1}$$

$$\mathbf{u} = \mathbf{C}\mathbf{H}(\mathbf{x} - \mathbf{v}) \ , \tag{1.2}$$

where $\mathbf{H}$ is the sensitivity matrix

$$\mathbf{H} = (I + \mathbf{G}\mathbf{C})^{-1} \ ,$$

and

$$\mathbf{x} = \mathbf{r} - \mathbf{G}_w\mathbf{w} \ . \tag{1.3}$$

We assume the following minimization criterion:

$$\begin{aligned} \mathcal{J} &= \mathcal{E}\left\{e^{\mathrm{T}}(t)Q_e e(t) + u^{\mathrm{T}}(t)Q_u u(t)\right\} \\ &= \mathrm{Tr}\left\{Q_e\Phi_{ee}(0)\right\} + \mathrm{Tr}\left\{Q_u\Phi_{uu}(0)\right\} \\ &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \left(\mathrm{Tr}\left\{Q_e\mathbf{\Phi}_{ee}\right\} + \mathrm{Tr}\left\{Q_u\mathbf{\Phi}_{uu}\right\}\right)\mathrm{d}s \ , \end{aligned} \tag{1.4}$$

where Q_e and Q_u are some positive semidefinite weighting matrices, for an unknown stabilizing controller.

We assume that the stochastic signals r, w, and v are pairwise stochastically independent and with spectral densities that are proper and positive semidefinite on the imaginary axis. We also allow shape-deterministic inputs r, w, and v (or parts of them) [1], [16].

We have

$$\mathbf{\Phi}_{ee} = \mathbf{H}\mathbf{\Phi}_{xx}\mathbf{H}^{\sim} + (I - \mathbf{H})\mathbf{\Phi}_{vv}(I - \mathbf{H})^{\sim} \ , \tag{1.5}$$

$$\mathbf{\Phi}_{xx} = \mathbf{\Phi}_{rr} + \mathbf{G}_w\mathbf{\Phi}_{ww}\mathbf{G}_w^{\sim} \ ,$$

$$\mathbf{\Phi}_{uu} = \mathbf{CH}(\mathbf{\Phi}_{xx} + \mathbf{\Phi}_{vv})(\mathbf{CH})^{\sim} \ . \tag{1.6}$$

The criterion (1.4) depends on the controller transfer matrix $\mathbf{C}$ nonquadratically. As usual, we shall transform the problem into an LQ problem, so that the criterion depends on the unknown rational matrix (Youla–Kučera parameter) quadratically. In section 2 we formulate a general theorem, which presents a solution of the above problem, and in section 3 we present the proof of the theorem. In section 4 we present three applications of the theorem, i.e., state-space solutions of optimal control problems.

2. Formulation of main result. Throughout this paper we assume the following.

ASSUMPTION 1. *The rational matrix $\mathbf{\Phi}_{xx}$ is strict proper.*

There are two reasons for such an assumption: (1) The criterion (1.4) can be finite only if $\operatorname{Tr}\{Q_e\mathbf{H}\mathbf{\Phi}_{xx}\mathbf{H}^{\sim}\} \le O(s^{-2})$. If $\mathbf{\Phi}_{xx}$ is only proper (not strict proper), having in mind that $\mathbf{H}$ can be only proper when dealing with proper controllers and plants, this condition is a constraint on optimal $\mathbf{C}$. Such a constrained optimization problem is not a topic of this paper. (2) There is not a physical meaning stating the reference signal r to contain white noise or impulsive shape-deterministic signal, when dealing with strict proper $\mathbf{G}_w\mathbf{\Phi}_{ww}\mathbf{G}_w^{\sim}$, as we do in our applications in section 4.

PROPOSITION 2.1. *Given a proper transfer matrix $\mathbf{G}$ without unstable hidden modes* [17], *and symmetric positive semidefinite matrices Q_u and Q_e, if Assumption* 2 *below holds, then there exist proper and stable rational matrices $\mathbf{\Psi}$, $\mathbf{\Theta}$, $\mathbf{U}$, $\mathbf{V}$, $\mathbf{\Psi}_1$, $\mathbf{\Theta}_1$, $\mathbf{U}_1$, and $\mathbf{V}_1$, of which $\mathbf{\Theta}$, $\mathbf{\Theta}_1$, $\mathbf{U}$, and $\mathbf{U}_1$ are biproper, and $\mathbf{V}$ and $\mathbf{V}_1$ are strict proper, such that*

$$\mathbf{G} = \mathbf{\Theta}_1^{-1}\mathbf{\Psi}_1 = \mathbf{\Psi}\mathbf{\Theta}^{-1} \ , \tag{2.1}$$

$$\begin{bmatrix} \mathbf{U} & \mathbf{V} \\ -\mathbf{\Psi}_1 & \mathbf{\Theta}_1 \end{bmatrix} \begin{bmatrix} \mathbf{\Theta} & -\mathbf{V}_1 \\ \mathbf{\Psi} & \mathbf{U}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{\Theta} & -\mathbf{V}_1 \\ \mathbf{\Psi} & \mathbf{U}_1 \end{bmatrix} \begin{bmatrix} \mathbf{U} & \mathbf{V} \\ -\mathbf{\Psi}_1 & \mathbf{\Theta}_1 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \ , \tag{2.2}$$

$$\mathbf{\Psi}^{\sim}Q_e\mathbf{\Psi} + \mathbf{\Theta}^{\sim}Q_u\mathbf{\Theta} = I \ . \tag{2.3}$$

If $\mathbf{G}$ is strict proper, then $\mathbf{\Psi}$ and $\mathbf{\Psi}_1$ are strict proper.

ASSUMPTION 2. *Rational matrix $Q_u + \mathbf{G}^{\sim}Q_e\mathbf{G}$ is biproper and without zeros on the imaginary axis.*

Proposition 2.1 is proved in [9], but for the case of positive definite matrices Q_u and Q_e.

The system of Figure 1 is internally stable if and only if the four transfer matrices $\mathbf{H}$, $\mathbf{HG}$, $\mathbf{CH}$, and $\mathbf{CHG}$ are proper and stable. A controller that stabilizes the system is called a stabilizing controller.

To represent the class of all stabilizing controllers [15], we consider plant transfer matrix $\mathbf{G}$, satisfying Assumption 2.

PROPOSITION 2.2. *A controller* $\mathbf{C}$ *is stabilizing if and only if there exists a proper and stable rational matrix* $\mathbf{S}$ *such that*

$$\mathbf{C} = \mathbf{\Theta W} \left(I - \mathbf{\Psi W}\right)^{-1} \ , \tag{2.4}$$

$$\det \left(I - \mathbf{\Psi W}\right) \not\equiv 0 \ , \tag{2.5}$$

where

$$\mathbf{W} = \mathbf{V} + \mathbf{S\Theta}_1 \ . \tag{2.6}$$

By Proposition 2.2, transfer matrices $\mathbf{H}$ and $\mathbf{CH}$ can be expressed as

$$\mathbf{H} = I - \mathbf{\Psi W} \ , \quad \mathbf{CH} = \mathbf{\Theta W} \ . \tag{2.7}$$

Equations (1.1) and (1.2), having in mind (1.3), are

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} -\mathbf{H} & -\mathbf{HG}_w \\ \mathbf{CH} & \mathbf{CHG}_w \end{bmatrix} \begin{bmatrix} \mathbf{r} - \mathbf{v} \\ -\mathbf{w} \end{bmatrix} + \begin{bmatrix} \mathbf{r} - \mathbf{v} \\ 0 \end{bmatrix} \ ;$$

therefore, besides internal system stability, transfer matrices $\mathbf{HG}_w$ and $\mathbf{CHG}_w$ have to be also proper and stable. We satisfy that condition by assuming that the proper plant transfer matrix $\mathbf{G}_w$ is

$$\mathbf{G}_w = \mathbf{\Theta}_1^{-1} \mathbf{\Psi}_2 \tag{2.8}$$

for some proper stable rational matrix $\mathbf{\Psi}_2$. With (2.4), (2.6), and (2.8), it is easy to prove that transfer matrices $\mathbf{HG}_w$ and $\mathbf{CHG}_w$ are proper and stable.

Remark 2.1. If $\mathbf{G}$ is strict proper, by Proposition 2.2 we have that (1) every stabilizing controller is proper, and (2) the condition (2.5) is void. (See [15] also.)

In the following proposition we present conditions on the possible unstable inputs $\mathbf{r}$, $\mathbf{w}$, and $\mathbf{v}$ such that the outputs $\mathbf{e}$ and $\mathbf{u}$ are strict proper and stable. This result we apply when dealing with shape-deterministic inputs [1], [16], for example, when $\mathbf{r}$ or $\mathbf{w}$ includes step or ramp functions, the most important inputs in respect to tracking and disturbance rejection.

PROPOSITION 2.3. *Let* $\mathbf{C}$ *be a stabilizing controller of Proposition* 2.2*, parametrized by* $\mathbf{S}$*. A necessary and sufficient condition for* $\mathbf{u}$ *and* $\mathbf{e}$ *to be strict proper and stable is that* $\mathbf{\Theta}_1\mathbf{r} - \mathbf{\Psi}_2\mathbf{w}$ *and* $\mathbf{Wv}$ *are strict proper and stable.*

Since we do not impose constraints on optimal $\mathbf{W}$ (or on optimal $\mathbf{C}$) in this paper, and since we cannot count on cancellation of the unstable terms of $\mathbf{\Theta}_1\mathbf{r}$ and $\mathbf{\Psi}_2\mathbf{w}$ in $\mathbf{\Theta}_1\mathbf{r} - \mathbf{\Psi}_2\mathbf{w}$ because of uncertainty in $\mathbf{w}$, when dealing with shape-deterministic inputs r and/or w and/or v, we assume the following.

ASSUMPTION 3. *The rational functions* $\mathbf{\Theta}_1\mathbf{r}$*,* $\mathbf{\Psi}_2\mathbf{w}$*, and* $\mathbf{v}$ *are strict proper and stable.*

We also introduce the following.

ASSUMPTION 4. *There exists the spectral factorization*

$$\mathbf{\Phi}_{vv} + \mathbf{\Phi}_{xx} = \mathbf{\Phi}\mathbf{\Phi}^{\sim} \ ,$$

where $\mathbf{\Phi}$ *is nonsingular and minimum phase.*

If possible, we choose spectral factor $\mathbf{\Phi}$ so that the following holds.

ASSUMPTION 5. *Rational matrix* $\mathbf{\Theta}_1\mathbf{\Phi}$ *is stable and minimum phase.*

This condition is not hard to satisfy, at least for the purpose of our applications in section 4.

By Assumptions 3 and 4, rational matrix $\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}$ does not have poles on the imaginary axis.

Denote by τ the union of the sets of poles of $\mathbf{\Psi}^{\sim}$, $\mathbf{\Theta}^{\sim}$, and unstable poles of $\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}$. Introduce rational matrices $\mathbf{L}$, $\mathbf{L}_\sigma$, and $\mathbf{L}_\tau$ by

$$\mathbf{L} = \mathbf{\Psi}^{\sim} Q_e \mathbf{\Phi}_{xx} \mathbf{\Phi}^{-\sim} = \mathbf{L}_\sigma + \mathbf{L}_\tau \ , \quad \mathbf{L}_\sigma \stackrel{\text{def}}{=} \{\mathbf{L}\}_\sigma \ , \quad \mathbf{L}_\tau \stackrel{\text{def}}{=} \{\mathbf{L}\}_\tau \ . \tag{2.9}$$

THEOREM 2.4. *Suppose in addition to Assumptions* 1–5 *one has the following.*

ASSUMPTION 6. *No pole of* $\mathbf{\Phi}$ *belongs to* τ.

ASSUMPTION 7. $\mathbf{\Phi}^{-1} \leq O(s)$.

Then an optimal controller is given by (2.4)*, i.e.,*

$$\mathbf{C} = \mathbf{\Theta}\mathbf{W}\left(I - \mathbf{\Psi}\mathbf{W}\right)^{-1} \ , \tag{2.10}$$

where

$$\mathbf{W} = \mathbf{L}_\sigma \mathbf{\Phi}^{-1} \ , \tag{2.11}$$

under condition (2.5) *(which is void if* $\mathbf{G}$ *is strict proper).*

A consequence of Theorem 2.4 is the following.

COROLLARY 2.5. *Under Assumptions* 1–7*, if* v *is a white noise, then an optimal controller is given by* (2.10) *and* (2.11)*, where* $\mathbf{L}_\sigma$ *is given by*

$$\mathbf{L}_\sigma = \{\mathbf{\Psi}^{\sim} Q_e \mathbf{\Phi}\}_\sigma \ . \tag{2.12}$$

3. Proof of main result. We start with the following.

Proof of Proposition 2.1. Since $\mathbf{G}$ is proper, there exist matrices A, B, C, and D, with an (A, B) stabilizable pair and a (C, A) detectable pair, such that

$$\mathbf{G} = D + C(sI - A)^{-1}B \stackrel{\text{def}}{=} \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \ . \tag{3.1}$$

Define matrices

$$\begin{bmatrix} \mathbf{U} & \mathbf{V} \\ -\mathbf{\Psi}_1 & \mathbf{\Theta}_1 \end{bmatrix} = \left[\begin{array}{c|cc} A - FC & B - FD & F \\ \hline N_u K & N_u & 0 \\ -C & -D & I \end{array}\right] \ , \tag{3.2}$$

$$\begin{bmatrix} \mathbf{\Theta} & -\mathbf{V}_1 \\ \mathbf{\Psi} & \mathbf{U}_1 \end{bmatrix} = \left[\begin{array}{c|cc} A - BK & BN_u^{-1} & F \\ \hline -K & N_u^{-1} & 0 \\ C - DK & DN_u^{-1} & I \end{array}\right] \ , \tag{3.3}$$

where matrices N_u, K, and F are arbitrary with the properties that matrix N_u is nonsingular and matrices $A - BK$ and $A - FC$ are stable. It can be checked that identities (2.1) and (2.2) hold.

We shall choose matrices N_u and K so that (2.3) is satisfied.

By Assumption 2, matrix $Q_{u1} = Q_u + D^{\mathrm{T}} Q_e D$ is nonsingular, and rational matrix $Q_u + \mathbf{G}^{\sim} Q_e \mathbf{G}$ does not have zeros on the imaginary axis. The zeros of that rational

matrix are the same as the eigenvalues of the Hamiltonian matrix, associated with the following continuous time algebraic Riccati equation (CARE):

$$YA + A^{\mathrm{T}}Y - (YB + C^{\mathrm{T}}Q_eD)Q_{u1}^{-1}(B^{\mathrm{T}}Y + D^{\mathrm{T}}Q_eC) + C^{\mathrm{T}}Q_eC = 0 \ . \tag{3.4}$$

We shall use the following lemma [3].

LEMMA 3.1. *Consider CARE*

$$YA + A^{\mathrm{T}}Y - (YB + \Sigma_{12})\Sigma_{22}^{-1}(B^{\mathrm{T}}Y + \Sigma_{21}) + \Sigma_{11} = 0 \ , \tag{3.5}$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \Sigma^{\mathrm{T}} \ .$$

Sufficient conditions for existence of a stabilizing solution Y *of CARE* (3.5)*, i.e., such that matrix* $A - B\Sigma_{22}^{-1}(B^{\mathrm{T}}Y + \Sigma_{21})$ *is stable, are as follows:*

(i) $\Sigma \geq 0$*;*

(ii) *the Hamiltonian matrix associated with* (3.5) *does not have eigenvalues on the imaginary axis;*

(iii) *the matrix pair* (A, B) *is stabilizable.*

By Lemma 3.1, the conditions of the positive semidefiniteness of Q_e and Q_u, nonsingularity of Q_{u1}, absence of purely imaginary eigenvalues of the Hamiltonian matrix, and stabilizability of the pair (A, B) are sufficient for existence of a stabilizing solution Y of CARE (3.4), i.e., for the stability of matrix $A - BK$, where

$$K = Q_{u1}^{-1}(B^{\mathrm{T}}Y + D^{\mathrm{T}}Q_eC) \ . \tag{3.6}$$

Using (3.4), it is easy to check the following spectral factorization:

$$Q_u + \mathbf{G}^{\sim}Q_e\mathbf{G} = \mathbf{\Delta}^{\sim}\mathbf{\Delta} \ , \tag{3.7}$$

where

$$\mathbf{\Delta} = N_u(I + K\mathbf{\Omega}B) \ , \quad N_u^2 = Q_{u1} \ , \quad \mathbf{\Omega} = (sI - A)^{-1} \ . \tag{3.8}$$

Now define

$$\mathbf{\Theta} = \mathbf{\Delta}^{-1} = (I - K\mathbf{\Omega}_Y B)N_u^{-1} \ , \quad \mathbf{\Omega}_Y = (sI - A + BK)^{-1} \tag{3.9}$$

and

$$\mathbf{\Psi} = \mathbf{G}\mathbf{\Theta} = \left[D + (C - DK)\mathbf{\Omega}_Y B\right] N_u^{-1} \ . \tag{3.10}$$

These $\mathbf{\Psi}$ and $\mathbf{\Theta}$ coincide with $\mathbf{\Psi}$ and $\mathbf{\Theta}$ defined in (3.3). The identity (3.7) becomes (2.3). □

Proof of Proposition 2.3. Express, at first, $\mathbf{u}$ and $\mathbf{e}$:

$$\begin{aligned} \mathbf{u} &= \mathbf{C}\mathbf{H}(\mathbf{x} - \mathbf{v}) = \mathbf{\Theta}(\mathbf{V} + \mathbf{S}\mathbf{\Theta}_1)(\mathbf{x} - \mathbf{v}) \\ &= (\mathbf{V}_1\mathbf{\Theta}_1 + \mathbf{\Theta}\mathbf{S}\mathbf{\Theta}_1)(\mathbf{x} - \mathbf{v}) = (\mathbf{V}_1 + \mathbf{\Theta}\mathbf{S})\mathbf{\Theta}_1(\mathbf{x} - \mathbf{v}) \ , \\ \mathbf{e} &= \mathbf{H}(\mathbf{x} - \mathbf{v}) + \mathbf{v} = (I - \mathbf{\Psi}\mathbf{W})(\mathbf{x} - \mathbf{v}) + \mathbf{v} \\ &= (I - \mathbf{\Psi}\mathbf{V} - \mathbf{\Psi}\mathbf{S}\mathbf{\Theta}_1)(\mathbf{x} - \mathbf{v}) + \mathbf{v} = (\mathbf{U}_1 - \mathbf{\Psi}\mathbf{S})\mathbf{\Theta}_1(\mathbf{x} - \mathbf{v}) + \mathbf{v} \ . \end{aligned}$$

Further we have

$$\mathbf{Uu}-\mathbf{Ve}=\mathbf{U}(\mathbf{V}_1+\mathbf{\Theta S})\mathbf{\Theta}_1(\mathbf{x}-\mathbf{v})-\mathbf{V}(\mathbf{U}_1-\mathbf{\Psi S})\mathbf{\Theta}_1(\mathbf{x}-\mathbf{v})-\mathbf{Vv}$$

$$=\mathbf{S\Theta}_1(\mathbf{x}-\mathbf{v})-\mathbf{Vv}=\mathbf{S\Theta}_1\mathbf{x}-\mathbf{Wv}=\mathbf{S}(\mathbf{\Theta}_1\mathbf{r}-\mathbf{\Psi}_2\mathbf{w})-\mathbf{Wv}\ ,$$

$$\mathbf{\Psi}_1\mathbf{u}+\mathbf{\Theta}_1\mathbf{e}=\mathbf{\Psi}_1(\mathbf{V}_1+\mathbf{\Theta S})\mathbf{\Theta}_1(\mathbf{x}-\mathbf{v})+\mathbf{\Theta}_1(\mathbf{U}_1-\mathbf{\Psi S})\mathbf{\Theta}_1(\mathbf{x}-\mathbf{v})+\mathbf{\Theta}_1\mathbf{v}$$

$$=\mathbf{\Theta}_1(\mathbf{x}-\mathbf{v})+\mathbf{\Theta}_1\mathbf{v}=\mathbf{\Theta}_1\mathbf{x}=\mathbf{\Theta}_1\mathbf{r}-\mathbf{\Psi}_2\mathbf{w}\ ,$$

i.e.,

$$\begin{bmatrix}\mathbf{U} & \mathbf{V}\\ -\mathbf{\Psi}_1 & \mathbf{\Theta}_1\end{bmatrix}\begin{bmatrix}-\mathbf{u}\\ \mathbf{e}\end{bmatrix}=\begin{bmatrix}-\mathbf{S}(\mathbf{\Theta}_1\mathbf{r}-\mathbf{\Psi}_2\mathbf{w})+\mathbf{Wv}\\ \mathbf{\Theta}_1\mathbf{r}-\mathbf{\Psi}_2\mathbf{w}\end{bmatrix}\ .$$

Therefore, the necessity is proved. The sufficiency is obvious if we solve the above matrix system for $-\mathbf{u}$ and $\mathbf{e}$, using (2.2). □

Proof of Theorem 2.4. Rational matrix $\mathbf{L}$ (2.9) is strict proper. This fact easily follows by Assumptions 1 and 7.[2]

Introduce the rational matrix

$$\mathbf{K}=\mathbf{L}-\mathbf{V\Phi}=\mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}_{xx}\mathbf{\Phi}^{-\sim}-\mathbf{V\Phi}\ , \tag{3.11}$$

which is also strict proper, because $\mathbf{V}$ is strict proper.

To find another representation of $\mathbf{K}$ besides (3.11), which will be crucial in our proof, we use the equation

$$\mathbf{U\Theta}+\mathbf{V\Psi}=I \tag{3.12}$$

for $\mathbf{U}$ and $\mathbf{V}$, which is contained in (2.2). Having in mind that a particular solution of (3.12) for $\mathbf{U}$, $\mathbf{V}$, by (2.3), is $\mathbf{\Theta}^{\sim}Q_u$, $\mathbf{\Psi}^{\sim}Q_e$, the general solution of (3.12) is

$$\mathbf{U}=\mathbf{\Theta}^{\sim}Q_u-\mathbf{R\Psi}_1\ , \tag{3.13}$$

$$\mathbf{V}=\mathbf{\Psi}^{\sim}Q_e+\mathbf{R\Theta}_1\ , \tag{3.14}$$

where $\mathbf{R}$ is a rational matrix. Using the identities (3.13), (3.14), and (2.2), we obtain

$$\mathbf{R}=-\mathbf{\Psi}^{\sim}Q_e\mathbf{U}_1+\mathbf{\Theta}^{\sim}Q_u\mathbf{V}_1\ . \tag{3.15}$$

Conversely, it is easy to check that this $\mathbf{R}$ satisfies the pair of equations (3.13) and (3.14). So, we have

$$\mathbf{K}=\mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}_{xx}\mathbf{\Phi}^{-\sim}-\mathbf{V\Phi}=\mathbf{\Psi}^{\sim}Q_e\left(\mathbf{\Phi\Phi}^{\sim}-\mathbf{\Phi}_{vv}\right)\mathbf{\Phi}^{-\sim}-\left(\mathbf{\Psi}^{\sim}Q_e+\mathbf{R\Theta}_1\right)\mathbf{\Phi}$$

$$=-\mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}-\mathbf{R\Theta}_1\mathbf{\Phi}$$

$$=-\mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}-\left(-\mathbf{\Psi}^{\sim}Q_e\mathbf{U}_1+\mathbf{\Theta}^{\sim}Q_u\mathbf{V}_1\right)\mathbf{\Theta}_1\mathbf{\Phi}\ . \tag{3.16}$$

By (3.16), matrix $\mathbf{K}$ does not have purely imaginary poles.

[2]Rational matrix $\mathbf{L}$ is strict proper without Assumption 7 (see Remark 3.3).

Using matrix $\mathbf{K}$ defined in (3.11), and (1.5), (1.6), and (2.7), the integrand-function in the criterion (1.4) can be expressed as

$$\begin{aligned}
&\mathrm{Tr}\,\{Q_e\mathbf{\Phi}_{ee}\} + \mathrm{Tr}\,\{Q_u\mathbf{\Phi}_{uu}\} \\
&= \mathrm{Tr}\,\{Q_e\,[\mathbf{H}\mathbf{\Phi}_{xx}\mathbf{H}^{\sim} + (I-\mathbf{H})\mathbf{\Phi}_{vv}(I-\mathbf{H})^{\sim}]\} + \mathrm{Tr}\,\{Q_u\mathbf{CH}(\mathbf{\Phi}_{xx}+\mathbf{\Phi}_{vv})(\mathbf{CH})^{\sim}\} \\
&= \mathrm{Tr}\,\{Q_e\,[(I-\mathbf{\Psi}\mathbf{W})\mathbf{\Phi}_{xx}(I-\mathbf{W}^{\sim}\mathbf{\Psi}^{\sim}) + \mathbf{\Psi}\mathbf{W}\mathbf{\Phi}_{vv}\mathbf{W}^{\sim}\mathbf{\Psi}^{\sim}]\} \\
&\quad + \mathrm{Tr}\,\{Q_u\mathbf{\Theta}\mathbf{W}(\mathbf{\Phi}_{xx}+\mathbf{\Phi}_{vv})\mathbf{W}^{\sim}\mathbf{\Theta}^{\sim}\} \\
&= \mathrm{Tr}\,\{Q_e\,[\mathbf{\Phi}_{xx} - \mathbf{\Phi}_{xx}\mathbf{W}^{\sim}\mathbf{\Psi}^{\sim} - \mathbf{\Psi}\mathbf{W}\mathbf{\Phi}_{xx} + \mathbf{\Psi}\mathbf{W}(\mathbf{\Phi}_{xx}+\mathbf{\Phi}_{vv})\mathbf{W}^{\sim}\mathbf{\Psi}^{\sim}]\} \\
&\quad + \mathrm{Tr}\,\{Q_u\mathbf{\Theta}\mathbf{W}(\mathbf{\Phi}_{xx}+\mathbf{\Phi}_{vv})\mathbf{W}^{\sim}\mathbf{\Theta}^{\sim}\} \\
&= \mathrm{Tr}\,\{Q_e\,(\mathbf{\Phi}_{xx} - \mathbf{\Phi}_{xx}\mathbf{W}^{\sim}\mathbf{\Psi}^{\sim} - \mathbf{\Psi}\mathbf{W}\mathbf{\Phi}_{xx})\} \\
&\quad + \mathrm{Tr}\,\{(\mathbf{\Psi}^{\sim}Q_e\mathbf{\Psi} + \mathbf{\Theta}^{\sim}Q_u\mathbf{\Theta})\mathbf{W}(\mathbf{\Phi}_{xx}+\mathbf{\Phi}_{vv})\mathbf{W}^{\sim}\} \\
&= \mathrm{Tr}\,\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{\Phi}^{\sim}\mathbf{W}^{\sim}\} - \mathrm{Tr}\,\{\mathbf{W}\mathbf{\Phi}\mathbf{L}^{\sim}\} + \mathrm{Tr}\,\{\mathbf{W}\mathbf{\Phi}\mathbf{\Phi}^{\sim}\mathbf{W}^{\sim}\} \\
&= \mathrm{Tr}\,\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{L}^{\sim}\} + \mathrm{Tr}\,\left\{\left(\mathbf{W}-\mathbf{L}\mathbf{\Phi}^{-1}\right)\mathbf{\Phi}\mathbf{\Phi}^{\sim}\left(\mathbf{W}-\mathbf{L}\mathbf{\Phi}^{-1}\right)^{\sim}\right\} \\
&= \mathrm{Tr}\,\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{L}^{\sim}\} + \mathrm{Tr}\,\{(\mathbf{S}\mathbf{\Theta}_1\mathbf{\Phi}-\mathbf{K})\,(\mathbf{S}\mathbf{\Theta}_1\mathbf{\Phi}-\mathbf{K})^{\sim}\}\ , \qquad (3.17)
\end{aligned}$$

where identities (2.3) and $\mathbf{W}-\mathbf{L}\mathbf{\Phi}^{-1} = (\mathbf{S}\mathbf{\Theta}_1\mathbf{\Phi}-\mathbf{K})\mathbf{\Phi}^{-1}$ have been also used.

With respect to (3.17), we shall prove the following.

LEMMA 3.2. *Rational function* $\mathrm{Tr}\,\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{L}^{\sim}\}$ *is* $\leq O(s^{-2})$ *and does not have purely imaginary poles.*

Proof. The first property is obvious. Concerning the absence of purely imaginary poles, compute at first

$$\begin{aligned}
&\mathrm{Tr}\,\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{L}^{\sim}\} = \mathrm{Tr}\,\{Q_e(\mathbf{\Phi}\mathbf{\Phi}^{\sim} - \mathbf{\Phi}_{vv})\} - \mathrm{Tr}\,\{(\mathbf{K}+\mathbf{V}\mathbf{\Phi})(\mathbf{K}+\mathbf{V}\mathbf{\Phi})^{\sim}\} \\
&= \mathrm{Tr}\,\{(\mathbf{V}^{\sim}\mathbf{\Psi}^{\sim}Q_e + Q_e\mathbf{\Psi}\mathbf{V} - Q_e)\,\mathbf{\Phi}_{vv}\} - \mathrm{Tr}\,\{\mathbf{K}\mathbf{K}^{\sim}\} \\
&\quad + \mathrm{Tr}\,\{\mathbf{\Phi}^{\sim}\,(Q_e + \mathbf{V}^{\sim}\mathbf{R}\mathbf{\Theta}_1 + \mathbf{\Theta}_1^{\sim}\mathbf{R}^{\sim}\mathbf{V} - \mathbf{V}^{\sim}\mathbf{V})\,\mathbf{\Phi}\}\ , \qquad (3.18)
\end{aligned}$$

where expression (3.11) for $\mathbf{K}$ is used. The matrices in the first two traces of (3.18) do not have purely imaginary poles.

Concerning the matrix in the third trace of (3.18), we have

$$Q_e + \mathbf{V}^{\sim}\mathbf{R}\mathbf{\Theta}_1 + \mathbf{\Theta}_1^{\sim}\mathbf{R}^{\sim}\mathbf{V} - \mathbf{V}^{\sim}\mathbf{V} = \mathbf{Z} + \mathbf{\Theta}_1^{\sim}\mathbf{R}^{\sim}\mathbf{R}\mathbf{\Theta}_1\ , \qquad (3.19)$$

where identity (3.14) is used, and matrix $\mathbf{Z}$ by

$$\mathbf{Z} = Q_e - Q_e\mathbf{\Psi}\mathbf{\Psi}^{\sim}Q_e$$

is introduced. We shall prove that matrix $\mathbf{Z}$ can be represented as

$$\mathbf{Z} = \mathbf{\Theta}_1^{\sim}\hat{\mathbf{M}}\mathbf{\Theta}_1 \qquad (3.20)$$

for some matrix $\hat{\mathbf{M}}$ that does not have purely imaginary poles. Since matrices $\mathbf{\Theta}_1\mathbf{\Phi}$ and $\mathbf{R}$ (3.15) do not have purely imaginary poles, by (3.18) and (3.19) it will follow

that rational function $\mathrm{Tr}\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\{\mathbf{L}\mathbf{L}^\sim\}$ does not have purely imaginary poles, as required.

Define matrix $\hat{\mathbf{G}}$, which is analytic on the imaginary axis, by

$$\hat{\mathbf{G}} = \begin{bmatrix} Q_u^{\frac{1}{2}}\mathbf{\Theta} \\ Q_e^{\frac{1}{2}}\mathbf{\Psi} \end{bmatrix} . \tag{3.21}$$

By $\hat{\mathbf{G}}^\sim\hat{\mathbf{G}} = I$, it has full column rank on the imaginary axis. Define matrix $\hat{\mathbf{Z}}$, which is analytic on the imaginary axis, by

$$\hat{\mathbf{Z}} = I - \hat{\mathbf{G}}\hat{\mathbf{G}}^\sim .$$

We have, on the imaginary axis $s = j\omega$,

$$0 \leq \begin{bmatrix} I \\ \hat{\mathbf{G}}^\sim \end{bmatrix} \begin{bmatrix} I, \hat{\mathbf{G}} \end{bmatrix} = \begin{bmatrix} I & \hat{\mathbf{G}} \\ \hat{\mathbf{G}}^\sim & I \end{bmatrix} \cong \begin{bmatrix} I - \hat{\mathbf{G}}\hat{\mathbf{G}}^\sim & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{Z}} & 0 \\ 0 & I \end{bmatrix} ,$$

where $\cong$ stands for the congruency relation of hermitian matrices. Therefore, matrix $\hat{\mathbf{Z}}$ is positive semidefinite on the imaginary axis and of constant rank, and there exists its factorization

$$\hat{\mathbf{Z}} = \mathbf{Q}^\sim\mathbf{Q}$$

for some rational matrix $\mathbf{Q}$.

By $\hat{\mathbf{Z}}\hat{\mathbf{G}} = (I - \hat{\mathbf{G}}\hat{\mathbf{G}}^\sim)\hat{\mathbf{G}} = 0$, it follows that

$$\mathbf{Q}\hat{\mathbf{G}} = 0 . \tag{3.22}$$

We have

$$\mathbf{Z} = Q_e - Q_e\mathbf{\Psi}\mathbf{\Psi}^\sim Q_e = Q_e^{\frac{1}{2}} \left(I - [0, I]\hat{\mathbf{G}}\hat{\mathbf{G}}^\sim \begin{bmatrix} 0 \\ I \end{bmatrix} \right) Q_e^{\frac{1}{2}}$$

$$= Q_e^{\frac{1}{2}}[0, I]\hat{\mathbf{Z}} \begin{bmatrix} 0 \\ I \end{bmatrix} Q_e^{\frac{1}{2}} = Q_e^{\frac{1}{2}}[0, I]\mathbf{Q}^\sim\mathbf{Q} \begin{bmatrix} 0 \\ I \end{bmatrix} Q_e^{\frac{1}{2}} = Q_e^{\frac{1}{2}}\mathbf{Q}_2^\sim\mathbf{Q}_2 Q_e^{\frac{1}{2}} , \tag{3.23}$$

where matrix $\mathbf{Q}$ is partitioned on $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2]$,

It follows from (3.22), i.e.,

$$\mathbf{Q}_1 Q_u^{\frac{1}{2}}\mathbf{\Theta} + \mathbf{Q}_2 Q_e^{\frac{1}{2}}\mathbf{\Psi} = 0 ,$$

that

$$\mathbf{Q}_1 Q_u^{\frac{1}{2}} = -\mathbf{M}\mathbf{\Psi}_1 , \quad \mathbf{Q}_2 Q_e^{\frac{1}{2}} = \mathbf{M}\mathbf{\Theta}_1 \tag{3.24}$$

for some rational matrix $\mathbf{M}$, i.e.,

$$\mathbf{Q} \begin{bmatrix} Q_u^{\frac{1}{2}} & 0 \\ 0 & Q_e^{\frac{1}{2}} \end{bmatrix} = \mathbf{M} \begin{bmatrix} -\mathbf{\Psi}_1, & \mathbf{\Theta}_1 \end{bmatrix} .$$

By right multiplication of the above identity with the matrix $\left[\begin{smallmatrix} -\mathbf{V}_1 \\ \mathbf{U}_1 \end{smallmatrix} \right]$, using (2.2), we obtain

$$\mathbf{M} = \mathbf{Q} \begin{bmatrix} Q_u^{\frac{1}{2}} & 0 \\ 0 & Q_e^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} -\mathbf{V}_1 \\ \mathbf{U}_1 \end{bmatrix} .$$

Now define matrix $\hat{\mathbf{M}}$ by

$$\hat{\mathbf{M}} = \mathbf{M}^{\sim}\mathbf{M} = [-\mathbf{V}_1^{\sim}, \mathbf{U}_1^{\sim}] \begin{bmatrix} Q_u^{\frac{1}{2}} & 0 \\ 0 & Q_e^{\frac{1}{2}} \end{bmatrix} \hat{\mathbf{Z}} \begin{bmatrix} Q_u^{\frac{1}{2}} & 0 \\ 0 & Q_e^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} -\mathbf{V}_1 \\ \mathbf{U}_1 \end{bmatrix} .$$

Matrix $\hat{\mathbf{M}}$ does not have purely imaginary poles, and, by (3.23) and (3.24), identity (3.20) holds, as required. □

By (3.17), Lemma 3.2, and the fact that $\mathbf{K}$ is strict proper and without poles on the imaginary axis, criterion (1.4) is finite if and only if $\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is strict proper.

By (3.16), since $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is stable, the unstable poles of $\mathbf{K}$ are contained in the union of the sets of poles of $\boldsymbol{\Psi}^{\sim}$, $\boldsymbol{\Theta}^{\sim}$, and the unstable poles of $\boldsymbol{\Phi}_{vv}\boldsymbol{\Phi}^{-\sim}$, i.e., the complex numbers in τ.

So, if we extract the poles in τ off the rational matrix $\mathbf{K}$ by partial fraction expansion, we shall obtain a stable rational matrix. Therefore, the rational matrix

$$\mathbf{K}_\sigma \stackrel{\text{def}}{=} \{\mathbf{K}\}_\sigma = \{\mathbf{L} - \mathbf{V}\boldsymbol{\Phi}\}_\sigma = \mathbf{L}_\sigma - \mathbf{V}\boldsymbol{\Phi} , \tag{3.25}$$

where we have used Assumption 6, is stable.

Further on, since matrices $\mathbf{S}$, $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$, $\mathbf{K}_\sigma$, and $\mathbf{K}_\tau^{\sim}$ are stable and matrices $\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi}$, $\mathbf{K}_\sigma$, and $\mathbf{K}_\tau^{\sim}$ are strict proper, where $\mathbf{K}_\tau \stackrel{\text{def}}{=} \{\mathbf{K}\}_\tau$, we have

$$\int_{-i\infty}^{i\infty} (\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma)\, \mathbf{K}_\tau^{\sim} \mathrm{d}s = 0 . \tag{3.26}$$

The integrand (3.17) can be expressed as

$$\begin{aligned} &\mathrm{Tr}\,\{Q_e\boldsymbol{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{L}^{\sim}\} + \mathrm{Tr}\,\{(\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma)\,(\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma)^{\sim}\} \\ &\quad - \mathrm{Tr}\,\{(\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma)\,\mathbf{K}_\tau^{\sim}\} - \mathrm{Tr}\,\{\mathbf{K}_\tau\,(\mathbf{S}\boldsymbol{\Theta}_1\Phi - \mathbf{K}_\sigma)^{\sim}\} + \mathrm{Tr}\,\{\mathbf{K}_\tau\mathbf{K}_\tau^{\sim}\} . \end{aligned}$$

Having in mind (3.26), it remains to integrate

$$\begin{aligned} &\mathrm{Tr}\,\{Q_e\boldsymbol{\Phi}_{xx}\} - \mathrm{Tr}\,\{\mathbf{L}\mathbf{L}^{\sim}\} + \mathrm{Tr}\,\{\mathbf{L}_\tau\mathbf{L}_\tau^{\sim}\} \\ &\quad + \mathrm{Tr}\,\{(\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma)\,(\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma)^{\sim}\} , \end{aligned} \tag{3.27}$$

where the identity $\mathbf{K}_\tau = \mathbf{L}_\tau$ has been used.

The integral of the function (3.27) is minimal if we take

$$\mathbf{S}\boldsymbol{\Theta}_1\boldsymbol{\Phi} - \mathbf{K}_\sigma = 0 .$$

The parameter-matrix

$$\mathbf{S} = \mathbf{K}_\sigma\boldsymbol{\Phi}^{-1}\boldsymbol{\Theta}_1^{-1} \tag{3.28}$$

is indeed proper (by Assumption 7) and stable (because rational matrix $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is minimum phase).

Optimal $\mathbf{W}$ is also proper and stable, and, by (3.28) and (3.25),

$$\mathbf{W} = \mathbf{V} + \mathbf{S}\boldsymbol{\Theta}_1 = \mathbf{V} + \mathbf{K}_\sigma\boldsymbol{\Phi}^{-1} = \mathbf{V} + (\mathbf{L}_\sigma - \mathbf{V}\boldsymbol{\Phi})\,\boldsymbol{\Phi}^{-1} = \mathbf{L}_\sigma\boldsymbol{\Phi}^{-1} .$$

The finiteness of the criterion with optimal $\mathbf{C}$ follows by Lemma 3.2 and (3.27). □

Remark 3.1. By (3.27), the minimum of the criterion (1.4) is obtained by integrating $\mathrm{Tr}\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\{\mathbf{L}\mathbf{L}^\sim\} + \mathrm{Tr}\{\mathbf{L}_\tau\mathbf{L}_\tau^\sim\}$. To compute the minumum, we can transform this problem into computing an $\mathcal{H}_2$ norm of possibly unstable transfer matrices. In applying such an algorithm, care must be taken to prevent existence of purely imaginary poles in the integrand-functions. For instance, if shape-deterministic inputs are present, the two functions $\mathrm{Tr}\{Q_e\mathbf{\Phi}_{xx}\}$ and $\mathrm{Tr}\{\mathbf{L}\mathbf{L}^\sim\}$ may have purely imaginary poles (although the function $\mathrm{Tr}\{Q_e\mathbf{\Phi}_{xx}\} - \mathrm{Tr}\{\mathbf{L}\mathbf{L}^\sim\}$ does not).

Remark 3.2. Assumptions 1, 3, and 5 are consistent. In particular, if r, w, and v are shape-deterministic signals satisfying Assumption 3, then $\mathbf{\Phi}_{xx} = \mathbf{\Phi}_{rr} + \mathbf{G}_w\mathbf{\Phi}_{ww}\mathbf{G}_w^\sim = \mathbf{\Theta}_1^{-1}(\mathbf{\Theta}_1\mathbf{\Phi}_{rr}\mathbf{\Theta}_1^\sim + \mathbf{\Psi}_2\mathbf{\Phi}_{ww}\mathbf{\Psi}_2^\sim)\mathbf{\Theta}_1^{-\sim}$ is strict proper. Conversely, if $\mathbf{\Phi}_{xx}$ is strict proper, then $\mathbf{\Theta}_1\mathbf{r}$ and $\mathbf{\Psi}_2\mathbf{w}$ are strict proper. Further on, under Assumption 3, from $\mathbf{\Theta}_1\mathbf{\Phi}\mathbf{\Phi}^\sim\mathbf{\Theta}_1^\sim = \mathbf{\Theta}_1\mathbf{\Phi}_{vv}\mathbf{\Theta}_1^\sim + \mathbf{\Theta}_1\mathbf{\Phi}_{rr}\mathbf{\Theta}_1^\sim + \mathbf{\Psi}_2\mathbf{\Phi}_{ww}\mathbf{\Psi}_2^\sim$, it follows that $\mathbf{\Theta}_1\mathbf{\Phi}$ does not have purely imaginary poles.

Remark 3.3. If we relax the definition of internal stability of the system of Figure 1 so that we do not impose properness of the four transfer matrices $\mathbf{H}$, $\mathbf{HG}$, $\mathbf{CH}$, and $\mathbf{CHG}$, and if we relax Assumption 3 so that we do not impose strict properness of $\mathbf{\Theta}_1\mathbf{r}$, $\mathbf{\Psi}_2\mathbf{w}$, and $\mathbf{v}$ when we deal with shape-deterministic inputs, we can generalize Theorem 2.4 so that we avoid Assumption 7. In this new optimal control problem, parameter-matrix $\mathbf{S}$ in (3.28) need not be proper.

To prove the generalization, it suffices to prove that matrix $\mathbf{L}$ is strict proper without Assumption 7. By (2.9), it suffices to prove that $\mathbf{\Phi}_{xx}\mathbf{\Phi}^{-\sim}$ is strict proper. Without loss of generality, assume that $\mathbf{\Phi}_{vv} = \mathbf{U}_v\mathbf{U}_v^\sim$ and $\mathbf{\Phi}_{xx} = \mathbf{U}_x\mathbf{U}_x^\sim$, for some possibly nonsquare proper matrices $\mathbf{U}_v$ and $\mathbf{U}_x$. Then

$$\mathbf{\Phi}_{vv} + \mathbf{\Phi}_{xx} = \mathbf{U}_v\mathbf{U}_v^\sim + \mathbf{U}_x\mathbf{U}_x^\sim = [\mathbf{U}_v, \mathbf{U}_x]\begin{bmatrix}\mathbf{U}_v^\sim \\ \mathbf{U}_x^\sim\end{bmatrix} = \mathbf{\Phi}\mathbf{\Phi}^\sim \ .$$

It follows that $\mathbf{\Phi}^{-1}[\mathbf{U}_v, \mathbf{U}_x]$ is proper and in particular, that $\mathbf{\Phi}^{-1}\mathbf{U}_x$ is proper. By the fact that $\mathbf{\Phi}_{xx} = \mathbf{U}_x\mathbf{U}_x^\sim$ is strict proper, $\mathbf{U}_x^\sim$ is strict proper. As a consequence, $\mathbf{\Phi}^{-1}\mathbf{\Phi}_{xx} = (\mathbf{\Phi}^{-1}\mathbf{U}_x)\mathbf{U}_x^\sim$ is strict proper.

Remark 3.4. Another generalization of Theorem 2.4 can be made by considering set τ being of as few elements as possible, but including all unstable poles of $\mathbf{K}$. Such a decrease in the number of elements of τ is required because of Assumption 6, and it does not affect $\mathbf{L}_\sigma$. If we consider the minimal τ, i.e., consisting of all unstable poles of $\mathbf{K}$, matrices $\mathbf{U}_1$ and $\mathbf{V}_1$ will appear in the controller transfer matrix implicitly. To compromise we assume that τ is a union of the sets of the poles of $\hat{\mathbf{G}}^\sim$ (3.21) and the unstable poles of $\mathbf{\Psi}^\sim Q_e\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}$.

Remark 3.5. Concerning the satisfaction of Assumption 5, we could take a systematic approach to finding a stable and minimum phase factor $\mathbf{\Theta}_1\mathbf{\Phi}$ in the factorization $\mathbf{\Theta}_1\mathbf{\Phi}\mathbf{\Phi}^\sim\mathbf{\Theta}_1^\sim = \mathbf{\Theta}_1(\mathbf{\Phi}_{vv} + \mathbf{\Phi}_{rr})\mathbf{\Theta}_1^\sim + \mathbf{\Psi}_2\mathbf{\Phi}_{ww}\mathbf{\Psi}_2^\sim$, but we regard such an approach as more complex, at least when the plant is given by state-space equations, because matrices $\mathbf{\Theta}_1$ and $\mathbf{\Psi}_2$ will appear in the controller transfer matrix.

Proof of Corollary 2.5. We have

$$\begin{aligned}\mathbf{L}_\sigma &= \left\{\mathbf{\Psi}^\sim Q_e\mathbf{\Phi}_{xx}\mathbf{\Phi}^{-\sim}\right\}_\sigma = \left\{\mathbf{\Psi}^\sim Q_e\left(\mathbf{\Phi}\mathbf{\Phi}^\sim - \mathbf{\Phi}_{vv}\right)\mathbf{\Phi}^{-\sim}\right\}_\sigma \\ &= \left\{\mathbf{\Psi}^\sim Q_e\mathbf{\Phi}\right\}_\sigma - \left\{\mathbf{\Psi}^\sim Q_e\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}\right\}_\sigma = \left\{\mathbf{\Psi}^\sim Q_e\mathbf{\Phi}\right\}_\sigma \ ,\end{aligned}$$

because rational matrix $\mathbf{\Psi}^\sim Q_e\mathbf{\Phi}_{vv}\mathbf{\Phi}^{-\sim}$ has its poles in τ. □

Example 3.1. This example illustrates the possibilities we have of satisfying Assumption 5. Let $\mathbf{G}$ be given by (3.1), $\mathbf{\Phi}_{ww} = 0$, $\mathbf{\Phi}_{rr} = 0$, and $\mathbf{\Phi}_{vv} = M =$

Const > 0. We have at least two possibilities: to take the constant matrix $\mathbf{\Phi} = N$, $N^2 = M$, or to take the matrix $\mathbf{\Phi} = (I + C\mathbf{\Omega}XC^{\mathrm{T}}M^{-1})N$, where $\mathbf{\Omega} = (sI - A)^{-1}$ and X is the stabilizing solution of CARE:

$$AX + XA^{\mathrm{T}} - XC^{\mathrm{T}}M^{-1}CX = 0 \ ,$$

i.e., a solution such that matrix $\mathbf{\Omega}_X = (sI - A + XC^{\mathrm{T}}M^{-1}C)^{-1}$ is stable. To achieve the property of minimum phase and stability of $\mathbf{\Theta}_1\mathbf{\Phi}$, we take the second solution. Indeed, by (3.2), we have $\mathbf{\Theta}_1 = I - C\mathbf{\Omega}_F F$, $\mathbf{\Omega}_F = (sI - A + FC)^{-1}$ and

$$\mathbf{\Theta}_1\mathbf{\Phi} = \left[I - C\mathbf{\Omega}_F(F - XC^{\mathrm{T}}M^{-1})\right] N \ ,$$

$$(\mathbf{\Theta}_1\mathbf{\Phi})^{-1} = N^{-1}\left[I + C\mathbf{\Omega}_X(F - XC^{\mathrm{T}}M^{-1})\right] \ .$$

Therefore, $\mathbf{\Theta}_1\mathbf{\Phi}$ is minimum phase and stable. Optimal controller $\mathbf{C}$ can be easily found (see section 4).

Example 3.2. This example illustrates that Assumption 6 cannot be avoided in general. Let $\mathbf{G} = (1+s)/(1-s)$, $Q_e = Q_u = 1$, $\mathbf{\Phi}_{vv} = 0$, $\mathbf{\Phi}_{xx} = 1/(1-s^2)$. We have that

$$\mathbf{\Psi} = \frac{1}{\sqrt{2}} \ , \quad \mathbf{\Delta} = \mathbf{\Theta}^{-1} = \sqrt{2}\frac{1+s}{1-s} \ , \quad \mathbf{\Theta}_1 = \mathbf{\Theta} \ , \quad \mathbf{\Phi} = \frac{1}{1-s} \ ;$$

hence $\tau = \{1\}$. Since $\mathbf{\Theta}_1\mathbf{\Phi} = \frac{1}{\sqrt{2}(1+s)}$, Assumption 5 is satisfied. By $\mathbf{L} = \frac{1}{\sqrt{2}(1-s)}$, it follows that $\mathbf{L}_\sigma = 0$, so application of (2.10) and (2.11) gives $\mathbf{C} = 0$. This controller renders the system unstable; hence it is not optimal. However, Assumption 6 is not satisfied because the pole at 1 of $\mathbf{\Phi}$ belongs to τ.

4. Applications. The first purpose of this section is to illustrate the application of the simplified formulas (2.10) and (2.11), in particular, to show how the new operator $\{\cdot\}_\sigma$ can be computed, and to show that Assumptions 1–7 are not unrealistic, especially the most important assumption that rational matrix $\mathbf{\Theta}_1\mathbf{\Phi}$ is stable and minimum phase. The other purpose of this section is to solve three important optimal control problems explicitly.

We assume the plant transfer matrix $\mathbf{G}$ given by (3.1), and the plant transfer matrix

$$\mathbf{G}_w = C(sI - A)^{-1}B_w = C\mathbf{\Omega}B_w \ .$$

These $\mathbf{G}$ and $\mathbf{G}_w$ are in the required forms (2.1) and (2.8), correspondingly, with $\mathbf{\Psi}_2 = C(sI - A + FC)^{-1}B_w$.

We shall use the following form of optimal controller (see (2.10) and (2.11)):

$$\mathbf{C} = \left(\mathbf{\Delta} - \mathbf{L}_\sigma\mathbf{\Phi}^{-1}\mathbf{G}\right)^{-1}\mathbf{L}_\sigma\mathbf{\Phi}^{-1} \ , \tag{4.1}$$

where matrix $\mathbf{\Delta} = \mathbf{\Theta}^{-1}$ is defined in (3.8).

4.1. Output zeroing: Colored measurement noise is present. In this subsection we assume the output zeroing case, i.e., $r = 0$, and white disturbance noise with $\mathbf{\Phi}_{ww} = W \geq 0$. We have

$$\mathbf{\Phi}_{xx} = C\mathbf{\Omega}S\mathbf{\Omega}^{\sim}C^{\mathrm{T}} \ , \quad S = B_w W B_w^{\mathrm{T}} \ .$$

Also we assume colored measurement noise with the following covariance matrix

$$\mathbf{\Phi}_{vv} = \mathbf{\Omega}_v V \mathbf{\Omega}_v^{\sim} \ , \quad \mathbf{\Omega}_v = (sI - A_v)^{-1} \ ,$$

for some stable matrix A_v and positive semidefinite V.

To apply Theorem 2.4, we have to solve the following spectral factorization:

$$\mathbf{\Phi}_{vv} + \mathbf{\Phi}_{xx} = C\mathbf{\Omega} S \mathbf{\Omega}^{\sim} C^{\mathrm{T}} + \mathbf{\Omega}_v V \mathbf{\Omega}_v^{\sim} = \mathbf{\Phi}\mathbf{\Phi}^{\sim} \ ,$$

where $\mathbf{\Phi}$ is minimum phase.

Since the rational matrix $\mathbf{\Omega}_v$ is minimum phase already, we can search $\mathbf{\Phi}$ in the form

$$\mathbf{\Phi} = \mathbf{\Omega}_v \mathbf{\Phi}_1 \ ,$$

where $\mathbf{\Phi}_1$ is a minimum phase spectral factor in the factorization

$$\mathbf{\Omega}_v^{-1} C \mathbf{\Omega} S \mathbf{\Omega}^{\sim} C^{\mathrm{T}} \mathbf{\Omega}_v^{-\sim} + V = \mathbf{\Phi}_1 \mathbf{\Phi}_1^{\sim} \ . \tag{4.2}$$

Having in mind the identity

$$\mathbf{\Omega}_v^{-1} C \mathbf{\Omega} = C + Q\mathbf{\Omega} \ , \quad Q = CA - A_v C \ , \tag{4.3}$$

the problem (4.2) transforms into

$$M + CS\mathbf{\Omega}^{\sim} Q^{\mathrm{T}} + Q\mathbf{\Omega} S C^{\mathrm{T}} + Q\mathbf{\Omega} S \mathbf{\Omega}^{\sim} Q^{\mathrm{T}} = \mathbf{\Phi}_1 \mathbf{\Phi}_1^{\sim} \ , \tag{4.4}$$

where

$$M = V + CSC^{\mathrm{T}} \ .$$

Using Assumption 7, we deduce that matrix M is positive definite. Introduce the matrix N by $N^2 = M$, and consider the following CARE:

$$XA^{\mathrm{T}} + AX + S - (SC^{\mathrm{T}} + XQ^{\mathrm{T}}) M^{-1} (CS + QX) = 0 \ .$$

To prove the existence of a stabilizing solution X of CARE, it suffices to prove that the matrix pair $(A^{\mathrm{T}}, Q^{\mathrm{T}})$ is stabilizable (see Lemma 3.1). We have, identically

$$\begin{bmatrix} sI - A_v & -C \\ 0 & I \end{bmatrix} \begin{bmatrix} C \\ sI - A \end{bmatrix} = \begin{bmatrix} Q \\ sI - A \end{bmatrix} \ .$$

Since the eigenvalues of matrix A_v are stable and matrix pair (C, A) is detectable, the pencil $\left[\begin{smallmatrix} Q \\ sI - A \end{smallmatrix}\right]$ has full column rank for every s, $\Re(s) \geq 0$; i.e., matrix pair $(A^{\mathrm{T}}, Q^{\mathrm{T}})$ is stabilizable.

The spectral factor $\mathbf{\Phi}_1$ in (4.4) is

$$\mathbf{\Phi}_1 = N + Q\mathbf{\Omega} \left(SC^{\mathrm{T}} + XQ^{\mathrm{T}} \right) M^{-1} N \ .$$

To check Assumption 5, write at first,

$$\mathbf{\Phi} = \mathbf{\Omega}_v \mathbf{\Phi}_1 = \mathbf{\Omega}_v \left[N + Q\mathbf{\Omega} (SC^{\mathrm{T}} + XQ^{\mathrm{T}}) M^{-1} N \right] = \mathbf{\Omega}_v N_1 + C\mathbf{\Omega} N_2 \ ,$$

where $N_2 = (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}N$, $N_1 = N - CN_2$. We have, by (3.2),

$$\begin{aligned}\boldsymbol{\Theta}_1\boldsymbol{\Phi} &= (I - C\boldsymbol{\Omega}_F F)(\boldsymbol{\Omega}_v N_1 + C\boldsymbol{\Omega} N_2)\\ &= \boldsymbol{\Omega}_v N_1 + C\boldsymbol{\Omega} N_2 - C\boldsymbol{\Omega}_F F\boldsymbol{\Omega}_v N_1 - C\boldsymbol{\Omega}_F FC\boldsymbol{\Omega} N_2\\ &= \boldsymbol{\Omega}_v N_1 + C\boldsymbol{\Omega} N_2 - C\boldsymbol{\Omega}_F F\boldsymbol{\Omega}_v N_1 - C(\boldsymbol{\Omega} - \boldsymbol{\Omega}_F)N_2\\ &= \boldsymbol{\Omega}_v N_1 - C\boldsymbol{\Omega}_F F\boldsymbol{\Omega}_v N_1 + C\boldsymbol{\Omega}_F N_2\ ,\end{aligned}$$

so, $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is stable, because $\boldsymbol{\Omega}_F = (sI - A + FC)^{-1}$ is stable. To check that it is minimum phase, introduce the rational matrix $\boldsymbol{\Omega}_X = \left[sI - A + (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}Q\right]^{-1}$. The matrix identity

$$\boldsymbol{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}Q\boldsymbol{\Omega} = \boldsymbol{\Omega} - \boldsymbol{\Omega}_X \tag{4.5}$$

is easy to check. We have

$$\begin{aligned}(\boldsymbol{\Theta}_1\boldsymbol{\Phi})^{-1} = \boldsymbol{\Phi}^{-1}\boldsymbol{\Theta}_1^{-1} &= N^{-1}\left[I - Q\boldsymbol{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right]\boldsymbol{\Omega}_v^{-1}(I + C\boldsymbol{\Omega}F)\\ &= N^{-1}\left[I - Q\boldsymbol{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right](\boldsymbol{\Omega}_v^{-1} + CF + Q\boldsymbol{\Omega}F)\\ &= N^{-1}\left[I - Q\boldsymbol{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right](\boldsymbol{\Omega}_v^{-1} + CF) + N^{-1}Q\boldsymbol{\Omega}F - N^{-1}Q(\boldsymbol{\Omega} - \boldsymbol{\Omega}_X)F\\ &= N^{-1}\left[I - Q\boldsymbol{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right](\boldsymbol{\Omega}_v^{-1} + CF) + N^{-1}Q\boldsymbol{\Omega}_X F\ ,\end{aligned}$$

so, $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is minimum phase.

From the identity $\boldsymbol{\Phi}_{vv}\boldsymbol{\Phi}^{-\sim} = \boldsymbol{\Omega}_v V\boldsymbol{\Omega}_v^{\sim}\boldsymbol{\Omega}_v^{-\sim}\boldsymbol{\Phi}_1^{-\sim} = \boldsymbol{\Omega}_v V\boldsymbol{\Phi}_1^{-\sim}$, it follows that the set τ consists of the poles of $\boldsymbol{\Omega}_X^{\sim}$ and $\boldsymbol{\Omega}_Y^{\sim}$, where $\boldsymbol{\Omega}_Y$ is given in (3.9).

To find optimal controller $\mathbf{C}$, at first compute, using (4.5),

$$\begin{aligned}&(C^{\mathrm{T}} + \boldsymbol{\Omega}^{\sim}Q^{\mathrm{T}})\left[I - M^{-1}(CS + QX)\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}}\right]\\ &= C^{\mathrm{T}} - C^{\mathrm{T}}M^{-1}(CS + QX)\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}} + \boldsymbol{\Omega}^{\sim}Q^{\mathrm{T}} - \boldsymbol{\Omega}^{\sim}Q^{\mathrm{T}}M^{-1}(CS + QX)\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}}\\ &= C^{\mathrm{T}} - C^{\mathrm{T}}M^{-1}(CS + QX)\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}} + \boldsymbol{\Omega}^{\sim}Q^{\mathrm{T}} - (\boldsymbol{\Omega}^{\sim} - \boldsymbol{\Omega}_X^{\sim})Q^{\mathrm{T}}\\ &= C^{\mathrm{T}} + \left[I - C^{\mathrm{T}}M^{-1}(CS + QX)\right]\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}}\ .\end{aligned} \tag{4.6}$$

Further on, using (4.6) and the matrix identity

$$\boldsymbol{\Omega}S\left[I - C^{\mathrm{T}}M^{-1}(CS + QX)\right]\boldsymbol{\Omega}_X^{\sim} = \boldsymbol{\Omega}X + X\boldsymbol{\Omega}_X^{\sim}\ ,$$

compute

$$\begin{aligned}&\boldsymbol{\Omega}S(C^{\mathrm{T}} + \boldsymbol{\Omega}^{\sim}Q^{\mathrm{T}})\left[I - M^{-1}(CS + QX)\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}}\right]\\ &= \boldsymbol{\Omega}SC^{\mathrm{T}} + \boldsymbol{\Omega}S\left[I - C^{\mathrm{T}}M^{-1}(CS + QX)\right]\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}}\\ &= \boldsymbol{\Omega}SC^{\mathrm{T}} + (\boldsymbol{\Omega}X + X\boldsymbol{\Omega}_X^{\sim})Q^{\mathrm{T}}\\ &= \boldsymbol{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}}) + X\boldsymbol{\Omega}_X^{\sim}Q^{\mathrm{T}}\ .\end{aligned} \tag{4.7}$$

Using (3.10), (4.3), (4.7), and the matrix identity

$$\mathbf{\Omega}_Y^{\sim}(C^{\mathrm{T}} - K^{\mathrm{T}}D^{\mathrm{T}})Q_eC\mathbf{\Omega} = \mathbf{\Omega}_Y^{\sim}Y + Y\mathbf{\Omega}\ , \tag{4.8}$$

where Y, K, and $\mathbf{\Omega}_Y$ are given by (3.4), (3.6), and (3.9), we compute

$$\begin{aligned}
\mathbf{L} &= \mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}_{xx}\mathbf{\Phi}^{-\sim} \\
&= \left[DN_u^{-1} + (C - DK)\mathbf{\Omega}_Y BN_u^{-1}\right]^{\sim} Q_eC\mathbf{\Omega}S\mathbf{\Omega}^{\sim}C^{\mathrm{T}}\mathbf{\Omega}_v^{-\sim} \\
&\quad \cdot \left[I + Q\mathbf{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right]^{-\sim} N^{-1} \\
&= N_u^{-1}\left[D^{\mathrm{T}} + B^{\mathrm{T}}\mathbf{\Omega}_Y^{\sim}(C^{\mathrm{T}} - K^{\mathrm{T}}D^{\mathrm{T}})\right] Q_eC\mathbf{\Omega}S(C^{\mathrm{T}} + \mathbf{\Omega}^{\sim}Q^{\mathrm{T}}) \\
&\quad \cdot \left[I - M^{-1}(CS + QX)\mathbf{\Omega}_X^{\sim}Q^{\mathrm{T}}\right] N^{-1} \\
&= N_u^{-1}\left[D^{\mathrm{T}} + B^{\mathrm{T}}\mathbf{\Omega}_Y^{\sim}(C^{\mathrm{T}} - K^{\mathrm{T}}D^{\mathrm{T}})\right] Q_eC\left[\mathbf{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}}) + X\mathbf{\Omega}_X^{\sim}Q^{\mathrm{T}}\right] N^{-1} \\
&= N_u^{-1}\left[D^{\mathrm{T}}Q_eC\mathbf{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}}) + D^{\mathrm{T}}Q_eCX\mathbf{\Omega}_X^{\sim}Q^{\mathrm{T}}\right. \\
&\quad \left. + B^{\mathrm{T}}(\mathbf{\Omega}_Y^{\sim}Y + Y\mathbf{\Omega})(SC^{\mathrm{T}} + XQ^{\mathrm{T}}) + B^{\mathrm{T}}\mathbf{\Omega}_Y^{\sim}(C^{\mathrm{T}} - K^{\mathrm{T}}D^{\mathrm{T}})Q_eCX\mathbf{\Omega}_X^{\sim}Q^{\mathrm{T}}\right] N^{-1}\ .
\end{aligned}$$

Since the poles of $\mathbf{\Omega}_X^{\sim}$ and $\mathbf{\Omega}_Y^{\sim}$ belong to τ, we have

$$\mathbf{L}_\sigma = \left\{\mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}_{xx}\mathbf{\Phi}^{-\sim}\right\}_\sigma = N_uK\mathbf{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})N^{-1}\ .$$

To compute $\mathbf{W}$, we use the identity (4.5):

$$\begin{aligned}
\mathbf{W} &= \mathbf{L}_\sigma\mathbf{\Phi}^{-1} \\
&= N_uK\mathbf{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\left[I - Q\mathbf{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right]\mathbf{\Omega}_v^{-1} \\
&= N_uK\left[\mathbf{\Omega}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1} - (\mathbf{\Omega} - \mathbf{\Omega}_X)(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\right]\mathbf{\Omega}_v^{-1} \\
&= N_uK\mathbf{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\mathbf{\Omega}_v^{-1}\ .
\end{aligned}$$

Further on, using identities (3.8), (4.3), and (4.5), we obtain

$$\begin{aligned}
N_u^{-1}(\mathbf{\Delta} - \mathbf{WG}) &= I + K\mathbf{\Omega}B - K\mathbf{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\mathbf{\Omega}_v^{-1}(D + C\mathbf{\Omega}B) \\
&= I + K\mathbf{\Omega}B - K\mathbf{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}(\mathbf{\Omega}_v^{-1}D + CB + Q\mathbf{\Omega}B) \\
&= I + K\mathbf{\Omega}B - K\mathbf{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}(\mathbf{\Omega}_v^{-1}D + CB) - K(\mathbf{\Omega} - \mathbf{\Omega}_X)B \\
&= I + K\mathbf{\Omega}_X\left[B - (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}(\mathbf{\Omega}_v^{-1}D + CB)\right] \stackrel{\mathrm{def}}{=} I + K\mathbf{\Omega}_X\mathbf{H}_v\ .
\end{aligned}$$

Introduce the matrices

$$\begin{aligned}
E_1 &= I - (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}DK\ , \\
A_1 &= A - BK - (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}(Q + A_vDK - CBK)\ .
\end{aligned}$$

To achieve proper controller $\mathbf{C}$, assume that matrix E_1 is nonsingular; then we have, for optimal controller $\mathbf{C}$,

$$\begin{aligned}\mathbf{C} &= (\mathbf{\Delta} - \mathbf{WG})^{-1}\,\mathbf{W} \\ &= (I + K\mathbf{\Omega}_X\mathbf{H}_v)^{-1}\,K\mathbf{\Omega}_X(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\mathbf{\Omega}_v^{-1} \\ &= K\left[sI - A + (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}Q + \mathbf{H}_vK\right]^{-1}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\mathbf{\Omega}_v^{-1} \\ &= K\left(sI - E_1^{-1}A_1\right)^{-1}E_1^{-1}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\mathbf{\Omega}_v^{-1} \\ &= D_{\mathbf{c}} + K\left(sI - E_1^{-1}A_1\right)^{-1}B_{\mathbf{c}}\ ,\end{aligned}$$

where

$$D_{\mathbf{c}} = KE_1^{-1}(SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}\ ,\quad B_{\mathbf{c}} = E_1^{-1}\left[A_1D_{\mathbf{c}} - (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}A_v\right]\ .$$

The order of the controller is n.

Concerning Assumption 6, we shall give its interpretation for this particular application. The poles of $\mathbf{\Phi}$ are a union of the eigenvalues of A and eigenvalues of A_v, and the set τ is a union of the eigenvalues of $-(A - BK)$ and eigenvalues of $-\left[A - (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}Q\right]$. Therefore, Assumption 6 means that the unstable eigenvalues of A are not eigenvalues of $-(A - BK)$ and eigenvalues of $-\left[A - (SC^{\mathrm{T}} + XQ^{\mathrm{T}})M^{-1}Q\right]$.

4.2. Output zeroing: Measurement noise is absent. To apply Corollary 2.5 on case $r = 0$, $v = 0$, and white disturbance noise with $\mathbf{\Phi}_{ww} = W \geq 0$, the following spectral factorization requires

$$\mathbf{G}_wW\mathbf{G}_w^{\sim} = \mathbf{\Phi}\mathbf{\Phi}^{\sim}\ , \tag{4.9}$$

where $\mathbf{\Phi}$ is a minimum phase spectral factor. Since $\mathbf{G}_w$ is strict proper, the spectral factorization (4.9) is not obvious. Using Assumption 7, we deduce that matrix $CB_wWB_w^{\mathrm{T}}C^{\mathrm{T}}$ is positive definite. Without loss of generality, we assume matrix A in block-triangular form

$$A = \begin{bmatrix} A_1 & A_3 \\ 0 & A_2 \end{bmatrix}\ ,$$

where A_1 is a square matrix of dimension p, p being the number of outputs, and A_2 is a square matrix of dimension $n - p$. In addition, we assume the matrix C_1 is nonsingular, where

$$C = [C_1, C_2]\ ,\quad B_w = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}\ .$$

For reduction of the matrix A in the above form by orthogonal similarity transformation matrices, the QR-algorithm can be applied. We do not have additional requirements on the eigenvalues of the matrices A_1 and A_2. We have

$$\mathbf{G}_w = [C_1, C_2]\begin{bmatrix} \mathbf{\Omega}_1 & \mathbf{\Omega}_1A_3\mathbf{\Omega}_2 \\ 0 & \mathbf{\Omega}_2 \end{bmatrix}\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = C_1\mathbf{\Omega}_1B_1 + C_1\mathbf{\Omega}_1A_3\mathbf{\Omega}_2B_2 + C_2\mathbf{\Omega}_2B_2\ ,$$

where $\mathbf{\Omega}_1 = (sI - A_1)^{-1}$, $\mathbf{\Omega}_2 = (sI - A_2)^{-1}$. Introduce the matrices P and Q by $P = C_1^{-1}C_2$ and $Q = PA_2 - A_1P$. Then $\mathbf{\Omega}_1^{-1}C_1^{-1}C_2\mathbf{\Omega}_2 = P + Q\mathbf{\Omega}_2$ and

$$\mathbf{G}_w = C_1\mathbf{\Omega}_1 \left[B_1 + PB_2 + (A_3 + Q)\mathbf{\Omega}_2 B_2\right] \ .$$

The rational matrix $C_1\mathbf{\Omega}_1$ is minimum phase already; therefore it remains to factorize the rational matrix

$$\left[B_1 + PB_2 + (A_3 + Q)\mathbf{\Omega}_2 B_2\right] W \left[B_1 + PB_2 + (A_3 + Q)\mathbf{\Omega}_2 B_2\right]^{\sim} = \mathbf{\Phi}_1\mathbf{\Phi}_1^{\sim} \ ,$$

where $\mathbf{\Phi}_1$ is a minimum phase spectral factor. Since the matrix

$$M = (B_1 + PB_2)\, W\, (B_1 + PB_2)^{\mathrm{T}} = C_1^{-1}CB_wWB_w^{\mathrm{T}}C^{\mathrm{T}}C_1^{-\mathrm{T}}$$

is positive definite, we can introduce matrix N by $N^2 = M$, and the following CARE:

$$\left[B_2WB_1^{\mathrm{T}} + B_2WB_2^{\mathrm{T}}P^{\mathrm{T}} + X(A_3 + Q)^{\mathrm{T}}\right] M^{-1} \left[B_1WB_2^{\mathrm{T}} + PB_2WB_2^{\mathrm{T}} + (A_3 + Q)X\right]$$

$$- XA_2^{\mathrm{T}} - A_2X - B_2WB_2^{\mathrm{T}} = 0 \ .$$

We shall prove that it has a stabilizing solution X. For that purpose, having in mind Lemma 3.1, it suffices to prove that matrix pair $(A_2^{\mathrm{T}}, (A_3 + Q)^{\mathrm{T}})$ is stabilizable. The matrix pencil

$$\begin{bmatrix} C_1 & C_2 \\ sI - A_1 & -A_3 \\ 0 & sI - A_2 \end{bmatrix} \sim \begin{bmatrix} C_1 & 0 \\ sI - A_1 & -[A_3 + (sI - A_1)P] \\ 0 & sI - A_2 \end{bmatrix}$$

has full column rank, for every complex s, $\Re(s) \geq 0$, because the pair (C, A) is detectable, where $\sim$ stands for the strict equivalence relation of matrix pencils. Therefore, the matrix pencil

$$\begin{bmatrix} A_3 + (sI - A_1)P \\ sI - A_2 \end{bmatrix} \sim \begin{bmatrix} A_3 + PA_2 - A_1P \\ sI - A_2 \end{bmatrix} = \begin{bmatrix} A_3 + Q \\ sI - A_2 \end{bmatrix}$$

has full column rank for every s, $\Re(s) \geq 0$. Therefore matrix pair $(A_2^{\mathrm{T}}, (A_3 + Q)^{\mathrm{T}})$ is stabilizable.

The spectral factorization (4.9) is solved by

$$\mathbf{\Phi} = C_1\mathbf{\Omega}_1\mathbf{\Phi}_1 \ , \tag{4.10}$$

where

$$\mathbf{\Phi}_1 = N + (A_3 + Q)\mathbf{\Omega}_2 ZN \ , \tag{4.11}$$

$$Z = \left[B_2WB_1^{\mathrm{T}} + B_2WB_2^{\mathrm{T}}P^{\mathrm{T}} + X(A_3 + Q)^{\mathrm{T}}\right] M^{-1} \ .$$

It is easy to check that another form of $\mathbf{\Phi}$ is

$$\mathbf{\Phi} = C\mathbf{\Omega}U \ , \quad U = \begin{bmatrix} I - PZ \\ Z \end{bmatrix} N \ . \tag{4.12}$$

Having solved the factorization (4.9), we have to check Assumption 5. Concerning stability of $\mathbf{\Theta}_1\mathbf{\Phi}$, we have, by (3.2) and (4.12),

$$\mathbf{\Theta}_1\mathbf{\Phi} = (I - C\mathbf{\Omega}_F F)C\mathbf{\Omega}U = C\mathbf{\Omega}U - C\mathbf{\Omega}_F FC\mathbf{\Omega}U$$

$$= C\mathbf{\Omega}U - C(\mathbf{\Omega}_F - \mathbf{\Omega})U = C\mathbf{\Omega}_F U \ ,$$

where $\mathbf{\Omega}_F = (sI - A + FC)^{-1}$. Hence $\mathbf{\Theta}_1\mathbf{\Phi}$ is stable. Further on, $\mathbf{\Theta}_1\mathbf{\Phi}$ is minimum phase, because it has the same zeros as $C\mathbf{\Omega}U$.

Now, proceed with computation of $\mathbf{L}_\sigma$ in (2.12). We have, by (3.10), (4.12), (4.8), and (3.6),

$$\mathbf{L}_\sigma = \{\mathbf{\Psi}^{\sim} Q_e \mathbf{\Phi}\}_\sigma = \left\{N_u^{-1}\left[D^{\mathrm{T}} + B^{\mathrm{T}}\mathbf{\Omega}_Y^{\sim}(C^{\mathrm{T}} - K^{\mathrm{T}}D^{\mathrm{T}})\right] Q_e C\mathbf{\Omega}U\right\}_\sigma$$

$$= \left\{N_u^{-1}\left[D^{\mathrm{T}}Q_e C\mathbf{\Omega} + B^{\mathrm{T}}(\mathbf{\Omega}_Y^{\sim}Y + Y\mathbf{\Omega})\right] U\right\}_\sigma = N_u^{-1}(B^{\mathrm{T}}Y + D^{\mathrm{T}}Q_e C)\mathbf{\Omega}U$$

$$= N_u K\mathbf{\Omega}U \ ,$$

because the set τ consists of the poles of $\mathbf{\Omega}_Y^{\sim}$. Introduce matrix $\mathbf{\Omega}_X = [sI - A_2 + Z(A_3 + Q)]^{-1}$. We shall need the following three matrix identities:

$$\mathbf{\Omega}UN^{-1}(A_3 + Q)\mathbf{\Omega}_X = \mathbf{\Omega}C^{\perp} - C^{\perp}\mathbf{\Omega}_X \ ,$$

where

$$C^{\perp} = \begin{bmatrix} -P \\ I \end{bmatrix} \ ,$$

$$UN^{-1} - C^{\perp}Z = \begin{bmatrix} I \\ 0 \end{bmatrix} \ ,$$

and

$$\mathbf{\Omega}_X Z\mathbf{\Omega}_1^{-1} = Z + \mathbf{\Omega}_X R \ , \quad R = -ZA_1 + A_2 Z - Z(A_3 + Q)Z \ ,$$

which can be easily checked. We have

$$\mathbf{W} = \mathbf{L}_\sigma\mathbf{\Phi}^{-1} = N_u K\mathbf{\Omega}UN^{-1}\left[I + (A_3 + Q)\mathbf{\Omega}_2 Z\right]^{-1}\mathbf{\Omega}_1^{-1}C_1^{-1}$$

$$= N_u K\mathbf{\Omega}UN^{-1}\left[I - (A_3 + Q)\mathbf{\Omega}_X Z\right]\mathbf{\Omega}_1^{-1}C_1^{-1}$$

$$= N_u K\left[\mathbf{\Omega}UN^{-1} - \mathbf{\Omega}UN^{-1}(A_3 + Q)\mathbf{\Omega}_X Z\right]\mathbf{\Omega}_1^{-1}C_1^{-1}$$

$$= N_u K\left(\mathbf{\Omega}UN^{-1} - \mathbf{\Omega}C^{\perp}Z + C^{\perp}\mathbf{\Omega}_X Z\right)\mathbf{\Omega}_1^{-1}C_1^{-1}$$

$$= N_u K\left(\mathbf{\Omega}\begin{bmatrix} I \\ 0 \end{bmatrix} + C^{\perp}\mathbf{\Omega}_X Z\right)\mathbf{\Omega}_1^{-1}C_1^{-1}$$

$$= N_u K\left(\begin{bmatrix} I \\ 0 \end{bmatrix} + C^{\perp}\mathbf{\Omega}_X Z\mathbf{\Omega}_1^{-1}\right)C_1^{-1}$$

$$= N_u K\left(\begin{bmatrix} I \\ 0 \end{bmatrix} + C^{\perp}(Z + \mathbf{\Omega}_X R)\right)C_1^{-1}$$

$$= N_u K\left(U + C^{\perp}\mathbf{\Omega}_X RN\right)N^{-1}C_1^{-1} \ .$$

To find controller transfer matrix $\mathbf{C}$, we shall need the following two identities:

$$\mathbf{\Omega}_X R C_1^{-1} C \mathbf{\Omega} = \mathbf{\Omega}_X L - L\mathbf{\Omega} \ , \quad L = [Z,\ ZP - I] \ ,$$

and

$$I - UN^{-1}C_1^{-1}C + C^{\perp}L = 0 \ .$$

We have

$$\begin{aligned} & N_u^{-1}\left(\mathbf{\Delta} - \mathbf{WG}\right) \\ &= I + K\mathbf{\Omega}B - K\left(U + C^{\perp}\mathbf{\Omega}_X RN\right)N^{-1}C_1^{-1}(D + C\mathbf{\Omega}B) \\ &= I - KUN^{-1}C_1^{-1}D + K(I - UN^{-1}C_1^{-1}C)\mathbf{\Omega}B \\ &\quad - KC^{\perp}\mathbf{\Omega}_X RC_1^{-1}D - KC^{\perp}\mathbf{\Omega}_X RC_1^{-1}C\mathbf{\Omega}B \\ &= I - KUN^{-1}C_1^{-1}D - KC^{\perp}L\mathbf{\Omega}B - KC^{\perp}\mathbf{\Omega}_X RC_1^{-1}D - KC^{\perp}(\mathbf{\Omega}_X L - L\mathbf{\Omega})B \\ &= N_u^{-1}N_1 - KC^{\perp}\mathbf{\Omega}_X(RC_1^{-1}D + LB) \ , \end{aligned}$$

where $\mathbf{\Delta}$ is given by (3.8), and the matrix

$$\begin{aligned} N_1 &= N_u(I - KUN^{-1}C_1^{-1}D) \\ &= N_u^{-1}\left[Q_u + D^{\mathrm{T}}Q_e D - (B^{\mathrm{T}}Y + D^{\mathrm{T}}Q_e C)UN^{-1}C_1^{-1}D)\right] \\ &= N_u^{-1}(Q_u - B^{\mathrm{T}}YUN^{-1}C_1^{-1}D) \end{aligned}$$

has been introduced. To achieve proper controller $\mathbf{C}$, assume that N_1 is nonsingular and introduce the rational matrix

$$\mathbf{\Omega}_{XY} = \left[sI - A_2 + Z(A_3 + Q) - (RC_1^{-1}D + LB)N_1^{-1}N_u KC^{\perp}\right]^{-1} \ .$$

The last matrix identity we shall need is

$$\mathbf{\Omega}_{XY}(RC_1^{-1}D + LB)N_1^{-1}N_u KC^{\perp}\mathbf{\Omega}_X = \mathbf{\Omega}_{XY} - \mathbf{\Omega}_X \ .$$

We have

$$\begin{aligned} & \mathbf{C} = \left(\mathbf{\Delta} - \mathbf{WG}\right)^{-1}\mathbf{W} \\ &= N_1^{-1}\left[I + N_u KC^{\perp}\mathbf{\Omega}_{XY}(RC_1^{-1}D + LB)N_1^{-1}\right]N_u K(U + C^{\perp}\mathbf{\Omega}_X RN)N^{-1}C_1^{-1} \\ &= N_1^{-1}\big[N_u KU + N_u KC^{\perp}\mathbf{\Omega}_X RN + N_u KC^{\perp}\mathbf{\Omega}_{XY}(RC_1^{-1}D + LB)N_1^{-1}N_u KU \\ &\qquad + N_u KC^{\perp}(\mathbf{\Omega}_{XY} - \mathbf{\Omega}_X)RN\big]N^{-1}C_1^{-1} \\ &= N_1^{-1}N_u K\left\{U + C^{\perp}\mathbf{\Omega}_{XY}\left[RN + (RC_1^{-1}D + LB)N_1^{-1}N_u KU\right]\right\}N^{-1}C_1^{-1} \ . \end{aligned}$$

This is a state space realization of optimal controller $\mathbf{C}$. Note that its order is minimal, equal to $n - p$.

Concerning Assumption 6, we shall give its interpretation for this particular application. By (4.12), the poles of $\mathbf{\Phi}$ are the eigenvalues of A, and the set τ consists of the eigenvalues of $-(A - BK)$. Therefore, Assumption 6 means that the unstable eigenvalues of A are not eigenvalues of $-(A - BK)$.

Remark 4.1. In the above example we have assumed that rational matrix $\mathbf{G}_w W$ is right invertible. If it is not right invertible, under very general conditions, we can apply an observer of arbitrary pole distribution, in which the disturbance variable w is absent [12] (of minimal order $n - p$), and detect the state variable x arbitrarily fast.

4.3. Tracking: White measurement noise is present. In this subsection we assume white noises with $\mathbf{\Phi}_{vv} = M > 0$, $N^2 = M$, $\mathbf{\Phi}_{ww} = W \geq 0$, and $\mathbf{G}$ strict proper (i.e., $D = 0$ in (3.1)).

For some symmetric positive definite matrix H, let the spectral density of a shape-deterministic reference signal r be

$$\mathbf{\Phi}_{rr} = C_r \mathbf{\Omega}_r B_r H B_r^{\mathrm{T}} \mathbf{\Omega}_r^{\sim} C_r^{\mathrm{T}} \ , \quad \mathbf{\Omega}_r = (sI - A_r)^{-1} \ ,$$

where (A_r, B_r) and (C_r, A_r) are controllable and observable matrix pairs, correspondingly.

To solve the formulated optimal control problem, we shall apply Corollary 2.5.

By Assumption 1, the condition of stability of $\mathbf{\Theta}_1 \mathbf{r}$ has to be satisfied. We have

$$\begin{aligned} \mathbf{\Theta}_1 \mathbf{r} &= (I - C\mathbf{\Omega}_F F) C_r \mathbf{\Omega}_r B_r = C_r \mathbf{\Omega}_r B_r - C \mathbf{\Omega}_F F C_r \mathbf{\Omega}_r B_r \\ &= C_r \mathbf{\Omega}_r B_r - C(\mathbf{\Omega}_F U_r - U_r \mathbf{\Omega}_r) B_r = (C_r + C U_r) \mathbf{\Omega}_r B_r - C \mathbf{\Omega}_F U_r B_r \ , \end{aligned}$$

where U_r is a solution of the Sylvester equation

$$U_r A_r - (A - FC) U_r + F C_r = 0 \ . \tag{4.13}$$

If the pair (C, A) is observable, we can take the eigenvalues of $A - FC$ pairwise distinct of the eigenvalues of A_r; hence (4.13) has a solution for U_r. Since the pair (A_r, B_r) is controllable, from the condition of stability of $(C_r + CU_r)\mathbf{\Omega}_r B_r$, we obtain the condition of stability of $(C_r + CU_r)\mathbf{\Omega}_r$.

To find $\mathbf{\Phi}$, write

$$\begin{aligned} \mathbf{\Phi}_{vv} + \mathbf{\Phi}_{xx} &= M + \mathbf{\Phi}_{rr} + \mathbf{G}_w W \mathbf{G}_w^{\sim} \\ &\stackrel{\mathrm{def}}{=} M + \bar{C}(sI - \bar{A})^{-1} \bar{B} \bar{\Phi} \bar{B}^{\mathrm{T}} (-sI - \bar{A}^{\mathrm{T}})^{-1} \bar{C}^{\mathrm{T}} \\ &\stackrel{\mathrm{def}}{=} M + \bar{\mathbf{G}} \bar{\Phi} \bar{\mathbf{G}}^{\sim} = \mathbf{\Phi} \mathbf{\Phi}^{\sim} \ , \end{aligned}$$

where

$$\bar{C} = [C_r, C] \ , \quad \bar{A} = \begin{bmatrix} A_r & 0 \\ 0 & A \end{bmatrix} \ , \quad \bar{B} = \begin{bmatrix} B_r & 0 \\ 0 & B_w \end{bmatrix} \ , \quad \bar{\Phi} = \begin{bmatrix} H & 0 \\ 0 & W \end{bmatrix} \ ,$$

$$\bar{\mathbf{G}} = \bar{C}(sI - \bar{A})^{-1} \bar{B} \stackrel{\mathrm{def}}{=} \bar{C} \bar{\mathbf{\Omega}} \bar{B} \ .$$

We have

$$\mathbf{\Phi} = N + \bar{C} \bar{\mathbf{\Omega}} X \bar{C}^{\mathrm{T}} M^{-1} N \ , \tag{4.14}$$

and X is the stabilizing solution of CARE

$$\bar{A} X + X \bar{A}^{\mathrm{T}} - X \bar{C}^{\mathrm{T}} M^{-1} \bar{C} X + \bar{B} \bar{\Phi} \bar{B}^{\mathrm{T}} = 0 \ .$$

The stabilizing solution exists because, by Lemma 3.1, matrix pair (C_r, A_r) is observable and (C, A) is observable; i.e., matrix pair $(\bar{C}, \bar{A})$ is observable.

Consider the following Sylvester equation for the unknown matrix K_1:

$$K_1 A_r + (A - B Q_u^{-1} B^{\mathrm{T}} Y)^{\mathrm{T}} K_1 + C^{\mathrm{T}} Q_e C_r = 0 \ . \tag{4.15}$$

We shall prove that it has a solution. By (4.14), the poles of $\mathbf{\Phi}$ are a union of the eigenvalues of A and eigenvalues of A_r, and the set τ consists of the eigenvalues of $-(A-BK)$ and eigenvalues of $-(\hat{A}-X\hat{C}^{\mathrm{T}}M\hat{C})$. Therefore, Assumption 6 means that the unstable eigenvalues of A and unstable eigenvalues of A_r are not eigenvalues of $-(A-BK)$ and of $-(\hat{A}-X\hat{C}^{\mathrm{T}}M\hat{C})$. In particular, the eigenvalues of A_r and eigenvalues of $-(A-BK)^{\mathrm{T}}=-(A-BQ_u^{-1}B^{\mathrm{T}}Y)^{\mathrm{T}}$ are pairwise distinct; hence Sylvester equation (4.15) has a solution.

We shall need the matrix identity

$$\mathbf{\Omega}_Y^{\sim}C^{\mathrm{T}}Q_eC_r\mathbf{\Omega}_r=\mathbf{\Omega}_Y^{\sim}K_1+K_1\mathbf{\Omega}_r\ ,$$

where K_1 is the solution of Sylvester equation (4.15), and the identity (4.8) with $D=0$.

Compute

$$\begin{aligned}&\left\{\mathbf{\Omega}_Y^{\sim}C^{\mathrm{T}}Q_e\bar{C}\bar{\mathbf{\Omega}}\right\}_\sigma\\ &=\left[\mathbf{\Omega}_Y^{\sim}C^{\mathrm{T}}Q_eC_r\mathbf{\Omega}_r,\ \mathbf{\Omega}_Y^{\sim}C^{\mathrm{T}}Q_eC\mathbf{\Omega}\right]_\sigma=\left[\mathbf{\Omega}_Y^{\sim}K_1+K_1\mathbf{\Omega}_r,\ \mathbf{\Omega}_Y^{\sim}Y+Y\mathbf{\Omega}\right]_\sigma\\ &=[K_1\mathbf{\Omega}_r,\ Y\mathbf{\Omega}]=[K_1,Y]\begin{bmatrix}\mathbf{\Omega}_r & 0\\ 0 & \mathbf{\Omega}\end{bmatrix}=Z\bar{\mathbf{\Omega}}\ ,\end{aligned}$$

where $Z=[K_1,Y]$. Further we have

$$\begin{aligned}\mathbf{L}_\sigma=\{\mathbf{\Psi}^{\sim}Q_e\mathbf{\Phi}\}_\sigma&=\left\{N_u^{-1}B^{\mathrm{T}}\mathbf{\Omega}_Y^{\sim}C^{\mathrm{T}}Q_e\left(N+\bar{C}\bar{\mathbf{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}N\right)\right\}_\sigma\\ &=N_u^{-1}B^{\mathrm{T}}Z\bar{\mathbf{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}N\end{aligned}$$

and

$$\begin{aligned}\mathbf{W}=\mathbf{L}_\sigma\mathbf{\Phi}^{-1}&=N_u^{-1}B^{\mathrm{T}}Z\bar{\mathbf{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}\left(I+\bar{C}\bar{\mathbf{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}\right)^{-1}\\ &=N_u^{-1}B^{\mathrm{T}}Z\left(I+\bar{\mathbf{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}\bar{C}\right)^{-1}\bar{\mathbf{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}\\ &=N_u^{-1}B^{\mathrm{T}}Z\left(sI-\bar{A}+X\bar{C}^{\mathrm{T}}M^{-1}\bar{C}\right)^{-1}X\bar{C}^{\mathrm{T}}M^{-1}\\ &=N_u^{-1}B^{\mathrm{T}}Z\mathbf{\Omega}_XX\bar{C}^{\mathrm{T}}M^{-1}\ ,\end{aligned}$$

where $\mathbf{\Omega}_X=\left(sI-\bar{A}+X\bar{C}^{\mathrm{T}}M^{-1}\bar{C}\right)^{-1}$. It is easy to check that

$$\mathbf{\Omega}_XX\bar{C}^{\mathrm{T}}M^{-1}C\mathbf{\Omega}=L\mathbf{\Omega}-\mathbf{\Omega}_XL\ ,$$

where $L=\left[\begin{smallmatrix}0\\ I\end{smallmatrix}\right]$. Therefore

$$\begin{aligned}\mathbf{C}&=(\mathbf{\Delta}-\mathbf{WG})^{-1}\mathbf{W}\\ &=\left[N_u+N_uQ_u^{-1}B^{\mathrm{T}}Y\mathbf{\Omega}B-N_u^{-1}B^{\mathrm{T}}Z\mathbf{\Omega}_XX\bar{C}^{\mathrm{T}}M^{-1}C\mathbf{\Omega}B\right]^{-1}\mathbf{W}\\ &=\left[N_u+N_uQ_u^{-1}B^{\mathrm{T}}Y\mathbf{\Omega}B-N_u^{-1}B^{\mathrm{T}}Z(L\mathbf{\Omega}-\mathbf{\Omega}_XL)B\right]^{-1}\mathbf{W}\\ &=(I+Q_u^{-1}B^{\mathrm{T}}Z\mathbf{\Omega}_XLB)^{-1}Q_u^{-1}B^{\mathrm{T}}Z\mathbf{\Omega}_XX\bar{C}^{\mathrm{T}}M^{-1}\\ &=Q_u^{-1}B^{\mathrm{T}}Z(I+\mathbf{\Omega}_XLBQ_u^{-1}B^{\mathrm{T}}Z)^{-1}\mathbf{\Omega}_XX\bar{C}^{\mathrm{T}}M^{-1}\\ &=Q_u^{-1}B^{\mathrm{T}}Z(sI-\bar{A}+X\bar{C}^{\mathrm{T}}M^{-1}\bar{C}+LBQ_u^{-1}B^{\mathrm{T}}Z)^{-1}X\bar{C}^{\mathrm{T}}M^{-1}\ ,\end{aligned}$$

because $ZL = Y$. This is a state-space realization of optimal controller $\mathbf{C}$. Its order is $n + n_r$, where n_r is the dimension of the matrix A_r.

It remains to check Assumption 5. We have

$$\begin{aligned}\boldsymbol{\Theta}_1\boldsymbol{\Phi} &= (I - C\boldsymbol{\Omega}_F F)(N + \bar{C}\bar{\boldsymbol{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}N)\\ &= N + \bar{C}\bar{\boldsymbol{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}N - C\boldsymbol{\Omega}_F FN - C\boldsymbol{\Omega}_F F\bar{C}\bar{\boldsymbol{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}N\\ &= N + \bar{C}\bar{\boldsymbol{\Omega}}X\bar{C}^{\mathrm{T}}M^{-1}N - C\boldsymbol{\Omega}_F FN - C(\boldsymbol{\Omega}_F U - U\bar{\boldsymbol{\Omega}})X\bar{C}^{\mathrm{T}}M^{-1}N\ ,\end{aligned}$$

where $U = [U_r, -I]$. Further we have

$$\boldsymbol{\Theta}_1\boldsymbol{\Phi} = N + (C_r + CU_r)\boldsymbol{\Omega}_r[I, 0]X\bar{C}^{\mathrm{T}}M^{-1}N - C\boldsymbol{\Omega}_F(F + UX\bar{C}^{\mathrm{T}}M^{-1})N\ .$$

Since $(C_r + CU_r)\boldsymbol{\Omega}_r$ is stable, $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is stable. Further we have

$$\begin{aligned}(\boldsymbol{\Theta}_1\boldsymbol{\Phi})^{-1} &= \boldsymbol{\Phi}^{-1}\boldsymbol{\Theta}_1^{-1} = N^{-1}(I - \bar{C}\boldsymbol{\Omega}_X X\bar{C}^{\mathrm{T}}M^{-1})(I + C\boldsymbol{\Omega}F)\\ &= N^{-1}\left[I + C\boldsymbol{\Omega}F - \bar{C}\boldsymbol{\Omega}_X X\bar{C}^{\mathrm{T}}M^{-1} - \bar{C}(L\boldsymbol{\Omega} - \boldsymbol{\Omega}_X L)F\right]\\ &= N^{-1}\left[I - \bar{C}\boldsymbol{\Omega}_X(X\bar{C}^{\mathrm{T}}M^{-1} - LF)\right]\ .\end{aligned}$$

Therefore, $\boldsymbol{\Theta}_1\boldsymbol{\Phi}$ is minimum phase.

Remark 4.2. It can be proved that the property $\boldsymbol{\Theta}_1\mathbf{r}$ being stable is independent of the matrix F in (3.2) and (3.3), introduced to define $\boldsymbol{\Theta}_1$.

Conclusions. We have presented solutions to an LQ output tracking controller in the presence of plant disturbances with or without measurement noise, in which matrices of Youla–Kučera parametrization do not appear. Further research can be as follows:

• Although matrices of Youla–Kučera parametrization are not present in the optimal controller transfer matrix, they are still present in the expression for the minimal value of the criterion (see (3.18) and (3.19)). On the other hand, if we deal with the LQG control problem, the minimal value of the criterion can be easily expressed in explicit form, without matrices of Youla–Kučera parametrization. Can it be expressed in a similar explicit form in general?

• Generalize the application from subsection 4.1 to consider $\boldsymbol{\Phi}_{vv} = V_1\boldsymbol{\Omega}_v V\boldsymbol{\Omega}_v^{\sim}V_1^{\mathrm{T}}$ for some matrix V_1. The augmentation method of subsection 4.3 can be applied, and then spectral factorization of strict proper matrices of subsection 4.2 can be applied.

Acknowledgment. The author is grateful to a reviewer for many helpful remarks.

REFERENCES

[1] J. J. Bongiorno, Jr., and D. C. Youla, *Wiener-Hopf design of optimal decoupling one-degree-of-freedom controllers for plants with rectangular transfer matrices*, Internat. J. Control, 74 (2001), pp. 1393–1411.

[2] A. Casavola and E. Mosca, *Polynomial LQG regulator design for general system configurations*, IEEE Trans. Automat. Control, AC-38 (1993), pp. 359–363.

[3] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

[4] K. J. Hunt, M. Šebek, and V. Kučera, *Polynomial solution of the standard multivariable $\mathcal{H}_2$-optimal control problem*, IEEE Trans. Automat. Control, AC-39 (1994), pp. 1502–1507.
[5] V. Kučera, *Stochastic multivariable control: A polynomial equation approach*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 913–919.
[6] H. Kwakernaak, *$\mathcal{H}_2$-optimization—Theory and applications to robust control design*, in Proceedings of the IFAC Symposium on Robust Control Design, Prague, Czech Republic, 2000.
[7] H. Kwakernaak, *Polynomial solution of the standard $\mathcal{H}_2$ problem*, in Proceedings of the European Control Conference, Porto, Portugal, 2001.
[8] K. Park and J. J. Bongiorno, Jr., *A general theory for the Wiener-Hopf design of multivariable control systems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 619–626.
[9] K. Park, Y.-S. Cho, J.-S. Lee, and K. Kang, *Wiener-Hopf approach to derivation of rational doubly coprime factorizations*, Automatica J. IFAC, 35 (1999), pp. 725–728.
[10] K. Park and D. C. Youla, *Numerical calculation of the optimal three-degree-of-freedom Wiener-Hopf controller*, Internat. J. Control, 56 (1992), pp. 227–244.
[11] M. Šebek, *Polynomial design of stochastic tracking systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 468–470.
[12] J. Stefanovski, *Condensed forms of linear control system under output feedback*, Linear Algebra Appl., 328 (2001), pp. 1–55.
[13] A. A. Stoorvogel, *The singular $\mathcal{H}_2$ control problem*, Automatica J. IFAC, 28 (1992), pp. 627–631.
[14] K. Takaba and T. Katayama, *$\mathcal{H}_2$ output feedback control for descriptor systems*, Automatica J. IFAC, 34 (1998), pp. 841–850.
[15] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880–894.
[16] D. C. Youla and J. J. Bongiorno, Jr., *Wiener-Hopf design of optimal decoupling one-degree-of-freedom controllers*, Internat. J. Control, 73 (2000), pp. 1657–1670.
[17] D. C. Youla, H. A. Jabr, and J. J. Bongiorno, Jr., *Modern Wiener-Hopf design of optimal controllers.* II: *The multivariable case*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 319–338.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2035–2053

ON THE PARAMETRIZATION OF ALL REGULARLY IMPLEMENTING AND STABILIZING CONTROLLERS*

C. PRAAGMAN†, H. L. TRENTELMAN‡, AND R. ZAVALA YOE‡

Abstract. In this paper we deal with problems of controller parametrization in the context of behavioral systems. Given a full plant behavior, a subbehavior of the manifest plant behavior is called *regularly implementable* if it can be achieved as the controlled behavior resulting from the interconnection of the full plant behavior with a suitable controller behavior, in such a way that the controller does not impose restrictions that are already present in the plant. We establish a parametrization of *all* controllers that regularly implement a given behavior. We also obtain a parametrization of all stabilizing controllers.

Key words. linear differential system behaviors, implementability, stabilization, parametrization of controllers, Youla parametrization

AMS subject classifications. 93B05, 93B07, 93B50, 93B52, 93C05, 93C15, 93D15, 93D20.

DOI. 10.1137/050628696

1. Introduction.

An important issue in the behavioral approach to control is that of *implementability*. The concept of implementability has been successfully applied to resolve a number of important synthesis problems in the behavioral approach, in particular the synthesis of dissipative systems [11] and the behavioral versions of the problems of pole placement and stabilization [2]. The concept was also studied in [3], for nD behaviors in [7], and for general behaviors in [8]. A nice overview can also be found in [1]. Implementability deals with the issue which system behaviors can be achieved ("implemented") by interconnecting a given system with a controller, and is thus concerned with the limits of performance of a given plant. In the behavioral framework this is made precise as follows. Given is a system behavior (plant) with two types of variables: the variable w to be controlled, and the variable c (the control variable) on which we are allowed to put restrictions. A controller for our plant behavior is an additional system behavior, called controller behavior. Interconnecting the plant with the controller simply means requiring c to be an element of the controller behavior. The space of all w trajectories that are possible after interconnecting the plant behavior with the controller behavior forms the so-called *manifest controlled behavior*. A behavior is called implementable (w.r.t. the given plant behavior) if it can be achieved as manifest controlled behavior in this way. In the contexts of synthesis of dissipative systems, pole placement, and stabilization, an important role is also played by *regular implementability*. A given behavior is called regularly implementable if it can be achieved by a controller behavior that does not impose restrictions on the control variable that are already present in the plant; equivalently, the number of outputs of the associated full controlled behavior is equal to the sum of the number of outputs of the plant and the number of outputs of the controller. In [11], for a given plant behavior a characterization was given of all implementable behaviors, and

*Received by the editors April 7, 2005; accepted for publication (in revised form) August 4, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sicon/45-6/62869.html
†Institute of Economics and Econometrics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (C.Praagman@eco.rug.nl).
‡Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (h.l.trentelman@math.rug.nl, rzyoe@math.rug.nl).

in [2] a characterization was given of all regularly implementable behaviors. Once a given behavior is (regularly) implementable, it is important to know which controller behaviors implement it. In this paper we establish a parametrization of all controller behaviors that regularly implement a given behavior.

A controller is called a stabilizing controller if it regularly implements a stable behavior. In [2], conditions on the plant behavior were given for the existence of a stabilizing controller. Once a stabilizing controller exists, it is important to have a parametrization of all stabilizing controllers. In fact, a result of paramount importance in feedback control is the celebrated Youla parametrization of all stabilizing controllers; see [12] and [9]. In this paper we find a parametrization of all stabilizing controllers in the behavioral framework. In this framework, the parametrization problem was considered before in [4] for the so-called full interconnection case. Here, we resolve the general, partial interconnection case.

The outline of this paper is as follows. In section 2 we review the basic material on linear differential systems needed in this paper and some less known material on left and right minimal annihilators. In section 3 we formulate the main problems that we treat in this paper, the problems of parametrizing all regularly implementing controllers and all stabilizing controllers. We also give some motivating examples there. Section 4 deals with the special case of full interconnection. We review the basic facts on implementability by full interconnection and present a new condition for regular implementability. Next, we solve the problems of parametrizing all regularly implementing controllers and all stabilizing controllers, both for the full interconnection case. Section 5 solves the parametrization problem for the case that in the plant behavior the control variable is observable from the manifest variable. Next, in section 6, we reduce the general, nonobservable, case to the observable case by describing two reduction steps. Then, in section 7 we give a parametrization of all stabilizing controllers, first in the observable case, and next by reducing the general case to the observable case. Finally, in section 8, we provide a number of worked out examples to illustrate the theory of this paper.

In several places in this paper we denote by $\mathrm{col}(A, B)$ the matrix obtained by stacking A over B. A polynomial p is called Hurwitz if all its zeros are contained in the open left half complex plane $\mathbb{C}^- := \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) < 0\}$. A square polynomial matrix P is called Hurwitz if $\det(P)$ is Hurwitz. Finally, we denote $\mathbb{C}^+ := \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) \geq 0\}$.

2. Linear differential systems. In the behavioral approach to linear systems, a dynamical system is given by a triple $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathfrak{B})$, where $\mathbb{R}$ is the time axis, $\mathbb{R}^q$ is the signal space, and the behavior $\mathfrak{B}$ is a subset of $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ (the space of all infinitely often differentiable functions from $\mathbb{R}$ to $\mathbb{R}^q$) consisting of all solutions of a set of higher order, linear, constant coefficient differential equations. More precisely, there exists a real polynomial matrix R with q columns such that $\mathfrak{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0\}$. Any such dynamical system Σ is called a *linear differential system*. The set of all linear differential systems with q variables is denoted by $\mathfrak{L}^q$. Since the behavior $\mathfrak{B}$ of the system Σ is the central item, we will speak mostly about the system $\mathfrak{B} \in \mathfrak{L}^q$ (instead of $\Sigma \in \mathfrak{L}^q$). Henceforth, in this paper we will suppress the notation $\frac{\mathrm{d}}{\mathrm{d}t}$ and write Rw instead of $R(\frac{\mathrm{d}}{\mathrm{d}t})w$.

The behavioral approach makes a distinction between the behavior as the space of all solutions to a set of (differential) equations and the set of equations itself. A set of equations in terms of which the behavior is defined is called a *representation* of the behavior. If a behavior $\mathfrak{B}$ is represented by $Rw = 0$, then we call this a kernel representation of $\mathfrak{B}$ and often write $\mathfrak{B} = \ker(R)$.

Suppose R has p rows. Then the kernel representation is said to be *minimal* if every other kernel representation of $\mathfrak{B}$ has at least p rows. A given kernel representation $\mathfrak{B} = \ker(R)$ is minimal if and only if the polynomial matrix R has full row rank (see [5, Theorem 3.6.4]). The number of rows in any minimal kernel representation of $\mathfrak{B}$ is denoted by $\mathfrak{p}(\mathfrak{B})$. This number is called the *output cardinality* of $\mathfrak{B}$. It corresponds to the number of outputs in any input/output representation of $\mathfrak{B}$ (see [5, section 3.3]).

A linear differential system is *defined* as the solution space $\mathfrak{B}$ of a differential equation of the form $Rw = 0$. However, such a system can have other representations as well. One of these is the image representation. Let M be a real polynomial matrix with q rows and, say, l columns. If

$$\mathfrak{B} = \{w \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^q) \mid \text{ there exists } \ell \in \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^l) \text{ such that } w = M\ell\},$$

then we call $w = M\ell$ an image representation of the system behavior $\mathfrak{B}$ and often write $\mathfrak{B} = \mathrm{im}(M)$. The linear differential system $\mathfrak{B}$ has an image representation if and only if it is *controllable* (see [5, Theorem 6.6.1]). If $\mathfrak{B} = \ker(R)$, then $\mathfrak{B}$ is controllable if and only if the rank of the complex matrix $R(\lambda)$ is independent of λ for $\lambda \in \mathbb{C}$. If $\mathfrak{B}$ is a linear differential system, then we denote by $\mathfrak{B}_{\mathrm{cont}}$ the largest controllable subbehavior of $\mathfrak{B}$ (see [5, Theorem 5.2.14]). The system $\mathfrak{B}$ is *stabilizable* if and only if the rank of $R(\lambda)$ is independent of λ for $\lambda \in \mathbb{C}^+$ (see [5, Theorem 5.2.30]).

We now recall the concepts of minimal (left and right) annihilator. If M is a polynomial matrix, then the polynomial matrix R is called *a minimal left annihilator* (MLA) of M if $\mathrm{im}(M) = \ker(R)$. Since, for any M, $\mathrm{im}(M)$ is a linear differential behavior (see [5, section 6.6]), and since every linear differential behavior has a kernel representation $\ker(R)$, every M has an MLA. For a given polynomial matrix R, the polynomial matrix M is called *a minimal right annihilator* (MRA) of R if $\mathrm{im}(M) = (\ker(R))_{\mathrm{cont}}$. Given R, the controllable part $(\ker(R))_{\mathrm{cont}}$ of $\ker(R)$ admits an image representation $\mathrm{im}(M)$ (see [5, Theorem 6.6.1]). Thus every R has an MRA.

In many places in this paper, we will use, for a given M, an MLA of R *with full row rank*. If the given M has full row rank, say q, then for consistency we *define* such a full row rank MLA as the "void" matrix R with 0 rows and q columns. Likewise we often use, for a given R, an MRA with the property that $M(\lambda)$ has full column rank for all λ. If R has full column rank q, we define such an M to be the void matrix with q rows and 0 columns. In that case, if K is a given matrix with q columns, then KM is again void. Finally, we use the convention that if R is void, with zero rows and q columns, then a full column rank MRA is given by the $q \times q$ identity matrix I_q.

Note that if R is a full row rank MLA, say of M, then automatically $R(\lambda)$ has full row rank for all $\lambda \in \mathbb{C}$. Indeed, if $\ker(R) = \mathrm{im}(M)$, then $\ker(R)$ is controllable so its rank is independent of λ. As an immediate consequence of this, if R is a full row rank MLA, then there exists a polynomial matrix R' such that $\mathrm{col}(R, R')$ is unimodular.

Next, we review some facts on observability. Suppose $\mathfrak{B} \in \mathfrak{L}^q$ with system variable $w = (w_1, w_2)$, where w_1 and w_2 take values in $\mathbb{R}^{q_1}$ and $\mathbb{R}^{q_2}$, respectively, $q = q_1 + q_2$. We call w_2 *observable from* w_1 if $(w_1, w_2), (w_1, w_2') \in \mathfrak{B}$ implies $w_2 = w_2'$. We call w_2 *detectable from* w_1 if $(w_1, w_2), (w_1, w_2') \in \mathfrak{B}$ implies $\lim_{t\to\infty}(w_2(t) - w_2'(t)) = 0$. If $\mathfrak{B}$ is represented by $R_1 w_1 + R_2 w_2 = 0$, then w_2 is observable from w_1 if and only if $R_2(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$ ([5, Theorem 5.3.3]). Also, w_2 is detectable from w_1 if and only if $R_2(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}^+$ ([5, Theorem 5.3.17]).

We now review some facts on elimination. Again, let $\mathfrak{B} \in \mathfrak{L}^q$ with system variable $w = (w_1, w_2)$. Let P_{w_1} denote the projection onto the w_1-component. Then the set $P_{w_1}\mathfrak{B}$ of all w_1 for which there exists w_2 such that $(w_1, w_2) \in \mathfrak{B}$ is again a linear differential system. In this paper we denote $P_{w_1}\mathfrak{B}$ by $\mathfrak{B}_{w_1}$. We call $\mathfrak{B}_{w_1}$ the

system obtained by eliminating w_2 from $\mathfrak{B}$ ([5, section 6.2.2]). If $\mathfrak{B} = \ker(R_1 \;\; R_2)$, then a representation of $\mathfrak{B}_{w_1}$ is obtained as follows: choose a unimodular matrix V such that $VR_2 = \mathrm{col}(R_{12}, 0)$, with R_{12} full row rank, and conformably partition $VR_1 = \mathrm{col}(R_{11}, R_{21})$. Then $\mathfrak{B}_{w_1} = \ker(R_{21})$ (see [5, section 6.2.2]). This is due to the fact that R_{12} is a full row rank polynomial matrix, which therefore induces a surjective differential operator from $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^{q_2})$ to $\mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^r)$, with $r := \mathrm{rowdim}(R_{12})$.

Sometimes, system behaviors are represented by latent variable representations of the form $Rw = M\ell$, with *latent variable* ℓ. Of course, this equation represents the full behavior of all (w, ℓ) that satisfy the differential equation. The w-behavior $\mathfrak{B}$ obtained by eliminating ℓ from this full behavior is called the *manifest behavior* associated with this latent variable representation. On several occasions in this paper we need to compute the output cardinality $\mathtt{p}(\mathfrak{B})$ of this behavior in terms of the polynomial matrices R and M. It was shown in [2, Lemma 8], that $\mathtt{p}(\mathfrak{B}) = \mathrm{rank}(R, M) - \mathrm{rank}(M)$.

Finally, we recall some facts on autonomous systems. If the behavior $\mathfrak{B}$ has the property that $\mathtt{p}(\mathfrak{B}) = q$ (the number of variables; thus all variables are output), then we call $\mathfrak{B}$ *autonomous*. An autonomous system is called *stable* if $\lim_{t\to\infty} w(t) = 0$ for all $w \in \mathfrak{B}$ (see [5, section 7.2]).

3. Problem formulation. In this section we introduce the main problems that are considered in this paper. We first briefly review the relevant definitions on interconnection and implementability. For an extensive treatment, see [2]. Let $\mathcal{P}_{\mathrm{full}} \in \mathfrak{L}^{q+k}$ be a linear differential system, with system variable (w, c), where w takes its values in $\mathbb{R}^q$ and c in $\mathbb{R}^k$. The variables w should be interpreted as the variables to be controlled, and the variables c are those through which we can interconnect the plant to a controller and are called *the control variables*. Let $\mathcal{C} \in \mathfrak{L}^k$ be a controller behavior, with variable c.

DEFINITION 1. *The interconnection of* $\mathcal{P}_{\mathrm{full}}$ *and* $\mathcal{C}$ *through* c *is defined as the system behavior* $\mathcal{K}_{\mathrm{full}}(\mathcal{C}) \in \mathfrak{L}^{q+k}$*, given by* $\mathcal{K}_{\mathrm{full}}(\mathcal{C}) = \{(w, c) \mid (w, c) \in \mathcal{P}_{\mathrm{full}}$ *and* $c \in \mathcal{C}\}$*. This behavior is called* the full controlled behavior.

DEFINITION 2. *The interconnection of* $\mathcal{P}_{\mathrm{full}}$ *and* $\mathcal{C}$ *through* c *is called* regular *if the output cardinality of the full controlled behavior is the sum of the output cardinalities of the plant and the controller, i.e.,* $\mathtt{p}(\mathcal{K}_{\mathrm{full}}(\mathcal{C})) = \mathtt{p}(\mathcal{P}_{\mathrm{full}}) + \mathtt{p}(\mathcal{C})$.

This condition is equivalent to the following: $\mathcal{C}$ does not reimpose restrictions on $\mathcal{K}_{\mathrm{full}}(\mathcal{C})$ that are already present in $\mathcal{P}_{\mathrm{full}}$ (see also [7, Definition 3.1]).

DEFINITION 3. *The behavior* $(\mathcal{K}_{\mathrm{full}}(\mathcal{C}))_w \in \mathfrak{L}^q$ *that is obtained by eliminating* c *from* $\mathcal{K}_{\mathrm{full}}(\mathcal{C})$ *is called the* manifest controlled behavior.

Let $\mathcal{K} \in \mathfrak{L}^q$ be a given behavior, which should be interpreted as a "desired" behavior. A fundamental question is whether this $\mathcal{K}$ can be achieved as controlled behavior as stated in the following definition.

DEFINITION 4. *If there exists* $\mathcal{C} \in \mathfrak{L}^k$ *such that* $\mathcal{K} = (\mathcal{K}_{\mathrm{full}}(\mathcal{C}))_w$*, then* $\mathcal{K}$ *is called* implementable by partial interconnection (through c, w.r.t. $\mathcal{P}_{\mathrm{full}}$). *If there exists* $\mathcal{C} \in \mathfrak{L}^k$ *such that* $\mathcal{K} = (\mathcal{K}_{\mathrm{full}}(\mathcal{C}))_w$ *and* $\mathtt{p}(\mathcal{K}_{\mathrm{full}}(\mathcal{C})) = \mathtt{p}(\mathcal{P}_{\mathrm{full}}) + \mathtt{p}(\mathcal{C})$*, then we call* $\mathcal{K}$ regularly implementable by partial interconnection (through c, w.r.t. $\mathcal{P}_{\mathrm{full}}$).

Necessary and sufficient conditions for a given $\mathcal{K} \in \mathfrak{L}^q$ to be (regularly) implementable by partial interconnection have been obtained in [11] and [2]. We will review these conditions in section 5.

We now formulate the first main problem that we deal with in this paper. Let $\mathcal{P}_{\mathrm{full}} = \ker(R_1 \;\; R_2)$ be a minimal representation of the plant. Let $\mathcal{K} = \ker(K)$ be a minimal representation of the desired behavior. Then the problem is as follows: *give a*

parametrization, in terms of the polynomial matrices R_1, R_2 and K, of all *polynomial matrices C such that the controller* $\ker(C)$ *regularly implements $\mathcal{K}$.*

Example 5. Consider the plant behavior $\mathcal{P}_{\text{full}}$ with manifest variable $w = (w_1, w_2)$ and control variable $c = (c_1, c_2)$ represented by

$$\begin{aligned} w_1 + \dot{w}_2 + \dot{c}_1 + c_2 &= 0, \\ c_1 + c_2 &= 0. \end{aligned}$$

Clearly, $(\mathcal{P}_{\text{full}})_w = \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^2)$. For the desired behavior $\mathcal{K}$ we take $\mathcal{K} = \{(w_1, w_2) \mid w_1 + \dot{w}_2 = 0\}$. The following controller regularly implements $\mathcal{K}$ through c: $\mathcal{C} = \{(c_1, c_2) \mid \dot{c}_1 + c_2 = 0\}$. Also every controller represented by $kc_1 + c_2 = 0$, with $k \neq 1$, regularly implements $\mathcal{K}$. We would like to find a parametrization of *all* 1×2 polynomial matrices $C(\xi) = (C_1(\xi) \quad C_2(\xi))$ such that $\mathcal{C} = \ker(C_1 \quad C_2)$ regularly implements $\mathcal{K}$.

We now recall the definition of a stabilizing controller (see [2, section 3]).

DEFINITION 6. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$. The controller $\mathcal{C} \in \mathfrak{L}^k$ is said to stabilize $\mathcal{P}_{\text{full}}$ through c if the manifest controlled behavior $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w$ is stable and the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{C}$ is regular. The controller $\mathcal{C}$ is then called a* stabilizing controller.

The following result was shown in [2, Theorem 6].

PROPOSITION 7. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$. There exists a stabilizing controller $\mathcal{C}$ if and only if $(\mathcal{P}_{\text{full}})_w$ is stabilizable and in $\mathcal{P}_{\text{full}}$ w is detectable from c.*

We now formulate the second main problem that is solved in this paper. Let $\mathcal{P}_{\text{full}} = \ker(R_1 \quad R_2)$ be a minimal representation of the plant. Then the problem is as follows: *give a parametrization, in terms of the polynomial matrices R_1 and R_2, of* all *polynomial matrices C such that the controller* $\ker(C)$ *is a stabilizing controller.*

Example 8. Consider the full plant behavior $\mathcal{P}_{\text{full}}$ represented by

$$\begin{aligned} w_1 + \dot{w}_2 + \dot{c}_1 + c_2 &= 0, \\ w_2 + c_1 + c_2 &= 0, \\ \dot{c}_1 + c_1 + \dot{c}_2 + c_2 &= 0. \end{aligned}$$

A stabilizing controller is given by $\mathcal{C} = \{(c_1, c_2) \mid \dot{c}_2 + 2c_1 + c_2 = 0\}$. Indeed, by eliminating c from the full controlled behavior $\mathcal{K}_{\text{full}}(\mathcal{C})$ (as described in section 2) we find that $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w = \ker(R)$, with

$$R(\xi) = \begin{pmatrix} 0 & \xi + 1 \\ -1 & 2 \end{pmatrix},$$

which is Hurwitz. Yet another class of stabilizing controllers is represented by $C(\xi) = (\xi(\xi + 1) + k, \xi + 1 + k)$, $k \in \mathbb{R}$. We want to find a parametrization of *all* 1×2 polynomial matrices $C(\xi)$ such that $\ker(C)$ is a stabilizing controller.

For the special case of full interconnection (see section 4), the problem of parametrizing all stabilizing controllers was considered earlier in [4]. Of course, in the context of feedback stabilization this parametrization problem dates back to the famous result of Youla [12].

4. Implementability and controller parametrization: The full interconnection case. In this section we treat the full interconnection case. First, we briefly review some facts on implementability and present a new condition for regular implementability. Then we establish a parametrization of all controllers that regularly implements a given behavior. Finally, we parametrize all stabilizing controllers.

Let $\mathcal{P} \in \mathfrak{L}^q$ be a plant behavior. A controller for $\mathcal{P}$ is a system behavior $\mathcal{C} \in \mathfrak{L}^q$. The *full interconnection* of $\mathcal{P}$ and $\mathcal{C}$ is defined as the system with behavior $\mathcal{P} \cap \mathcal{C}$. This *controlled behavior* is again an element of $\mathfrak{L}^q$. The full interconnection is called *regular* if $\mathfrak{p}(\mathcal{P} \cap \mathcal{C}) = \mathfrak{p}(\mathcal{P}) + \mathfrak{p}(\mathcal{C})$.

Let $\mathcal{K} \in \mathfrak{L}^q$ be a given behavior, to be interpreted as a "desired" behavior. If $\mathcal{K}$ can be achieved as controlled behavior, i.e., if there exists $\mathcal{C} \in \mathfrak{L}^q$ such that $\mathcal{K} = \mathcal{P} \cap \mathcal{C}$, then we call $\mathcal{K}$ *implementable by full interconnection* (w.r.t. $\mathcal{P}$). If $\mathcal{K}$ can be achieved by regular interconnection, i.e., if there exists $\mathcal{C}$ such that $\mathcal{K} = \mathcal{P} \cap \mathcal{C}$ and $\mathfrak{p}(\mathcal{P} \cap \mathcal{C}) = \mathfrak{p}(\mathcal{P}) + \mathfrak{p}(\mathcal{C})$, then we call $\mathcal{K}$ *regularly implementable by full interconnection*. Obviously, a given $\mathcal{K} \in \mathfrak{L}^q$ is implementable by full interconnection w.r.t. $\mathcal{P}$ if and only if $\mathcal{K} \subseteq \mathcal{P}$. Indeed, if $\mathcal{K} \subseteq \mathcal{P}$, then with "controller" $\mathcal{C} = \mathcal{K}$ we have $\mathcal{K} = \mathcal{P} \cap \mathcal{C}$. Thus, if $\mathcal{P} = \ker(R)$ and $\mathcal{K} = \ker(K)$ are minimal representations, then $\mathcal{K}$ is implementable w.r.t. $\mathcal{P}$ if and only if there exists a polynomial matrix F such that $R = FK$. The property of *regular* implementability turns out to be equivalent with the existence of such a polynomial matrix F with, in addition, $F(\lambda)$ full row rank for all λ, as stated in the following theorem.

THEOREM 9. *Let $\mathcal{P}, \mathcal{K} \in \mathfrak{L}^q$. Let $\mathcal{P} = \ker(R)$ and $\mathcal{K} = \ker(K)$ be minimal representations. Then the following statements are equivalent:*

1. *$\mathcal{K}$ is regularly implementable w.r.t. $\mathcal{P}$ by full interconnection.*
2. *There exists a polynomial matrix F with $F(\lambda)$ full row rank for all $\lambda \in \mathbb{C}$, such that $R = FK$.*
3. *$\mathcal{K} + \mathcal{P}_{\text{cont}} = \mathcal{P}$.*

Here, $\mathcal{P}_{\text{cont}}$ denotes the controllable part of $\mathcal{P}$.

Proof. The equivalence of statements 1 and 3 was proven in [2, Lemma 7]. We will only prove the equivalence of 1 and 2 here.

($1 \Rightarrow 2$) Let C be such that $\binom{R}{C} w = 0$ is a minimal representation of $\mathcal{K}$. Then there exists a unimodular U such that $\operatorname{col}(R, C) = UK$. This implies $R = FK$, with F consisting of the upper rows of U.

($2 \Rightarrow 1$) Assume $R = FK$. Let V be such that $\operatorname{col}(F, V)$ is unimodular. Define $C = VK$. Then $\binom{R}{C} w = 0$ is a minimal representation of $\mathcal{K}$, and thus $\mathcal{K}$ is regularly implemented by the controller $\mathcal{C} = \ker(C)$. ∎

From the above, for a given regularly implementable $\mathcal{K}$ it is easy to obtain a controller that regularly implements it. Indeed, if $R = FK$ with $F(\lambda)$ full row rank for all λ, let V be such that $\operatorname{col}(F, V)$ is unimodular. Then clearly the controller $\ker(VK)$ does the job. Note that this construction requires the representation $\mathcal{K} = \ker(K)$ to be minimal. In the subsequent development, in particular in section 5, we will need an expression in terms of K for a regularly implementing controller in the general case that K is *not* full row rank. This issue is dealt with in the following lemma.

LEMMA 10. *Let $\mathcal{P}, \mathcal{K} \in \mathfrak{L}^q$. Let K be such that $\mathcal{K} = \ker(K)$ (not necessarily minimal), and let R be such that $\mathcal{P} = \ker(R)$ is a minimal representation. Construct a polynomial matrix W as follows:*

1. *Let M be an MRA of R such that $M(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}$;*
2. *Let Q be a full row rank MLA of KM;*
3. *Let W be a polynomial matrix such that $\operatorname{col}(Q, W)$ is unimodular.*

Then $\mathcal{K}$ is regularly implementable by full interconnection w.r.t. $\mathcal{P}$ if and only if

$$\begin{pmatrix} R \\ WK \end{pmatrix} w = 0 \tag{1}$$

is a minimal representation of $\mathcal{K}$. A controller that regularly implements $\mathcal{K}$ is then represented by the minimal representation $WKw = 0$.

Proof. Factor $R = DR_1$, with D square and nonsingular and $R_1(\lambda)$ full row rank for all λ. Then $\mathcal{P}_{\text{cont}} = \ker(R_1)$ (see [5, Theorem 5.2.14]). Let M^+ be a polynomial left inverse of M and R_1^+ a polynomial right inverse of R_1. Define $S := M^+(I - R_1^+R_1)$. By direct verification we then have $\left(\begin{smallmatrix} R_1 \\ S \end{smallmatrix}\right)(R_1^+ \;\; M) = I_q$. It follows that $R_1^+R_1 + MS = I_q$. We claim that there exists a polynomial matrix T such that $TR = QK$. In order to prove this, we show that $\ker(R) \subseteq \ker(QK)$. Indeed, let w be such that $Rw = 0$. Since $\mathcal{P} = \mathcal{K} + \mathcal{P}_{\text{cont}}$, there exist $w_1 \in \mathcal{K}$ and $w_2 \in \mathcal{P}_{\text{cont}}$ such that $w = w_1 + w_2$. Hence, since $w_2 \in \ker(R_1)$ and $QKM = 0$, $QKw = QKw_2 = QK(R_1^+R_1 + MS)w_2 = 0$. Next, note that

$$\begin{pmatrix} I_p & 0 \\ -T & Q \\ 0 & W \end{pmatrix} \begin{pmatrix} R \\ K \end{pmatrix} = \begin{pmatrix} R \\ 0 \\ WK \end{pmatrix}.$$

The leftmost matrix in this equation is unimodular. Thus we have that ($w \in \mathcal{K}$) if and only if ($Kw = 0$ and $Rw = 0$), which in turn is equivalent to ($WKw = 0$ and $Rw = 0$). This proves that (1) is indeed a kernel representation of $\mathcal{K}$.

Finally, we show that the representation (1) is minimal. Indeed,

$$\begin{pmatrix} R \\ WK \end{pmatrix} \begin{pmatrix} R_1^+ & M \end{pmatrix} = \begin{pmatrix} D & 0 \\ WKR_1^+ & WKM \end{pmatrix}.$$

It is easily seen that, by construction, WKM has full row rank. Since D is nonsingular, we conclude that $\text{col}(R, WK)$ must have full row rank as well. ∎

We now establish, for a given plant $\mathcal{P} \in \mathfrak{L}^q$ and a given regularly implementable behavior $\mathcal{K} \in \mathfrak{L}^q$, a parametrization of all controllers $\mathcal{C} \in \mathfrak{L}^q$ that regularly implement $\mathcal{K}$ by full interconnection. This problem was considered before in [4] for the case where the plant behavior $\mathcal{P}$ is controllable and the given subbehavior $\mathcal{K}$ is autonomous. Here, we will establish a parametrization for arbitrary $\mathcal{P}$ and arbitrary (regularly implementable) $\mathcal{K}$.

THEOREM 11. *Let $\mathcal{P} \in \mathfrak{L}^q$, with minimal representation $\mathcal{P} = \ker(R)$. Let $\mathcal{K} \in \mathfrak{L}^q$ be regularly implementable by full interconnection, and let $\mathcal{K} = \ker(K)$. Construct a polynomial matrix W as in Lemma 10. Then for any $\mathcal{C} \in \mathfrak{L}^q$, $\mathcal{C} = \ker(C)$, the following statements are equivalent:*

1. *$\mathcal{C} = \ker(C)$ is a minimal representation, and $\mathcal{C}$ regularly implements $\mathcal{K}$.*
2. *There exist a polynomial matrix F and a unimodular polynomial matrix U such that $C = FR + UWK$.*

Proof. (2 ⇒ 1) First note that since $\mathcal{K}$ is regularly implementable, by Lemma 10 the polynomial matrix $\text{col}(R, WK)$ has full row rank. Since

$$\begin{pmatrix} I_p & 0 \\ F & U \end{pmatrix} \begin{pmatrix} R \\ WK \end{pmatrix} = \begin{pmatrix} R \\ FR + UWK \end{pmatrix}, \tag{2}$$

this implies that also $C = FR + UWK$ has full row rank, so $Cw = 0$ is a minimal representation of $\mathcal{C}$. It also follows from (2) that $\mathcal{C}$ implements $\mathcal{K}$. Clearly, the interconnection of $\mathcal{P}$ and $\mathcal{C}$ is regular.

(1 ⇒ 2) Assume that C has full row rank and that $\mathcal{C}$ regularly implements $\mathcal{K}$. Then both

$$\begin{pmatrix} R \\ C \end{pmatrix} w = 0 \quad \text{and} \quad \begin{pmatrix} R \\ WK \end{pmatrix} w = 0$$

are minimal representations of $\mathcal{K}$. Consequently, there exists a unimodular polynomial matrix $V = \left(\begin{smallmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{smallmatrix}\right)$ such that $V\mathrm{col}(R, WK) = \mathrm{col}(R, C)$. This implies $R = V_{11}R + V_{12}WK$. Since $\mathrm{col}(R, WK)$ has full row rank, this yields $V_{11} = I_p$ and $V_{12} = 0$. It follows that V_{22} is unimodular. We also have $C = V_{21}R + V_{22}WK$. This completes the proof of the theorem. ∎

Note that since $\mathrm{col}(R, WK)$ has full row rank, the linear map $(F, U) \mapsto FR + UWK$ is one-one, so different parameters (F, U) yield different controllers C.

To conclude this section, we parametrize all controllers that stabilize a given plant behavior by full interconnection. The problem of stabilization by full interconnection is formulated as follows. Let $\mathcal{P} \in \mathfrak{L}^q$ be a given plant behavior. Find a controller behavior $\mathcal{C} \in \mathfrak{L}^q$ such that the controlled behavior $\mathcal{K} = \mathcal{P} \cap \mathcal{C}$ is autonomous and stable and the interconnection is regular. It was proved in [10] that such a stabilizing controller $\mathcal{C}$ exists if and only if $\mathcal{P}$ is stabilizable.

Let $\mathcal{P} = \ker(R)$ be a minimal representation. Assume that $\mathcal{P}$ is stabilizable; equivalently, $R(\lambda)$ has full row rank for all $\lambda \in \mathbb{C}^+ = \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) \geq 0\}$. The following theorem yields a parametrization of all stabilizing controllers.

THEOREM 12. *Let $\mathcal{P} \in \mathfrak{L}^q$ be stabilizable. Let $\mathcal{P} = \ker(R)$ be a minimal representation, and let R_1 be such that $\ker(R_1)$ is a minimal representation of the controllable part $\mathcal{P}_{\mathrm{cont}}$ of $\mathcal{P}$. Let C_0 be such that $\mathrm{col}(R_1, C_0)$ is unimodular. Then for any $\mathcal{C} \in \mathfrak{L}^q$ with $\mathcal{C} = \ker(C)$ the following statements are equivalent:*

1. *$\mathcal{P} \cap \mathcal{C}$ is autonomous and stable, the interconnection is regular, and the representation $\mathcal{C} = \ker(C)$ is minimal.*
2. *There exist a polynomial matrix F and a Hurwitz polynomial matrix D such that $C = FR + DC_0$.*

Proof. By combining [5, Theorems 5.2.14 and 5.2.30], note that $R = D_1R_1$, with D_1 Hurwitz. ($2 \Rightarrow 1$) Assume that $C = FR + DC_0$. We have

$$\begin{pmatrix} R \\ C \end{pmatrix} = \begin{pmatrix} D_1 & 0 \\ FD_1 & D \end{pmatrix} \begin{pmatrix} R_1 \\ C_0 \end{pmatrix}.$$

This implies that $\mathrm{col}(R, C)$ has full row rank, so the interconnection of $\mathcal{P}$ and $\mathcal{C}$ is regular. Also, for some nonzero constant c, $\det \mathrm{col}(R, C) = c\det(D_1)\det(D)$, so the interconnection is autonomous and stable.

($1 \Rightarrow 2$) Since $\mathcal{P} \cap \mathcal{C}$ is stable and the interconnection is regular, $\mathrm{col}(R, C)$ is Hurwitz. Also, $\mathrm{col}(R_1, C_0)$ is unimodular, so there exist polynomial matrices F_{11}, F_{12}, F_{21}, and F_{22} such that

$$\begin{pmatrix} R \\ C \end{pmatrix} = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix} \begin{pmatrix} R_1 \\ C_0 \end{pmatrix}.$$

This implies that $F_{11} = D_1$ and $F_{12} = 0$. The left-hand side of the above equation is Hurwitz, so F_{22} must be Hurwitz. The proof is completed by taking $F = F_{21}$ and $D = F_{22}$. ∎

The above result generalizes the result from [4] for controllable $\mathcal{P}$. If, in the above, we assume that $\mathcal{P}$ is controllable, then we can take $R = R_1$, and we recover the parametrization obtained in [4].

Remark 13. In the special case that the plant $\mathcal{P}$ to be stabilized is given together with an input/output partition $w = (y, u)$, our parametrization result of Theorem 12 specializes to the well-known Youla parametrization of all stabilizing controllers. For simplicity, assume that $\mathcal{P}$ is controllable. Assume that, in $\mathcal{P}$, G is the transfer matrix

from u to y. Let $P^{-1}Q$ be a left coprime factorization of G. Then $\mathcal{P} = \ker(P \quad -Q)$. Choose polynomial matrices X and Y such that

$$\begin{pmatrix} P & -Q \\ X & Y \end{pmatrix}$$

is unimodular. According to Theorem 12, a parametrization of all stabilizing controllers $\ker(Q_c \quad P_c)$ is given by $(Q_c \quad P_c) = F(P \quad -Q) + D(X \quad Y)$, where F is arbitrary polynomial and D is Hurwitz. In transfer matrix form this yields $C := -P_c^{-1}Q_c = -(DY - FQ)^{-1}(DX + FP) = -(Y - D^{-1}FQ)^{-1}(X + D^{-1}FP)$. Finally, denote $D^{-1}F$ by T, and let T vary over all proper stable rational matrices to obtain the original Youla parametrization $C = -(Y - TQ)^{-1}(X + TP)$ (see [12]).

5. All controllers that regularly implement a given behavior: The observable case. We now turn to the partial interconnection case. Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$, with system variable (w, c), where w takes its values in $\mathbb{R}^q$ and c in $\mathbb{R}^k$. In this section and the next we study the problem of parametrizing, for a given regularly implementable $\mathcal{K} \in \mathfrak{L}^q$, all controllers $\mathcal{C} \in \mathfrak{L}^k$ that regularly implement $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$. We first assume that in the full plant behavior $\mathcal{P}_{\text{full}}$, c is observable from w. Starting from this assumption, in the present section we establish a parametrization. Then, in the next section we will lift the observability assumption and describe a parametrization for the general case.

Implementability by *partial* interconnection was already defined in section 3. Necessary and sufficient conditions for a given $\mathcal{K} \in \mathfrak{L}^q$ to be (regularly) implementable by partial interconnection can be given in terms of the manifest plant behavior and hidden behavior associated with the full plant behavior $\mathcal{P}_{\text{full}}$, which are defined as follows.

DEFINITION 14. *The* manifest plant behavior *is the behavior* $(\mathcal{P}_{\text{full}})_w \in \mathfrak{L}^q$ *obtained from* $\mathcal{P}_{\text{full}}$ *by eliminating* c. *The* hidden behavior $\mathcal{N}$ *consists of those* w *trajectories that appear in* $\mathcal{P}_{\text{full}}$ *with* c *equal to zero, i.e.,* $\mathcal{N} = \{w \mid (w, 0) \in \mathcal{P}_{\text{full}}\}$.

Conditions for implementability and regular implementability were obtained in [11, Theorem 1] and [2, Theorem 4], respectively, as follows.

PROPOSITION 15.

1. $\mathcal{K} \in \mathfrak{L}^q$ *is implementable by partial interconnection through* c *w.r.t.* $\mathcal{P}_{\text{full}}$ *if and only if* $\mathcal{N} \subseteq \mathcal{K} \subseteq (\mathcal{P}_{\text{full}})_w$.
2. $\mathcal{K} \in \mathfrak{L}^q$ *is regularly implementable by partial interconnection through* c *w.r.t.* $\mathcal{P}_{\text{full}}$ *if and only if* $\mathcal{N} \subseteq \mathcal{K} \subseteq (\mathcal{P}_{\text{full}})_w$ *and* $\mathcal{K}$ *is regularly implementable w.r.t.* $(\mathcal{P}_{\text{full}})_w$ *by full interconnection.*

For a given $\mathcal{K} \in \mathfrak{L}^q$, an important role will be played by the subbehavior $\mathcal{L}_{\text{full}}(\mathcal{K})$ of $\mathcal{P}_{\text{full}}$ defined as the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{K}$ through w:

$$\mathcal{L}_{\text{full}}(\mathcal{K}) := \{(w, c) \in \mathcal{P}_{\text{full}} \mid w \in \mathcal{K}\}. \tag{3}$$

The proof of the following lemma is straightforward and is left to the reader.

LEMMA 16. *If* $\mathcal{K} \subseteq (\mathcal{P}_{\text{full}})_w$, *then* $(\mathcal{L}_{\text{full}}(\mathcal{K}))_w = \mathcal{K}$.

Now let $\mathcal{K} \in \mathfrak{L}^q$ be implementable through c w.r.t. $\mathcal{P}_{\text{full}}$. We first consider the problem of finding one controller $\mathcal{C} \in \mathfrak{L}^k$ that implements $\mathcal{K}$. We will derive a representation of one such controller in terms of representations of $\mathcal{P}_{\text{full}}$ and $\mathcal{K}$. Let $\mathcal{P}_{\text{full}} = \ker(R_1 \quad R_2)$ and $\mathcal{K} = \ker(K)$. Then clearly the behavior $\mathcal{L}_{\text{full}}(\mathcal{K})$ defined by (3) is represented by

$$\begin{pmatrix} R_1 & R_2 \\ K & 0 \end{pmatrix} \begin{pmatrix} w \\ c \end{pmatrix} = 0.$$

Note that the hidden behavior $\mathcal{N}$ is equal to $\ker(R_1)$. Since $\mathcal{N} \subseteq \mathcal{K}$ there exists a polynomial matrix F such that $K = FR_1$. Now define a controller behavior $\mathcal{C}^* \in \mathfrak{L}^k$ by

$$\mathcal{C}^* := \ker(FR_2). \tag{4}$$

This controller indeed implements $\mathcal{K}$, as shown in the following lemma.

LEMMA 17. $\mathcal{K} = (\mathcal{K}_{\text{full}}(\mathcal{C}^*))_w$.

Proof. We have

$$\begin{pmatrix} R_1 & R_2 \\ 0 & FR_2 \end{pmatrix} = \begin{pmatrix} I & 0 \\ F & -I \end{pmatrix} \begin{pmatrix} R_1 & R_2 \\ K & 0 \end{pmatrix}.$$

Hence $\mathcal{L}_{\text{full}}(\mathcal{K})$ is equal to the full controlled behavior $\mathcal{K}_{\text{full}}(\mathcal{C}^*)$. The conclusion then follows from Lemma 16. ∎

The following lemma states that if c is observable from w and if a given subbehavior of the manifest plant behavior is obtained by elimination of c from a subbehavior of $\mathcal{P}_{\text{full}}$, then this subbehavior of $\mathcal{P}_{\text{full}}$ is unique, as shown in the following lemma.

LEMMA 18. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$ with system variable (w, c). Assume that c is observable from w. Let $\mathcal{K}^1_{\text{full}}, \mathcal{K}^2_{\text{full}} \in \mathfrak{L}^{q+k}$ be subbehaviors of $\mathcal{P}_{\text{full}}$. Then we have $(\mathcal{K}^1_{\text{full}})_w = (\mathcal{K}^2_{\text{full}})_w$ if and only if $\mathcal{K}^1_{\text{full}} = \mathcal{K}^2_{\text{full}}$.*

Proof. Assume $(w, c) \in \mathcal{K}^1_{\text{full}}$. Then $w \in (\mathcal{K}^1_{\text{full}})_w = (\mathcal{K}^2_{\text{full}})_w$, so there exists c' such that $(w, c') \in \mathcal{K}^2_{\text{full}}$. Thus, $(0, c - c') = (w, c) - (w, c') \in \mathcal{P}_{\text{full}}$, so $c = c'$. It follows that $(w, c) \in \mathcal{K}^2_{\text{full}}$. ∎

LEMMA 19. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$ with system variable (w, c). Let $\mathcal{K} \in \mathfrak{L}^q$ be implementable through c w.r.t. $\mathcal{P}_{\text{full}}$. Let $\mathcal{C}$ be a controller such that $\mathcal{K} = (\mathcal{K}_{\text{full}}(\mathcal{C}))_w$. Then we have the following:*

1. *$\mathcal{K}_{\text{full}}(\mathcal{C}) \subseteq \mathcal{L}_{\text{full}}(\mathcal{K})$;*
2. *if c is observable from w, then $\mathcal{K}_{\text{full}}(\mathcal{C}) = \mathcal{L}_{\text{full}}(\mathcal{K})$.*

Proof. (1) If $(w, c) \in \mathcal{K}_{\text{full}}(\mathcal{C})$, then $w \in (\mathcal{K}_{\text{full}}(\mathcal{C}))_w = \mathcal{K}$. Also, $(w, c) \in \mathcal{P}_{\text{full}}$. It follows that $(w, c) \in \mathcal{L}_{\text{full}}(\mathcal{K})$. (2) By Lemma 16 $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w = \mathcal{K} = (\mathcal{L}_{\text{full}}(\mathcal{K}))_w$. If c is observable from w, then this implies $\mathcal{K}_{\text{full}}(\mathcal{C}) = \mathcal{L}_{\text{full}}(\mathcal{K})$. ∎

For the special case that, in $\mathcal{P}_{\text{full}}$, c is observable from w, the following theorem reduces the problem of parametrizing all controllers that regularly implement $\mathcal{K}$ via interconnection through c w.r.t. $\mathcal{P}_{\text{full}}$ to that of parametrizing all controllers that regularly implement $(\mathcal{L}_{\text{full}}(\mathcal{K}))_c$ via *full interconnection* w.r.t. $(\mathcal{P}_{\text{full}})_c$ (see also [6, Prop. 1]).

THEOREM 20. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$ with system variable (w, c). Assume that c is observable from w. Let $\mathcal{K} \in \mathfrak{L}^q$ be regularly implementable through c. Let $\mathcal{L}_{\text{full}}(\mathcal{K})$ be the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{K}$ through w. Let $\mathcal{C} \in \mathfrak{L}^k$. Then the following two statements are equivalent:*

1. *$\mathcal{C}$ regularly implements $\mathcal{K}$ by interconnection through c.*
2. *$\mathcal{C}$ regularly implements $(\mathcal{L}_{\text{full}}(\mathcal{K}))_c$ via full interconnection w.r.t. $(\mathcal{P}_{\text{full}})_c$.*

Proof. Let $\mathcal{P}_{\text{full}} = \ker(R_1 \;\; R_2)$ be a minimal representation. Let V be unimodular such that $VR_1 = \text{col}(R_{11}, 0)$ with R_{11} full row rank. Partition $VR_2 = \text{col}(R_{12}, R_{22})$. Then $R_{22}c = 0$ is a minimal representation of $(\mathcal{P}_{\text{full}})_c$. $\mathcal{K}$ is implementable, so (using Lemma 19), there exists a polynomial matrix, say C^* (with C^* any polynomial matrix such that $\mathcal{C}^* = \ker(C^*)$ with $\mathcal{C}^*$ given by (4)), such that

$$\begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & C^* \end{pmatrix} \begin{pmatrix} w \\ c \end{pmatrix} = 0$$

is a kernel representation of $\mathfrak{L}_{\text{full}}(\mathcal{K})$. Hence a kernel representation of $(\mathfrak{L}_{\text{full}}(\mathcal{K}))_c$ is given by

$$\begin{pmatrix} R_{22} \\ C^* \end{pmatrix} c = 0. \tag{5}$$

$(1 \Rightarrow 2)$ Assume that $\mathcal{C}$ regularly implements $\mathcal{K}$; i.e., $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w = \mathcal{K}$ and the interconnection is regular. Let $\mathcal{C} = \ker(C)$ be a minimal representation of $\mathcal{C}$. Then the polynomial matrix

$$\begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & C \end{pmatrix} \tag{6}$$

has full row rank and (by Lemma 19) represents $\mathfrak{L}_{\text{full}}(\mathcal{K})$. It is then immediate that $(\mathcal{P}_{\text{full}})_c \cap \mathcal{C} = (\mathfrak{L}_{\text{full}}(\mathcal{K}))_c$. Obviously, $\operatorname{col}(R_{22}, C)$ has full row rank, so the full interconnection of $(\mathcal{P}_{\text{full}})_c$ and $\mathcal{C}$ is regular.

$(2 \Rightarrow 1)$ Conversely, assume that $\mathcal{C}$ regularly implements $(\mathfrak{L}_{\text{full}}(\mathcal{K}))_c$ w.r.t. $(\mathcal{P}_{\text{full}})_c$ by full interconnection, and that $\mathcal{C} = \ker(C)$ is a minimal representation. Then $\binom{R_{22}}{C} c = 0$ is a minimal kernel representation of $(\mathfrak{L}_{\text{full}}(\mathcal{K}))_c$. Since (5) is also a kernel representation, it is easily seen that $\mathcal{K}_{\text{full}}(\mathcal{C}) = \mathfrak{L}_{\text{full}}(\mathcal{K})$, which, by Lemma 16, implies $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w = (\mathfrak{L}_{\text{full}}(\mathcal{K}))_w = \mathcal{K}$. In addition, (6) has full row rank so the interconnection is regular. ∎

Finally, we arrive at the main result of this section. We establish, for the case that in $\mathcal{P}_{\text{full}}$ c is observable from w, a parametrization of all controllers that regularly implement a given $\mathcal{K}$ w.r.t. $\mathcal{P}_{\text{full}}$. The idea is to compute representations of $(\mathcal{P}_{\text{full}})_c$ and $(\mathfrak{L}_{\text{full}}(\mathcal{K}))_c$ and to parametrize all controllers that regularly implement $(\mathfrak{L}_{\text{full}}(\mathcal{K}))_c$ by full interconnection w.r.t. $(\mathcal{P}_{\text{full}})_c$ using Theorem 11. Then, by Theorem 20, this yields a parametrization of all controllers that regularly implemement $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$.

THEOREM 21. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$ with system variable (w, c) and c observable from w. Let $\mathcal{P}_{\text{full}} = \ker(R_1 \;\; R_2)$ be a minimal representation. Let $\mathcal{K} \in \mathfrak{L}^q$ be regularly implementable through c w.r.t. $\mathcal{P}_{\text{full}}$. Let $\mathcal{K} = \ker(K)$ be a minimal representation. Construct polynomial matrices V_1, V_2, F_1, and W as follows:*

1. *Let V_2 be a full row rank MLA of R_1.*
2. *Choose V_1 such that $\operatorname{col}(V_1, V_2)$ is unimodular.*
3. *Let M be an MRA of $V_2 R_2$ with $M(\lambda)$ full column rank for all λ.*
4. *Let F_1 be such that $K = F_1 V_1 R_1$.*
5. *Let Q be a full row rank MLA of $F_1 V_1 R_2 M$.*
6. *Choose W such that $\operatorname{col}(Q, W)$ is unimodular.*

Then for any $\mathcal{C} \in \mathfrak{L}^k$ with $\mathcal{C} = \ker(C)$, the following statements are equivalent:

1. *$\mathcal{C} = \ker(C)$ is a minimal representation and $\mathcal{C}$ regularly implements $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$.*
2. *There exist a polynomial matrix G and a unimodular U such that*

$$C = (UWF_1V_1 + GV_2)R_2.$$

Proof. V_2 is a full row rank MLA of R_1, so the unimodular matrix $V = \operatorname{col}(V_1, V_2)$ satisfies $VR_1 = \operatorname{col}(V_1R_1, 0)$ with V_1R_1 full row rank. Also $VR_2 = \operatorname{col}(V_1R_2, V_2R_2)$. Obviously, $(\mathcal{P}_{\text{full}})_c = \ker(V_2R_2)$ is a minimal representation, and the hidden behavior $\mathcal{N}$ is represented by $\ker(V_1R_1)$. Let F_1 be such that $K = F_1V_1R_1$ (such F_1 exists since

$\mathcal{N} \subseteq \mathcal{K}$). Then the controller $\ker(F_1V_1R_2)$ implements $\mathcal{K}$, so by Lemma 19 $\mathcal{L}_{\text{full}}(\mathcal{K})$ is represented by

$$\begin{pmatrix} V_1R_1 & V_1R_2 \\ 0 & V_2R_2 \\ 0 & F_1V_1R_2 \end{pmatrix} \begin{pmatrix} w \\ c \end{pmatrix} = 0;$$

thus $(\mathcal{L}_{\text{full}}(\mathcal{K}))_c = \ker\left(\begin{smallmatrix} V_2R_2 \\ F_1V_1R_2 \end{smallmatrix}\right)$. Now, M is an MRA of V_2R_2 so

$$\begin{pmatrix} V_2R_2 \\ F_1V_1R_2 \end{pmatrix} M = \begin{pmatrix} 0 \\ F_1V_1R_2M \end{pmatrix}.$$

Since Q is a full row rank MLA of $F_1V_1R_2M$, clearly $\left(\begin{smallmatrix} I & 0 \\ 0 & Q \end{smallmatrix}\right)$ is a full row rank MLA of $\left(\begin{smallmatrix} V_2R_2 \\ F_1V_1R_2M \end{smallmatrix}\right)$. If we choose W such that $\operatorname{col}(Q, W)$ is unimodular, then the matrix

$$\begin{pmatrix} I & 0 \\ 0 & Q \\ 0 & W \end{pmatrix}$$

is unimodular. By applying Theorem 11, a parametrization of all controllers that regularly implement $(\mathcal{L}_{\text{full}}(\mathcal{K}))_c$ by full interconnection w.r.t. $(\mathcal{P}_{\text{full}})_c$ is then given by $C = GV_2R_2 + U(0 \quad W)\left(\begin{smallmatrix} V_2R_2 \\ F_1V_1R_2 \end{smallmatrix}\right) = (GV_2 + UWF_1V_1)R_2$, where G ranges over all polynomial matrices and U ranges over all unimodular polynomial matrices, of course of suitable dimensions. Finally, by Theorem 20, the same parametrization holds for all controllers $\ker(C)$ that regularly implement $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$. ∎

6. Parametrization of all regular controllers: The nonobservable case. We now treat the nonobservable case. Consider the system $\mathcal{P}_{\text{full}}$ represented by $R_1w + R_2c = 0$. We no longer assume that c is observable from w, but show that the general case can be reduced to the observable case. This reduction requires two steps. First, we reduce the general case to the case that R_2 has full column rank, and next reduce the latter to the case that $R_2(\lambda)$ has full column rank for all λ, i.e., the observable case.

1. *Reduction to the case that R_2 has full column rank.* Let V be a unimodular matrix such that $R_2 = \begin{pmatrix} \tilde{R}_2 & 0 \end{pmatrix} V$, with $\tilde{R}_2$ full column rank k'. Define $\mathcal{P}'_{\text{full}} \in \mathfrak{L}^{q+k'}$ as the system (with control variable c') represented by $R_1w + \tilde{R}_2c' = 0$.
2. *Reduction to the observable case.* Assume now that in $\mathcal{P}_{\text{full}}$ the matrix R_2 has full column rank. Let L be a square, nonsingular polynomial matrix such that $R_2 = \tilde{R}_2L$, with $\tilde{R}_2(\lambda)$ full column rank for all $\lambda \in \mathbb{C}$. Using the Smith form of R_2 it is easily seen that this is always possible. Define $\mathcal{P}'_{\text{full}}$ as the system (with control variable c') represented by $R_1w + \tilde{R}_2c' = 0$. In the system $\mathcal{P}'_{\text{full}}$, c' is observable from w.

As it will turn out, in both reduction steps, $\mathcal{K} \in \mathfrak{L}^q$ is regularly implementable through c w.r.t. $\mathcal{P}_{\text{full}}$ if and only if it is regularly implementable through c' w.r.t. $\mathcal{P}'_{\text{full}}$. Also, every controller that regularly implements $\mathcal{K}$ w.r.t. $\mathcal{P}'_{\text{full}}$ will turn out to lead to a *set* of controllers that implement $\mathcal{K}$ w.r.t. $\mathcal{P}_{\text{full}}$. In the following two subsections, we will treat the two reduction steps separately.

6.1. Reduction to the case that R_2 has full column rank. In this subsection the parametrization problem for the original plant $\mathcal{P}_{\text{full}}$ is reduced to the parametrization problem for a plant $\mathcal{P}'_{\text{full}}$ in which the R_2-matrix has full column rank. In the following, let V be a unimodular matrix such that $R_2 = (\tilde{R}_2 \quad 0)V$, with $\tilde{R}_2$ full column rank $k' = \text{rank}(R_2)$. Define $\mathcal{P}'_{\text{full}}$ as the system represented by $R_1 w + \tilde{R}_2 c' = 0$.

THEOREM 22. *Let $\mathcal{K} \in \mathfrak{L}^q$. Then $\mathcal{K}$ is regularly implementable through c w.r.t. $\mathcal{P}_{\text{full}}$ if and only if $\mathcal{K}$ is regularly implementable through c' w.r.t. $\mathcal{P}'_{\text{full}}$. Let $\mathcal{C} \in \mathfrak{L}^k$, with $\mathcal{C} = \ker(C)$ a minimal representation. Then the following two statements are equivalent:*

1. *The controller $\mathcal{C}$ regularly implements $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$.*
2. *There exist a polynomial matrix C_{11}, polynomial matrices C_{12} and C_{21} of full row rank, and a unimodular matrix U such that*

$$C = U \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & 0 \end{pmatrix} V \tag{7}$$

and such that the controller $\mathcal{C}_{21} = \ker(C_{21})$ regularly implements $\mathcal{K}$ through c' w.r.t. $\mathcal{P}'_{\text{full}}$.

Proof. We first prove the equivalence of statements 1 and 2. Partition $CV^{-1} = (C_1 \quad C_2)$ with the number of columns of C_1 equal to $k' = \text{rank}(R_2)$. Choose a unimodular matrix U such that $U^{-1}C_2 = \text{col}(C_{12}, 0)$ with C_{12} full row rank. Partition $U^{-1}C_1 = \text{col}(C_{11}, C_{21})$. Then we have

$$\mathcal{K}_{\text{full}}(\mathcal{C}) = \{(w, V^{-1}c') \mid (w, c') \in \ker(M)\}, \text{ where } M = \begin{pmatrix} R_1 & \tilde{R}_2 & 0 \\ 0 & C_{11} & C_{12} \\ 0 & C_{21} & 0 \end{pmatrix}.$$

Define $\mathcal{K}'_{\text{full}}(\mathcal{C}_{21}) := \{(w, c_1) \mid R_1 w + \tilde{R}_2 c_1 = 0 \text{ and } C_{21} c_1 = 0\}$, the full controlled behavior of $\mathcal{P}'_{\text{full}}$ using the controller $\mathcal{C}_{21} := \ker(C_{21})$. Using the full row rank of C_{12} we then have $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w = (\ker(M))_w = (\mathcal{K}'_{\text{full}}(\mathcal{C}_{21}))_w$. From this we conclude that $\mathcal{C}$ implements $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$ if and only if $\mathcal{C}_{21}$ implements $\mathcal{K}$ through c' w.r.t. $\mathcal{P}'_{\text{full}}$. Furthermore, again by full row rank of C_{12}, regularity of either of the interconnections implies the same for the other one.

Finally, the statement that $\mathcal{K}$ is regularly implementable through c w.r.t. $\mathcal{P}_{\text{full}}$ if and only if $\mathcal{K}$ is regularly implementable through c' w.r.t. $\mathcal{P}'_{\text{full}}$ follows immediately from the equivalence of statements 1 and 2. ∎

6.2. Reduction to the observable case. In the previous subsection it was shown that our parametrization problem can be reduced to a problem for a plant behavior with R_2-matrix of full row rank. In the present subsection we reduce the full column rank case to the observable case. Let $\mathcal{P}_{\text{full}} = \ker(R_1 \quad R_2)$ be a minimal representation, with R_2 full column rank. Let L be square and nonsingular such that $R_2 = \tilde{R}_2 L$, with $\tilde{R}_2(\lambda)$ full column rank for all λ. Let $\mathcal{P}'_{\text{full}}$ be the (observable) system represented by $R_1 w + \tilde{R}_2 c' = 0$.

THEOREM 23. *Let $\mathcal{K} \in \mathfrak{L}^q$. Then $\mathcal{K}$ is regularly implementable through c w.r.t. $\mathcal{P}_{\text{full}}$ if and only if $\mathcal{K}$ is regularly implementable through c' w.r.t. $\mathcal{P}'_{\text{full}}$. Let $\mathcal{C} \in \mathfrak{L}^k$ with $\mathcal{C} = \ker(C)$ a minimal representation. Then the following two statements are equivalent:*

1. *The controller $\mathcal{C}$ regularly implements $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$.*

2. *The controller* $\mathcal{C}'$ *represented in latent variable representation (with latent variable* ℓ*) by*

$$\begin{pmatrix} I \\ 0 \end{pmatrix} c' = \begin{pmatrix} L \\ C \end{pmatrix} \ell \tag{8}$$

regularly implements $\mathcal{K}$ *through* c' *w.r.t.* $\mathcal{P}'_{\text{full}}$.

Proof. We first prove the equivalence of statements 1 and 2. The manifest controlled behavior resulting from the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{C}$ is equal to

$$(\mathcal{K}_{\text{full}}(\mathcal{C}))_w = \{w \mid \text{ there exists } c \text{ such that } R_1 w + \tilde{R}_2 L c = 0,\ Cc = 0\},$$

which, since L is nonsingular, equals

$$\{w \mid \text{ there exists } c', c \text{ such that } R_1 w + \tilde{R}_2 c' = 0,\ c' = Lc,\ Cc = 0\}.$$

The latter is equal to $(\mathcal{K}'_{\text{full}}(\mathcal{C}'))_w$, the manifest controlled behavior resulting from the interconnection of $\mathcal{P}'_{\text{full}}$ and $\mathcal{C}'$. Thus, $\mathcal{C}$ implements $\mathcal{K}$ w.r.t. $\mathcal{P}_{\text{full}}$ if and only if $\mathcal{C}'$ implements $\mathcal{K}$ w.r.t. $\mathcal{P}'_{\text{full}}$. Next, we prove that the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{C}$ is regular if and only if the interconnection of $\mathcal{P}'_{\text{full}}$ and $\mathcal{C}'$ is regular. Note that $\mathcal{K}'_{\text{full}}(\mathcal{C}')$ has latent variable representation

$$\begin{pmatrix} R_1 & \tilde{R}_2 \\ 0 & I \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ c' \end{pmatrix} = \begin{pmatrix} 0 \\ L \\ C \end{pmatrix} \ell.$$

Hence the output cardinality of $\mathcal{K}'_{\text{full}}(\mathcal{C}')$ equals

$$\mathfrak{p}(\mathcal{K}'_{\text{full}}(\mathcal{C}')) = \operatorname{rank} \begin{pmatrix} R_1 & \tilde{R}_2 & 0 \\ 0 & I & L \\ 0 & 0 & C \end{pmatrix} - \operatorname{rank} \begin{pmatrix} 0 \\ L \\ C \end{pmatrix}$$

(see [2, Lemma 8]). Using elementary row and column operations and the fact that L is nonsingular, this can be shown to be equal to $\operatorname{rank} \left(\begin{smallmatrix} R_1 & R_2 \\ 0 & C \end{smallmatrix}\right)$, which equals $\mathfrak{p}(\mathcal{K}_{\text{full}}(\mathcal{C}))$. Also, $\mathfrak{p}(\mathcal{C}') = \operatorname{rank} \left(\begin{smallmatrix} I & L \\ 0 & C \end{smallmatrix}\right) - \operatorname{rank} \left(\begin{smallmatrix} L \\ C \end{smallmatrix}\right) = \operatorname{rank}(C) = \mathfrak{p}(\mathcal{C})$. Finally, $\mathfrak{p}(\mathcal{P}_{\text{full}}) = \operatorname{rank}(R_1 \ \ R_2) = \operatorname{rank}(R_1 \ \ \tilde{R}_2) = \mathfrak{p}(\mathcal{P}'_{\text{full}})$. This proves our claim.

Again, the first statement of the theorem follows immediately from the equivalence of statements 1 and 2. ∎

According to this theorem, the controller $\mathcal{C} = \ker(C)$ works for $\mathcal{P}_{\text{full}}$ if and only if the controller $L\ker(C)$ (with control variable c') works for the observable system $\mathcal{P}'_{\text{full}}$. What we are looking for here is a parametrization of *all* such polynomial matrices C. The next theorem reduces the parametrization of these C's to the parametrization of all polynomial matrices C' such that $\ker(C')$ regularly implements $\mathcal{K}$ through c' w.r.t. $\mathcal{P}'_{\text{full}}$, which was already established in Theorem 21.

THEOREM 24. *Let* $\mathcal{P}_{\text{full}} = \ker(R_1 \ \ R_2)$ *be a minimal representation, with* R_2 *full column rank. Let* L *be square and nonsingular such that* $R_2 = \tilde{R}_2 L$*, with* $\tilde{R}_2(\lambda)$ *full column rank for all* λ*. Let* $\mathcal{P}'_{\text{full}}$ *be the (observable) system represented by* $R_1 w + \tilde{R}_2 c' = 0$*. Let* $\mathcal{C} \in \mathfrak{L}^k$ *with minimal representation* $\mathcal{C} = \ker(C)$*. Then for every* $\mathcal{K} \in \mathfrak{L}^q$ *that is regularly implementable through* c *w.r.t.* $\mathcal{P}_{\text{full}}$ *the following two statements are equivalent:*

1. *The controller* $\mathcal{C}$ *regularly implements* $\mathcal{K}$ *through* c *w.r.t.* $\mathcal{P}_{\text{full}}$.

2. *There exists a square, nonsingular polynomial matrix X and a full row rank polynomial matrix C' such that $C = X^{-1}C'L$, where $(C'(\lambda) \;\; X(\lambda))$ has full row rank for all $\lambda \in \mathbb{C}$ and the controller $\mathcal{C}' = \ker(C')$ regularly implements $\mathcal{K}$ through c' w.r.t. $\mathcal{P}'_{\text{full}}$.*

Proof. $(1 \Rightarrow 2)$ Let $\mathcal{C} = \ker(C)$ regularly implement $\mathcal{K}$ through c w.r.t. $\mathcal{P}_{\text{full}}$. By Theorem 23, the controller $L\ker(C)$ regularly implements $\mathcal{K}$ through c' w.r.t. $\mathcal{P}'_{\text{full}}$. Let

$$\begin{pmatrix} A & B \\ Y & X \end{pmatrix} \begin{pmatrix} L \\ C \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix}, \tag{9}$$

where the leftmost matrix is unimodular and D has full row rank. Obviously, this can always be done. We claim that $L\ker(C) = \ker(Y)$. Indeed, $(c' \in L\ker(C)) \Leftrightarrow$ (there exists ℓ such that $c' = L\ell$ and $C\ell = 0$) $\Leftrightarrow$ (there exists ℓ such that $\left(\begin{smallmatrix} A & B \\ Y & X \end{smallmatrix}\right)\left(\begin{smallmatrix} c' \\ 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} D \\ 0 \end{smallmatrix}\right)\ell$) $\Leftrightarrow (Yc' = 0)$. The last equivalence follows from the fact that D has full row rank, so it induces a surjective differential operator. Next, we prove that Y has full row rank. Let p be a polynomial row vector such that $pY = 0$. Since $YL + XC = 0$ and C has full row rank, we obtain $pX = 0$. Since $(Y \;\; X)$ has full row rank, this yields $p = 0$. In the same way it can be proved that X has full row rank.

Now define $C' := -Y$. Then the controller $\mathcal{C}' = \ker(C')$ regularly implements $\mathcal{K}$ through c' w.r.t. $\mathcal{P}'_{\text{full}}$. Of course, coming from a unimodular matrix, $(C'(\lambda) \;\; X(\lambda))$ has full row rank for all $\lambda \in \mathbb{C}$. We show that X is nonsingular. Since it has full row rank, it suffices to show that it is square. This follows immediately from (9):

$$\operatorname{rowdim}(X) = \operatorname{rowdim}(C) + \operatorname{rowdim}(L) - \operatorname{rank}(\operatorname{col}(L, C)).$$

Since L is nonsingular, $\operatorname{rank}(\operatorname{col}(L, C)) = \operatorname{rank}(L)$, so $\operatorname{rowdim}(X) = \operatorname{rowdim}(C)$. Of course, also $\operatorname{coldim}(X) = \operatorname{rowdim}(C)$, so X is square. Finally, we have $C = X^{-1}C'L$.

$(2 \Rightarrow 1)$ We will prove that $L\ker(C) = \ker(C')$. The implication will then follow from Theorem 23. Clearly, $-C'L + XC = 0$. Let A and B be polynomial matrices such that $\left(\begin{smallmatrix} A & B \\ -C' & X \end{smallmatrix}\right)$ is unimodular. Let D be defined by

$$\begin{pmatrix} A & B \\ -C' & X \end{pmatrix} \begin{pmatrix} L \\ C \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix}.$$

By the same argument as was used in the first part of this proof, it suffices to prove that D has full row rank. Indeed, using the fact that A is square, we have that $\operatorname{rank}(D) = \operatorname{rank}(\operatorname{col}(D, 0)) = \operatorname{rank}(\operatorname{col}(L, C)) = \operatorname{rank}(L) = \operatorname{rowdim}(L) = \operatorname{coldim}(A) = \operatorname{rowdim}(A) = \operatorname{rowdim}(D)$. ∎

Thus, for any given full row rank C' that works for the observable system $\mathcal{P}'_{\text{full}}$, a *set* of polynomial matrices C that work for $\mathcal{P}_{\text{full}}$ is obtained by dividing $C'L$ by those nonsingular polynomial matrices X that have the properties that $(C'(\lambda) \;\; X(\lambda))$ has full row rank for all λ, and the quotient $X^{-1}C'L$ is a polynomial matrix again.

7. All stabilizing controllers: The partial interconnection case. In this section we return to the stabilization problem. Whereas in Theorem 12 we gave a parametrization in the full interconnection case, we now solve the problem of parametrizing, for a given plant $\mathcal{P}_{\text{full}} = \ker(R_1 \;\; R_2)$, all stabilizing controllers (see section 3, Definition 6) for the case of partial interconnection. This is done along the same lines as the parametrization of all regularly implementing controllers: we first establish a parametrization under the condition that in $\mathcal{P}_{\text{full}}$ c is observable from

w. Then we lift the assumption and treat the general case. As already mentioned in section 3, necessary and sufficient conditions for the existence of a stabilizing controller for $\mathcal{P}_{\text{full}}$ are that $(\mathcal{P}_{\text{full}})_w$ is stabilizable and that w is detectable from c (see Proposition 7). For the observable case the following lemma is instrumental.

LEMMA 25. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$ with system variable (w, c). Assume that c is observable from w. Assume that $(\mathcal{P}_{\text{full}})_w$ is stabilizable and that w is detectable from c. Let $\mathcal{C} \in \mathfrak{L}^k$. Then the following two statements are equivalent:*

1. *$\mathcal{C}$ stabilizes $\mathcal{P}_{\text{full}}$ through c.*
2. *$\mathcal{C}$ stabilizes $(\mathcal{P}_{\text{full}})_c$ by full interconnection.*

Proof. $(1 \Rightarrow 2)$ $\mathcal{K} := (\mathcal{K}_{\text{full}}(\mathcal{C}))_w$ is stable and the interconnection is regular. Let $\mathcal{L}_{\text{full}}(\mathcal{K})$ be the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{K}$ through w. Then, by Lemma 19, $\mathcal{K}_{\text{full}}(\mathcal{C}) = \mathcal{L}_{\text{full}}(\mathcal{K})$. According to Theorem 20, $\mathcal{C}$ regularly implements $(\mathcal{L}_{\text{full}}(\mathcal{K}))_c$ by full interconnection with $(\mathcal{P}_{\text{full}})_c$. We claim that $(\mathcal{L}_{\text{full}}(\mathcal{K}))_c$ is stable. Indeed, let $c \in (\mathcal{L}_{\text{full}}(\mathcal{K}))_c$. There exists w such that $(w, c) \in \mathcal{L}_{\text{full}}(\mathcal{K}) \subseteq \mathcal{P}_{\text{full}}$. Let R_2^+ be a polynomial left-inverse of R_2. Then we have $c = -R_2^+ R_1 w$. Hence $c(t) \to 0$ $(t \to \infty)$ (note that the components of w are products of polynomials and stable exponentials).

$(2 \Rightarrow 1)$ Let $\mathcal{P}_{\text{full}} = \ker\left(\begin{smallmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{smallmatrix}\right)$ be a minimal representation, with R_{11} full row rank (see also the proof of Theorem 20). Then $(\mathcal{P}_{\text{full}})_c = \ker(R_{22})$ is a minimal representation. Represent $\mathcal{C} = \ker(C)$ minimally. Then $\operatorname{col}(R_{22}, C)$ is Hurwitz. Using this, together with the fact that R_{11} has full row rank, it is immediate that the interconnection of $\mathcal{P}_{\text{full}}$ and $\mathcal{C}$ is regular. We now prove that $(\mathcal{K}_{\text{full}}(\mathcal{C}))_w$ is stable. Let $w \in (\mathcal{K}_{\text{full}}(\mathcal{C}))_w$. There exists c such that $(w, c) \in \mathcal{K}_{\text{full}}(\mathcal{C})$ so $R_{11}w + R_{12}c = 0$ and $\left(\begin{smallmatrix} R_{22} \\ C \end{smallmatrix}\right)c = 0$. Thus, the components of c are products of polynomials and stable exponentials. Since w is detectable from c, $R_{11}(\lambda)$ has full column rank for all $\lambda \in \mathbb{C}^+$. This implies that $w(t) \to 0$ $(t \to \infty)$. ∎

The following theorem then gives a parametrization of all stabilizing controllers for the observable case.

COROLLARY 26. *Let $\mathcal{P}_{\text{full}} \in \mathfrak{L}^{q+k}$ satisfy the assumptions of Lemma 25. Let $\mathcal{P}_{\text{full}} = \ker(R_1 \;\; R_2)$ be a minimal representation. Construct polynomial matrices V_2, S, and C_0 as follows:*

1. *Let V_2 be a full row rank MLA of R_1.*
2. *Factorize $V_2 R_2 = TS$ with T square, nonsingular and $S(\lambda)$ full row rank for all $\lambda \in \mathbb{C}$.*
3. *Let C_0 be such that $\operatorname{col}(S, C_0)$ is unimodular.*

Then for any $\mathcal{C} \in \mathfrak{L}^k$ with $\mathcal{C} = \ker(C)$ the following statements are equivalent:

1. *$\mathcal{C}$ stabilizes $\mathcal{P}_{\text{full}}$ through c and the representation $\mathcal{C} = \ker(C)$ is minimal.*
2. *There exist a polynomial matrix F and a Hurwitz polynomial matrix D such that $C = FS + DC_0$.*

Proof. This is an immediate Corollary of Theorem 12 and Lemma 25. ∎

Thus we have obtained a parametrization of all stabilizing controllers for the observable case. In order to arrive at a parametrization for the general case, we can perform the same two reduction steps as in section 6. We will describe both steps separately now; the proofs are left to the reader.

The first step concerns the reduction of a general $\mathcal{P}_{\text{full}}$ to a full plant behavior $\mathcal{P}'_{\text{full}}$ with R_2-matrix full column rank. Let V be a unimodular matrix such that $R_2 = (\tilde{R}_2 \;\; 0)V$, with $\tilde{R}_2$ full column rank. Let $\mathcal{P}'_{\text{full}}$ be represented by $R_1 w + \tilde{R}_2 c' = 0$.

COROLLARY 27. *$(\mathcal{P}_{\text{full}})_w$ is stabilizable if and only if $(\mathcal{P}'_{\text{full}})_w$ is stabilizable, and in $\mathcal{P}_{\text{full}}$, w is detectable from c if and only in $\mathcal{P}'_{\text{full}}$, w is detectable from c'. Furthermore,*

if $\mathcal{C} \in \mathfrak{L}^k$ with minimal representation $\mathcal{C} = \ker(C)$, then the following two statements are equivalent:

1. *The controller $\mathcal{C}$ stabilizes $\mathcal{P}_{\text{full}}$ through c.*
2. *There exist a polynomial matrix C_{11}, polynomial matrices C_{12} and C_{21} of full row rank, and a unimodular matrix U such that*

$$C = U \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & 0 \end{pmatrix} V, \tag{10}$$

and such that the controller $\mathcal{C}_{21} = \ker(C_{21})$ stabilizes $\mathcal{P}'_{\text{full}}$ through c'.

The next step concerns the reduction of a full plant behavior $\mathcal{P}_{\text{full}}$ with full column rank R_2-matrix to a behavior $\mathcal{P}'_{\text{full}}$ in which the control variable c' is observable from w. Let L be square, nonsingular, such that $R_2 = L\tilde{R}_2$, with $\tilde{R}_2(\lambda)$ full column rank for all λ. Let $\mathcal{P}'_{\text{full}}$ be represented by $R_1 w + \tilde{R}_2 c' = 0$.

COROLLARY 28. *$(\mathcal{P}_{\text{full}})_w$ is stabilizable if and only if $(\mathcal{P}'_{\text{full}})_w$ is stabilizable, and in $\mathcal{P}_{\text{full}}$, w is detectable from c if and only if in $\mathcal{P}'_{\text{full}}$, w is detectable from c'. Furthermore, for $\mathcal{C} \in \mathfrak{L}^k$ with minimal representation $\mathcal{C} = \ker(C)$, the following two statements are equivalent:*

1. *The controller $\mathcal{C}$ stabilizes $\mathcal{P}_{\text{full}}$ through c.*
2. *There exist a square, nonsingular polynomial matrix X and a full row rank polynomial matrix C' such that $C = X^{-1}C'L$, where $(C'(\lambda) \quad X(\lambda))$ has full row rank for all $\lambda \in \mathbb{C}$ and the controller $\mathcal{C}' = \ker(C')$ stabilizes $\mathcal{P}'_{\text{full}}$ through c'.*

8. Worked-out examples. In order to illustrate the theory developed in this paper, we now present some worked-out examples. The examples are those that were already presented in the problem formulation in section 3.

Example 29. Let $\mathcal{P}_{\text{full}}$ with manifest variable $w = (w_1, w_2)$ and control variable $c = (c_1, c_2)$ be represented by

$$\begin{aligned} w_1 + \dot{w}_2 + \dot{c}_1 + c_2 &= 0, \\ c_1 + c_2 &= 0. \end{aligned}$$

Clearly, $(\mathcal{P}_{\text{full}})_w = \mathfrak{C}^\infty(\mathbb{R}, \mathbb{R}^2)$. For $\mathcal{K}$ take the behavior represented by $w_1 + \dot{w}_2 = 0$. $\mathcal{K}$ is regularly implementable through (c_1, c_2) w.r.t. $\mathcal{P}_{\text{full}}$. We have

$$R_1(\xi) = \begin{pmatrix} 1 & \xi \\ 0 & 0 \end{pmatrix} \text{ and } R_2(\xi) = \begin{pmatrix} \xi & 1 \\ 1 & 1 \end{pmatrix}.$$

R_2 has full column rank. Factorize $R_2 = \tilde{R}_2 L$ with $\tilde{R}_2 = I_2$, the 2×2 identity matrix, and $L = R_2$. The resulting system $\mathcal{P}'_{\text{full}}$ represented by $R_2 w + \tilde{R}_2 c' = 0$ is observable. We first parametrize all controllers that regularly implement $\mathcal{K}$ w.r.t. $\mathcal{P}'_{\text{full}}$. For this, we perform the steps described in Theorem 21: $V_2 = (0, 1)$, $V_1 = (1, 0)$, $V_2\tilde{R}_2 = (0, 1)$, so $M = \text{col}(1, 0)$. Next, $V_1 R_1 = K = (1, \xi)$, so $F_1 = 1$. We have $F_1 V_1 \tilde{R}_2 M = 1$. Thus Q, as full row rank MLA of 1, is void. We take $W = 1$. A parametrization of all full row rank controller representations C' that regularly implement $\mathcal{K}$ w.r.t. $\mathcal{P}'_{\text{full}}$ is given by $C'(\xi) = (u, g(\xi))$, with $0 \neq u \in \mathbb{R}$ and g an arbitrary polynomial with real coefficients.

Next we parametrize all controllers C that regularly implement $\mathcal{K}$ w.r.t. the original full plant behavior $\mathcal{P}_{\text{full}}$. According to Theorem 24, for any choice of $u \neq 0$ and polynomial g, we should find all nonzero polynomials $x(\xi)$ that divide $C'L =$

$(u\xi + g(\xi), u + g(\xi))$ such that $(x(\lambda), u, g(\lambda)) \neq 0$ for all $\lambda \in \mathbb{C}$. Since $u \neq 0$, this constraint is automatically satisfied. Thus we need only compute all common factors $x(\xi)$ of the polynomials $u\xi + g(\xi)$ and $u + g(\xi)$. If $x(\xi)$ is such a common factor, then it must also divide the difference $u(\xi - 1)$. Hence there are two possibilities:

1. $g(1) \neq -u$. In this case $u\xi + g(\xi)$ and $u + g(\xi)$ are coprime. The only common factor is then $x(\xi) = 1$.
2. $g(1) = -u$. In this case $x(\xi) = \xi - 1$ is the only common factor.

Thus we find that a parametrization of all controllers that regularly implement $\mathcal{K}$ for $\mathcal{P}_{\text{full}}$ is given by $C(\xi) = (u\xi + g(\xi), u + g(\xi))$, $u \neq 0$ and g arbitrary polynomial, or $C(\xi) = (\frac{u\xi+g(\xi)}{\xi-1}, \frac{u+g(\xi)}{\xi-1})$, $u \neq 0$ and g arbitrary polynomial such that $g(1) = -u$. Since $g(1) = -u$ if and only if there exists a polynomial h such that $g(\xi) = -u + h(\xi)(\xi - 1)$, the latter is equivalent to $C(\xi) = (u + h(\xi), h(\xi))$, $u \neq 0$ and h arbitrary polynomial.

Example 30. Consider the full plant behavior $\mathcal{P}_{\text{full}}$ represented by

$$\begin{aligned} w_1 + \dot{w}_2 + \dot{c}_1 + c_2 &= 0, \\ w_2 + c_1 + c_2 &= 0, \\ \dot{c}_1 + c_1 + \dot{c}_2 + c_2 &= 0. \end{aligned}$$

We will parametrize all controllers $C(\frac{\mathrm{d}}{\mathrm{d}t})c = 0$ that stabilize $\mathcal{P}_{\text{full}}$ through c. We have

$$R_1 = \begin{pmatrix} 1 & \xi \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \; R_2 = \begin{pmatrix} \xi & 1 \\ 1 & 1 \\ \xi + 1 & \xi + 1 \end{pmatrix}, \; \tilde{R}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & \xi + 1 \end{pmatrix}, \; L = \begin{pmatrix} \xi & 1 \\ 1 & 1 \end{pmatrix}.$$

In $\mathcal{P}'_{\text{full}}$, represented by $R_1 w + \tilde{R}_2 c' = 0$, c' is observable from w. We first parametrize all controllers $C'(\frac{\mathrm{d}}{\mathrm{d}t})c' = 0$ that stabilize $\mathcal{P}'_{\text{full}}$. Performing the steps of Corollary 26, we obtain $V_2 = (0, 0, 1)$, $V_2\tilde{R}_2 = TS$ with $T(\xi) = \xi + 1$ and $S = (0, 1)$. Choose $C_0 = (1, 0)$. The required parametrization is then $C'(\xi) = (d(\xi), f(\xi))$ with d an arbitrary Hurwitz polynomial and f an arbitrary polynomial. We compute $C'(\xi)L(\xi) = (\xi d(\xi) + f(\xi), d(\xi) + f(\xi))$. A parametrization for the original plant $\mathcal{P}_{\text{full}}$ is obtained by computing, for any choice of d and f, all nonzero common factors $x(\xi)$ of the polynomials $\xi d(\xi) + f(\xi)$ and $d(\xi) + f(\xi)$ with the property that $(x(\lambda), d(\lambda), f(\lambda)) \neq 0$ for all λ. Let d and f be given, with d Hurwitz. If $x(\xi)$ is a common factor, then it is also a common factor of $(\xi - 1)d(\xi)$. Hence the following possibilities occur:

1. $x(\xi) = c$, constant, unequal to zero. These $x(\xi)$'s satisfy the requirements.
2. $x(\xi) = c(\xi - 1)$, with $c \neq 0$, equivalently, $d(1) + f(1) = 0$. Since d is Hurwitz, $d(1) \neq 0$, so we have $(x(1), d(1), f(1)) \neq 0$, and the rank condition holds.
3. $x(\xi)$ is not constant and divides $d(\xi)$. In this case there is λ such that $x(\lambda) = 0$ and $d(\lambda) = 0$. However, then also $f(\lambda) = 0$, violating the rank condition.

By applying Corollary 28, we conclude that a parametrization of all stabilizing controllers for $\mathcal{P}_{\text{full}}$ is given by $C(\xi) = (\xi d(\xi) + f(\xi), d(\xi) + f(\xi))$, with d Hurwitz polynomial, f arbitrary polynomial, or $C(\xi) = \frac{1}{\xi-1}(\xi d(\xi) + f(\xi), d(\xi) + f(\xi))$, with d Hurwitz polynomial and f polynomial such that $d(1) + f(1) = 0$.

Acknowledgment. The authors wish to thank one of the anonymous referees for contributions that have led to major improvements of parts of this paper, in particular of section 6.2.

REFERENCES

[1] M. N. Belur, *Control in a Behavioral Context*, Doctoral dissertation, University of Groningen, The Netherlands, 2003.

[2] M. N. Belur and H. L. Trentelman, *Stabilization, pole placement and regular implementability*, IEEE Trans. Automat. Control, 47 (2002), pp. 735–744.

[3] A. Agung Julius, J. C. Willems, M. N. Belur, and H. L. Trentelman, *The canonical controllers and regular interconnection*, Systems Control Lett., 54 (2005), pp. 787–797.

[4] M. Kuijper, *Why do stabilizing controllers stabilize?*, Automatica J. IFAC, 31 (1995), pp. 621–625.

[5] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, Berlin, 1997.

[6] P. Rocha, *Canonical controllers and regular implementation of nD behaviors*, in Proceedings of the 16th IFAC World Congress, Prague, 2005.

[7] P. Rocha and J. Wood, *Trajectory control and interconnection of* 1*D and nD systems*, SIAM J. Control Optim., 40 (2001), pp. 107–134.

[8] A. J. van der Schaft, *Achievable behavior of general systems*, Systems Control Lett., 49 (2003), pp. 141–149.

[9] M. Vidyasagar, *Control System Synthesis, A Factorization Approach*, The MIT Press, Cambridge, MA, 1985.

[10] J. C. Willems, *On interconnections, control and feedback*, IEEE Trans. Automat. Control, 42 (1997), pp. 326–339.

[11] J. C. Willems and H. L. Trentelman, *Synthesis of dissipative systems using quadratic differential forms*. I, IEEE Trans. Automat. Control, 47 (2002), pp. 53–69.

[12] D. C. Youla, H. A. Jabr, and J. J. Bongiorno, *Modern Wiener-Hopf design of optimal controllers*. II. *The multivariable case*, IEEE Trans. Automat. Control, 21 (1976), pp. 319–338.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2054–2064

ON THE INTERSECTION OF A CLARKE CONE WITH A BOLTYANSKII CONE*

ALBERTO BRESSAN†

Abstract. We provide an example of two closed sets $S_1, S_2 \subset \mathbb{R}^4$ such that $S_1 \cap S_2 = \{0\}$. Yet, at the origin, a Boltyanskii tangent cone C_1 to S_1 and the Clarke tangent cone C_2 to S_2 are strongly transversal. This settles a question originally proposed by H. Sussmann.

Key words. optimal control, necessary conditions, tangent cones

AMS subject classifications. 49K15, 49K05

DOI. 10.1137/060654852

1. Introduction.

Let $t \mapsto x^*(t)$ and $t \mapsto u^*(t)$ be, respectively, a trajectory and an optimal control for the Mayer problem

$$\begin{aligned} \text{minimize} \qquad & \varphi\big(x(T)\big) \\ \text{subject to} \qquad & \dot{x}(t) = f\big(t,\, x(t),\, u(t)\big) \\ \text{and} \qquad & x(0) = \bar{x}\,, \qquad x(T) \in S\,. \end{aligned}$$

Here $x \in \mathbb{R}^m$ is the state variable, while $u \in U \subseteq \mathbb{R}^m$ is the control variable.

A set of necessary conditions for optimality of the pair (x^*, u^*) is provided by the celebrated Pontryagin maximum principle (PMP), which provides one of the cornerstones of the mathematical theory of control [6]. In essence, these necessary conditions can be related to a separation property: On one hand, we have a set S_1 of "reachable points," i.e., points $x\big(T, u\big)$ that can be reached at the terminal time T by means of admissible controls $t \mapsto u(t) \in U$. On the other hand, we can consider a set S_2 of "profitable points," i.e., points that satisfy the terminal constraint $x \in S$ and achieve a lower cost $\varphi(x) \leq \varphi\big(x^*(T)\big)$, with equality holding only if $x = x^*(T)$.

Necessary conditions for optimality are then obtained by constructing a tangent cone C_1 to the set S_1 at the terminal point $x^*(T)$, and a tangent cone C_2 to S_2, also at $x^*(T)$. If these cones are transversal, under suitable assumptions one can conclude that the intersection $S_1 \cap S_2$ is nontrivial; i.e., it contains points other than $x^*(T)$. Hence, the pair (x^*, u^*) is not optimal. Reversing the argument, the optimality of the trajectory-control pair (x^*, u^*) implies that the tangent cones C_1 and C_2 are weakly separated. This provides an alternative way to state the PMP.

Subsequent work on the PMP has been aimed at deriving high order necessary conditions [4, 5] and at weakening the regularity assumptions [1, 2, 3, 7, 8, 9, 10, 11]. At present, two essentially different versions of the PMP can be found in the literature: one is based on the use of "Clarke tangent cones" [1, 2, 3], while the other relies on the concept of "Boltyanskii approximating cones" [7, 8, 9, 10, 11]. The precise definitions of these cones will be recalled in section 2. The following statements on the nontrivial

*Received by the editors March 22, 2006; accepted for publication (in revised form) August 30, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sicon/45-6/65485.html
†Department of Mathematics, Penn State University, University Park, PA 16802 (bressan@math.psu.edu).

intersection are well known in the literature and provide the key ingredients in the two approaches.

(I) Assume $\bar{x} \in S_1 \cap S_2$ and let C_1 and C_2 be the Clarke tangent cones to S_1 and S_2, respectively, at the point $\bar{x}$. If C_1 and C_2 are strongly transversal, then the sets S_1 and S_2 have a nontrivial intersection. Indeed, $\bar{x}$ belongs to the closure of the set $(S_1 \cap S_2) \setminus \{\bar{x}\}$.

(II) Assume $\bar{x} \in S_1 \cap S_2$ and let C_1 and C_2 be Boltyanskii approximating cones to S_1 and S_2, respectively, at the point $\bar{x}$. If C_1 and C_2 are strongly transversal, then the sets S_1 and S_2 have a nontrivial intersection. Indeed, $\bar{x}$ belongs to the closure of the set $(S_1 \cap S_2) \setminus \{\bar{x}\}$.

It is interesting to speculate whether these two approaches can be combined into a single, more general PMP. To achieve this goal, one should introduce a new definition of "generalized tangent cone" such that the following hold:

(i) The Clarke tangent cone to a closed set S and all Boltyanskii approximating cones to S at a given point $\bar{x} \in S$ are "generalized tangent cones" in the sense of this new definition.

(ii) If two closed sets S_1 and S_2 have "generalized tangent cones" C_1, C_2 at a point $\bar{x}$, and if these two cones are strongly transversal, then the sets themselves have a nontrivial intersection. Namely, $\bar{x}$ belongs to the closure of $(S_1 \cap S_2) \setminus \{\bar{x}\}$.

In particular, if such a notion of generalized tangent cone did exist, the following question posed by H. Sussmann (see Conjecture 3.6.4 in [9]) should have a positive answer.

Let $n \geq 1$ and let S_1, S_2 be closed subsets of $\mathbb{R}^n$ such that $0 \in S_1 \cap S_2$. Assume that

- *C_1 is a Boltyanskii approximating cone to S_1 at 0;*
- *C_2 is the Clarke tangent cone to S_2 at 0;*
- *C_1 and C_2 are strongly transversal.*

Does this imply that 0 belongs to the closure of $(S_1 \cap S_2) \setminus \{0\}$?

In the present paper, however, we prove that this question has a negative answer in every space dimension $n \geq 4$. For the general theory of necessary conditions, our result implies that there are at least two noncomparable maximum principles, which cannot be unified by means of any new concept of tangent cone.

The remainder of the paper is organized as follows. In section 2 we review some basic definitions and discuss an example in space dimension $n = 3$. Our main counterexample is then given in section 3, where we construct two closed sets $S_1, S_2 \subset R^4$. At the origin, the Clarke tangent cone to S_1 and a Boltyanskii tangent cone to S_2 are strongly transversal. Yet, the two sets have a trivial intersection, namely, $S_1 \cap S_2 = \{0\}$.

Based on the present counterexample, the forthcoming paper by Sussmann [12] exhibits an optimal control problem and an optimal trajectory for which the usual conclusions of the PMP cannot be achieved if the Clarke tangent cone to the terminal set is used instead of a Boltyanskii approximating cone.

2. Preliminary analysis. For the reader's convenience, we first recall some basic definitions.

DEFINITION 1. *A nonempty set $C \subseteq \mathbb{R}^n$ is a cone if, whenever $v \in C$ and $r \geq 0$, it follows that $rv \in C$.*

DEFINITION 2. *Let $S \subseteq \mathbb{R}^n$. The Clarke tangent cone to a point $x \in S$ is the set of all vectors $v \in \mathbb{R}^n$ such that the following holds. For every sequence $x_k \to x$ with*

$x_k \in S$ for all $k \geq 1$, there exist a sequence $v_k \to v$ such that

$$\liminf_{h \to 0+} \frac{d\big(x_k + hv_k\,,\ S\big)}{h} = 0\,.$$

DEFINITION 3. *Given a set $S \subseteq \mathbb{R}^n$ and a point $x \in S$, we say that a closed convex cone $C \subseteq \mathbb{R}^n$ is a Boltyanskii approximating cone to S at x if there exists a continuous map $F : C \mapsto S$ such that*

$$\lim_{v \to 0,\, v \in C} \frac{F(v) - x - v}{|v|} = 0\,.$$

DEFINITION 4. *Two convex cones $C_1, C_2 \subseteq \mathbb{R}^n$ are strongly transversal if $C_1 \cap C_2 \neq \{0\}$ and, moreover,*

$$C_1 - C_2 \doteq \big\{v_1 - v_2\,;\quad v_1 \in C_1\,,\quad v_2 \in C_2\big\} = \mathbb{R}^n\,.$$

Before giving a counterexample in dimension $n = 4$, it is useful to see what can happen in dimension $n = 3$.

Consider $\mathbb{R}^3$ with coordinates t, x, y. Let $\phi : \mathbb{R} \mapsto [0, 1]$ be a $\mathcal{C}^\infty$ function such that

$$\begin{array}{ll} \phi(\xi) > 0 & \text{if}\ \ 0 < \xi < 1\,, \\ \phi(\xi) = 0 & \text{if}\ \ \xi \leq 0\ \ \text{or}\ \ \xi \geq 1\,. \end{array} \tag{1}$$

Consider a sequence of times t_m decreasing to zero, say $t_m = m^{-1}$. Moreover, choose a positive sequence ε_m decreasing to zero and a sequence of integers $N_m \to \infty$ such that

$$\frac{\varepsilon_m}{t_m} \to 0\,, \qquad\qquad \frac{\varepsilon_m}{t_{m-1} - t_m} \to 0\,, \tag{2}$$

$$t_{m-1} - t_m < \frac{t_m}{2N_m}\,. \tag{3}$$

We wish to construct two sets S_1, S_2 whose tangent cones at the origin will be

$$\begin{aligned} C_1 &= \big\{(t, x, y)\,;\quad t \geq 0,\quad x = 0\,,\quad |y| \leq t\,\big\}\,, \\ C_2 &= \big\{(t, x, y)\,;\quad t \geq 0,\quad y = 0\,,\quad |x| \leq t\big\}\,. \end{aligned}$$

To define S_2 we proceed as follows. For each $m \geq 1$ we divide the segment $[-t_m,\ t_m]$ into N_m equal parts. This is achieved by setting

$$x_{m,k} = -t_m + \frac{2kt_m}{N_m}$$

so that

$$-t_m = x_{m,0} < x_{m,1} < \cdots < x_{m,N_m} = t_m\,.$$

For $t \in [t_m,\ t_{m-1}]$ and $j = 1, \ldots, N_m$, define the spatial interval

$$I_{m,j}(t) \doteq \big[x_{m,j-1} - (t - t_m),\quad x_{m,j} + (t - t_m)\big]\,.$$

Notice that each of these closed intervals becomes wider as time increases. However, if the time interval $[t_m,\ t_{m-1}]$ is sufficiently short, so that (3) holds, then disjoint

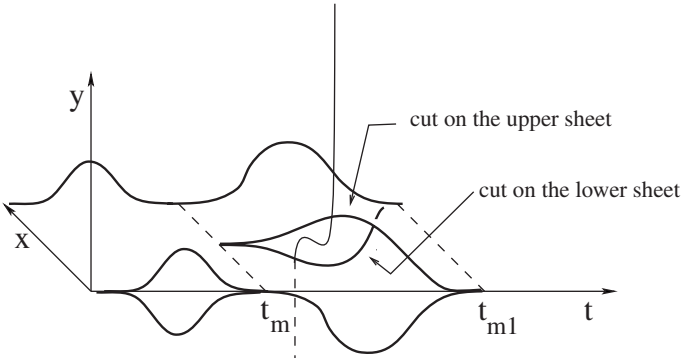


FIG. 1. *For $t \neq t_m$, a nearly vertical thread can pass through both sheets.*

intervals will not overlap for $t \in [t_m,\ t_{m-1}]$. More precisely

$$I_{m,j}(t) \cap I_{m,\ell}(t) = \emptyset \qquad \text{whenever} \quad \ell \geq j+2\,.$$

As S_2 we take the set

$$S_2 \doteq \bigcup_{m\geq 1} \left[\left\{ (t,x,y)\,;\ t\in[t_m,\ t_{m-1}]\,,\ x\in I_{m,j}\,,\ y = \varepsilon_m \phi\Big(\frac{t-t_m}{t_{m-1}-t_m}\Big) \quad j \text{ even} \right\} \right.$$
$$\left. \cup \left\{ (t,x,y)\,;\ t\in[t_m,\ t_{m-1}]\,,\ x\in I_{m,j}\,,\ y = -\varepsilon_m \phi\Big(\frac{t-t_m}{t_{m-1}-t_m}\Big) \quad j \text{ odd} \right\} \right].$$

If the numbers ε_m converge to zero fast enough, so that (2) holds, then the Clarke tangent cone to S_2 to the origin is precisely C_2.

We now consider whether is it possible to slightly deform the cone C_1, by means of a map

$$\varphi : (t,0,y) \mapsto \big(t,\ x(t,y), y\big)$$

with

$$\big|x(t,y)\big| = o\big(|t|\big)\,,$$

so that the image

$$S_1 \doteq \varphi(C_1)$$

intersects S_2 only at the origin.

Of course this is not possible. Indeed, at each time $t = t_m$ (with m sufficiently large) there must be an intersection. However, we notice that we could choose φ so that intersections occur only in arbitrarily small neighborhoods of the points $(t_m, 0, 0)$. Indeed, consider a time $\tau \notin \{t_1, t_2, t_3, \ldots\}$, say with $t_m < \tau < t_{m-1}$. The intersection of S_2 with the plane $\{t = \tau\}$ then consists of two sets of segments, at heights

$$y = \pm\varepsilon_m\, \phi\Big(\frac{\tau - t_m}{t_{m-1}-t_m}\Big).$$

One can thus deform a vertical line on the same plane so that no intersection occurs (Figure 1).

The construction performed in the previous example amounted to the creation of "pockets," which could be used by a vertical thread in order to cross both sheets

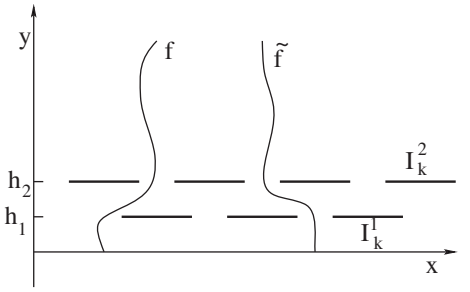


FIG. 2. *The functions $x = f(y)$ and $x = \tilde{f}(y)$ are not homotopic (avoiding the segments).*

without intersection. In dimension $n = 3$ this can be only partially successful, since the crossings at times $t = t_m$ ($m \geq 1$) cannot be avoided.

Performing a similar construction in dimension $n = 4$, however, one can completely remove any intersection (except at the origin), thus answering in the negative the question posed by Sussmann. The crucial difference between dimensions $n = 3$ and $n = 4$ is illustrated in the following remarks, which we regard as obvious.

Remark 1. Fix $\delta > \delta' > 0$ and $h_1 < h_2$. On the x-y plane, consider two countable sets of disjoint segments, all of length $2\delta'$:

$$I^1_k \doteq \Big\{(x,y)\,;\quad y = h_1\,,\quad \big|x - (2k)\delta\big| \leq \delta'\Big\}\,,$$
$$I^2_k \doteq \Big\{(x,y)\,;\quad y = h_2\,,\quad \big|x - (2k+1)\delta\big| \leq \delta'\Big\}\,.$$

If $x = f(y)$, $x = \tilde{f}(y)$ are two continuous functions whose graphs do not intersect any of the segments I^i_k, in general it is not possible to construct a homotopy preserving this property (Figure 2). In other words, there exists no continuous map $F : [0,1] \times \mathbb{R} \mapsto \mathbb{R}$ such that

$$F(0,x) = f(x)\,, \qquad\qquad F(1,x) = \tilde{f}(x),$$

and the graph of F does not intersect any of the segments I^i_k.

Remark 2. Fix $\delta > \delta' > 0$ and $h_1 < h_2 < h_3 < h_4$. In x-y-z space, consider four countable sets of disjoint squares, parallel to the x-y plane, located at four different heights $z = h_i$. These will all have side length $2\delta'$ and centers at points in the lattice $\delta\,\mathbb{Z}^2$:

$$Q^1_{jk} \doteq \Big\{(x,y,z)\,;\quad z = h_1\,,\quad \big|x - (2j)\delta\big| \leq \delta'\,,\quad \big|y - (2k)\delta\big| \leq \delta'\Big\}\,,$$
$$Q^2_{jk} \doteq \Big\{(x,y,z)\,;\quad z = h_2\,,\quad \big|x - (2j+1)\delta\big| \leq \delta'\,,\quad \big|y - (2k)\delta\big| \leq \delta'\Big\}\,,$$
$$Q^3_{jk} \doteq \Big\{(x,y,z)\,;\quad z = h_3\,,\quad \big|x - (2j)\delta\big| \leq \delta'\,,\quad \big|y - (2k+1)\delta\big| \leq \delta'\Big\}\,,$$
$$Q^4_{jk} \doteq \Big\{(x,y,z)\,;\quad z = h_4\,,\quad \big|x - (2j+1)\delta\big| \leq \delta'\,,\quad \big|y - (2k+1)\delta\big| \leq \delta'\Big\}\,.$$

Let now $(x,y) = f(z)$ and $(x,y) = \tilde{f}(z)$ be two continuous functions, whose graphs do not intersect any of the squares Q^i_{jk}. Then they are homotopic (Figure 3), i.e. we can now construct a continuous map $F : [0,1] \times \mathbb{R} \mapsto \mathbb{R}^2$ such that

$$F(0,z) = f(z)\,, \qquad\qquad F(1,z) = \tilde{f}(z),$$

and the graph of F does not intersect any of the squares Q^i_{jk}.

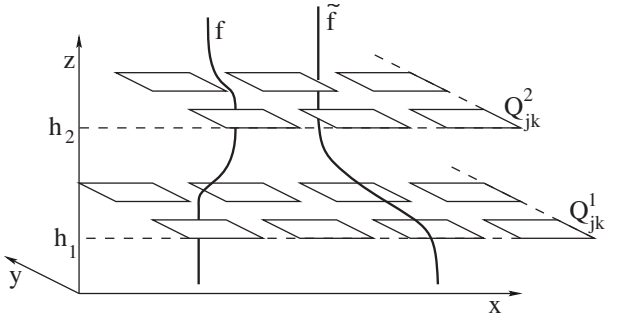


FIG. 3. *The functions $(x, y) = f(z)$ and $(x, y) = \tilde{f}(z)$ are homotopic (avoiding the squares).*

3. The four-dimensional counterexample. In the space $\mathbb{R}^4$ with coordinates (t, x, y, z) we will construct two closed sets S_1, S_2, having the origin as their only point in common. A Boltyanskii tangent cone to S_1 at the origin will be

$$C_1 \doteq \big\{(t, x, y, z)\,;\quad t \geq 0\,,\quad |z| \leq t\,,\quad x = y = 0\big\}\,,$$

while the Clarke tangent cone to S_2 at the origin will be

$$C_2 \doteq \big\{(t, x, y, z)\,;\quad t \geq 0\,,\quad z = 0\big\}\,.$$

The strong transversality of the two cones is thus clear.

We begin by constructing the set S_2. Define the decreasing sequence of times

$$t_m \doteq \frac{1}{m}\,. \tag{4}$$

Each time slice

$$S_2(\tau) \doteq S_2 \cap \{t = \tau\}$$

will be a subset of the x-y-z space $\mathbb{R}^3$ consisting of a finite union of square patches.

In addition to the function ϕ in (1), let $\tilde{\phi} : \mathbb{R} \mapsto [0, 1]$ be a $\mathcal{C}^\infty$ function such that

$$\begin{cases} \tilde{\phi}(\xi) > 0 & \text{if} \quad -1 < \xi < 1\,, \\ \tilde{\phi}(\xi) = 0 & \text{if} \quad \xi \leq -1 \quad \text{or} \quad \xi \geq 1\,. \end{cases} \tag{5}$$

Moreover, define the function

$$\delta(t, x) = t^6 \cdot \tilde{\phi}\Big(\frac{x}{t^{4/3}} - 2\cos\frac{\pi}{t}\Big)\,. \tag{6}$$

This will determine the separation in height between the four layers of squares. Notice that

$$\delta > 0 \qquad \Longleftrightarrow \qquad \Big|x - 2t^{4/3}\cos\frac{\pi}{t}\Big| < t^{4/3}. \tag{7}$$

Moreover,

$$\cos\frac{\pi}{t_m} = \cos m\pi = \begin{cases} 1 & \text{if} \quad m \text{ is even}, \\ -1 & \text{if} \quad m \text{ is odd}. \end{cases}$$

The region where $\delta > 0$ is depicted in Figure 4.

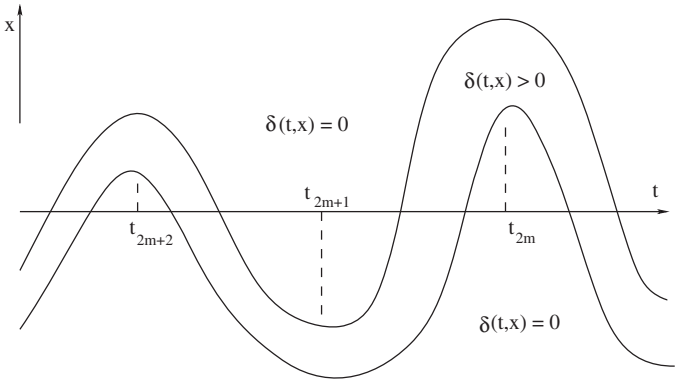


FIG. 4. *The region in the t-x plane where the sheets are separated.*

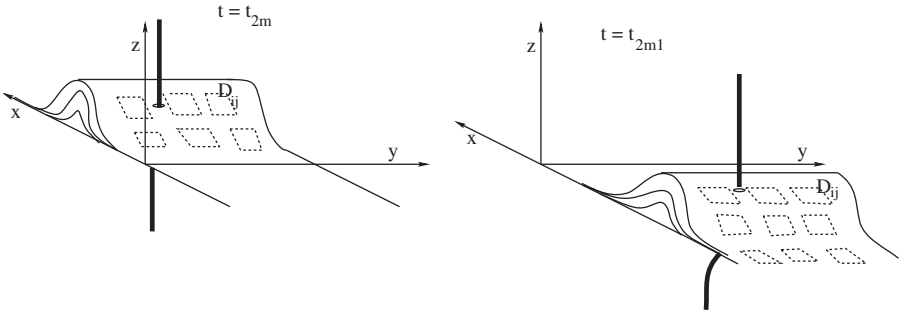


FIG. 5. *A thread crossing four sheets of fabric, without ever touching the square patches on them.*

Before writing down lengthy formulas, we explain the underlying idea in plain words.

KNITTING PROBLEM. *Consider four sheets of fabric spread over a horizontal table, the x-y plane (see Figure* 5*). At a given time t, the heights of the four sheets are given by the smooth functions*

$$z = i \cdot \delta(t, x), \qquad i = 0, 1, 2, 3,$$

where $\delta(t, x)$ *is as in* (6)*. On each of these sheets we single out finitely many square patches:* $A_{ij}(t)$ *on the first sheet,* $B_{ij}(t)$ *on the second,* C_{ij} *on the third, and* D_{ij} *on the fourth. Any two distinct* A_{ij} *do not intersect, and the same is true for* B_{ij}*,* C_{ij}*, and* D_{ij}*.*

We now want to pass an (almost vertical) thread through the four sheets of cloth, moving the thread continuously in time, without ever touching any of the square patches A_{ij}*,* B_{ij}*,* C_{ij}*, and* D_{ij}*. The crossing cannot take place on regions where the four sheets stick together (i.e., where* $\delta(t, x) = 0$*), because the vertical projections of all these patches cover the x-y plane. However, we can easily pass our thread through the four sheets, as long as the crossing takes place over a region of the x-y plane where the four sheets are separated, i.e., where* $\delta(t, x) > 0$*.*

We now specify details. At every time t, the projection of the set S_2^τ on the x-y plane is the entire plane. In x-y-z space, the time slice $S_2(\tau)$ consists of countably many square patches. The ones having centers at points with $x > 0$ will be treated separately from the ones having centers at points with $x < 0$.

(i) *Squares with centers with $x > 0$.* To define these squares, assume $t_{2m+1} \le \tau \le t_{2m-1}$. Divide the upper half-plane $\{(x,y)\,;\quad x \ge 0\}$ into squares of equal size, each with side of length t_{2m+1}/N_{2m+1}. The centers of these squares are located at the points

$$(x_i, y_j) \doteq \left(\frac{(i-\frac{1}{2})t_{2m+1}}{N_{2m+1}}\ ,\ \frac{(j-\frac{1}{2})t_{2m+1}}{N_{2m+1}} \right), \qquad i \ge 1\,,\quad j \in \mathbb{Z}\,.$$

We split these squares into four classes, depending on whether i, j are even or odd. For $\tau \in [t_{2m+1},\ t_{2m-1}]$ we then define

$$\begin{aligned}
A_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad & |x-x_i| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\\
& |y-y_j| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\quad z = 0\Big\}\,,\\
B_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad & |x-x_i| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\\
& |y-y_j| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\quad z = \delta(\tau,x)\Big\}\,,\\
C_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad & |x-x_i| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\\
& |y-y_j| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\quad z = 2\delta(\tau,x)\Big\}\,,\\
D_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad & |x-x_i| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\\
& |y-y_j| \le \frac{t_{2m+1}}{2N_{2m+1}} + (\tau - t_{2m+1})\,,\quad z = 3\delta(\tau,x)\Big\}\,,
\end{aligned}$$

and then

$$S_2^{(\mathrm{upper})}(\tau) \doteq \left(\bigcup_{\substack{i\ \text{even}\\ j\ \text{even}}} A_{ij}(\tau) \right) \cup \left(\bigcup_{\substack{i\ \text{odd}\\ j\ \text{even}}} B_{ij}(\tau) \right) \cup \left(\bigcup_{\substack{i\ \text{even}\\ j\ \text{odd}}} C_{ij}(\tau) \right) \cup \left(\bigcup_{\substack{i\ \text{odd}\\ j\ \text{odd}}} D_{ij}(\tau) \right).$$

We stress that here the union ranges over all couples of indices (i,j) with $i \ge 1$.

(ii) *Squares with centers with $x < 0$.* To define these squares, we now assume $\tau \in [t_{2m},\ t_{2m-2}]$. We divide the lower half-plane $\{(x,y)\,;\quad x \le 0\}$ into squares of equal size, each with side of length t_{2m}/N_{2m}. The centers of these squares are located at the points

$$(x_i, y_j) \doteq \left(\frac{(i-\frac{1}{2})t_{2m}}{N_{2m}}\ ,\ \frac{(j-\frac{1}{2})t_{2m}}{N_{2m}} \right), \qquad i \le 0\,,\quad j \in \mathbb{Z}\,.$$

These will not be confused with the previous ones because now $i \le 0$. We define

$$\begin{aligned}
A_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad & |x-x_i| \le \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,\\
& |y-y_j| \le \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,\quad z = 0\Big\}\,,
\end{aligned}$$

$$B_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad |x-x_i| \leq \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,$$
$$|y-y_j| \leq \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,\quad z = \delta(\tau,x)\Big\}\,,$$
$$C_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad |x-x_i| \leq \frac{t_{2m+1}}{2N_{2m}} + (\tau - t_{2m})\,,$$
$$|y-y_j| \leq \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,\quad z = 2\delta(\tau,x)\Big\}\,,$$
$$D_{ij}(\tau) = \Big\{(\tau,x,y,z)\,;\quad |x-x_i| \leq \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,$$
$$|y-y_j| \leq \frac{t_{2m}}{2N_{2m}} + (\tau - t_{2m})\,,\quad z = 3\delta(\tau,x)\Big\}$$

and then

$$S_2^{(\text{lower})}(\tau) \doteq \left(\bigcup_{\substack{i\ \text{even}\\ j\ \text{even}}} A_{ij}(\tau)\right) \cup \left(\bigcup_{\substack{i\ \text{odd}\\ j\ \text{even}}} B_{ij}(\tau)\right) \cup \left(\bigcup_{\substack{i\ \text{even}\\ j\ \text{odd}}} C_{ij}(\tau)\right) \cup \left(\bigcup_{\substack{i\ \text{odd}\\ j\ \text{odd}}} D_{ij}(\tau)\right).$$

In this case, the union ranges over all couples of indices (i,j) with $i \leq 0$.

Finally, we define

$$S_2(0) \doteq \Big\{(0,x,y,0)\,;\quad x,y \in \mathbb{R}\Big\}\,,$$
$$S_2(\tau) \doteq S_2^{(\text{upper})}(\tau) \cup S_2^{(\text{lower})}(\tau) \qquad \text{for}\quad \tau > 0\,,$$
$$S_2 \doteq \bigcup_{\tau \geq 0} S_2(\tau)\,.$$

Next, we claim that there exists a continuous function $\varphi : C_1 \mapsto \mathbb{R}^2$, say $(x,y) = \varphi(t,0,0,z)$, such that

$$\varphi(t,0,0,z) = o\big(|t|\big) \qquad \text{as } t \to 0\,,\quad (t,0,0,z) \in C_1\,,$$

and, moreover, the graph

$$S_1 \doteq \Big\{(t,x,y,z)\,;\quad (x,y) = \varphi(t,0,0,z)\,,\quad (t,0,0,z) \in C_1\Big\}$$

intersects the set S_2 only at the origin.

Intuitively, at each time $\tau > 0$ the one-dimensional set $S_1^\tau \doteq S_1 \cap \{t = \tau\}$ gives the position of the thread, which should not touch any of the square patches $A_{ij}(\tau)$, $B_{ij}(\tau)$, $C_{ij}(\tau)$, $D_{ij}(\tau)$ contained in $S_2(\tau)$.

In view of Remark 2, such a continuous function φ exists provided that

• the width of the region where $\delta(\tau,\cdot) > 0$ is much larger than the sides of the single paths $A_{ij}(\tau)$, $B_{ij}(\tau)$, $C_{ij}(\tau)$, $D_{ij}(\tau)$.

• square patches on the same sheet remain disjoint, i.e.,

$$\text{[length of a side]} < \text{[distance between two centers]}.$$

Since patches grow in time at unit rate, this last condition is a consequence of the inequalities

$$t_{2m-1} - t_{2m+1} < \frac{t_{2m-1}}{5N_{2m-1}}\,,\qquad t_{2m-2} - t_{2m} < \frac{t_{2m}}{5N_{2m}} \qquad (m \geq 1). \tag{8}$$

Finally, in order for the set S_2 to have C_2 as a Clarke tangent cone at the origin, we need

$$\frac{\partial}{\partial t}\,\delta(t,x)\to 0 \qquad \frac{\partial}{\partial x}\,\delta(t,x)\to 0 \qquad \text{as } \ t\to 0. \tag{9}$$

On the other hand, in order for the set S_1 to have C_1 as a Boltyanskii tangent cone at the origin, we need the following:

• The maximum size of the patches $A_{ij}(t)$, $B_{ij}(t)$, $C_{ij}(t)$, $D_{ij}(t)$ is $o(t)$ as $t\to 0$. Moreover,

$$\sup\ \big\{|x|\,;\ \ \delta(t,x)>0\big\}=o(t) \qquad \text{as } t\to 0.$$

It is not difficult to make choices such that all the above conditions are satisfied. Indeed,

$$t_{m-2}-t_m=\mathcal{O}(t_m^2)=o(t_m^{2-\varepsilon}).$$

Therefore, if we choose N_m so that

$$[\text{size of patches at time } t_m]\approx\frac{t_m}{N_m}\approx t_m^{3/2},$$

then conditions (8) will certainly hold. To fix the ideas, we will choose

$$\begin{cases} N_m \ \doteq m^{1/2} & \text{if } m \text{ is even}\,,\\ N_m \ \doteq 9m^{1/2} & \text{if } m \text{ is odd}\,. \end{cases} \tag{10}$$

By these choices, as $t\to 0$,

• the sizes of the patches are $\mathcal{O}(t^{3/2})$.

• the region where $\delta(t,\cdot)>0$, where the crossing can occur, is at a distance of $O(t^{4/3})$ from the x-axis.

Hence, we can construct the function $(x,y)=\varphi(t,z)$ such that

$$\big|\varphi(t,z)\big|=\mathcal{O}(t^{4/3}).$$

This guarantees that C_1 is indeed a Boltyanskii tangent cone to S_1 at the origin.

Finally, the factor t^6 on the right-hand side of (6) guarantees that the derivatives of $\delta(\cdot,\cdot)$ w.r.t. both t and x vanish as $t\to 0$. Hence, C_2 is the Clarke tangent cone to S_2. This completes the analysis.

Remark 3. Some additional words of explanation might be helpful. With reference to Figure 5, what is going on is the following. At time $t=t_m$, the four sheets are sticking together in the half-plane where $x\le 0$, but are separated in some region where $x>0$, (Figure 5, left).

We split the lower half-plane $\{x\le 0\}$ into four classes of square patches, of size length $\ell(t_{2m})=t_{2m}/N_{2m}=\mathcal{O}(t_{2m}^{2/3})$. These patches keep sticking together for $t>t_{2m}$ until δ becomes positive.

At time $t=t_{2m-1}$, all four sheets now coincide in the upper half-plane $\{x\ge 0\}$, but are separated in some region where $x<0$ (Figure 5, right). It is here that the thread can cross all sheets without intersecting any of the square patches.

The reader should notice that, on the interval $[t_{2m-1}\,,\ t_{2m-3}]$, sheets in the upper half-plane are split into square patches which are different from those defined during the previous time interval $[t_{2m+1}\,,\ t_{2m-1}]$. This different definition, however, does not introduce any discontinuity at the time $t=t_{2m-1}$. Indeed, since $\delta(t_{2m-1},x)=0$ for $x\ge 0$, all four sheets in the upper half-plane coincide. Therefore, it does not make

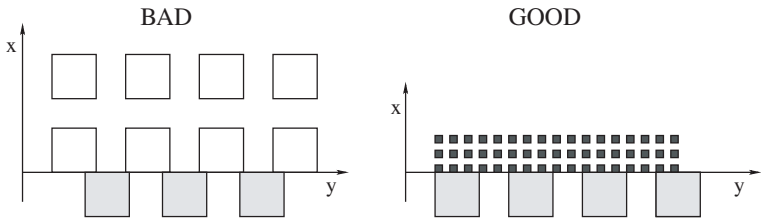


FIG. 6. *Choosing upper and lower squares of different sizes, we ensure that the complement set is connected.*

any difference to represent the set

$$S_2^{(\text{upper})}(t_{2m-1}) \doteq \Big\{(t_{2m-1}, x, y, z)\,;\quad x \geq 0\,,\quad z = 0\Big\}$$

as a union of squares of side length $\ell(t_{2m-1})$ rather than $\ell(t_{2m+1}) + 2(t_{2m-1} - t_{2m+1})$.

Remark 4. The factor "9" in (10) is motivated by a small technical detail. On the upper and lower half-planes $\{x > 0\}$ and $\{x < 0\}$ we usually have patches of different sizes, and centers on different lattice points, on one single sheet. Indeed, the sizes of upper squares are defined at the times t_{2m+1}, while the sizes of lower squares are determined at the times t_{2m}, for $m \geq 1$. In principle, on one of the sheets we may have a configuration as shown in Figure 6, left. This would prevent the thread from moving freely from the upper to the lower half-plane. This can be easily prevented by requiring that the lower squares be considerably larger that the upper ones, as in Figure 6, right.

REFERENCES

[1] F. H. CLARKE, *The maximum principle under minimal hypotheses*, SIAM J. Control Optim., 14 (1976), pp. 1078–1091.

[2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley Interscience, New York, 1983.

[3] F. H. CLARKE, *Necessary conditions in dynamic optimization*, Mem. Amer. Math. Soc., 173 (2005), no. 816.

[4] H. W. KNOBLOCH, *Higher Order Necessary Conditions in Optimal Control Theory*, Springer-Verlag, Berlin, 1981.

[5] A. J. KRENER, *The high order maximum principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.

[6] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.

[7] H. J. SUSSMANN, *A strong version of the maximum principle under weak hypotheses*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Orlando, FL, 1994, pp. 1950–1956.

[8] H. J. SUSSMANN, *A strong maximum principle for systems of differential inclusions*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 1809–1814.

[9] H. J. SUSSMANN, *The geometrization of the maximum principle*, paper written for the Proceedings of the COSY Workshop on ontrol of Nonlinear and Uncertain Systems, London, February 6-8, 1998 (unpublished; available online at http://www.math/rutgers.edu/∼sussman/).

[10] H. J. SUSSMANN, *New theories of set-valued differentials and new versions of the maximum principle of optimal control theory*, in Nonlinear Control in the Year 2000, Vol. 2, A. Isidori, F. Lamnabhi-Lagarrigue, and W. Respondek, eds., Springer, London, 2000, pp. 487–526.

[11] H. J. SUSSMANN, *Generalized differentials, variational generators, and the maximum principle with state constraints*, in Nonlinear and Optimal Control Theory, G. Stefani and P. Nistri, eds., C.I.M.E. Lecture Notes in Math., Springer-Verlag, Berlin, to appear.

[12] H. J. SUSSMANN, *On the validity of the transversality condition for different concepts of tangent cone to a set*, in Proceedings of the 4th IEEE Conference on Decision and Control, 2006, to appear.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2065–2093

ON THE INTERNAL MODEL STRUCTURE FOR INFINITE-DIMENSIONAL SYSTEMS: TWO COMMON CONTROLLER TYPES AND REPETITIVE CONTROL*

EERO IMMONEN†

Abstract. In this paper we shall extend and amplify the recent theory of controllers employing the so-called internal model structure. For exponentially stable closed loop control systems this structure has been shown in another paper to be necessary and sufficient for robust output regulation, also in infinite-dimensional spaces. Here we shall derive conditions under which two controller types occurring frequently in applications have the internal model structure. Under these conditions robust output regulation is achieved—also in infinite-dimensional spaces—if the closed loop system is sufficiently stable. In the case that the closed loop system is only strongly (but not exponentially) stable, it is sometimes possible to obtain conditional robustness. This means that asymptotic tracking/disturbance rejection is not destroyed by small perturbations so long as closed loop stability also persists. Our results allow for infinite-dimensional plants, controllers, and exogenous systems, and as an application of such a general setting we shall consider generalized repetitive control for exponentially stable infinite-dimensional SISO systems.

Key words. robust output regulation, infinite-dimensional linear systems, regulator equations, internal model principle, internal model structure

AMS subject classifications. 37N35, 47A48, 93B50, 93B51, 93B52, 93B55, 93C25

DOI. 10.1137/050638916

1. Introduction.

Although regulating the output of a given (linear time-invariant) plant is one of the central problems in control theory, this output regulation theory is rather useless in practice unless the controllers provide a degree of robustness. Robustness means fault tolerance; here we treat it so that small perturbations in some of the parameters of the plant, the controller, and the exogenous reference/disturbance signal generator should not affect closed loop stability and asymptotic tracking of the reference signals in the presence of disturbances. Such robust controllers have been subject to a considerable amount of research during the past three decades. The problem was solved for finite-dimensional systems by Davison [10], Davison and Goldenberg [11], Francis [13], Francis and Wonham [14], Sebakhy and Wonham [35], Wonham [40] and others in the 1970s. It seems that although many authors independently evolved their own design techniques, these methods can be quite generally grouped into two distinct categories: those employing estimates of the state of a coupled system containing the plant and the exosystem (see, e.g., [13]), and those employing dynamic error augmentation (see, e.g., [10]). Moreover, these two categories arise as special instances of the general robust output regulation paradigm for finite-dimensional systems: the celebrated internal model principle due to Francis and Wonham [14]. This principle asserts, roughly, that an error feedback controller which stabilizes the closed loop system also achieves robust output regulation if and only if the controller utilizes a suitably reduplicated copy of the maximal cyclic component of the exogenous system operator. For a survey of these (and related) results and the differences between the above two controller types the reader should consult

*Received by the editors August 25, 2005; accepted for publication (in revised form) July 6, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sicon/45-6/63891.html
†Institute of Mathematics, Tampere University of Technology, PL 553, 33101 Tampere, Finland (Eero.Immonen@tut.fi).

the article of Kwatny and Kalnitsky [27].

During the past few decades several authors have extended the theory cited above for infinite-dimensional plants and finite-dimensional exogenous systems. The finite-dimensional results of Davison [10] and Davison and Goldenberg [11] have been generalized by Pohjolainen [30, 31], Hämäläinen and Pohjolainen [17, 18], Ukai and Iwazumi [36], and others (see [19] for more detailed information). On the other hand, the finite-dimensional results of Francis [13] have been generalized to this setting, e.g., by Schumacher [34] and Byrnes et al. [6]. Interestingly enough, the existing infinite-dimensional generalizations of the theory of Francis do not seem to cover robustness issues. Moreover, while robustness does play a central role in some of the above generalizations of the results of Davison, none of these authors has studied the internal model principle for infinite-dimensional systems. In the author's view this is probably due to the fact that the principle, although easy to formulate in vague terms, is notoriously dependent on purely finite-dimensional concepts when formulated in precise terms [14].

In the papers cited above the exogenous signal generator is finite-dimensional. This means that the above theory is applicable in the case of simple reference/disturbance signals such as constants and sinusoids. However, the use of infinite-dimensional exogenous systems allows for the asymptotic tracking of arbitrary bounded uniformly continuous reference signals in the presence of disturbances of a similar type [4, 5, 21, 23, 24, 22]. The repetitive control scheme, which is a frequency domain–based error feedback control technique introduced by Hara et al. [15] (see also [39, 42, 43]) for the asymptotic tracking of periodic signals for finite-dimensional systems, furnishes a widespread example of the application of an infinite-dimensional exogenous system—the p-periodic signal generator. As a matter of fact, Yamamoto and Hara [43] have also proved an analogue of the internal model principle for repetitive control systems: stable asymptotic tracking necessitates the internal model $\frac{1}{1-e^{-sp}}$ of the exogenous signals in the closed loop system.

Only rather recently has the internal model principle been generalized for infinite-dimensional systems in the spirit of its original state space formulation [20]. This generalization (see section 3 below) employs purely operator theoretic concepts and is thus valid regardless of the dimensions of the state spaces of the plant, the controller, and the exogenous system. The core of this infinite-dimensional internal model principle lies in the observation that *any* closed loop error feedback control system, when properly stabilized, already contains the error zeroing dynamics. An appropriate choice of the controller's parameters resulting in the so-called internal model structure (see Definition 3.1) then realizes the desired dynamical behavior of output regulation. In particular, in finite dimensions, an error feedback controller containing a suitably controllable reduplication of the exogenous system dynamics has this internal model structure [20].

Whereas [20] presented a general robustness theory for infinite-dimensional systems without any regard to the choice of the controller's parameters, in this paper our purpose is to extend and amplify those results by establishing conditional robustness results for one member in each of the two aforementioned distinct regulator categories (for infinite-dimensional systems). Here conditional robustness means that asymptotic tracking and disturbance rejection are not destroyed by small perturbations so long as closed loop stability also persists; this is specified precisely in Definition 2.5. As a member of the first category we introduce (following [6, 22] closely) an infinite-dimensional generalization of a finite-dimensional controller of Francis [13]; we call this "type I controller" (see subsection 3.1 and in particular (3.13)). As a member of

the second category we introduce an infinite-dimensional generalization of the finite-dimensional controller (originally due to Davison [10]) studied in [16]; we call this "type II controller" (cf. subsection 3.2). We shall show that under certain assumptions both type I and type II controllers have the internal model structure. An application of the abstract results in [20] then immediately yields sufficient conditions for conditionally robust output regulation for these controllers. An important feature in our results is that they describe *how* these controllers achieve conditionally robust output regulation of bounded uniformly continuous exogenous signals for infinite-dimensional systems. Moreover, it turns out that we can design controllers achieving conditionally robust output regulation without relying on any purely finite-dimensional concepts, such as maximal cyclic components and rational canonical decompositions, so long as sufficient closed loop stability is achieved. To the author's knowledge this has not been possible before.

Although not all of our results are new—some finite-dimensional results of Davison and his coworkers have actually been generalized[1] for plants in the Callier–Desoer algebra [17] and subsequently for plants which are well-posed infinite-dimensional systems [32]—their presentation within the context of the infinite-dimensional internal model principle is new. Furthermore, our results explain in a novel way the function of the internal model principle in the robust output regulation of finite-dimensional systems. In particular, our framework provides a natural interpretation and generalization for the somewhat unintuitive (finite-dimensional) process of incorporating reduplications of the maximal cyclic component of the exosystem generator in the controller. On the other hand, the existing infinite-dimensional generalizations [6, 22, 34] of the theory of Francis [13] do not cover robustness issues; our results will cover robustness of the controllers in [6, 22] (type I).

One particularly interesting application for which the conditional robustness theory of this paper is suitable is repetitive control. In the classical repetitive control literature, e.g., [15, 42, 43], one invariably requires *exponential* stability of the closed loop system containing the plant and the infinite-dimensional internal model of the exogenous signals. Unfortunately this requirement can never be satisfied for any strictly proper finite-dimensional plant (see section V of [43] and Proposition 2 of [15]), so the scope of the classical repetitive control theory is somewhat limited. In this paper we shall apply the conditional robustness theory in a generalized repetitive control problem, where only *strong* closed loop stability and periodic tracking/rejection are required, and where the plant is a (possibly infinite-dimensional) exponentially stable single input–single output (SISO) system. We shall show that generalized repetitive control is possible provided that the corresponding feedforward regulator equations are solvable and provided that the reference/disturbance signals are smooth enough. In this SISO case the feedforward regulator equations can be solved using the methods of [24], provided that there are no transmission zeros at the complex frequencies $\frac{2\pi ni}{p}$, $n \in \mathbb{Z}$, of the exogenous signals; this is illustrated in an example in section 4. Moreover, conditional robustness allows us to dispense with certain severe stabilization problems present in a related earlier work [22]. A consequence of our results is that generalized repetitive control in the above sense is possible even for exponentially stable strictly proper finite-dimensional SISO plants (see section 4 for more details).

1.1. Notation. For Banach spaces E and F, $\mathcal{L}(E, F)$ denotes the space of bounded linear operators $E \to F$. The resolvent set of a closed linear operator

[1] The exosystems utilized in [17, 32] are still finite-dimensional, however.

$A : E \to F$ is denoted by $\rho(A)$, whereas $\sigma(A)$ denotes its spectrum. The point spectrum of A is denoted by $\sigma_p(A)$. $R(\lambda, A)$ denotes (whenever it exists) the resolvent operator $(\lambda I - A)^{-1}$. If $\widetilde{E}$ is a subspace of E, then $A|_{\widetilde{E}}$ denotes the restriction of A to $\widetilde{E}$. If $A : E \to E$ generates a strongly continuous (or C_0-) semigroup in E, then this semigroup is in general denoted by $T_A(t)$. We say that $T_A(t)$ is strongly stable if $\lim_{t\to\infty}\|T_A(t)x\| = 0$ for each $x \in E$. $T_A(t)$ is exponentially stable if $\|T_A(t)\| \le Me^{-\omega t}$ for some $M, \omega > 0$ and all $t \ge 0$. For the sake of brevity in these circumstances we sometimes say that A is strongly/exponentially stable. The inner product is denoted by $\langle\cdot,\cdot\rangle_E$, where the subscript E is occasionally included to clarify the space on which the inner product is defined.

2. The output regulation problem and robustness. In this section we recall the error feedback output regulation problem studied under various assumptions, e.g., in [6, 10, 13, 14, 22]. A solution of this problem amounts to designing a controller which employs error feedback to achieve stability of the closed loop (consisting of the plant and the controller) and asymptotic tracking of the reference signals under disturbances. Here asymptotic tracking means $e(t) = y(t) - y_{ref}(t) \to 0$ as $t \to \infty$, where $y(t)$ is the output of the plant and $y_{ref}(t)$ is the reference signal. In order to be more precise, we need to first define the plant and the exogenous system that generates the reference and disturbance signals.

2.1. The plant. We consider a plant described by the following (possibly infinite-dimensional) control system:

$$\dot{z}(t) = Az(t) + Bu(t) + \mathcal{U}_{dist}(t), \quad z(0) \in Z, \quad t \ge 0, \tag{2.1a}$$

$$y(t) = Cz(t) + Du(t), \quad t \ge 0. \tag{2.1b}$$

Here A generates a C_0-semigroup $T_A(t)$, $t \ge 0$, in a complex Banach space Z. The continuous input $u : \mathbb{R}_+ \to H$ and continuous output $y : \mathbb{R}_+ \to H$ take values in a complex Banach space H. The control operator $B \in \mathcal{L}(H, Z)$, the observation operator $C \in \mathcal{L}(Z, H)$, and the feedthrough operator $D \in \mathcal{L}(H)$. The continuous function $\mathcal{U}_{dist}$ is a disturbance (to be defined shortly). Equation (2.1a) is to be considered in the mild sense (see Curtain and Zwart [8]).

2.2. The exogenous system. It is a common (and mathematically convenient) assumption in control literature that the reference and disturbance signals in output regulation problems are generated by a so-called exogenous system. A common choice for an exogenous signal generator is a linear matrix differential equation with observation. Such an exosystem can handle the generation of simple signals—e.g., constants and sinusoids—but unfortunately signals which occur in practice are often more complex than this. Following the approach in [20, 21, 22, 23, 24], we shall next construct a possibly infinite-dimensional exogenous system which can generate general bounded uniformly continuous reference functions.

Let $\mathcal{H}$ be a Banach space which is continuously embedded in $BUC(\mathbb{R}, H)$, the space of bounded uniformly continuous H-valued functions on $\mathbb{R}$ endowed with the sup-norm ($\mathcal{H}$ need not have the sup-norm). In symbols, let $\mathcal{H} \hookrightarrow BUC(\mathbb{R}, H)$. We let $T_S(t)$ denote the shift C_0-group in $BUC(\mathbb{R}, H)$ defined as $T_S(t)f = f(\cdot + t)$; its infinitesimal generator is $S = \frac{d}{dx}$ with a suitable domain of definition $\mathcal{D}(S) \subset BUC(\mathbb{R}, H)$. If $\mathcal{H}$ is invariant for $T_S(t)$ and the restrictions $T_S(t)|_{\mathcal{H}}$ to $\mathcal{H}$ constitute an isometric C_0-group, we denote this by $\mathcal{H} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$. In this case the generator of $T_S(t)|_{\mathcal{H}}$ is denoted by $S|_{\mathcal{H}}$.

DEFINITION 2.1 (the exogenous system). *Let $\mathcal{H} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$, let $Q \in \mathcal{L}(\mathcal{H}, H)$ be an observation operator, and let $P \in \mathcal{L}(\mathcal{H}, Z)$ be a disturbance operator. The exogenous system generating the reference signals y_{ref} and the disturbance signals $\mathcal{U}_{dist}$ is defined on the state space $\mathcal{H}$ as*

$$\dot{w}(t) = S|_{\mathcal{H}} w(t), \quad w(0) = w_0 \in \mathcal{H}, \tag{2.2a}$$

$$y_{ref}(t) = Qw(t), \quad t \in \mathbb{R}, \tag{2.2b}$$

$$\mathcal{U}_{dist}(t) = Pw(t), \quad t \in \mathbb{R}, \tag{2.2c}$$

where (2.2a) *is to be considered in the mild sense* (*see Curtain and Zwart* [8]).

One particularly good choice is $Q = \delta_0$, where δ_0 is the point evaluation at the origin in $\mathcal{H}$, i.e., $\delta_0 f = f(0)$ for $f \in \mathcal{H}$. Clearly δ_0 is linear and $\|\delta_0 f\|_H = \|f(0)\|_H \le \|f\|_\infty \le c\|f\|_{\mathcal{H}}$ for some $c \ge 0$ (because $\mathcal{H} \hookrightarrow BUC(\mathbb{R}, H)$) and so $\delta_0 \in \mathcal{L}(\mathcal{H}, H)$. Moreover, $\delta_0 T_S(t)|_{\mathcal{H}} f = f(x+t)|_{x=0} = f(t)$ for every $f \in \mathcal{H}$ and $t \in \mathbb{R}$. Consequently, for $Q = \delta_0$, every reference signal y_{ref} in $\mathcal{H}$ (and only those) can be obtained from (2.2b) by choosing $w(0) = y_{ref}$. Although the choice $Q = \delta_0$ is sufficient for output regulation purposes [22], it turns out that for robustness considerations it is convenient to let Q vary in $\mathcal{L}(\mathcal{H}, H)$.

2.3. The error feedback output regulation problem and robustness. Our primary concern in this paper is to study robustness issues for the following output regulation problem [6, 10, 13, 14, 22].

DEFINITION 2.2. *The task in the error feedback regulation problem (EFRP) is to design an error feedback controller of the form*

$$\dot{x}(t) = Fx(t) + Ge(t), \quad x(0) \in X, \quad t \ge 0, \tag{2.3a}$$

$$u(t) = Jx(t) \tag{2.3b}$$

on some Banach state space X, where F generates a C_0-semigroup, $G \in \mathcal{L}(H, X)$, and $J \in \mathcal{L}(X, H)$. The controller must satisfy the following requirements:

1. *In the closed loop system for $t \ge 0$,*

$$\dot{z}(t) = Az(t) + BJx(t) + Pw(t), \tag{2.4a}$$

$$\dot{x}(t) = GCz(t) + (F + GDJ)x(t) - GQw(t), \tag{2.4b}$$

$$\dot{w}(t) = S|_{\mathcal{H}} w(t), \tag{2.4c}$$

$$e(t) = y(t) - y_{ref}(t) = Cz(t) + DJx(t) - Qw(t), \tag{2.4d}$$

   the semigroup $T_{\mathcal{A}}(t)$ generated by $\mathcal{A} = \left(\begin{smallmatrix} A & BJ \\ GC & F+GDJ \end{smallmatrix}\right)$ on $Z \times X$ is strongly stable (or, for brevity, $\mathcal{A}$ is strongly stable on $Z \times X$).
2. *The tracking error $e(t) \to 0$ as $t \to \infty$ for any initial conditions $z(0) \in Z$, $x(0) \in X$, and $w(0) \in \mathcal{H}$.*

The foundations of this paper are laid on the following result, first proved in [22].

THEOREM 2.3. *Let P and Q in the exosystem be fixed. Assume that F, G, and J in the controller* (2.3) *have been chosen such that the operator $\mathcal{A} = \left(\begin{smallmatrix} A & BJ \\ GC & F+GDJ \end{smallmatrix}\right)$ generates a strongly stable C_0-semigroup $T_{\mathcal{A}}(t)$ on $Z \times X$. If in addition there exist $\Pi \in \mathcal{L}(\mathcal{H}, Z)$ and $\Lambda \in \mathcal{L}(\mathcal{H}, X)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(A)$ and $\Lambda(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(F)$, and the following regulator equations are satisfied:*

$$A\Pi + BJ\Lambda + P = \Pi S|_{\mathcal{H}} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{2.5a}$$

$$F\Lambda = \Lambda S|_{\mathcal{H}} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{2.5b}$$

$$C\Pi + DJ\Lambda = Q \quad \text{in } \mathcal{H}, \tag{2.5c}$$

then with this triplet (F, G, J) the EFRP is solvable.

Explicit operators F, G, J which satisfy the conditions of Theorem 2.3 will be provided later in this paper (see also [6, 22]).

In our robustness analysis some (but not necessarily all) of the parameters A, B, C, D, G, J, P, and Q of the plant, the controller, and the exogenous system are subject to perturbations. Throughout this paper we assume that all perturbations are bounded, additive, and linear. For example, upon perturbation A may become $A + \Delta_A$, where $\Delta_A \in \mathcal{L}(Z)$. Furthermore, we assume that all perturbations to any one of the above operators are independent of the possible perturbations to the other operators in the above list. We then treat robustness in the following sense.

DEFINITION 2.4. *Output regulation is robust with respect to a set* $\Omega \subset \{A, B, C, D, G, J, P, Q\}$ *of parameters of the plant, the controller, and the exogenous system if the two conditions in Definition* 2.2 *hold in spite of sufficiently small perturbations in the elements of* Ω.

DEFINITION 2.5. *Output regulation is conditionally robust with respect to a set* $\Omega \subset \{A, B, C, D, G, J, P, Q\}$ *of parameters of the plant, the controller, and the exogenous system if all such small perturbations in the elements of* Ω *which preserve closed loop stability* (*i.e., condition* 1 *in Definition* 2.2) *also preserve asymptotic tracking* (*i.e., condition* 2 *in* Definition 2.2).

We point out that perturbations to F and $S|_{\mathcal{H}}$ are not permitted in our framework; this is in accordance with the corresponding finite-dimensional theory [13]. However, by varying $\mathcal{H}$ one can investigate robustness of output regulation for different classes of reference signals. On the other hand, perturbations to the parameters P and Q of the exogenous system never destroy closed loop stability. Hence such perturbations need not always be small for conditional robustness (see, e.g., Theorem 3.2). Finally, we point out that occasionally we must place additional structural requirements for the perturbations even for conditional robustness results (see section 4).

Remark 2.6. A potentially useful result in the study of conditional robustness is the following one due to Casarino and Piazzera [7]: Assuming that $\mathcal{A}$ generates a strongly stable C_0-semigroup on $Z \times X$, a sufficient condition that the perturbed closed loop operator $\mathcal{A} + \Delta_{\mathcal{A}}$, where $\Delta_{\mathcal{A}} \in \mathcal{L}(Z \times X)$,[2] still generates a strongly stable C_0-semigroup on $Z \times X$ is

$$\sup_{t \geq 0} \int_0^t \|\Delta_{\mathcal{A}} T_{\mathcal{A}}(s)\theta\| ds \leq C\|\theta\| \quad \forall \theta \in \mathcal{D}(\mathcal{A}), \quad 0 < C < 1. \tag{2.6}$$

Remark 2.7. If $T_{\mathcal{A}}(t)$ is exponentially stable, then bounded perturbations $\Delta_{\mathcal{A}}$ for which $\|\Delta_{\mathcal{A}}\|$ is sufficiently small do not destroy exponential stability of the closed loop system. Consequently for exponentially stable closed loop systems conditionally robust output regulation implies robust output regulation.

3. Conditionally robust regulation for two controllers. In this section we shall prove conditional robustness results for infinite-dimensional systems by utilizing type I and type II controllers. As mentioned in section 1, type I and type II controllers are representative members of the following two general regulator categories, respectively: those employing estimates of the state of the coupled system containing the plant and the exosystem (type I) and those employing dynamic error augmentation (type II). Our aim here is to (i) illustrate the abstract robustness theory—in particular the internal model structure—developed in [20]; (ii) prove new robustness

[2] In [7] the additive perturbation $\Delta_{\mathcal{A}}$ may actually be of the more general Miyadera–Voigt type.

results for control systems in infinite-dimensional spaces; and (iii) unify some existing robustness results using the abstract setting. To our knowledge conditional robustness (in the sense of Definition 2.5) of these two controller types has not been studied before.

We first recall some definitions from [20]. The linear Sylvester operator $\mathcal{T}_{F,S|_{\mathcal{H}}}$ is defined on a subspace of $\mathcal{L}(\mathcal{H},X)$ by

$$\mathcal{D}(\mathcal{T}_{F,S|_{\mathcal{H}}}) = \{\, \Lambda \in \mathcal{L}(\mathcal{H},X) \mid \Lambda(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(F), \exists M \in \mathcal{L}(\mathcal{H},X) : \tag{3.1}$$
$$Mf = F\Lambda f - \Lambda S|_{\mathcal{H}} f \ \forall f \in \mathcal{D}(S|_{\mathcal{H}}) \,\},$$

$$\mathcal{T}_{F,S|_{\mathcal{H}}}\Lambda = M. \tag{3.2}$$

Definition 3.1 (internal model structure [20]). *The controller* (2.3) *has the internal model structure if the following implication holds:*

$$\forall \Lambda \in \mathcal{D}(\mathcal{T}_{F,S|_{\mathcal{H}}}), \forall \Delta \in \mathcal{L}(\mathcal{H},H) : \Lambda S|_{\mathcal{H}} = F\Lambda + G\Delta \text{ in } \mathcal{D}(S|_{\mathcal{H}}) \implies \Delta = 0. \tag{3.3}$$

The central idea behind the internal model structure is, roughly stated, the following. Suppose that F, G, and J in (2.3) have been fixed in such a way that the controller has the internal model structure. If, in addition, the closed loop system is "sufficiently" stable, then there always exist $\Pi \in \mathcal{L}(\mathcal{H},Z)$ and $\Lambda \in \mathcal{L}(\mathcal{H},X)$ such that the first regulator equation (2.5a) is satisfied *and* such that $\Lambda S|_{\mathcal{H}} = F\Lambda + G(C\Pi + DJ\Lambda - Q)$ in $\mathcal{D}(S|_{\mathcal{H}})$. This operator equation combines the second and third regulator equations (2.5b)–(2.5c), but the internal model structure ensures that we can separate them and thus satisfy all of the regulator equations (2.5). As a consequence of this, sufficient closed loop stability and the internal model structure together imply output regulation by Theorem 2.3. A glance at the appendix of [14] immediately reveals that this idea generalizes the finite-dimensional arguments of Francis and Wonham.

In [20] the author showed that under *exponential* closed loop stability (and certain additional assumptions) the internal model structure is both necessary and sufficient for *robust* output regulation. Moreover, the following result was essentially proved in [20]. It provides sufficient conditions for closed loop stability and internal model structure to yield (conditionally) robust output regulation. We choose to prove the result also here to emphasize the effect of the internal model structure: *Any* sufficiently stable closed loop error feedback control system already contains the error zeroing dynamics. If, in addition, the controller has the internal model structure, then the desirable dynamical behavior of output regulation is realized. It is remarkable that here the regulator equations (2.5) do not have to be explicitly solved as in, e.g., [22], where nonrobust output regulation was studied.

Theorem 3.2. *Let F, G, and J be chosen such that the closed loop operator $\mathcal{A}$ generates a strongly stable C_0-semigroup on $Z \times X$ and such that the controller has the internal model structure. Let A, B, C, D, and J be subject to perturbations and let $\mathcal{A}^p$, with $\mathcal{D}(\mathcal{A}^p) = \mathcal{D}(\mathcal{A})$, denote the perturbed closed loop operator. Then the following hold:*

1. *If for given $P \in \mathcal{L}(\mathcal{H},Z)$ and $Q \in \mathcal{L}(\mathcal{H},H)$ in the exosystem there exists $\Pi \in \mathcal{L}(\mathcal{H}, Z \times X)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A}^p)$ and $\Pi S|_{\mathcal{H}} = \mathcal{A}^p\Pi + \left(\begin{smallmatrix} P \\ -GQ \end{smallmatrix}\right)$ in $\mathcal{D}(S|_{\mathcal{H}})$, and if $\mathcal{A}^p$ is still strongly stable, then the EFRP is solved for these P, Q and the above perturbed closed loop parameters.*
2. *If for every $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z \times X)$ there exists a unique $\Pi \in \mathcal{L}(\mathcal{H}, Z \times X)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A})$ and $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$, then the controller*

achieves conditionally robust output regulation with respect to the above perturbed closed loop parameters for all $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$.

3. *If* $\mathcal{A}$ *is exponentially stable, then the controller achieves robust output regulation with respect to the above perturbed closed loop parameters for all* $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$.

Proof.

1. Let $P \in \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{H}, H)$ be given, and set $\mathcal{P} = \left(\begin{smallmatrix} P \\ -GQ \end{smallmatrix}\right)$. Let $A^p = A + \Delta_A, B^p = B + \Delta_B, C^p = C + \Delta_C, D^p = D + \Delta_D, J^p = J + \Delta_J$, where the perturbations are bounded and linear on suitable spaces. By our assumption, there exists an operator $\left(\begin{smallmatrix} \Pi_1 \\ \Pi_2 \end{smallmatrix}\right) \in \mathcal{L}(\mathcal{H}, Z \times X)$ such that $\left(\begin{smallmatrix} \Pi_1 \\ \Pi_2 \end{smallmatrix}\right) S|_{\mathcal{H}} = \mathcal{A}^p \left(\begin{smallmatrix} \Pi_1 \\ \Pi_2 \end{smallmatrix}\right) + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$, i.e.,

$$\begin{pmatrix} \Pi_1 \\ \Pi_2 \end{pmatrix} S|_{\mathcal{H}} = \begin{pmatrix} A^p & B^p J^p \\ GC^p & F + GD^p J^p \end{pmatrix} \begin{pmatrix} \Pi_1 \\ \Pi_2 \end{pmatrix} + \begin{pmatrix} P \\ -GQ \end{pmatrix} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}). \tag{3.4}$$

But this shows that $\Pi_2 S|_{\mathcal{H}} = F\Pi_2 + G(C^p\Pi_1 + D^p J^p \Pi_2 - Q)$ in $\mathcal{D}(S|_{\mathcal{H}})$, and since the controller has the internal model structure, ultimately $C^p\Pi_1 + D^p J^p \Pi_2 = Q$. Consequently the following regulator equations are satisfied:

$$A^p\Pi_1 + B^p J^p \Pi_2 + P = \Pi_1 S|_{\mathcal{H}} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{3.5a}$$

$$F\Pi_2 = \Pi_2 S|_{\mathcal{H}} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{3.5b}$$

$$C^p\Pi_1 + D^p J^p \Pi_2 = Q \quad \text{in } \mathcal{H}. \tag{3.5c}$$

According to Theorem 2.3 solvability of (3.5) and strong closed loop stability together imply output regulation.

2. Let $P \in \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{H}, H)$ be arbitrary, and set $\mathcal{P} = \left(\begin{smallmatrix} P \\ -GQ \end{smallmatrix}\right)$. Let $A^p = A + \Delta_A, B^p = B + \Delta_B, C^p = C + \Delta_C, D^p = D + \Delta_D, J^p = J + \Delta_J$, where the perturbations are bounded and linear on suitable spaces. For conditional robustness, we assume that $\mathcal{A}^p$ is still the generator of a strongly stable C_0-semigroup. Define the linear Sylvester operator $\mathcal{T}_{\mathcal{A}, S|_{\mathcal{H}}}$ by

$$\mathcal{D}(\mathcal{T}_{\mathcal{A}, S|_{\mathcal{H}}}) = \{\Gamma \in \mathcal{L}(\mathcal{H}, Z \times X) \mid \Gamma(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A}), \exists Y \in \mathcal{L}(\mathcal{H}, Z \times X) :$$

$$Yf = \mathcal{A}\Gamma f - \Gamma S|_{\mathcal{H}} f \; \forall f \in \mathcal{D}(S|_{\mathcal{H}}) \}, \tag{3.6}$$

$$\mathcal{T}_{\mathcal{A}, S|_{\mathcal{H}}} \Gamma = Y. \tag{3.7}$$

By our assumptions, $\mathcal{T}_{\mathcal{A}, S|_{\mathcal{H}}}$ is a boundedly invertible closed operator [2]; consequently so is $\mathcal{T}_{\mathcal{A}^p, S|_{\mathcal{H}}}$ as long as the above perturbations to A, B, C, D, and J are small enough [26, p. 196]. Hence there exists a unique operator $\left(\begin{smallmatrix} \Pi_1 \\ \Pi_2 \end{smallmatrix}\right) \in \mathcal{L}(\mathcal{H}, Z \times X)$ such that $\left(\begin{smallmatrix} \Pi_1 \\ \Pi_2 \end{smallmatrix}\right) S|_{\mathcal{H}} = \mathcal{A}^p \left(\begin{smallmatrix} \Pi_1 \\ \Pi_2 \end{smallmatrix}\right) + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$. The result now follows by item 1.

3. By [38] exponential stability of $\mathcal{A}$ implies that for every $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z \times X)$ there exists a unique $\Pi \in \mathcal{L}(\mathcal{H}, Z \times X)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A})$ and $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ [37, 38]. Hence the conditions of item 2 are satisfied, and so we obtain conditionally robust output regulation with respect to A, B, C, D, and J for all P and Q. However, now all sufficiently small bounded perturbations to the closed loop operator $\mathcal{A}$ result in an operator $\mathcal{A}^p$, which still generates an exponentially stable semigroup. We thus deduce that the controller achieves robust output regulation with respect to A, B, C, D, and J for every P and Q. ∎

Remark 3.3. The above proof is based on solving the regulator equations (2.5) for the perturbed operators. Since these equations do not depend on G, in items 1–3 of Theorem 3.2 we also obtain (conditional) robustness with respect to G, because of the assumption that perturbations to the parameters can be considered independently of each other. From now on, perturbations to G are always justified by a similar reasoning and the assumption about independence of the perturbations.

Unique solvability of the Sylvester-type operator equation $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ for each $\mathcal{P}$, which in Theorem 3.2 is shown to be crucial for conditional robustness, has been studied, e.g., in [2, 3, 37, 38]. We emphasize that in many instances this unique solvability can be deduced without explicitly calculating the solution operator Π. As mentioned above, this is the case, for example, whenever $\mathcal{A}$ is exponentially stable. Unfortunately, if $\mathcal{A}$ is only strongly stable, then this conclusion cannot be drawn. In fact, if $\mathcal{A}$ is strongly stable such that $\sigma(\mathcal{A}) \cap \sigma(S|_{\mathcal{H}}) \neq \emptyset$ (which is clearly possible), then the operator equation $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ does not have a unique solution for every $\mathcal{P}$ [38]. On the other hand, in the case of finite-dimensional spaces Z and X, strong stability of $\mathcal{A}$ is equivalent to exponential stability of $\mathcal{A}$. Consequently, in finite dimensions, as soon as closed loop stability is achieved, the internal model structure guarantees robust output regulation without additional assumptions.

3.1. Conditional robustness results for type I controllers. In this subsection we shall explicitly specify the parameters of type I controllers in (3.13) and we shall show *how* one achieves (conditionally) robust output regulation using such controllers. The internal model structure will play a decisive role in our analysis. Before proving the results, however, we shall provide the reader with some essential background on the construction of type I controllers.

Type I controllers—as we call them in this paper—were introduced in a slightly different form in the finite-dimensional setup by Francis [13], where he presented the synthesis algorithm (SA) and its robust counterpart, the structurally stable synthesis algorithm (SSSA), for the solution of the output regulation problem. In the original paper, Francis [13] did not allow for perturbations in C and Q, but it has been shown in [20] that if the (finite-dimensional) EFRP of this paper is solved using the SSSA, then output regulation is robust with respect to C for an arbitrary Q as well. Byrnes et al. [6] have extended Francis's SA for infinite-dimensional plants, and this theory has subsequently been generalized for infinite-dimensional exogenous systems in [22]. However, as neither of the articles [6, 22] discusses robustness issues, the results of this subsection are new even for finite-dimensional exogenous systems.

In the original finite-dimensional setting of SSSA, robust controllers for the solution of the output regulation problem are constructed by employing $\dim(H)$-fold reduplications of the maximal cyclic component of the exosystem generator [13]. Here we shall employ a slightly different—and perhaps more intuitive—strategy in which none of the spaces has to be finite-dimensional; our idea is to embed the generator of a suitable translation group in the controller. In order to do this, we first embed the state space $\mathcal{H}$ of the exosystem continuously in a possibly larger space $\mathcal{G} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$ (we shall motivate this selection of $\mathcal{G}$ shortly). For every isolated point $i\omega \in \sigma(S|_{\mathcal{G}})$ we can then define the spectral projection by $P^{\mathcal{G}}_{i\omega} = \frac{1}{2\pi i} \oint_{i\omega} R(\lambda, S|_{\mathcal{G}}) d\lambda$ (and similarly for the isolated points of $\sigma(S|_{\mathcal{H}})$). We shall make the following standing assumptions, which hold throughout this subsection:

1. There is no feedthrough, i.e., $D = 0$.
2. The spectra $\sigma(S|_{\mathcal{H}}) = \sigma(S|_{\mathcal{G}})$ are discrete.

3. There exist $P \in \mathcal{L}(\mathcal{G}, Z)$ and $Q \in \mathcal{L}(\mathcal{G}, H)$ and $G = \left(\begin{smallmatrix} G_1 \\ G_2 \end{smallmatrix}\right) \in \mathcal{L}(H, Z \times \mathcal{G})$ such that $\left(\begin{smallmatrix} A & P \\ 0 & S|_{\mathcal{G}} \end{smallmatrix}\right) - \left(\begin{smallmatrix} G_1 \\ G_2 \end{smallmatrix}\right)\left(\begin{smallmatrix} C & -Q \end{smallmatrix}\right)$ generates a strongly stable C_0-semigroup on $Z \times \mathcal{G}$.
4. There exists $K \in \mathcal{L}(Z, H)$ such that $A + BK$ generates an exponentially stable C_0-semigroup on Z.
5. There exist $\Pi \in \mathcal{L}(\mathcal{G}, Z)$, such that $\Pi(\mathcal{D}(S|_{\mathcal{G}})) \subset \mathcal{D}(A)$, and $\Gamma \in \mathcal{L}(\mathcal{G}, H)$ satisfying the regulator equations

$$\Pi S|_{\mathcal{G}} = A\Pi + B\Gamma + P \quad \text{in } \mathcal{D}(S|_{\mathcal{G}}), \tag{3.8a}$$

$$Q = C\Pi \quad \text{in } \mathcal{G} \tag{3.8b}$$

6. For every $i\omega \in \sigma(S|_{\mathcal{G}})$ the operator $P^{\mathcal{G}}_{i\omega} G_2 : H \to \operatorname{ran}(P^{\mathcal{G}}_{i\omega})$ is injective.

Assumptions 1–5 are natural infinite-dimensional generalizations of those in the SSSA of Francis [13]; see also [22] for a discussion on how to satisfy assumption 3. However, assumption 6 requires additional justification. To this end we shall temporarily assume that $H = \mathbb{C}^N$ for some $N \in \mathbb{N}$. Then $\mathcal{G} \subset BUC(\mathbb{R}, \mathbb{C}^N)$ and $P^{\mathcal{G}}_{i\omega}$ maps $\mathcal{G}$ to $\{\, ae^{i\omega\cdot} \mid a \in \mathbb{C}^N \,\}$, which is an N-dimensional linear space. Since $G_2 : \mathbb{C}^N \to \mathcal{G}$, it is evident that $P^{\mathcal{G}}_{i\omega} G_2$ can be represented by an $N \times N$ matrix. In fact, each $f \in \mathbb{C}^N$ has the representation $f = \sum_{n=1}^N \langle f, e_n \rangle e_n$ in terms of the natural orthornormal basis for $\mathbb{C}^N$. Then, upon defining $\psi_n = G_2 e_n \in \mathcal{G}$ for each $1 \le n \le N$, we have

$$P^{\mathcal{G}}_{i\omega} G_2 f = \sum_{n=1}^N \langle f, e_n \rangle P^{\mathcal{G}}_{i\omega} \psi_n = \begin{pmatrix} P^{\mathcal{G}}_{i\omega}\psi_1 & \cdots & P^{\mathcal{G}}_{i\omega}\psi_N \end{pmatrix} \begin{pmatrix} \langle f, e_1 \rangle \\ \vdots \\ \langle f, e_N \rangle \end{pmatrix} \tag{3.9}$$

$$= e^{i\omega\cdot} \begin{pmatrix} \hat{\psi}_{11}(\omega) & \cdots & \hat{\psi}_{1N}(\omega) \\ \vdots & \vdots & \vdots \\ \hat{\psi}_{N1}(\omega) & \cdots & \hat{\psi}_{NN}(\omega) \end{pmatrix} \begin{pmatrix} \langle f, e_1 \rangle \\ \vdots \\ \langle f, e_N \rangle \end{pmatrix}, \tag{3.10}$$

where for each $1 \le n \le N$,

$$P^{\mathcal{G}}_{i\omega} \psi_n = \begin{pmatrix} \hat{\psi}_{1n}(\omega) \\ \vdots \\ \hat{\psi}_{Nn}(\omega) \end{pmatrix} e^{i\omega\cdot} \in \{\, ae^{i\omega\cdot} \mid a \in \mathbb{C}^N \,\} = \operatorname{ran}(P^{\mathcal{G}}_{i\omega}). \tag{3.11}$$

Consequently in order to meet assumption 6, the $N \times N$ matrix in (3.10) must be nonsingular for every $i\omega \in \sigma(S|_{\mathcal{G}})$. This requirement can often be met in practice because the functions $\psi_n = G_2 e_n$ can be chosen freely modulo assumption 3. In the SISO case ($N = 1$) it reduces to $\hat{\psi}_{11}(\omega) \neq 0$ for each $i\omega \in \sigma(S|_{\mathcal{G}})$. If the plant is SISO, if $\mathcal{G}$ is a Hilbert space, and if $S|_{\mathcal{G}}$ is a Riesz spectral operator[3] (see section 4 for a concrete example of such a situation), then $G_2 u = gu$ for some $g \in \mathcal{G}$ and all $u \in \mathbb{C}$. Moreover, in this case for each $i\omega \in \sigma(S|_{\mathcal{G}})$ we have $P^{\mathcal{G}}_{i\omega} G_2 = P^{\mathcal{G}}_{i\omega} g = \langle g, \phi_\omega \rangle_{\mathcal{G}} \theta_\omega$ for some eigenvector ϕ_ω of $S|_{\mathcal{G}}$; thus whenever $\langle g, \phi_\omega \rangle_{\mathcal{G}} \neq 0$ for each $i\omega \in \sigma(S|_{\mathcal{G}})$, the operator $P_{i\omega} G_2$ is injective. This condition, on the other hand, reduces to approximate controllability of the pair $(S|_{\mathcal{G}}, G_2)$ by Theorem 4.2.3 in [8].

We next drop the assumption $\dim(H) < \infty$ and explicitly specify the parameters F, G, and J of type I controllers in (3.13). Under assumptions 1–6 we can use an

[3] A Riesz-spectral operator on $\mathcal{G}$ is a closed operator whose spectrum consists of simple eigenvalues $\{\, \lambda_n \mid n \in \mathbb{Z} \,\}$, such that $\overline{\{\, \lambda_n \mid n \in \mathbb{Z} \,\}}$ is totally disconnected and such that the corresponding eigenvectors ϕ_ω constitute a Riesz basis (with biorthogonal vectors θ_ω) on $\mathcal{G}$ [8].

observer to generate an estimate $\Theta(t) = \left(\begin{smallmatrix}\theta_1(t)\\ \theta_2(t)\end{smallmatrix}\right)$ of the state $\left(\begin{smallmatrix}z(t)\\ w(t)\end{smallmatrix}\right)$ of the system

$$\begin{pmatrix}\dot z(t)\\ \dot w(t)\end{pmatrix} = \begin{pmatrix}A & P\\ 0 & S|_{\mathcal G}\end{pmatrix}\begin{pmatrix}z(t)\\ w(t)\end{pmatrix} + \begin{pmatrix}B\\ 0\end{pmatrix}u(t), \quad z(0)\in Z, \quad w(0)\in\mathcal G, \tag{3.12a}$$

$$e(t) = \begin{pmatrix}C & -Q\end{pmatrix}\begin{pmatrix}z(t)\\ w(t)\end{pmatrix} \tag{3.12b}$$

and then apply the control $u(t) = J\Theta(t) = J_1\theta_1(t) + J_2\theta_2(t)$, where $J_1 \in \mathcal L(Z,H)$ and $J_2 \in \mathcal L(\mathcal G, H)$ are chosen so that they solve a corresponding feedforward output regulation problem. This is the main idea in the construction of type I controllers [6, 13, 22]; the following new results validate this procedure and show that we obtain (conditionally) robust output regulation.

THEOREM 3.4. *Under assumptions* 1–6 *the dynamic type* I *controller*

$$\begin{aligned} F &= \begin{pmatrix}A+BK-G_1C & P+B(\Gamma-K\Pi)+G_1Q\\ -G_2C & S|_{\mathcal G}+G_2Q\end{pmatrix},\\ J &= \begin{pmatrix}K & \Gamma-K\Pi\end{pmatrix}, \quad G = \begin{pmatrix}G_1\\ G_2\end{pmatrix}\end{aligned} \tag{3.13}$$

on the state space $X = Z\times\mathcal G$ *has the internal model structure.*

Proof. Assume that $\Delta\in\mathcal L(\mathcal H,H)$ and $\Lambda = \left(\begin{smallmatrix}\Lambda_1\\ \Lambda_2\end{smallmatrix}\right) \in \mathcal D(\mathcal T_{F,S|_{\mathcal H}}) \subset \mathcal L(\mathcal H, X)$ are such that $\Lambda S|_{\mathcal H} = F\Lambda + G\Delta$ in $\mathcal D(S|_{\mathcal H})$. Then

$$\begin{pmatrix}\Lambda_1\\ \Lambda_2\end{pmatrix}S|_{\mathcal H} = \begin{pmatrix}A+BK-G_1C & P+B(\Gamma-K\Pi)+G_1Q\\ -G_2C & S|_{\mathcal G}+G_2Q\end{pmatrix}\begin{pmatrix}\Lambda_1\\ \Lambda_2\end{pmatrix} + \begin{pmatrix}G_1\\ G_2\end{pmatrix}\Delta \quad \text{in } \mathcal D(S|_{\mathcal H}), \tag{3.14}$$

whence $\Lambda_2 S|_{\mathcal H} = S|_{\mathcal G}\Lambda_2 + G_2(Q\Lambda_2 - C\Lambda_1 + \Delta)$ in $\mathcal D(S|_{\mathcal H})$.

By our assumptions and Gelfand's theorem [1, Cor. 4.4.8] the spectra $\sigma_p(S|_{\mathcal G}) = \sigma(S|_{\mathcal G}) = \sigma(S|_{\mathcal H}) = \sigma_p(S|_{\mathcal H})$. Moreover, as we define the spectral projections $P^{\mathcal H}_{i\omega} = \frac{1}{2\pi i}\oint_{i\omega} R(\lambda, S|_{\mathcal H})d\lambda$, it is true that $S|_{\mathcal F}P^{\mathcal F}_{i\omega} = i\omega P^{\mathcal F}_{i\omega} = P^{\mathcal F}_{i\omega}S|_{\mathcal F}$ for each $i\omega\in\sigma(S|_{\mathcal F})$, for $\mathcal F = \mathcal G, \mathcal H$. Applying projections on both sides of the equation $\Lambda_2 S|_{\mathcal H} = S|_{\mathcal G}\Lambda_2 + G_2(Q\Lambda_2 - C\Lambda_1 + \Delta)$ in $\mathcal D(S|_{\mathcal H})$ yields

$$0 = P^{\mathcal G}_{i\omega}[\Lambda_2 S|_{\mathcal H} - S|_{\mathcal G}\Lambda_2 - G_2(Q\Lambda_2 - C\Lambda_1+\Delta)]P^{\mathcal H}_{i\omega} = P^{\mathcal G}_{i\omega}G_2(Q\Lambda_2 - C\Lambda_1+\Delta)P^{\mathcal H}_{i\omega}. \tag{3.15}$$

But since $P^{\mathcal G}_{i\omega}G_2$ is injective, we must have $(Q\Lambda_2 - C\Lambda_1+\Delta)P^{\mathcal H}_{i\omega} = 0$ for every $i\omega\in\sigma(S|_{\mathcal H})$. Now since $T_S(t)|_{\mathcal H}$ is an isometric C_0-group, by [38, pp. 399–401] there exists a sequence $(\mathcal H_n)_{n\in\mathbb N}\subset\mathcal H$ of closed translation invariant (i.e., $T_S(t)|_{\mathcal H}$-invariant) subspaces such that $\mathcal H_n\subset\mathcal H_{n+1}$ for every $n\in\mathbb N$, $\sigma(S|_{\mathcal H_n})\subset\sigma(S|_{\mathcal H})$, $S_n = S|_{\mathcal H_n}\in\mathcal L(\mathcal H)$, and $\overline{\cup_{n\in\mathbb N}\mathcal H_n} = \mathcal H$. Moreover, we can choose $\mathcal H_n$ as the maximal invariant subspace on which $S|_{\mathcal H}$ is bounded and has a spectrum contained in (say) $[-in, in]$. Hence $S_n = \sum_{k\in I_n} i\omega_k P^{\mathcal H}_{i\omega_k}$ for some finite index set I_n and every $n\in\mathbb N$. Moreover, $\mathcal H_n = \operatorname{span}\{\, P^{\mathcal H}_{i\omega_k}f \mid f\in\mathcal H, k\in I_n\,\}$. Consequently $Q\Lambda_2 - C\Lambda_1+\Delta = 0$ on $\cup_{n\in\mathbb N}\mathcal H_n$. An extension by continuity gives $Q\Lambda_2 - C\Lambda_1+\Delta = 0$ on $\mathcal H$. Hence also $\Lambda_2 S|_{\mathcal H} = S|_{\mathcal G}\Lambda_2$ in $\mathcal D(S|_{\mathcal H})$.

On the other hand, by (3.14) we also have

$$\Lambda_1 S|_{\mathcal H} = (A+BK-G_1C)\Lambda_1 + (P+B(\Gamma-K\Pi)+G_1Q)\Lambda_2 + G_1\Delta \tag{3.16}$$

$$= (A+BK)\Lambda_1 + (P+B(\Gamma-K\Pi))\Lambda_2 + G_1(Q\Lambda_2 - C\Lambda_1+\Delta) \tag{3.17}$$

$$= (A+BK)\Lambda_1 + (P+B(\Gamma-K\Pi))\Lambda_2 \quad \text{in } \mathcal D(S|_{\mathcal H}). \tag{3.18}$$

Consequently $\left(\begin{smallmatrix}\Lambda_1\\ \Lambda_2\end{smallmatrix}\right)S|_{\mathcal{H}} = \left(\begin{smallmatrix}A+BK & B(\Gamma-K\Pi)+P\\ 0 & S|_{\mathcal{G}}\end{smallmatrix}\right)\left(\begin{smallmatrix}\Lambda_1\\ \Lambda_2\end{smallmatrix}\right)$ in $\mathcal{D}(S|_{\mathcal{H}})$. But since by our assumptions the regulator equations (3.8) are satisfied, we have

$$\begin{pmatrix} I & \Pi \\ 0 & I \end{pmatrix}\begin{pmatrix} A+BK & 0 \\ 0 & S|_{\mathcal{G}} \end{pmatrix}\begin{pmatrix} I & -\Pi \\ 0 & I \end{pmatrix} = \begin{pmatrix} A+BK & B(\Gamma-K\Pi)+P \\ 0 & S|_{\mathcal{G}} \end{pmatrix} \tag{3.19}$$

so that $\left(\begin{smallmatrix}I & -\Pi\\ 0 & I\end{smallmatrix}\right)\left(\begin{smallmatrix}\Lambda_1\\ \Lambda_2\end{smallmatrix}\right)S|_{\mathcal{H}} = \left(\begin{smallmatrix}A+BK & 0\\ 0 & S|_{\mathcal{G}}\end{smallmatrix}\right)\left(\begin{smallmatrix}I & -\Pi\\ 0 & I\end{smallmatrix}\right)\left(\begin{smallmatrix}\Lambda_1\\ \Lambda_2\end{smallmatrix}\right)$ in $\mathcal{D}(S|_{\mathcal{H}})$. Therefore $(\Lambda_1 - \Pi\Lambda_2)S|_{\mathcal{H}} = (A+BK)(\Lambda_1-\Pi\Lambda_2)$ in $\mathcal{D}(S|_{\mathcal{H}})$. But $A+BK$ is exponentially stable, so that the only operator $M \in \mathcal{L}(\mathcal{H}, Z)$ satisfying $MS|_{\mathcal{H}} = (A+BK)M$ in $\mathcal{D}(S|_{\mathcal{H}})$ is $M=0$. We thus have $\Lambda_1 = \Pi\Lambda_2$ in $\mathcal{H}$.

Finally, recall from the above that $Q\Lambda_2 - C\Lambda_1 + \Delta = 0$ in $\mathcal{H}$, so that by the regulator equations (3.8) we have $Q\Lambda_2 - C\Pi\Lambda_2 + \Delta = Q\Lambda_2 - Q\Lambda_2 + \Delta = \Delta = 0$. This shows that the controller has the internal model structure. ∎

COROLLARY 3.5. *Let assumptions* 1–6 *hold, and consider the type* I *controller* (3.13). *If for every* $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z\times X)$ *there exists a unique operator* $M \in \mathcal{L}(\mathcal{H}, Z\times X)$ *such that* $M(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A})$ *and* $MS|_{\mathcal{H}} = \mathcal{A}M + \mathcal{P}$ *in* $\mathcal{D}(S|_{\mathcal{H}})$, *then the controller* (3.13) *solves the EFRP. Moreover, output regulation is conditionally robust with respect to* $A, B, C, G,$ *and* J *regardless of* $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$ *in the exosystem.*

Proof. Under assumptions 1–6 the controller (3.13) has the internal model structure by Theorem 3.4. We shall verify that the closed loop system is strongly stable and appeal to Theorem 3.2. Applying a similarity transform U given as

$$U = \begin{pmatrix} I & 0 & 0 \\ I & -I & 0 \\ 0 & 0 & -I \end{pmatrix} \tag{3.20}$$

on $Z\times Z\times\mathcal{G}$ to the closed loop operator $\mathcal{A}$ we obtain the operator $\widetilde{\mathcal{A}} = U\mathcal{A}U$ having the expression

$$\widetilde{\mathcal{A}} = \begin{pmatrix} A+BK & -BK & -B(\Gamma-K\Pi) \\ 0 & A-G_1C & P+G_1Q \\ 0 & -G_2C & S|_{\mathcal{G}}+G_2Q \end{pmatrix} := \begin{pmatrix} A+BK & \Delta \\ 0 & \mathcal{A}_s \end{pmatrix}. \tag{3.21}$$

By our assumption, $\mathcal{A}_s$ generates a strongly stable C_0-semigroup on X, and $A+BK$ generates an exponentially stable C_0-semigroup on Z. Clearly the C_0-semigroup generated by $\widetilde{\mathcal{A}}$ on $Z\times X$ is given by

$$T_{\widetilde{\mathcal{A}}}(t) = \begin{pmatrix} T_{A+BK}(t) & \int_0^t T_{A+BK}(t-s)\Delta T_{\mathcal{A}_s}(s)ds \\ 0 & T_{\mathcal{A}_s}(t) \end{pmatrix}. \tag{3.22}$$

Consequently $T_{\widetilde{\mathcal{A}}}(t)$ is strongly stable if

$$\lim_{t\to\infty}\int_0^t T_{A+BK}(t-s)\Delta T_{\mathcal{A}_s}(s)xds = 0 \quad \forall x\in X. \tag{3.23}$$

But (3.23) holds by Proposition 5.6.1 in [1]. This proves that also $T_{\mathcal{A}}(t)$ is strongly stable. ∎

COROLLARY 3.6. *Let assumptions* 1–6 *hold, and assume that the closed loop operator* $\mathcal{A}$ *employing the type* I *controller* (3.13) *generates an exponentially stable* C_0*-semigroup on* $Z\times X$. *Then output regulation is robust with respect to* $A, B, C, G,$ *and* J *regardless of* $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$ *in the exosystem.*

In [6] (Theorem IV.2) Byrnes et al. have proved a complete characterization for the existence of a dynamic controller achieving (possibly nonrobust) output regulation for a given finite-dimensional exogenous system with fixed operators P and Q. They also employed controllers analogous to (3.13), but in their setting the closed loop system is exponentially stable; hence Corollary 3.6 above reveals that if G_2 is chosen appropriately, then the controller does not need to be changed if the operators P and Q in the exosystem are changed. Moreover, Corollary 3.6 proves that in this case output regulation is in fact robust with respect to certain parameters.

We emphasize that in the case of finite-dimensional spaces $\mathcal{G}$ and H, the differential operator $S|_{\mathcal{G}}$ can be represented by a diagonal matrix S_0. Moreover, it is easy to see that S_0 is similar to a matrix containing a $\dim(H)$-fold copy of the maximal cyclic component of S_0. Hence the above robustness results are in accordance with the corresponding finite-dimensional theory [13]. Moreover, they provide a natural interpretation and generalization for the somewhat unintuitive process of incorporating copies of maximal cyclic components of S_0 into the controller, which in the finite-dimensional setting is well known to guarantee robust output regulation under closed loop stability [10, 13, 14, 27, 40].

We also point out that it is convenient to allow for $\mathcal{H} \hookrightarrow \mathcal{G}$, because then $\mathcal{H}$ and $\mathcal{G}$ may also have different topologies. This turns out to be important in the verification of the conditions of Theorem 3.2. In fact, we shall see in Section 4 that although it may not be possible to solve $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ uniquely for every $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z \times X)$, as required in Corollary 3.5, sometimes it is possible to solve $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ for every $\mathcal{P} \in \mathcal{L}(\mathcal{G}, Z \times X) \subset \mathcal{L}(\mathcal{H}, Z \times X)$. This anomalous situation occurs, however, only if the closed loop system is not exponentially stable. If $\mathcal{A}$ generates an exponentially stable C_0-semigroup, then the equation $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ has a unique solution regardless of $\mathcal{H} \hookrightarrow BUC(\mathbb{R}, H)$ and $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z \times X)$ [38].

We conclude this subsection with an example of error feedback control for a one-dimensional heat equation. The example below is essentially Example VI.2 of [6] (which is why we omit some details). Byrnes et al. [6] have constructed a controller which achieves output regulation of a sinusoidal reference signal in the disturbance-free case; below we shall show that their controller is also robust. In the following example we have $\mathcal{G} = \mathcal{H}$ and $S|_{\mathcal{G}} = S|_{\mathcal{H}}$.

Example 3.7. Consider a disturbance-free controlled one-dimensional heat equation on the interval $[0, 1]$ with Neumann boundary conditions: $\frac{\partial z(x,t)}{\partial t} = \frac{\partial^2 z(x,t)}{\partial^2 x} + Bu(t)$, $\frac{\partial z(0,t)}{\partial x} = \frac{\partial z(1,t)}{\partial x} = 0$, $z(x, 0) = \psi(x)$. The output is given as $y(t) = Cz(t)$. The bounded control operator $B : \mathbb{C} \to L^2(0, 1)$ is defined by $Bu = b(x)u$, with $b(x) = 2\chi_{[\frac{1}{2},1]}(x)$. Here $\chi_{[\epsilon,\delta]}(x)$ denotes the characteristic function of the interval $[\epsilon, \delta]$. The bounded observation operator $C : L^2(0, 1) \to \mathbb{C}$ is defined by $C\psi = \int_0^1 c(x)\psi(x)dx$, with $c(x) = 2\chi_{[0,\frac{1}{2}]}(x)$.

It is well known how to put this system in the form (2.1) with $D = 0$, $P = 0$ [6, 8, 23]. In [6] (Example VI.2) Byrnes et al. designed a dynamic error feedback controller (2.3) such that the plant can asymptotically track the reference signal $y_{ref}(t) = M \sin(\gamma t)$, with $M, \gamma > 0$. They used the exosystem $\dot{w}_0(t) = Sw_0(t)$ with $S = \left(\begin{smallmatrix} 0 & \gamma \\ -\gamma & 0 \end{smallmatrix}\right)$, $w_0(0) = \left(\begin{smallmatrix} 0 \\ M \end{smallmatrix}\right)$, $Q_0 = \left(\begin{smallmatrix} 1 & 0 \end{smallmatrix}\right)$ to generate $y_{ref}(t)$. After an application of the similarity transformation $U = \left(\begin{smallmatrix} i & -i \\ 1 & 1 \end{smallmatrix}\right)$ this exosystem becomes

(3.24)

$$\dot{w}(t) = S|_{\mathcal{H}} w(t) = \begin{pmatrix} \dot{w}_1(t) \\ \dot{w}_2(t) \end{pmatrix} = \begin{pmatrix} -i\gamma & 0 \\ 0 & i\gamma \end{pmatrix} \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix}, \quad Q = \begin{pmatrix} i & -i \end{pmatrix}, \quad w(0) = \begin{pmatrix} \frac{1}{2}M \\ \frac{1}{2}M \end{pmatrix},$$

which is now isomorphic to the desired form (2.2) on the state space $\mathcal{H} = \operatorname{span}\{e^{-i\gamma\cdot}\} \oplus \operatorname{span}\{e^{i\gamma\cdot}\}$. For example, in this basis the above initial condition is $w(0) = \frac{1}{2}M(e^{-i\gamma\cdot} + e^{i\gamma\cdot})$.

For the above plant Byrnes et al. [6] have solved the regulator equations $\Pi_0 S = A\Pi_0 + B\Gamma_0$ and $C\Pi_0 = Q_0$ for Π_0 and Γ_0. Since $S|_{\mathcal{H}} = U^{-1}SU$ and $Q = Q_0 U$, we readily see that $\Pi = \Pi_0 U$ and $\Gamma = \Gamma_0 U$ solve the regulator equations (3.8). We may then let

$$
\begin{aligned}
F &= \begin{pmatrix} A + BK - G_1 C & B(\Gamma - K\Pi) + G_1 Q \\ -G_2 C & S|_{\mathcal{H}} + G_2 Q \end{pmatrix}, \\
&= \begin{pmatrix} K & \Gamma - K\Pi \end{pmatrix}, \quad \text{and} \quad G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix},
\end{aligned} \tag{3.25}
$$

where $G_2 = U^{-1}G_2^0$, and $G_0 = \left(\begin{smallmatrix} G_1 \\ G_2^0 \end{smallmatrix}\right)$ is the same exponentially stabilizing output injection for the pair $\left(\left(\begin{smallmatrix} A & 0 \\ 0 & S \end{smallmatrix}\right), \left(\begin{smallmatrix} C & -Q_0 \end{smallmatrix}\right)\right)$ and K is the same exponentially stabilizing state feedback for the pair (A, B) used in [6]. These choices lead to an exponentially stable closed loop system operator $\mathcal{A}$, which is also similar to the one in Example VI.2 of [6]. In fact,

$$
\begin{aligned}
&\begin{pmatrix} A & BJ \\ GC & F \end{pmatrix} \\
&= \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & U^{-1} \end{pmatrix} \begin{pmatrix} A & BK & B(\Gamma_0 - K\Pi_0) \\ G_1 C & A + BK - G_1 C & B(\Gamma_0 - K\Pi_0) + G_1 Q_0 \\ G_2^0 C & -G_2^0 C & S + G_2^0 Q_0 \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & U \end{pmatrix}.
\end{aligned} \tag{3.26}
$$

According to Corollary 3.6, in order to obtain robust output regulation, we need only verify that $P_{i\omega}^{\mathcal{H}} G_2$ is injective for $\omega = \pm\gamma$. But since Byrnes et al. [6] chose $G_2^0 = \left(\begin{smallmatrix} -3 \\ -3 \end{smallmatrix}\right)$, we obtain $G_2 = U^{-1}G_2^0 = \frac{3}{2}\left(\begin{smallmatrix} i-1 \\ -i-1 \end{smallmatrix}\right)$. Since neither component of G_2 is zero, $P_{i\omega}^{\mathcal{H}} G_2$ is injective for $\omega = \pm\gamma$. Consequently, output regulation is robust with respect to A, B, C, G, and J regardless of $P \in \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{H}, \mathbb{C})$ in the exosystem. Observe that these operators P, Q need not coincide with those inside F (cf. (3.13)), and that robustness allows for asymptotic tracking *and* disturbance rejection (only tracking is considered in [6]).

3.2. Conditional robustness results for type II controllers. In this subsection we shall explicitly specify the parameters of type II controllers in (3.32), and we shall prove conditional robustness results for such controllers by utilizing the concept of internal model structure. Before proving any results, however, we shall provide the reader with some background on the construction of type II controllers.

Recall that a type II controller is a specific (possibly infinite-dimensional) representative member of the general regulator category in which dynamic error augmentation is employed. Type II controllers—as we call them in this paper—were originally introduced in a slightly different form in the finite-dimensional setup by Davison [10]. Davison's arguments, which proved robustness of the controllers, were based on calculations of the steady state time response of the closed loop system, and they have subsequently been somewhat simplified by Huang [16]. A key element in the proofs is that the controllers employ $\dim(H)$-fold reduplications of the exosystem generator in a suitably transformed form. Davison's robust output regulation theory has also been generalized for infinite-dimensional plants and finite-dimensional exogenous

systems, e.g., in [18, 31, 36]. However, to the author's knowledge type II controllers for infinite-dimensional exogenous systems have not been studied before.

Type II controllers are often designed in two phases [16]. In the first phase, a type II *state feedback* controller is designed in order to achieve robust output regulation. In the second phase, the parameters of this state feedback controller are utilized in the design of a dynamic type II controller which does not employ state feedback. In order to illustrate this procedure, let us consider the controller

$$\dot{x}(t) = Fx(t) + Ge(t), \quad x(0) \in X, \quad t \geq 0, \tag{3.27a}$$

$$u(t) = K_1 z(t) + Jx(t), \tag{3.27b}$$

which also employs a state feedback $K_1 \in \mathcal{L}(Z, H)$ from the plant. Let $\mathcal{H} \hookrightarrow \mathcal{G} \overset{s}{\hookrightarrow} BUC(\mathbb{R}, H)$, and assume the following throughout this subsection:

(a) The spectra $\sigma(S|_{\mathcal{H}}) = \sigma(S|_{\mathcal{G}})$ are discrete.

(b) There exist $K_1 \in \mathcal{L}(H, Z)$, $K_2 \in \mathcal{L}(\mathcal{G}, H)$, and $G_0 \in \mathcal{L}(H, \mathcal{G})$ such that $\mathcal{A}_0 = \left(\begin{smallmatrix} A+BK_1 & BK_2 \\ G_0(C+DK_1) & S|_{\mathcal{G}}+G_0DK_2 \end{smallmatrix}\right)$ generates a strongly stable C_0-semigroup on $Z \times \mathcal{G}$.

(c) For every $i\omega \in \sigma(S|_{\mathcal{G}})$ the operator $P^{\mathcal{G}}_{i\omega}G_0 : H \to \operatorname{ran}(P^{\mathcal{G}}_{i\omega})$ is injective (see subsection 3.1 for the notation).

Under the above assumptions we can specify $X = \mathcal{G}, F = S|_{\mathcal{G}}, G = G_0$, and $J = K_2$ in (3.27). We thus obtain the following controller, which we call the type II state feedback controller:

$$\dot{x}(t) = S|_{\mathcal{G}}x(t) + G_0e(t), \quad x(0) \in \mathcal{G}, \tag{3.28a}$$

$$u(t) = K_1 z(t) + K_2 x(t). \tag{3.28b}$$

We shall next establish conditions under which this state feedback controller solves the EFRP for a plant, where A is replaced with $A + BK_1$ and C is replaced with $C + DK_1$. Of course, this replacement procedure is just a compact way of showing that the state feedback controller (3.28) achieves closed loop stability and asymptotic tracking of the reference signals in the presence of disturbances for the plant (2.1).

THEOREM 3.8. *Let assumptions* (a)–(c) *hold, and consider a plant in which A is replaced with $A + BK_1$ and C is replaced with $C + DK_1$. Set $X = \mathcal{G}$ with $F = S|_{\mathcal{G}}$, $G = G_0$, $J = K_2$. Then the resulting dynamic controller (F, G, J) has the internal model structure.*

Proof. The proof that $\Lambda S|_{\mathcal{H}} = S|_{\mathcal{G}}\Lambda + G_0\Delta$ in $\mathcal{D}(S|_{\mathcal{H}})$ implies $\Delta = 0$ is carried out analogously to the proof of Theorem 3.4 using Gelfand's theorem, the injectivity of $P^{\mathcal{G}}_{i\omega}G_0$, and the decomposition of $\mathcal{H}$ to the closed translation-invariant subspaces $\mathcal{H}_n$, with $\mathcal{H} = \overline{\cup_{n\in\mathbb{N}}\mathcal{H}_n}$. ∎

COROLLARY 3.9. *Let assumptions* (a)–(c) *hold, and consider a plant in which A is replaced with $A + BK_1$ and C is replaced with $C + DK_1$. If, in addition, for every $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z\times X)$ there exists a unique $\Pi \in \mathcal{L}(\mathcal{H}, Z\times X)$ such that $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A}_0)$ and $\Pi S|_{\mathcal{H}} = \mathcal{A}_0\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$, then the controller of Theorem* 3.8 *achieves a conditionally robust output regulation with respect to the operators A, B, C, D, K_1, K_2, and G_0 for all $P \in \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{H}, H)$.*

Proof. Let $P \in \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{H}, H)$ be arbitrary, and set $\mathcal{P} = \left(\begin{smallmatrix} P \\ -G_0Q \end{smallmatrix}\right)$. As in the proof of Theorem 3.2 we deduce that (after possibly applying small bounded perturbations to the parameters A, B, C, D, K_1, K_2) there exists a unique operator

$\left(\begin{smallmatrix}\Pi\\ \Lambda\end{smallmatrix}\right) \in \mathcal{L}(\mathcal{H}, Z\times\mathcal{G})$ such that $\left(\begin{smallmatrix}\Pi\\ \Lambda\end{smallmatrix}\right)S|_{\mathcal{H}} = \mathcal{A}_0\left(\begin{smallmatrix}\Pi\\ \Lambda\end{smallmatrix}\right) + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$, i.e.,

$$\begin{pmatrix}\Pi\\ \Lambda\end{pmatrix} S|_{\mathcal{H}} = \begin{pmatrix} A+BK_1 & BK_2\\ G_0(C+DK_1) & S|_{\mathcal{G}}+G_0DK_2\end{pmatrix}\begin{pmatrix}\Pi\\ \Lambda\end{pmatrix} + \begin{pmatrix} P\\ -G_0Q\end{pmatrix} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}). \tag{3.29}$$

But this shows that $\Pi S|_{\mathcal{H}} = (A+BK_1)\Pi + BK_2\Lambda + P$ and $\Lambda S|_{\mathcal{H}} = S|_{\mathcal{G}}\Lambda + G_0[(C+DK_1)\Pi + DK_2\Lambda - Q]$ in $\mathcal{D}(S|_{\mathcal{H}})$. Since by Theorem 3.8 the controller has the internal model structure, we must have that $(C+DK_1)\Pi + DK_2\Lambda - Q = 0$ in $\mathcal{H}$ and $\Lambda S|_{\mathcal{H}} = S|_{\mathcal{G}}\Lambda$ in $\mathcal{D}(S|_{\mathcal{H}})$. The result now follows from Theorem 2.3 because by our assumptions the closed loop operator $\mathcal{A} = \mathcal{A}_0$ generates a strongly stable C_0-semigroup and because the regulator equations (2.5) have a solution. ∎

COROLLARY 3.10. *Let assumptions* (a)–(c) *hold, and consider a plant in which* A *is replaced with* $A+BK_1$ *and* C *is replaced with* $C+DK_1$. *If, in addition,* $\mathcal{A}_0$ *is exponentially stable, then the controller of Theorem* 3.8 *achieves a robust output regulation with respect to the operators* A, B, C, D, K_1, K_2, *and* G_0 *for all* $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$.

Before turning to the construction of dynamic controllers (2.3), which do not involve state feedback from the plant, namely the type II controllers, we show how the operators K_1, K_2, and G_0 above could be chosen such that assumption 2 is satisfied.

PROPOSITION 3.11. *Assume that there exist* $K \in \mathcal{L}(H, Z)$, $K_2 \in \mathcal{L}(\mathcal{G}, H)$, *and* $L \in \mathcal{L}(H, \mathcal{G})$ *such that* either $A+BK$ *generates an exponentially stable* C_0*-semigroup on* Z *and* $S|_{\mathcal{G}} + LK_2$ *generates a strongly stable* C_0*-semigroup on* $\mathcal{G}$ or $A+BK$ *generates a strongly stable* C_0*-semigroup on* Z *and* $S|_{\mathcal{G}} + LK_2$ *generates an exponentially stable* C_0*-semigroup on* $\mathcal{G}$. *If there exist* $G_0 \in \mathcal{L}(H, \mathcal{G})$ *and* $\Pi \in \mathcal{L}(Z, \mathcal{G})$ *such that* $\Pi(\mathcal{D}(A)) \subset \mathcal{D}(S|_{\mathcal{G}})$ *and*

$$S|_{\mathcal{G}}\Pi = \Pi(A+BK) + G_0(C+DK) \quad \text{in } \mathcal{D}(A), \tag{3.30a}$$

$$L = \Pi B + G_0 D \quad \text{in } H, \tag{3.30b}$$

then for $K_1 = K + K_2\Pi$ *the operator* $\mathcal{A}_0 = \left(\begin{smallmatrix} A+BK_1 & BK_2\\ G_0(C+DK_1) & S|_{\mathcal{G}}+G_0DK_2\end{smallmatrix}\right)$ *generates a strongly stable* C_0*-semigroup on* $Z\times\mathcal{G}$.

Proof. With the above choices we have $\mathcal{A}_0 = \left(\begin{smallmatrix} A+B(K+K_2\Pi) & BK_2\\ G_0(C+D[K+K_2\Pi]) & S|_{\mathcal{G}}+G_0DK_2\end{smallmatrix}\right)$. Then a direct calculation shows that

$$\begin{pmatrix} I & 0\\ \Pi & I\end{pmatrix}\begin{pmatrix} A+B(K+K_2\Pi) & BK_2\\ G_0(C+D[K+K_2\Pi]) & S|_{\mathcal{G}}+G_0DK_2\end{pmatrix}\begin{pmatrix} I & 0\\ -\Pi & I\end{pmatrix} = \begin{pmatrix} A+BK & BK_2\\ 0 & S|_{\mathcal{G}}+LK_2\end{pmatrix}, \tag{3.31}$$

which shows (see Theorem 5.1.2 and Proposition 5.6.4 in [1]) that $\mathcal{A}_0$ generates a strongly stable C_0-semigroup. ∎

Remark 3.12. If in Proposition 3.11 both $A+BK$ and $S|_{\mathcal{G}} + LK_2$ generate exponentially stable C_0-semigroups, then the closed loop system operator $\mathcal{A}_0$ is also exponentially stable. In fact, if we let $\|T_{A+BK}(t)\| \le M_1 e^{-\omega_1 t}$ and $\|T_{S|_{\mathcal{G}}+LK_2}(t)\| \le M_2 e^{-\omega_2 t}$ for all $t \ge 0$ and if, without loss of generality, we assume $\omega_2 \ge \omega_1 > 0$, then for all $y \in \mathcal{G}$ we have $\|\int_0^t T_{A+BK}(t-s)BK_2T_{S|_{\mathcal{G}}+LK_2}(s)yds\| \le M_1M_2\|BK_2\|\|y\|e^{-\omega_1 t}$ $\int_0^t e^{(\omega_1-\omega_2)s}ds$, which decays to 0 exponentially fast as $t\to\infty$.

We emphasize that state feedback makes it possible to treat plants in which (A, B) is only strongly stabilizable. For an alternative approach for strong stabilization of the closed loop system using Riccati equations we refer the reader to [9].

Now that we have presented type II state feedback controllers (3.28) which achieve (conditionally) robust output regulation, we turn to the construction of a dynamic

type II controller (2.3) which does not involve state feedback from the plant. In addition to assumptions (a)–(c) we shall need the following assumption:

(d) There exists $L \in \mathcal{L}(H,Z)$ such that $A-LC$ generates an exponentially stable C_0-semigroup on Z.

A dynamic type II controller (2.3) employing the parameters of the above state feedback controller is specified by the following parameters:

$$F=\begin{pmatrix} A+BK_1-L(C+DK_1) & (B-LD)K_2 \\ 0 & S|_{\mathcal{G}} \end{pmatrix}, \quad J=\begin{pmatrix} K_1 & K_2 \end{pmatrix}, \quad \text{and} \quad G=\begin{pmatrix} L \\ G_0 \end{pmatrix}. \tag{3.32}$$

It is clear that under the above assumptions the type II controller (3.32) has the internal model structure. The following result presents conditions under which we also achieve conditionally robust output regulation.

THEOREM 3.13. *Let assumptions* (a)–(d) *hold and assume that for every* $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z\times X)$ *there exists a unique* $\Pi \in \mathcal{L}(\mathcal{H}, Z\times X)$ *such that* $\Pi(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(\mathcal{A})$ *and* $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ *in* $\mathcal{D}(S|_{\mathcal{H}})$. *Then the type* II *controller* (3.32) *solves the EFRP. Moreover, output regulation is conditionally robust with respect to* $A, B, C, D, G,$ *and* J *regardless of* $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$.

Proof. Since clearly the controller (3.32) has the internal model structure, according to Theorem 3.2 it is sufficient to prove that the closed loop system is strongly stable. With $F, G,$ and J as in (3.32), the closed loop system operator becomes

$$\mathcal{A} = \begin{pmatrix} A & BJ \\ GC & F+GDJ \end{pmatrix} = \begin{pmatrix} A & BK_1 & BK_2 \\ LC & A+BK_1-LC & BK_2 \\ G_0C & G_0DK_1 & S|_{\mathcal{H}}+G_0DK_2 \end{pmatrix}. \tag{3.33}$$

Applying the similarity transform U given as

$$U = \begin{pmatrix} I & -I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, \tag{3.34}$$

we see that $\mathcal{A}$ is similar to the operator $\widetilde{\mathcal{A}} = U\mathcal{A}U^{-1}$ having the expression

$$\widetilde{\mathcal{A}} = \begin{pmatrix} A-LC & 0 & 0 \\ LC & A+BK_1 & BK_2 \\ G_0C & G_0(C+DK_1) & S|_{\mathcal{H}}+G_0DK_2 \end{pmatrix}, \tag{3.35}$$

which by our assumptions generates a strongly stable C_0-semigroup on $Z\times X$ (use, e.g., Proposition 5.6.1 in [1]). This proves that also the similar closed loop semigroup $T_{\mathcal{A}}(t)$ is strongly stable. □

COROLLARY 3.14. *Let assumptions* (a)–(d) *hold and assume that the closed loop operator* $\mathcal{A}$ *generates an exponentially stable* C_0*-semigroup on* $Z\times X$. *Then the type* II *controller* (3.32) *solves the EFRP. Moreover, output regulation is robust with respect to* $A, B, C, D, G,$ *and* J *regardless of* $P \in \mathcal{L}(\mathcal{H}, Z)$ *and* $Q \in \mathcal{L}(\mathcal{H}, H)$.

Remark 3.15. To the author's knowledge, besides that of Ukai and Iwazumi [36], no work on type II controllers allowing for strongly stable closed loop operator has been reported, even for finite-dimensional exosystems. Ukai and Iwazumi essentially constructed type II controllers for infinite-dimensional systems on reflexive spaces and finite-dimensional exosystems, and in their work the closed loop system does not have to be exponentially stable.

We conclude this subsection with an example of regulation of finite-dimensional signals for a weakly damped string of unit length. In this case the plant (and hence also the closed loop system) is only strongly stable. The obtained type II state feedback controller is finite-dimensional, and it achieves conditionally robust output regulation.

Example 3.16. A model for a weakly damped string of unit length, with clamped ends, is (see [29])

$$\frac{\partial^2 v(x,t)}{\partial t^2} + M\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v(x,t)}{\partial x^2} \quad \text{for } t \geq 0 \quad \text{and} \quad 0 \leq x \leq 1, \tag{3.36a}$$

$$v(0,t) = v(1,t) = 0 \quad \text{for } t \geq 0. \tag{3.36b}$$

Here M is a damping operator which will be defined shortly. We define the operator U by $Uv = -\frac{\partial^2 v}{\partial x^2}$ with $\mathcal{D}(U) = \{\, v \in H^2(0,1) \mid v(0) = v(1) = 0 \,\}$, where $H^2(0,1)$ denotes the standard Sobolev space on the unit interval. It can be shown that U has eigenvalues $\lambda_k = k^2\pi^2$, $k = 1, 2, \ldots$, and the corresponding eigenvectors $\phi_k(x) = \sqrt{2}\sin(k\pi x)$ constitute an orthonormal basis in $L^2(0,1)$ [29].

The damping operator M is defined by $Mv = \epsilon\langle m, v\rangle_{L^2(0,1)} m$, where $\epsilon > 0$ and

$$m = \sum_{k=1}^{\infty} \gamma_k \phi_k \tag{3.37}$$

with γ_k satisfying $0 < |\gamma_k| \leq \frac{d}{\sqrt{\lambda_k}}$ (for example, $\gamma_k = \frac{1}{k}$) for some $d > 0$. Next we define the Hilbert space $Z = L^2(0,1) \times L^2(0,1)$ with the natural inner product $\langle \cdot, \cdot \rangle$, and introduce

$$z(t) = \begin{pmatrix} U^{\frac{1}{2}} v(x,t) \\ \frac{\partial v(x,t)}{\partial t} \end{pmatrix}, \quad \text{and} \quad A = \begin{pmatrix} 0 & U^{\frac{1}{2}} \\ -U^{\frac{1}{2}} & M \end{pmatrix}. \tag{3.38}$$

Then (3.36) can be rewritten as $\dot{z}(t) = Az(t)$, $z(0) \in Z$, and it can be shown (see [29] and the references therein) that A generates a strongly (but not exponentially) stable C_0-semigroup $T_A(t)$ on Z. Moreover, all eigenvalues ν_k of A have negative real part, and the corresponding eigenvectors $(\psi_k)_k$ form a Riesz basis in Z. Consequently $\sigma(A) \cap i\mathbb{R} = \emptyset$, because for every $\omega \in \mathbb{R}$ and every index k the elements $R(i\omega, A)\psi_k = \int_0^\infty e^{-i\omega t} T_A(t)\psi_k dt = \int_0^\infty e^{(-i\omega+\nu_k)t}\psi_k dt$ exist in Z.

Consider then application of distributed control and observation to system (3.36) in the following sense. For a control operator $B \in \mathcal{L}(\mathbb{C}, Z)$ and an observation operator $C \in \mathcal{L}(Z, \mathbb{C})$ the SISO plant is given as

$$\dot{z}(t) = Az(t) + Bu(t), \quad z(0) \in Z, \quad t \geq 0, \tag{3.39a}$$

$$y(t) = Cz(t), \quad t \geq 0. \tag{3.39b}$$

Let the exogenous system (i.e., $\mathcal{H}$) be finite-dimensional, let $(\theta_n)_{n\in I}$ be the (finite) orthonormal basis for $\mathcal{H}$ corresponding to the eigenvectors of $S|_{\mathcal{H}}$, and assume that $S|_{\mathcal{H}} + LK_2$ is exponentially stable for some operators $L \in \mathcal{L}(\mathbb{C}, \mathcal{H})$ and $K_2 \in \mathcal{L}(\mathcal{H}, \mathbb{C})$. Assume in addition that $G_0 \in \mathcal{L}(\mathbb{C}, \mathcal{H})$ can be chosen such that $G_0 u = g_0 u$ for all $u \in \mathbb{C}$, with $g_0 \in \mathcal{H}$ and $\langle g_0, \theta_n\rangle_{\mathcal{H}} \neq 0$ for each $n \in I$, and such that the following regulator equations are satisfied:

$$S|_{\mathcal{H}}\Pi = \Pi A + G_0 C \quad \text{in } \mathcal{D}(A), \tag{3.40a}$$

$$\Pi B = L \quad \text{in } H \tag{3.40b}$$

for some $\Pi \in \mathcal{L}(Z, \mathcal{H})$ (we point out that by taking adjoints these equations can be solved for the adjoint operator of Π using the methods of [6]). In particular, by the skew-adjointness of $S|_{\mathcal{H}}$, the operators Π and G_0 must be chosen such that for every $n \in I$

$$\Pi^* \theta_n = R(-iw_n, A^*) C^* G_0^* \theta_n, \tag{3.41a}$$

$$L^* \theta_n = B^* R(-iw_n, A^*) C^* G_0^* \theta_n. \tag{3.41b}$$

If, in particular, $L = \delta_0^*$ as in the above, and if $B^* R(-iw_n, A^*) C^* \neq 0$ for all $n \in I$ (i.e., there are no transmission zeros on $\sigma(S|_{\mathcal{H}})$), then suitable operators Π and G_0, with $\langle g_0, \theta_n \rangle_{\mathcal{H}} \neq 0$ for each $n \in I$, are readily found by taking operator adjoints and by using the fact that θ_n is an orthonormal basis for $\mathcal{H}$.

A suitable state feedback controller is then given on the state space $X = \mathcal{H}$ by

$$\dot{x}_0(t) = S|_{\mathcal{H}} x_0(t) + G_0 e(t), \quad x_0(0) \in \mathcal{H}, \tag{3.42a}$$

$$u(t) = K_2 \Pi z(t) + K_2 x_0(t). \tag{3.42b}$$

In fact, the resulting closed loop system operator $\mathcal{A}_0 = \left(\begin{smallmatrix} A + BK_2\Pi & BK_2 \\ G_0 C & S|_{\mathcal{H}} \end{smallmatrix} \right)$ satisfies

$$\begin{pmatrix} I & 0 \\ \Pi & I \end{pmatrix} \begin{pmatrix} A + BK_2\Pi & BK_2 \\ G_0 C & S|_{\mathcal{H}} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\Pi & I \end{pmatrix} = \begin{pmatrix} A & BK_2 \\ 0 & S|_{\mathcal{H}} + LK_2 \end{pmatrix}, \tag{3.43}$$

and hence it is strongly stable on $Z \times X$. Moreover, $\sigma(\mathcal{A}_0) \cap i\mathbb{R} = \emptyset$, because $\sigma(A) \cap i\mathbb{R} = \emptyset$ and $\sigma(S|_{\mathcal{H}} + LK_2) \cap i\mathbb{R} = \emptyset$, so that by the finite-dimensionality of the exogenous system for every $\mathcal{P} \in \mathcal{L}(\mathcal{H}, Z \times X)$ there exists a unique $Y \in \mathcal{L}(\mathcal{H}, Z \times X)$ such that $\operatorname{ran}(Y) \subset \mathcal{D}(\mathcal{A}_0)$ and $YS|_{\mathcal{H}} = \mathcal{A}_0 Y + \mathcal{P}$ in $\mathcal{H}$ [2, 38]. Corollary 3.9 shows that the controller (3.42) achieves conditionally robust output regulation with respect to A, B, C, Π, K_2, and G_0 for all $P \in \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{H}, \mathbb{C})$.

4. Generalized repetitive control for exponentially stable SISO systems. In this section we shall focus on an important application, namely, generalized repetitive control, in which the goal is to achieve strong closed loop stability and asymptotic tracking/rejection of p-periodic reference and disturbance signals for some fixed $p > 0$ and for infinite-dimensional plants. Although such exogenous signals are rather simple, they must be generated using an infinite-dimensional linear exogenous system (2.2); otherwise we would be bound to deal with trigonometric polynomials only. Consequently, the well-known results on output regulation for finite-dimensional exogenous systems, e.g., those in [6], do not apply.

From [15, 22, 39, 42, 43] it is evident that one of the biggest obstacles in the design of repetitive controllers is that the closed loop system is often difficult to stabilize. In fact, Hara et al. [15] proved that repetitive controllers can never be designed (in the classical sense) for strictly proper finite-dimensional plants (e.g., A is exponentially stable and $D = 0$), because the requirement for exponential closed loop stability cannot be satisfied. In [22] the authors solved generalized repetitive control problems also for strictly proper infinite-dimensional systems using type I controllers. However, certain rather severe stabilizability problems are also present in [22]: In particular, for a *fixed* $P \in \mathcal{L}(\mathcal{G}, Z)$ and a *fixed* $Q \in \mathcal{L}(\mathcal{G}, H)$ the existence of $G = \left(\begin{smallmatrix} G_1 \\ G_2 \end{smallmatrix} \right) \in \mathcal{L}(H, Z \times \mathcal{G})$, for which $\left(\begin{smallmatrix} A & P \\ 0 & S|_{\mathcal{G}} \end{smallmatrix} \right) - \left(\begin{smallmatrix} G_1 \\ G_2 \end{smallmatrix} \right) \left(C \ -Q \right)$ generates a strongly stable C_0-semigroup on $Z \times \mathcal{G}$, is not trivial for an infinite-dimensional $\mathcal{G}$. This was required in [22] (see assumption 3 in subsection 3.1 and the discussion in [22]).

In this section we shall show that the above stabilizability problem of [22] can be partially solved using the conditional robustness theory of this paper. We shall design a type I controller for generalized repetitive control in such a way that the closed loop system is stabilized for certain simple operators P and Q contained in the operator F of (2.3) and (3.13) (see also assumption 3 of subsection 3.1). The conditional robustness theory then guarantees that the *same fixed* controller also achieves output regulation for *different* operators P and Q in the exosystem. Thus, robustness allows us—to a degree—to dispense with the problem in [22] of stabilizing the closed loop control system. Moreover, as opposed to [22], here the disturbances can be genuinely unmeasurable: Apart from the fact that we require sufficient continuity for the disturbance operator P (see Theorem 4.6 below), this operator P in the exosystem need not be explicitly known.

We assume the following throughout this section.

ASSUMPTION 4.1. *The plant is a SISO system, i.e., $H = \mathbb{C}$, the operator A generates an exponentially stable C_0-semigroup, $D = 0$, and $\mathcal{G} = H^\alpha_{per}(0,p)$ for some $\alpha > \frac{1}{2}$ (this is the standard Sobolev space of α-differentiable p-periodic functions* [25]*).*

We remark that the assumption that there is no feedthrough in the plant, i.e., $D = 0$, is deliberate: We aim to show that, contrary to the view in [15], conditionally robust regulation of periodic signals is possible even in this case if we do not require exponential closed loop stability but only strong closed loop stability. This is a key difference between the classical repetitive control problems [15, 39, 43] and the generalized repetitive control problems that we study.

We begin the controller design process by stabilizing the exogenous system for an appropriate Q. Recall from [24, 25] that $\mathcal{G} = H^\alpha_{per}(0,p)$ is a Hilbert space; its inner product is denoted by $\langle\cdot,\cdot\rangle$ and it is defined by $\langle f,g\rangle = \sum_{n\in\mathbb{Z}}(1+\omega_n^2)^\alpha \hat{f}_n\overline{\hat{g}_n}$, where $\hat{f}_n$ and $\hat{g}_n$ are ordinary L^2-Fourier coefficients and $\omega_n = \frac{2\pi n}{p}$. Moreover, every $f \in H^\alpha_{per}(0,p)$, $\alpha > \frac{1}{2}$, has the uniformly convergent Fourier series expansion $f(t) = \sum_{n\in\mathbb{Z}} \hat{f}_n e^{i\omega_n t} = \sum_{n\in\mathbb{Z}}\langle f,\phi_n\rangle\phi_n$, where the Fourier coefficients satisfy $\sum_{n\in\mathbb{Z}}(1+\omega_n^2)^\alpha|\hat{f}_n|^2 < \infty$, and $(\phi_n)_{n\in\mathbb{Z}}$ denotes the orthonormal basis of (weighted) exponentials $c_n e^{i\omega_n\cdot} = \frac{e^{i\omega_n\cdot}}{(\sqrt{1+\omega_n^2})^\alpha}$ in $H^\alpha_{per}(0,p)$.

LEMMA 4.2. *The space $\mathcal{G} \overset{s}{\hookrightarrow} BUC(\mathbb{R},\mathbb{C})$ and $T_S(t)|_{\mathcal{G}} f = \sum_{n\in\mathbb{Z}} e^{i\omega_n t}\langle f,\phi_n\rangle\phi_n$ for every $t \in \mathbb{R}$ and every $f \in \mathcal{G}$. Moreover, $S|_{\mathcal{G}}$ is a Riesz-spectral operator on $\mathcal{G}$ and $S|_{\mathcal{G}} = \sum_{n\in\mathbb{Z}} i\omega_n\langle\cdot,\phi_n\rangle\phi_n$, with $\mathcal{D}(S|_{\mathcal{G}}) = \{\, f \in \mathcal{G} \mid \sum_{n\in\mathbb{Z}}\omega_n^2|\langle f,\phi_n\rangle|^2 < \infty \,\}$.*

Proof. It is clear that $\mathcal{G} \hookrightarrow BUC(\mathbb{R},\mathbb{C})$ (see, e.g., Theorem 3.195 in [25]) and that the left shift operators $T_S(t)$ are isometries on $\mathcal{G}$ leaving this space invariant. The (semi)group property and strong continuity of $T_S(t)|_{\mathcal{G}}$ are easy to verify, so that $\mathcal{G} \overset{s}{\hookrightarrow} BUC(\mathbb{R},\mathbb{C})$. Since the elements of the orthonormal basis (ϕ_n) for $\mathcal{G}$ satisfy $\phi_n(\cdot) = c_n e^{i\omega_n\cdot}$ for $c_n = \frac{1}{(\sqrt{1+\omega_n^2})^\alpha}$ and every $n \in \mathbb{Z}$, it is also evident that $T_S(t)|_{\mathcal{G}}\phi_n = e^{i\omega_n t}\phi_n$ for each $t \in \mathbb{R}$. Hence $T_S(t)|_{\mathcal{G}} f = T_S(t)|_{\mathcal{G}}\sum_{n\in Z}\langle f,\phi_n\rangle\phi_n = \sum_{n\in Z}\langle f,\phi_n\rangle T_S(t)|_{\mathcal{G}}\phi_n = \sum_{n\in Z} e^{i\omega_n t}\langle f,\phi_n\rangle\phi_n$ for each $f \in \mathcal{G}$. That $S|_{\mathcal{G}}$ is a Riesz-spectral operator is evident; the spectrum of $S|_{\mathcal{G}}$ consists of the simple eigenvalues $\{\, i\omega_n \mid n \in \mathbb{Z} \,\}$ (whose closure is obviously totally disconnected). Moreover, the corresponding eigenvectors ϕ_n constitute an orthonormal (hence Riesz) basis in $\mathcal{G}$. The expression of $S|_{\mathcal{G}}$ follows from Theorem 2.3.5 in [8]. ∎

We are now ready to prove the following stabilization result, where δ_0 denotes the (bounded and linear) point evaluation operator on $\mathcal{G}$ centered at the origin.

LEMMA 4.3. *Let $\gamma > \alpha + \frac{1}{2}$. Then there exists $L \in \mathcal{L}(\mathbb{C}, \mathcal{G})$ such that*

1. *$S|_{\mathcal{G}} + L\delta_0$ generates a strongly stable C_0-semigroup on $\mathcal{G}$;*
2. *the resolvent satisfies $\|R(i\omega_n, S|_{\mathcal{G}} + L\delta_0)\| \le C'(\sqrt{1+\omega_n^2})^{\gamma}$ for some $C' > 0$ and every $n \in \mathbb{Z}$;*
3. *there exists a unique $l \in \mathcal{G}$ such that $Lu = lu$ for every $u \in \mathbb{C}$ and $\langle l, \phi_n \rangle \neq 0$ for every $n \in \mathbb{Z}$.*

Proof.

1. We shall first employ the theory of Xu and Sallet [41] to find $L \in \mathcal{L}(\mathbb{C}, \mathcal{G})$ such that $S|^*_{\mathcal{G}} + \delta_0^* L^*$ generates a strongly stable C_0-semigroup (here and elsewhere the superscript * denotes the operator adjoint). To this end, we need to show that the standing assumptions $H1 - H3$ in [41] are satisfied.
   First of all, $H1$ amounts to verifying that $S|^*_{\mathcal{G}}$ has compact resolvents and a simple spectrum. Clearly whenever $\lambda > 0$ we have $R(\lambda, S|_{\mathcal{G}}) = \sum_{n\in\mathbb{Z}} \frac{\langle \cdot, \phi_n\rangle \phi_n}{\lambda - i\omega_n}$. Since $\omega_n = \frac{2\pi n}{p}$, the resolvent $R(\lambda, S|_{\mathcal{G}})$ is a uniform limit of finite-rank operators and hence is compact. Moreover, the spectrum of $S|_{\mathcal{G}}$ consists of the simple eigenvalues $i\omega_n$, $n \in \mathbb{Z}$. Since $S|^*_{\mathcal{G}} = -S|_{\mathcal{G}}$, hypothesis $H1$ in [41] is satisfied.
   Second, $H2$ amounts to showing that $\mathcal{D}(S|_{\mathcal{G}})$ is a Hilbert space in the graph norm and that δ_0^* (when interpreted as an input element) satisfies $\delta_0^* \in \mathcal{D}(S|_{\mathcal{G}})'$ (the topological dual of $\mathcal{D}(S|_{\mathcal{G}})$). The first assertion is evident, and the second assertion follows from the boundedness of δ_0^*.
   Finally, in order to meet assumption $H3$ in [41] the eigenvectors ϕ_n of $S|^*_{\mathcal{G}}$ need to form a Riesz basis in $\mathcal{G}$, $\langle \phi_n, \delta_0^* \rangle \neq 0$ for each $n \in \mathbb{Z}$, and there must exist a positive constant M such that

$$\sum_{n\in\mathbb{Z}} \left| \frac{\langle \phi_n, \delta_0^* \rangle}{\lambda + i\omega_n} \right|^2 \le M \quad \text{and} \quad \sup_{m\in\mathbb{Z}} \sum_{n\neq m} \left| \frac{\langle \phi_n, \delta_0^* \rangle}{-i\omega_m + i\omega_n} \right|^2 \le M \tag{4.1}$$

   whenever the distance from λ to $\sigma(S|_{\mathcal{H}})$ is at least $\frac{2\pi}{3p}$. Now the eigenvectors ϕ_n of $S|^*_{\mathcal{G}} = -S|_{\mathcal{G}}$ constitute an orthonormal (hence Riesz) basis in $\mathcal{G}$. Moreover, since $\phi_n(\cdot) = c_n e^{i\omega_n \cdot}$, with $c_n = \frac{1}{(\sqrt{1+\omega_n^2})^{\alpha}}$, for each $f \in \mathcal{G}$ we have $\delta_0 f = f(0) = \sum_{n\in\mathbb{Z}} \langle f, \phi_n \rangle \phi_n(0) = \sum_{n\in\mathbb{Z}} c_n \langle f, \phi_n \rangle = \langle f, \sum_{n\in\mathbb{Z}} c_n \phi_n \rangle$, so that the adjoint of δ_0 satisfies $\delta_0^* u = u \sum_{n\in\mathbb{Z}} c_n \phi_n$ for every $u \in \mathbb{C}$. Obviously then $\langle \phi_n, \delta_0^* \rangle = c_n \neq 0$ for every $n \in \mathbb{Z}$. Moreover, for λ as in the above, we have

$$\sum_{n\in\mathbb{Z}} \left| \frac{\langle \phi_n, \delta_0^* \rangle}{\lambda + i\omega_n} \right|^2 \le \frac{9p^2}{4\pi^2} \sum_{n\in\mathbb{Z}} |c_n|^2 = \frac{9p^2}{4\pi^2} \sum_{n\in\mathbb{Z}} (1+\omega_n^2)^{-\alpha} = M < \infty \tag{4.2}$$

   and

$$\sup_{m\in\mathbb{Z}} \sum_{n\neq m} \left| \frac{\langle \phi_n, \delta_0^* \rangle}{-i\omega_m + i\omega_n} \right|^2 = \sup_{m\in\mathbb{Z}} \sum_{n\neq m} \left| \frac{c_n}{\frac{2\pi i}{p}(n-m)} \right|^2 \le \frac{p^2}{4\pi^2} \sum_{n\in\mathbb{Z}} (1+\omega_n^2)^{-\alpha} < M. \tag{4.3}$$

   In conclusion, the assumptions $H1$–$H3$ in [41] are satisfied.

Let $\gamma > \alpha + \frac{1}{2}$, and set $\mu_n = -i\omega_n - \frac{1}{(\sqrt{1+\omega_n^2})^\gamma}$ for every $n \in \mathbb{Z}$. We next design a feedback $L^* \in \mathcal{L}(\mathcal{G}, \mathbb{C})$ such that it assigns the spectrum $\sigma(S|_{\mathcal{G}}^* - \delta_0^* L^*) = \{\, \mu_n \mid n \in \mathbb{Z} \,\} \subset \{\, z \in \mathbb{C} \mid \Re(z) < 0 \,\}$. If we can accomplish this, then $S|_{\mathcal{G}}^* - \delta_0^* L^*$ generates a strongly stable C_0-semigroup on $\mathcal{G}$. In fact, the operator $S|_{\mathcal{G}}^* + \delta_0^* L^*$ is regular spectral [41]; i.e., it has compact resolvents and its eigenvectors $(\theta_n)_{n\in\mathbb{Z}}$, with $(S|_{\mathcal{G}}^* + \delta_0^* L^*)\theta_n = \mu_n \theta_n$ for every $n \in \mathbb{Z}$, constitute a Riesz basis in $\mathcal{G}$. Hence the C_0-semigroup $T_{S|_{\mathcal{G}}^* + \delta_0^* L^*}(t)$ is uniformly bounded; strong stability follows from the Arendt–Batty–Lyubich–Vu theorem (see, e.g., Theorem V.2.21 in [12]) because $S|_{\mathcal{G}}^* + \delta_0^* L^*$ has no spectrum on $i\mathbb{R}$. In order to show that such an L^* indeed exists, we aim to apply Theorem 1 in [41]. To this end, we observe that the assigned spectral points μ_n satisfy the necessary condition (3) on page 522 of [41]:

$$\sum_{n\in\mathbb{Z}} \left| \frac{\mu_n + i\omega_n}{\langle \phi_n, \delta_0^* \rangle} \right|^2 = \sum_{n\in\mathbb{Z}} \left| \frac{1}{c_n} \frac{1}{\left(\sqrt{1+\omega_n^2}\right)^\gamma} \right|^2 = \sum_{n\in\mathbb{Z}} \left| \left(\sqrt{1+\omega_n^2}\right)^{\alpha-\gamma} \right|^2 = \sum_{n\in\mathbb{Z}} (1+\omega_n^2)^{\alpha-\gamma} < \infty.$$

Theorem 1 in [41] now implies that the claim follows for the particular choice of $L^* = \langle \cdot, l \rangle$, where $l \in \mathcal{G}$ is given by

(4.4)
$$l = \sum_{n\in\mathbb{Z}} l_n \phi_n, \quad \text{where } \overline{l_n} = -\frac{\left(\sqrt{1+\omega_n^2}\right)^\alpha}{\left(\sqrt{1+\omega_n^2}\right)^\gamma} \prod_{k=-\infty, k\neq n}^{\infty} \frac{-i\omega_n + i\omega_k + \left(\sqrt{1+\omega_k^2}\right)^{-\gamma}}{-i\omega_n + i\omega_k}.$$

Our final task in item 1 is to show that the adjoint $S|_{\mathcal{G}} + L\delta_0$ is also the generator of a strongly stable C_0-semigroup on $\mathcal{G}$. Clearly $S|_{\mathcal{G}} + L\delta_0$ generates at least a weakly stable C_0-semigroup on $\mathcal{G}$. Since $S|_{\mathcal{G}}$ has compact resolvents, so has $S|_{\mathcal{G}} + L\delta_0$ (see [12, p. 159]). For such semigroup generators, weak stability implies strong stability (see Proposition 3.21 in [28]).

2. According to Theorem 1 in [41] the operator $S|_{\mathcal{G}}^* + \delta_0^* L^*$ above is regular spectral; i.e., it has compact resolvents and its eigenvectors $(\theta_n)_{n\in\mathbb{Z}}$, with $(S|_{\mathcal{G}}^* + \delta_0^* L^*)\theta_n = \mu_n \theta_n$ for every $n \in \mathbb{Z}$, constitute a Riesz basis in $\mathcal{G}$. We remark that the eigenvalues $(\mu_n)_{n\in\mathbb{Z}}$ of $S|_{\mathcal{G}}^* + \delta_0^* L^*$ need not be simple in general. Let $(\psi_n)_{n\in\mathbb{Z}}$ denote the sequence in $\mathcal{G}$ which is biorthogonal to $(\theta_n)_{n\in\mathbb{Z}}$ [8]. Whenever $\lambda \neq \mu_n$ for each $n \in \mathbb{Z}$ we have $(\lambda I - S|_{\mathcal{G}}^* - \delta_0^* L^*)\theta_n = (\lambda - \mu_n)\theta_n$, whence $R(\lambda, S|_{\mathcal{G}}^* + \delta_0^* L^*)\theta_n = \frac{1}{\lambda - \mu_n}\theta_n$, and so for every such $\lambda \in \mathbb{C}$,

$$R(\lambda, S|_{\mathcal{G}}^* + \delta_0^* L^*) f = \sum_{n\in\mathbb{Z}} \frac{\langle f, \psi_n \rangle \theta_n}{\lambda - \mu_n} \quad \forall f \in \mathcal{G}. \tag{4.5}$$

For every $m \in \mathbb{Z}$ we may now estimate using the well-known properties of Riesz bases [8] as follows:

(4.6)

$$\|R(-i\omega_m, S|_{\mathcal{G}}^* + \delta_0^* L^*)f\|^2 = \left\| \sum_{n\in\mathbb{Z}} \frac{\langle f, \psi_n\rangle \theta_n}{-i\omega_m - \mu_n} \right\|^2 \le D' \sum_{n\in\mathbb{Z}} \left| \frac{\langle f, \psi_n\rangle}{-i\omega_m - \mu_n} \right|^2$$

$$= D' \left[\sum_{n\in I_m} \left| \frac{\langle f, \psi_n\rangle}{-i\omega_m - \mu_m} \right|^2 + \sum_{n\notin I_m} \left| \frac{\langle f, \psi_n\rangle}{-i\omega_m - \mu_n} \right|^2 \right] \tag{4.7}$$

$$\le D' \left[(1+\omega_m^2)^\gamma \sum_{n\in\mathbb{Z}} |\langle f, \psi_n\rangle|^2 + c' \sum_{n\in\mathbb{Z}} |\langle f, \psi_n\rangle|^2 \right] \tag{4.8}$$

$$\le C'^2 (1+\omega_m^2)^\gamma \|f\|^2, \tag{4.9}$$

where I_m is the finite (cf. Corollary IV.1.19 in [12]) multiplicity of the eigenvalue μ_m, and c', D', C' are positive constants. This shows that $\|R(-i\omega_n, S|_{\mathcal{G}}^* + \delta_0^* L^*)\| \le C' \left(\sqrt{1+\omega_n^2}\right)^\gamma$ for each $n \in \mathbb{Z}$. But according to Lemmas A.3.65 and A.3.60 in [8], $\|R(i\omega_n, S|_{\mathcal{G}} + L\delta_0)\| = \|R(-i\omega_n, S|_{\mathcal{G}}^* + \delta_0^* L^*)\| \le C' \left(\sqrt{1+\omega_n^2}\right)^\gamma$ for each $n \in \mathbb{Z}$.

3. If we choose L^* according to (4.4), then evidently $Lu = lu$ for the unique $l \in \mathcal{G}$. Moreover, $\langle \phi_n, l\rangle = \overline{l_n}$ for each $n \in \mathbb{Z}$. Now by Theorem 1 in [41] we have

$$\lim_{M,N\to\infty} \frac{\mu_n + i\omega_n}{c_n} \prod_{k=-M, k\ne n}^{N} \frac{-i\omega_n + i\omega_k + \frac{1}{\left(l\sqrt{1+\omega_k^2}\right)^\gamma}}{-i\omega_n + i\omega_k} \tag{4.10}$$

$$= \frac{\mu_n + i\omega_n}{c_n} \lim_{M,N\to\infty} \prod_{k=-M, k\ne n}^{N} \left[1 + \frac{1}{i(\omega_k - \omega_n)\left(\sqrt{1+\omega_k^2}\right)^\gamma} \right]$$

$$= \overline{l_n}. \tag{4.11}$$

Since

$$\left| \prod_{k=-M, k\ne n}^{N} \left[1 + \frac{1}{i(\omega_k - \omega_n)\left(\sqrt{1+\omega_k^2}\right)^\gamma} \right] \right|$$

$$= \prod_{k=-M, k\ne n}^{N} \left| 1 - \frac{i}{(\omega_k - \omega_n)\left(\sqrt{1+\omega_k^2}\right)^\gamma} \right| > 1$$

for every $M, N \in \mathbb{N}$, the above infinite product cannot converge to 0. Hence we have $l_n \ne 0$ for each $n \in \mathbb{Z}$. □

Remark 4.4. The stabilizing operator L is actually constructed in the above proof; see (4.4).

Remark 4.5. Strong stability is the best we can achieve in Lemma 4.3. In fact, since $S|_{\mathcal{G}}$ generates an isometric C_0-group on an infinite-dimensional space, $S|_{\mathcal{G}} + \Delta$ is not exponentially stable for any compact perturbation $\Delta \in \mathcal{L}(\mathcal{G})$ (cf. Corollary 3.58 in [28]).

The following is the main result of this section.

THEOREM 4.6. *Let $\mathcal{G}$ and L be as in Lemma 4.3. Let $\beta > \gamma + \alpha + \epsilon$ for arbitrary $\epsilon > \frac{1}{2}$ and set $\mathcal{H} = H^\beta_{per}(0, p)$. Assume that there exist $\Pi \in \mathcal{L}(\mathcal{G}, Z)$ and $\Gamma \in \mathcal{L}(\mathcal{G}, \mathbb{C})$ such that $\Pi(\mathcal{D}(S|_{\mathcal{G}})) \subset \mathcal{D}(A)$ and the following (feedforward) regulator equations are*

satisfied:

$$A\Pi + B\Gamma = \Pi S|_{\mathcal{G}} \quad \text{in } \mathcal{D}(S|_{\mathcal{G}}), \tag{4.12a}$$

$$C\Pi = \delta_0 \quad \text{in } \mathcal{G}. \tag{4.12b}$$

Let $X = Z \times \mathcal{G}$, *and set* $F = \left(\begin{smallmatrix} A & B\Gamma \\ -LC & S|_{\mathcal{G}} + L\delta_0 \end{smallmatrix}\right)$, $J = (0\ \Gamma)$, *and* $G = \left(\begin{smallmatrix} 0 \\ L \end{smallmatrix}\right)$. *Then this type* I *controller solves the EFRP for every* $P \in \mathcal{L}(\mathcal{G}, Z)$ *and each* $Q \in \mathcal{L}(\mathcal{G}, \mathbb{C})$.

Proof. Consider the closed loop operator

$$\mathcal{A} = \begin{pmatrix} A & BJ \\ GC & F \end{pmatrix} = \begin{pmatrix} A & 0 & B\Gamma \\ 0 & A & B\Gamma \\ LC & -LC & S|_{\mathcal{G}} + L\delta_0 \end{pmatrix}. \tag{4.13}$$

Applying the similarity transform U given as

$$U = \begin{pmatrix} I & 0 & 0 \\ I & -I & 0 \\ 0 & 0 & -I \end{pmatrix} \tag{4.14}$$

on $Z \times X = Z \times Z \times \mathcal{G}$ to $\mathcal{A}$ we obtain the operator $\widetilde{\mathcal{A}} = U\mathcal{A}U$ having the expression

$$\widetilde{\mathcal{A}} = \begin{pmatrix} A & 0 & -B\Gamma \\ 0 & A & 0 \\ 0 & -LC & S|_{\mathcal{G}} + L\delta_0 \end{pmatrix} := \begin{pmatrix} A & \Delta \\ 0 & \mathcal{A}_s \end{pmatrix}. \tag{4.15}$$

By Lemma 4.3, $S|_{\mathcal{G}} + L\delta_0$ generates a strongly stable C_0-semigroup on $\mathcal{G}$. Since in addition A generates an exponentially stable C_0-semigroup on Z, the semigroup generated by $\mathcal{A}_s$ on $Z \times \mathcal{G}$ is strongly stable (use, e.g., Proposition 5.6.1 in [1]). The same reasoning shows that the C_0-semigroup generated by $\widetilde{\mathcal{A}}$ on $Z \times X$ is strongly stable. Consequently so is the similar semigroup generated by $\mathcal{A}$ on $Z \times X$.

Let $\mathcal{P} \in \mathcal{L}(\mathcal{G}, Z \times Z \times \mathcal{G}) \subset \mathcal{L}(\mathcal{H}, Z \times Z \times \mathcal{G})$ be arbitrary. We show that the equation $\Pi S|_{\mathcal{H}} = \widetilde{\mathcal{A}}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ has a solution $\Pi \in \mathcal{L}(\mathcal{H}, Z \times Z \times \mathcal{G})$ and apply Theorem 3.2 to obtain the desired result. Now $\Pi S|_{\mathcal{H}} = \widetilde{\mathcal{A}}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ reads

$$\begin{pmatrix} \Pi_1 \\ \Pi_2 \\ \Pi_3 \end{pmatrix} S|_{\mathcal{H}} = \begin{pmatrix} A & 0 & -B\Gamma \\ 0 & A & 0 \\ 0 & -LC & S|_{\mathcal{G}} + L\delta_0 \end{pmatrix} \begin{pmatrix} \Pi_1 \\ \Pi_2 \\ \Pi_3 \end{pmatrix} + \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} \quad \text{in } \mathcal{D}(S|_{\mathcal{H}}), \tag{4.16}$$

where Π_i, $i = 1, 2, 3$, denote the components of Π and P_i, $i = 1, 2, 3$, denote the components of $\mathcal{P}$. Now $P_2 \in \mathcal{L}(\mathcal{G}, Z) \subset \mathcal{L}(\mathcal{H}, Z)$, and by exponential stability of A there exists a unique $\Pi_2 \in \mathcal{L}(\mathcal{G}, Z)$ for which $\Pi_2 S|_{\mathcal{G}} = A\Pi_2 + P_2$ in $\mathcal{D}(S|_{\mathcal{G}})$. Since $\mathcal{H} \hookrightarrow \mathcal{G}$, the same operator $\Pi_2 \in \mathcal{L}(\mathcal{H}, Z)$ satisfies $\Pi_2 S|_{\mathcal{H}} = A\Pi_2 + P_2$ in $\mathcal{D}(S|_{\mathcal{H}})$. It suffices to show that $\Pi_3 \in \mathcal{L}(\mathcal{H}, \mathcal{G})$ exists such that $\Pi_3 S|_{\mathcal{H}} = (S|_{\mathcal{G}} + L\delta_0)\Pi_3 + P_3 - LC\Pi_2$ in $\mathcal{D}(S|_{\mathcal{H}})$, because then exponential stability of A also guarantees the existence and uniqueness of $\Pi_1 \in \mathcal{L}(\mathcal{H}, Z)$ such that $\Pi_1 S|_{\mathcal{H}} = A\Pi_1 + P_1 - B\Gamma\Pi_3$ in $\mathcal{D}(S|_{\mathcal{H}})$.

Let $\Delta = P_3 - LC\Pi_2 \in \mathcal{L}(\mathcal{G})$. Recall that (ϕ_n) is an orthonormal basis for $\mathcal{G}$. By Lemmas 4.2 and 4.3, $S|_{\mathcal{H}}\phi_n = S|_{\mathcal{G}}\phi_n = i\omega_n\phi_n$ for every $n \in \mathbb{Z}$ and $\sigma(S|_{\mathcal{G}} + L\delta_0) \cap i\mathbb{R} = \emptyset$. Hence we have for each $n \in \mathbb{Z}$ that $\Pi_3 S|_{\mathcal{H}}\phi_n = (S|_{\mathcal{G}} + L\delta_0)\Pi_3\phi_n + \Delta\phi_n$ if and only if $\Pi_3\phi_n = R(i\omega_n, S|_{\mathcal{G}} + L\delta_0)\Delta\phi_n$. Define the linear operator $\Pi_3 : \mathcal{D}(\Pi_3) \subset \mathcal{G} \to \mathcal{G}$ by

$$\Pi_3 f = \sum_{n \in \mathbb{Z}} \langle f, \phi_n \rangle R(i\omega_n, S|_{\mathcal{G}} + L\delta_0)\Delta\phi_n \quad \forall f \in \mathcal{D}(\Pi_3) = \mathcal{H}. \tag{4.17}$$

Now for some $C', k > 0$, and every $f \in \mathcal{H}$ we have by the Schwartz inequality and Lemma 4.3 that

$$\|\Pi_3 f\|_{\mathcal{G}} \le \sum_{n\in\mathbb{Z}} |\langle f, \phi_n\rangle| \|R(i\omega_n, S|_{\mathcal{G}} + L\delta_0)\|_{\mathcal{L}(\mathcal{G})} \|\Delta\|_{\mathcal{L}(\mathcal{G})} \|\phi_n\|_{\mathcal{G}} \tag{4.18}$$

$$= \|\Delta\|_{\mathcal{L}(\mathcal{G})} \sum_{n\in\mathbb{Z}} |\hat{f}_n| \Big(\sqrt{1+\omega_n^2}\Big)^{\alpha} \|R(i\omega_n, S|_{\mathcal{G}} + L\delta_0)\|_{\mathcal{L}(\mathcal{G})} \tag{4.19}$$

$$\le C' \|\Delta\|_{\mathcal{L}(\mathcal{G})} \sum_{n\in\mathbb{Z}} |\hat{f}_n| \Big(\sqrt{1+\omega_n^2}\Big)^{\alpha} \Big(\sqrt{1+\omega_n^2}\Big)^{\gamma} \tag{4.20}$$

$$\le C' \|\Delta\|_{\mathcal{L}(\mathcal{G})} \sqrt{\sum_{n\in\mathbb{Z}} |\hat{f}_n|^2 (1+\omega_n^2)^{\beta}} \sqrt{\sum_{n\in\mathbb{Z}} \frac{(1+\omega_n^2)^{\gamma+\alpha}}{(1+\omega_n^2)^{\beta}}} < k\|f\|_{\mathcal{H}}, \tag{4.21}$$

whence $\Pi_3 \in \mathcal{L}(\mathcal{H}, \mathcal{G})$. Using the closedness of $S|_{\mathcal{G}} + L\delta_0$ and arguments similar to the above it is straightforward to show that $\Pi_3(\mathcal{D}(S|_{\mathcal{H}})) \subset \mathcal{D}(S|_{\mathcal{G}} + L\delta_0)$ and that $\Pi_3 S|_{\mathcal{H}} f = (S|_{\mathcal{G}} + L\delta_0)\Pi_3 f + (P_3 - LC\Pi_2)f$ for each $f \in \mathcal{D}(S|_{\mathcal{H}})$.

By the above, the equation $\Pi S|_{\mathcal{H}} = \widetilde{\mathcal{A}}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ has a solution for every $\mathcal{P} \in \mathcal{L}(\mathcal{G}, Z \times Z \times \mathcal{G})$. But then also the equation $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ has a solution for every such $\mathcal{P}$, because $\widetilde{\mathcal{A}} = U\mathcal{A}U$ and U is boundedly invertible. Now for each $P \in \mathcal{L}(\mathcal{G}, Z) \subset \mathcal{L}(\mathcal{H}, Z)$ and $Q \in \mathcal{L}(\mathcal{G}, \mathbb{C}) \subset \mathcal{L}(\mathcal{H}, \mathbb{C})$ we have $\mathcal{P} = \left(\begin{smallmatrix} P \\ -GQ \end{smallmatrix}\right) \in \mathcal{L}(\mathcal{G}, Z \times X)$, and so condition 1 in Theorem 3.2 is satisfied for $\mathcal{A}^p = \mathcal{A}$.

In order to be able to apply Theorem 3.2, we still need to verify that the controller has the internal model structure. This is easy to do by mimicking the proof of Theorem 3.4 together with property 3 in Lemma 4.3. □

Remark 4.7. It is crucial in the above proof that we appeal only to Theorem 3.2 to achieve output regulation if the closed loop system contains a copy of the exogenous system operator, which is stabilized on a larger space $\mathcal{G} = H^{\alpha}_{per}(0, p)$ than the space $\mathcal{H} = H^{\beta}_{per}(0, p)$ on which we require output regulation. This peculiarity is purely a consequence of the lack of exponential closed loop stability.

Remark 4.8. It has been shown in [24] that the solvability of the regulator equations (4.12) is a necessary and sufficient condition for the existence of a simpler feedforward controller achieving output regulation in the disturbance-free case. The results of [21, 24] also immediately show that under the assumptions of this section and under the assumption that the plant does not have transmission zeros on $\sigma(S|_{\mathcal{G}})$ the regulator equations (4.12) have solutions $\Pi \in \mathcal{L}(\mathcal{G}, Z)$ and $\Gamma \in \mathcal{L}(\mathcal{G}, \mathbb{C})$ if and only if the periodic signals in $\mathcal{G}$ are smooth enough to satisfy $\sum_{n\in\mathbb{Z}} |H(i\omega_n)|^{-2}(1 + \omega_n^2)^{-\alpha} < \infty$. Explicit solutions of the regulator equations (4.12) for this case have been constructed in [24]; see also Example 4.10.

Remark 4.9. It is easy to see from the above proof that for each $\mathcal{P} \in \mathcal{L}(\mathcal{G}, Z \times X)$ the solution of $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ in $\mathcal{D}(S|_{\mathcal{H}})$ is actually unique. Based on the arguments in Theorem 3.2 we suspect that the above controller also possesses a degree of conditional robustness with respect to some perturbations in the parameters of the plant and the controller. However, we cannot conclude conditional robustness directly using Theorem 3.2. This is because $0 \in \overline{\sigma(S|_{\mathcal{G}} + L\delta_0) - \sigma(S|_{\mathcal{H}})}$, which, by [2], implies that the equation $\Pi_3 S|_{\mathcal{H}} = (S|_{\mathcal{G}} + L\delta_0)\Pi_3 + \Delta$ in $\mathcal{D}(S|_{\mathcal{H}})$ does not have a unique solution for each $\Delta \in \mathcal{L}(\mathcal{H}, \mathcal{G})$. Hence we can solve only the equation $\Pi S|_{\mathcal{H}} = \mathcal{A}\Pi + \mathcal{P}$ for $\mathcal{P} \in \mathcal{L}(\mathcal{G}, Z \times X) \subset \mathcal{L}(\mathcal{H}, Z \times X)$.

We conclude this section with an example which illustrates the use of Theorem 4.6.

Example 4.10. Let $a > 0$, $r \neq 0$, $\tau_1 > \tau_2 > 0$, and consider the scalar delay

differential equation

$$\dot{x}(t) = -ax(t) - b\big[x(t-\tau_1) + x(t-\tau_2)\big] + u(t) + \mathcal{U}_{dist}(t), \tag{4.22a}$$

$$y(t) = rx(t), \quad t \geq 0, \tag{4.22b}$$

which was also studied in [22, 24] for possibly nonrobust controllers. As in [22, 24], our aim here is to study asymptotic tracking of reference signals in the Sobolev spaces $\mathcal{H} = H^\beta_{per}(0,p)$ for some $p > 0$ and $\beta > \frac{1}{2}$. However, here we can utilize Theorem 4.6 to also obtain a degree of conditional robustness for large enough values of β, i.e., for sufficiently smooth reference signals.

Taking into account initial conditions for $x(\cdot)$, the pair (4.22) can be formulated as a plant of the form (2.1) in which $D = 0$ [8]. Moreover, it can be shown (see, e.g., [8, Lem. 4.3.9]) that the transfer function $H(s) = CR(s,A)B$ of this plant is given by

$$H(s) = \frac{r}{s + a + b(e^{-s\tau_1} + e^{-s\tau_2})} \tag{4.23}$$

for those $s \in \mathbb{C}$ at which the denominator is not equal to zero.

The semigroup generated by A is exponentially stable if and only if $s+a+b(e^{-s\tau_1} + e^{-s\tau_2}) \neq 0$ for all $s \in \{\, z \in \mathbb{C} \mid \Re(z) \geq 0 \,\}$ [8, Thm. 5.1.7]. Ruan and Wei [33] have given a complete characterization (in terms of a, b, τ_1, and τ_2) of those instances in which all roots of equation $s + a + b(e^{-s\tau_1} + e^{-s\tau_2}) = 0$ have negative real parts. In their characterization, the parameter b lies on an interval (b_0^-, b_0^+). We assume that the semigroup generated by A is exponentially stable. By the above discussion, we then have $i\omega_n = i\frac{2\pi n}{p} \in \rho(A)$ and $H(i\omega_n) \neq 0$ for every $n \in \mathbb{Z}$.

Let $\mathcal{G} = H^\alpha_{per}(0,p)$ for $\alpha = \frac{5}{3}$ (we need $\alpha > \frac{3}{2}$ by the results of [24]). We can now easily solve the feedforward regulator equations (4.12) for $\Pi \in \mathcal{L}(\mathcal{G}, Z)$ and $\Gamma \in \mathcal{L}(\mathcal{G}, \mathbb{C})$ using the method of [24]. In fact, clearly $\sum_{n=-\infty}^{\infty} |H(i\omega_n)^{-1}|^2(1+\omega_n^2)^{-\alpha} < \infty$. Then by the Schwartz inequality, exponential stability of A, and the Riemann–Lebesgue lemma (cf. [24]) the following operators are in $\mathcal{L}(\mathcal{G}, \mathbb{C})$ and $\mathcal{L}(\mathcal{G}, Z)$, respectively:

$$\Gamma y = \sum_{n\in\mathbb{Z}} y_n \frac{\delta_0 \phi_n}{H(i\omega_n)} = \sum_{n\in\mathbb{Z}} \frac{\hat{y}_n}{H(i\omega_n)} \quad \forall y \in \mathcal{G}, \tag{4.24}$$

$$\Pi y = \sum_{n\in\mathbb{Z}} y_n R(i\omega_n, A)B\Gamma\phi_n = \sum_{n\in\mathbb{Z}} \hat{y}_n \frac{R(i\omega_n, A)B}{H(i\omega_n)} \quad \forall y \in \mathcal{G}. \tag{4.25}$$

Here $(\phi_n)_{n\in\mathbb{Z}}$ is the natural orthonormal basis of weighted exponentials for $\mathcal{G}$, y_n is the nth Fourier coefficient of $y \in \mathcal{G}$ with respect to this basis, and $\hat{y}_n$ is the nth L^2-Fourier coefficient of y (recall that $\sum_{n=-\infty}^{\infty} |\hat{y}_n|^2(1+\omega_n^2)^\alpha < \infty$). It is straight-forward to verify using the unique Fourier series description of the elements in $\mathcal{G}$ that these two operators Π and Γ also satisfy the regulator equations (4.12) (see [24] for more details).

Let the stabilizing operator L be as in Lemma 4.3 for $\gamma = \alpha+1$. Then by Theorem 4.6 the dynamic type I controller (2.3) with $X = Z \times \mathcal{G}$ and with parameters

$$F = \begin{pmatrix} A & B\Gamma \\ -LC & S|_{\mathcal{G}} + L\delta_0 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & \Gamma \end{pmatrix}, \quad \text{and} \quad G = \begin{pmatrix} 0 \\ L \end{pmatrix} \tag{4.26}$$

achieves the asymptotic tracking of all reference signals in $\mathcal{H} = H^\beta_{per}(0,p)$ for (say) $\beta > 5$ (take $Q = \delta_0 \in \mathcal{L}(\mathcal{H}, \mathbb{C}) \cap \mathcal{L}(\mathcal{G}, \mathbb{C})$ in Theorem 4.6 and argue as in subsection

2.2 to see this). Moreover, here we simultaneously also achieve rejection of possibly unknown (i.e., P need not be known) p-periodic disturbances, whereas only tracking was achieved in [22].

5. Conclusions and future work. In this paper we showed that two controller types encountered frequently in practice achieve conditionally robust output regulation by employing the so-called internal model structure, a concept introduced recently in another paper. Our results are valid also for infinite-dimensional plants, controllers, and exogenous systems, and we have presented an example of such a general situation, namely, repetitive control of an exponentially stable SISO plant. The general conditional robustness theory for the two controller types is new for infinite-dimensional systems, but it also unifies and simplifies the treatment of these controllers in the finite-dimensional case.

We have restricted our attention to the case in which the plant has bounded control and observation operators. In the future it would be interesting to see if the results of this paper can also be generalized for unbounded control and observation operators in the plant. This would allow for boundary control and observation. On the other hand, it is well known that unbounded feedback can yield better stabilizability than bounded feedback; in particular, it is sometimes possible to uniformly shift an infinite number of eigenvalues from the imaginary axis to the open left half plane. Consequently, by allowing for unbounded control/observation operators also in the controller it might be possible to obtain better closed loop stability properties and thus better robustness properties. This would be especially useful in repetitive control, where exponential stabilization of the closed loop system using only compact feedback operators is impossible. Other interesting directions for future research include unbounded perturbations in the system parameters, and quantitative robustness analysis describing how small the perturbations must be in order to guarantee conditionally robust output regulation.

Acknowledgments. The author is indebted to Prof. Seppo Pohjolainen and Dr. Timo Hämäläinen for numerous helpful discussions concerning this article. The author also thanks the anonymous referees for their constructive criticism and useful suggestions for improving the paper.

REFERENCES

[1] W. Arendt, C. Batty, M. Hieber, and F. Neubrander, *Vector-Valued Laplace Transforms and Cauchy Problems*, Birkhäuser Verlag, Basel, 2001.

[2] W. Arendt, F. Räbiger, and A. Sourour, *Spectral properties of the operator equation $AX + XB = Y$*, Quart. J. Math. Oxford Ser. (2), 45 (1994), pp. 133–149.

[3] R. Bhatia and P. Rosenthal, *How and why to solve the operator equation $AX - XB = Y$*, Bull. London Math. Soc., 29 (1997), pp. 1–21.

[4] C. Byrnes, D. Gilliam, V. Shubov, and J. Hood, *An example of output regulation for a distributed parameter system with infinite dimensional exosystem*, in Electronic Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems, http://www.nd.edu/∼mtns/papers/22618_2.pdf (2002).

[5] C. Byrnes, D. Gilliam, V. Shubov, and J. Hood, *Examples of output regulation for distributed parameter systems with infinite dimensional exosystem*, in Proceedings of the 40th IEEE Conference on Decision and Control, Vol. 1, 2001, pp. 547–548.

[6] C. Byrnes, I. Laukó, D. Gilliam, and V. Shubov, *Output regulation for linear distributed parameter systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 2236–2252.

[7] V. Casarino and S. Piazzera, *On the stability of asymptotic properties of perturbed C_0-semigroups*, Forum Math., 13 (2001), pp. 91–107.

[8] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
[9] R. Curtain and H. Zwart, *Riccati equations and normalized coprime factorizations for strongly stabilizable infinite-dimensional systems*, Systems Control Lett., 28 (1996), pp. 11–22.
[10] E. Davison, *The robust control of a servomechanism problem for linear time-invariant multivariable systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 25–34.
[11] E. Davison and A. Goldenberg, *Robust control of a general servomechanism problem: The servo compensator*, Automatica, 11 (1975), pp. 461–471.
[12] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.
[13] B. A. Francis, *The linear multivariable regulator problem*, SIAM J. Control Optim., 15 (1977), pp. 486–505.
[14] B. Francis and W. Wonham, *The internal model principle of control theory*, Automatica, 12 (1976), pp. 457–465.
[15] S. Hara, Y. Yamamoto, T. Omata, and M. Nakano, *Repetitive control system: A new type servo system for periodic exogenous signals*, IEEE Trans. Automat. Control, 33 (1988), pp. 659–668.
[16] J. Huang, *A simple proof of the robust linear regulator*, in Proceedings of the American Control Conference, IEEE, Piscataway, NJ, 1995, pp. 3382–3383.
[17] T. Hämalainen and S. Pohjolainen, *A finite-dimensional robust controller for systems in the CD-algebra*, IEEE Trans. Automat. Control, 45 (2000), pp. 421–431.
[18] T. Hämäläinen and S. Pohjolainen, *Robust control and tuning problem for distributed parameter systems*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 479–500.
[19] T. Hämäläinen, *Robust Low-Gain Regulation of Stable Infinite-Dimensional Systems*, Doctoral dissertation, Tampere University of Technology, Tampere, Finland, 2005.
[20] E. Immonen, *Robust state space output regulation theory for linear infinite-dimensional systems*, SIAM J. Control Optim., submitted.
[21] E. Immonen, *Some properties of infinite-dimensional systems capable of asymptotically tracking bounded uniformly continuous signals*, Math. Control Signals Systems, 18 (2006), pp. 323–344.
[22] E. Immonen and S. Pohjolainen, *Feedback and feedforward output regulation of bounded uniformly continuous signals for infinite-dimensional systems*, SIAM J. Control Optim., 45 (2006), pp. 1714–1735.
[23] E. Immonen and S. Pohjolainen, *Output regulation of periodic signals for DPS: An infinite-dimensional signal generator*, IEEE Trans. Automat. Control, 50 (2005), pp. 1799–1804.
[24] E. Immonen and S. Pohjolainen, *What periodic signals can an exponentially stabilizable linear feedforward control system asymptotically track?*, SIAM J. Control Optim., 44 (2006), pp. 2253–2268.
[25] R. Iorio and V. Iorio, *Fourier Analysis and Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 2001.
[26] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[27] H. Kwatny and K. Kalnitsky, *On alternative methodologies for the design of robust linear multivariable regulators*, IEEE Trans. Automat. Control, 23 (1978), pp. 930–933.
[28] Z-H. Luo, B-Z. Guo, and O. Morgul, *Stability and Stabilization of Infinite Dimensional Systems with Applications*, Springer-Verlag, London, 1999.
[29] J. Oostveen, *Strongly Stabilizable Infinite-Dimensional Systems*, Ph.D. Thesis, University of Groningen, Groningen, Germany, 1999.
[30] S. Pohjolainen, *Robust controller for systems with exponentially stable strongly continuous semigroups*, J. Math. Anal. Appl., 111 (1985), pp. 622–636.
[31] S. Pohjolainen, *Robust multivariable PI-controller for infinite dimensional systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 17–30.
[32] R. Rebarber and G. Weiss, *Internal model based tracking and disturbance rejection for stable well-posed systems*, Automatica, 39 (2003), pp. 1555–1569.
[33] S. Ruan and J. Wei, *On The zeros of transcendental functions with applications to stability of delay differential equations with two delays*, Dynam. Contin. Discrete Impuls. Systems, 10 (2003), pp. 863–874.
[34] J. Schumacher, *Finite-dimensional regulators for a class of infinite-dimensional systems*, Systems Control Lett., 3 (1983), pp. 7–12.
[35] O. Sebakhy and W. Wonham, *A design procedure for multivariable regulators*, Automatica, 12 (1976), pp. 467–478.

[36] H. Ukai and T. Iwazumi, *General servomechanism problems for distributed parameter systems*, Internat. J. Control, 42 (1985), pp. 1195–1212.

[37] Q. P. Vu, *The operator equation $AX - XB = C$ with unbounded operators A and B and related abstract Cauchy problems*, Math. Z., 208 (1991), pp. 567–588.

[38] Q. P. Vu and E. Schüler, *The operator equation $AX-XB = C$, admissibility, and asymptotic behavior of differential equations*, J. Differential Equations, 145 (1998), pp. 394–419.

[39] G. Weiss, and M. Häfele, *Repetitive control of MIMO systems using H^∞ design*, Automatica, 35 (1999), pp. 1185–1199.

[40] W. Wonham, *Linear Multivariable Control. A Geometric Approach*, Springer-Verlag, New York, 1979.

[41] C.-Z. Xu and G. Sallet, *On spectrum and Riesz basis assignment of infinite-dimensional linear systems by bounded linear feedbacks*, SIAM J. Control Optim., 34 (1996), pp. 521–541.

[42] Y. Yamamoto, *Learning control and related problems in infinite-dimensional systems*, in Essays on Control: Perspectives in the Theory and Its Applications, H. Trentelman and J. Willems, eds., Birkhäuser Boston, MA, 1993, pp. 191–222.

[43] Y. Yamamoto and S. Hara, *Relationships between internal and external stability for infinite-dimensional systems with applications to a servo problem*, IEEE Trans. Automat. Control, 33 (1988), pp. 1044–1052.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2094–2118

WEIGHTED ADMISSIBILITY AND WELLPOSEDNESS OF LINEAR SYSTEMS IN BANACH SPACES*

BERNHARD H. HAAK† AND PEER CHR. KUNSTMANN†

Abstract. We study linear control systems in infinite-dimensional Banach spaces governed by analytic semigroups. For $p \in [1, \infty]$ and $\alpha \in \mathbb{R}$ we introduce the notion of L^p-admissibility of type α for unbounded observation and control operators. Generalizing earlier work by Le Merdy [*J. London Math. Soc.* (2), 67 (2003), pp. 715–738] and Haak and Le Merdy [*Houston J. Math.*, 31 (2005), pp. 1153–1167], we give conditions under which L^p-admissibility of type α is characterized by boundedness conditions which are similar to those in the well-known Weiss conjecture. We also study L^p-wellposedness of type α for the full system. Here we use recent ideas due to Prüss and Simonett [*Arch. Math. (Basel)*, 82 (2004), pp. 415–431]. Our results are illustrated by a controlled heat equation with boundary control and boundary observation where we take Lebesgue and Besov spaces as state space. This extends the considerations in [C. I. Byrnes et al., *J. Dynam. Control Systems*, 8 (2002), pp. 341–370] to non-Hilbertian settings and to $p \neq 2$.

Key words. control theory, linear systems, admissibility, H^∞-calculus, square-function estimates

AMS subject classifications. 93C05, 47D06, 47A60, 47A10

DOI. 10.1137/060656139

1. Introduction and main theorems.

We are concerned with linear control systems of the form

$$\left\{ \begin{array}{lcll} x'(t) + Ax(t) & = & Bu(t) & (t > 0), \\ x(0) & = & x_0, & \\ y(t) & = & Cx(t) & (t > 0), \end{array} \right. \tag{1.1}$$

where $-A$ is the generator of a strongly continuous semigroup $(T(\cdot))$ in a Banach space X. The function $x(\cdot)$ takes values in X, and the functions $u(\cdot)$ and $y(\cdot)$ take values in Banach spaces U and Y, respectively. The control operator B is an unbounded operator from U to X, and the observation operator C is an unbounded operator from X to Y. We refer, e.g., to [25, 28, 32, 31].

A commonly used minimal assumption on B and C is that C is bounded $X_1(A) \to Y$ and B is bounded $U \to X_{-1}(A)$, where $X_1(A)$ denotes the domain $D(A)$ of A equipped with the graph norm, and $X_{-1}(A)$ denotes the completion of $(X, \|R(\lambda_0, A)\cdot\|_X)$ with λ_0 in the resolvent set $\rho(A)$ of A (cf., e.g., [9, sect. II.5]). Note that for $\lambda_0 \in \rho(A)$, the norm $\|(\lambda_0 - A)\cdot\|_X$ on $X_1(A)$ is equivalent to the graph norm of A. The semigroup $T(\cdot)$ has an extension to a strongly continuous semigroup $T_{-1}(\cdot)$ on $X_{-1} := X_{-1}(A)$ whose generator A_{-1} is an extension of A (cf. [9, sect. II.5]). These extensions are needed to give a precise meaning to compositions involving the operator B.

Now let $\mathcal{X} := C([0, \infty), X)$ and, for each $\tau > 0$, $\mathcal{X}_\tau = C([0, \tau], X)$. Suppose that we are given spaces $\mathcal{Y}$ of functions $\mathbb{R}_+ \to Y$ and $\mathcal{U}$ of functions $\mathbb{R}_+ \to U$ with restrictions $\mathcal{Y}_\tau$, $\mathcal{U}_\tau$ to $[0, \tau]$, $\tau > 0$, respectively. Then the system (1.1) is called

*Received by the editors April 3, 2006; accepted for publication (in revised form) July 12, 2006; published electronically January 12, 2007. This research was supported in part by DFG project "H^∞–Kalkül und seine Anwendungen auf partielle Differentialgleichungen" under contract WE 2847/1-1.
http://www.siam.org/journals/sicon/45-6/65613.html

†Mathematisches Institut I, Universität Karlsruhe, Englerstraße 2, 76128 Karlsruhe, Germany (Bernhard.Haak@math.uni-karlsruhe.de, Peer.Kunstmann@math.uni-karlsruhe.de).

wellposed if, for each $\tau > 0$, state and output of the system (1.1) depend continuously on initial state and input, i.e., if the mapping

$$X \times \mathcal{U}_\tau \to X \times \mathcal{Y}_\tau, \quad (x_0, u(\cdot)) \mapsto (x(\tau), y(\cdot))$$

is continuous. Since for $t > 0$ the solution to (1.1) has the (formal) representation

$$\big(x(t), y(t)\big) = \left(T(t)x_0 + \int_0^t T(s)Bu(t-s)\,ds,\ CT(t)x_0 + C\int_0^t T(s)Bu(t-s)\,ds \right),$$

and $T(\cdot)$ is strongly continuous, this means continuity of the three maps

$$\Psi_\tau : \begin{cases} X \to \mathcal{Y}_\tau, \\ x_0 \mapsto CT(\cdot)x_0, \end{cases} \quad \Phi_\tau : \begin{cases} \mathcal{U}_\tau \to X, \\ u \mapsto T_{-1}(\cdot)B * u, \end{cases} \quad \text{and} \quad \mathcal{F}_\tau : \begin{cases} \mathcal{U}_\tau \to \mathcal{Y}_\tau, \\ u \mapsto CT_{-1}(\cdot)B * u, \end{cases}$$

where $*$ denotes convolution. Requiring continuity of these maps leads to the notions of *admissibility* for the observation operator C, for the control operator B, respectively, and the notion of *wellposedness* for the input-output map $\mathcal{F}_\tau$ of system (1.1).

In case of uniformly exponentially stable semigroups and $\mathcal{Y}_\tau = L^p([0,\tau], Y)$ and $\mathcal{U}_\tau = L^p([0,\tau], U)$, one may also consider wellposedness on $\mathbb{R}_+ = (0,\infty)$, i.e., the spaces $\mathcal{Y} = L^p(\mathbb{R}_+, Y)$ and $\mathcal{U} = L^p(\mathbb{R}_+, U)$ and continuity of the maps

$$\Psi : \begin{cases} X \to \mathcal{Y}, \\ x_0 \mapsto CT(\cdot)x_0, \end{cases} \quad \Phi : \begin{cases} \mathcal{U}_\tau \to X, \\ u \mapsto \displaystyle\int_0^\infty T_{-1}(t)Bu(t)\,dt, \end{cases} \quad \text{and} \quad \mathcal{F} : \begin{cases} \mathcal{U} \to \mathcal{Y}, \\ u \mapsto CT_{-1}(\cdot)B * u. \end{cases}$$

Notice that in the second mapping the convolution is replaced by an integration of the semigroup against the right-hand side (cf. [31] and Remark 1.4).

If X, Y, and U are Hilbert spaces, it is natural to take $\mathcal{Y} = L^2(\mathbb{R}_+, Y)$ and $\mathcal{U} = L^2(\mathbb{R}_+, U)$ (cf. [15, 31, 32]). In the Banach space case one can consider $\mathcal{Y} = L^p(\mathbb{R}_+, Y)$ and $\mathcal{U} = L^p(\mathbb{R}_+, U)$, where $p \in [1,\infty]$ (cf. [8, 15, 27, 31]). For the case $p = 2$, the notion of *α-admissibility* for observation and control operators was introduced in [12] meaning that the L^2-space on $\mathbb{R}_+$ with values in Y or U, respectively, is taken with respect to a polynomial weight on $\mathbb{R}_+$. In this paper, we will extend that notion for observation operators from L^2-norms to L^p-norms with $p \in [1,\infty]$. For control operators our definition differs from that given in [12] (cf. Remarks 1.2 and 1.4). Furthermore, we will introduce and study the new notion of *L^p-wellposedness of type α* for the system (1.1). For a short discussion on our motivation we refer to Theorem 1.16 and Remark 1.18.

The main basic question in modeling a given system in order to obtain a wellposed system of the form (1.1) is, of course, how to check for admissibility of the operators C and B and the wellposedness of the input-output map $\mathcal{F}$. For Hilbert spaces X, Y, U, the well-known *Weiss conjecture* (cf. [15, 33]) relates L^2-admissibility on $\mathbb{R}_+$ of C and B to boundedness of

$$W_C := \{\lambda^{1/2} C(\lambda + A)^{-1} :\ \lambda > 0\} \subseteq B(X, Y) \qquad \text{and}$$
$$W_B := \{\lambda^{1/2} (\lambda + A_{-1})^{-1} B :\ \lambda > 0\} \subseteq B(U, X),$$

respectively. Assuming that $T(\cdot)$ is bounded analytic and that $A^{1/2}$ is L^2-admissible for A, boundedness of W_C actually characterizes L^2-admissibility of C. (See [20] for this result and a detailed discussion of the Weiss conjecture.) In [12], this characterization was extended to cover the case of L^2-admissibility of type α for a certain range of α.

For $\omega \in (0,\pi)$, let $S(\omega) := \{z \in \mathbb{C}\setminus\{0\} : |\arg z| < \omega\}$ and let $S(0) := (0,\infty)$. We recall that a *sectorial operator A of type* $\omega \in [0,\pi)$ in a Banach space X is a closed linear operator A satisfying $\sigma(A) \subseteq \overline{S(\omega)}$ and, for any $\nu \in (\omega,\pi)$,

$$\sup\{\|\lambda R(\lambda,A)\| : |\arg\lambda| \geq \nu\} < \infty.$$

Observe that $-A$ generates a bounded analytic semigroup in X if and only if A is a densely defined sectorial operator in X of type $< \pi/2$. Moreover, we recall that a sectorial operator with dense range is actually injective (cf. [22, Thm. 3.8]). Occasionally we shall use fractional powers of sectorial operators or their functional calculus. We refer to, e.g., [13, 19, 21].

L^p-admissibility of type α. In this paper, we assume that A is a densely defined sectorial operator of type $< \pi/2$ with dense range, and we characterize L^p–admissibility of type α for observation and control operators, the range of values of α depending on $p \in [1,\infty]$. For a fixed $p \in [1,\infty]$ we denote by p' the dual exponent given by $1/p + 1/p' = 1$. To formulate our first definition we introduce, for any $k \in \mathbb{N}$ and a given generator $-A$ of a bounded strongly continuous semigroup $T(\cdot)$ in X, the spaces $X_k := X_k(A)$, which are $\mathcal{D}(A^k)$ equipped with the norm $\|(\mathrm{Id}+A)^k\cdot\|_X$, and $X_{-k} := X_{-k}(A)$, which is the completion of X with respect to the norm $\|(\mathrm{Id}+A)^{-k}\cdot\|_X$. Observe that $1 \in \rho(-A)$ and that replacing $\mathrm{Id}+A$ by $\lambda_0 + A$ in these expressions leads to equivalent norms whenever $\lambda_0 \in \rho(-A)$ (again, we refer to [9, sect. II.5]). The semigroup $T(\cdot)$ has an extension to a bounded strongly continuous semigroup $T_{-k}(\cdot)$ on X_{-k} whose generator A_{-k} is an extension of the operator A.

We denote, for $\alpha \in \mathbb{R}$ and intervals $I \subseteq \mathbb{R}_+$,

$$L^p_\alpha(I,X) := \big\{f : I \to X :\ t \mapsto t^\alpha f(t) \in L^p(I,X)\big\}.$$

The space $L^p_\alpha(\mathbb{R}_+,X)$ is abbreviated $L^p_\alpha(X)$.

Definition 1.1. *Let X,U,Y be Banach spaces and for some $k \in \mathbb{N}$ let $C \in B(X_k,Y)$ and $B \in B(U,X_{-k})$. Given $p \in [1,\infty]$ and a bounded analytic semigroup $T(\cdot)$ on X,*

(a) *C is called* finite-time L^p-admissible of type $\alpha > -1/p$ *if for any $\tau > 0$ there is a constant $M_\tau > 0$ such that for all $x \in X_k$ we have*

$$\big\|t \mapsto CT(t)x\big\|_{L^p_\alpha([0,\tau],Y)} \leq M_\tau \|x\|_X; \tag{1.2}$$

(b) *B is called* finite-time L^p-admissible of type $\alpha < 1/p'$ *(or $\alpha \leq 0$ for $p = 1$) if for any $\tau > 0$ there is a constant $K_\tau > 0$ such that for all $u \in L^p_\alpha((0,\tau),U)$ we have*

$$\big\|T_{-k}B * u\big\|_{L^\infty((0,\tau),X)} \leq K_\tau \|u\|_{L^p_\alpha((0,\tau),U)}. \tag{1.3}$$

If the above estimates hold with constants M_τ and K_τ that can be chosen independently of $\tau > 0$, the operators B and C are called (infinite-time) L^p-admissible of type α.

Remark 1.2. In part (b) of the definition, inequality (1.3) means the estimate $\operatorname{ess\,sup}_{t\in(0,\tau)} \|\Phi_t(u)\|_X \leq K_\tau \|u\|_{L^p_\alpha((0,\tau),U)}$. This seems to be better suited when dealing with weighted spaces than requiring just $\|\Phi_\tau(u)\|_X \leq K_\tau \|u\|_{L^p_\alpha((0,\tau),U)}$.

Notice that our definition differs from that in [12]: for observation operators, L^2-admissibility of type α is called 2α-admissibility there. For control operators, our definition is different from the notion studied in [12]. This is due to the application

to nonlinear problems in Theorem 1.16. See also Remark 1.4 for a study of both notions.

LEMMA 1.3. *The notion of finite-time L^p-admissibility of type α for A is independent of the underlying interval $[0,\tau]$ in the definition.*

If the semigroup is uniformly exponentially stable and $\alpha \geq 0$ or $p = \infty$, finite-time L^p-admissibility of type α for control operators is equivalent to (infinite-time) L^p-admissibility of type α. For observation operators this is true for all α. Moreover, an observation operator C is (infinite-time) L^p-admissible of type α of A if and only if

$$\left(\int_0^\infty \|t^\alpha CT(t)x\|^p \, dt\right)^{1/p} \leq M_{\tau,\alpha}\|x\| \tag{1.4}$$

holds.

Notice that the requirement $\alpha \geq 0$ or $p = \infty$ in the above equivalence assertion for control operators is necessary. We provide a short counterexample in case $\alpha < 0$ in Example 3.1.

Remark 1.4 (dualization). Let X, Y be reflexive Banach spaces and $p \in (1,\infty)$ and assume that $C \in B(X_1, Y)$ is L^p-admissible of type α for A on $\mathbb{R}_+$. Then, for $u \in L^{p'}_{-\alpha}(\mathbb{R}_+, Y')$,

$$\begin{aligned}\big\langle t \mapsto CT(t)x, u\big\rangle_{L^p_\alpha \times L^{p'}_{-\alpha}} &= \int_0^\infty \big\langle t^\alpha CT(t)x, t^{-\alpha}u(t)\big\rangle_{Y\times Y'} \, dt \\ &= \left\langle x \int_0^\infty T(t)'C'u(t)\, dt X \times X' \right\rangle.\end{aligned}$$

One can therefore consider the following dual condition to (1.4), which was introduced in [12]:

$$\left\| \int_0^\infty T(t)Bu(t)\, dt \right\|_X \leq K\|u\|_{L^p_{-\alpha}(\mathbb{R}_+,U)}, \tag{1.5}$$

where the integral is considered as a Pettis integral in X_{-k} taking values in X. Notice that in case $\alpha \neq 0$ the reflection $R_\tau u = u(\tau - \cdot)$ is not bounded on $L^p_\alpha([0,\tau])$. Therefore, one cannot hope that (1.5) might be equivalent to infinite-time L^p-admissibility of type α unless $\alpha = 0$. We shall see in Theorem 1.8 that condition (1.5) is indeed a stronger notion than L^p-admissibility of type α. However, for $\alpha > 0$, L^p-admissibility of an observation operator C *implies* L^p-admissibility of C' on X' by Theorem 1.8.

L^p_*-estimates and the real interpolation method. As mentioned above, the crucial condition in [20] (besides $T(\cdot)$ being bounded and analytic) for the Weiss conjecture to hold was that $A^{1/2}$ is admissible for A, i.e., the existence of a constant $L > 0$ such that

$$\int_0^\infty \big\|A^{1/2}T(t)x\big\|^2 \, dt = \int_0^\infty \big\|(tA)^{1/2}T(t)x\big\|^2 \, \tfrac{dt}{t} = \big\|\psi(tA)x\big\|^2_{L^2(\mathbb{R}_+,dt/t,X)} \leq L\|x\|^2_X,$$

where $\psi(z) = z^{1/2}e^{-z}$. In our situation this corresponds to

$$\begin{aligned}\int_0^\infty \big\|t^\alpha A^{\alpha+1/p}T(t)x\big\|^p \, dt &= \int_0^\infty \big\|(tA)^{\alpha+1/p}T(t)x\big\|^p \, \tfrac{dt}{t} \\ &= \big\|\psi(tA)x\big\|^p_{L^p(\mathbb{R}_+,dt/t,X)} \leq \widetilde{L}\|x\|^p_X\end{aligned} \tag{1.6}$$

for $\psi(z) = z^{\alpha+1/p}\exp(-z)$. We denote $L^p_*(\mathbb{R}_+, X) := L^p(\mathbb{R}_+, dt/t, X)$. It is known that the property of a sectorial operator A of type ω on a Banach space X to satisfy an estimate $\|\psi(\cdot A)x\|_{L^p_*(\mathbb{R}_+,Y)} \le L\|x\|_X$ does not depend upon the particular choice of the function $\psi \in H^\infty_0(S(\nu))\backslash\{0\}$, $\nu > \omega$, where

$$H^\infty_0(S(\nu)) = \big\{ f \in H^\infty(S(\nu)) :\ \exists c, s > 0 :\ \big|f(z)\big| \le c\,\tfrac{|z|^s}{1+|z|^{2s}} \big\};$$

see [22, Thm. 5], [2, Thm. 4.1], [14, Thm. 4.3], or [13, Thm. 6.4.3]. Therefore, we say that A *satisfies* L^p_**-estimates on* X if (1.6) holds.

Let A be a densely defined sectorial operator of type ω with dense range on X. Let $\mathbb{X} := (X, \|A(I+A)^{-2}\cdot\|)^\sim$. For some holomorphic function ψ on $S(\nu)$, $\nu > \omega$ and $\theta \in (-1, 1)$ such that $z^{-\theta}\psi(z) \in H^\infty_0(S(\nu))\backslash\{0\}$ consider the space

$$X_{\theta,\psi,p} := \big\{ x \in \mathbb{X} :\ t^{-\theta}\psi(tA)x \in L^p_*(\mathbb{R}_+, X) \big\}.$$

Since $X_{\theta,\psi,p}$ does not depend on $\psi \in H^\infty_0(S(\nu))\backslash\{0\}$ (see above), we write for short $X_{\theta,p} := X_{\theta,\psi,p}$. Resorting to the space $\mathbb{X}$ in this setting allows explicitly that $X_{\theta,p}$ may be a larger space than X. For the background of this construction we refer, e.g., to [18], [11, sect. 2] or [13, Chap. 6.3]. Notice that in this terminology L^p_*-estimates for A read as $X \hookrightarrow X_{0,p}$. These spaces are strongly connected to real interpolation spaces between the homogeneous spaces $\dot X_{-1}$ and $\dot X_1$. Here, for $\sigma \in \mathbb{Z}$, we denote by $\dot X_\sigma := \dot X_\sigma(A)$ the completion of the space $\mathcal{D}(A^\sigma)$ with respect to the homogeneous norm $\|A^\sigma\cdot\|_X$. We refer to [17], [19, sect. 15], [11], or [14] for more details on these spaces. Observe that—due to our assumptions—the space $\mathcal{D}(A^\sigma)$ is dense in X and that $\dot X_\sigma = X_\sigma$ in case $0 \in \rho(A)$. With these notations the following result holds true. (See [2] for the case $p = 2$ and [13, Thm. 6.4.6], [14, Thm. 5.2] for the general case.)

THEOREM 1.5. *Let X be a Banach space and A be a densely defined sectorial operator of type ω with dense range on X. Then*

$$\big(\dot X_{-1}(A), \dot X_1(A)\big)_{\theta,p} = X_{2\theta-1,p}$$

for all $\theta \in (0,1)$ and $p \in [1,\infty]$.

Remark 1.6. In the situation of the above theorem, A has L^p_*-estimates if and only if $X \hookrightarrow \big(\dot X_{-1}(A), \dot X_1(A)\big)_{1/2,p}$. This can be used to establish L^p_*-estimates for given operators in concrete cases, as we will show in section 2.

We start our main results with the following characterizations of L^p-admissibility of type α for observation and control operators that extend the results in [12].

THEOREM 1.7. *Let $p \in [1,\infty]$ and A be a densely defined sectorial operator of type $\omega < \pi/2$ with dense range on X. Let $C \in B(X_k, Y)$ be an observation operator for some $k \ge 1$. Let $\alpha \in (-1/p, k - 1/p)$ and consider the set*

$$W_C := \big\{ \lambda^{k-\alpha-1/p}\, C(\lambda + A)^{-k} :\ \lambda > 0 \big\} \subseteq B(X, Y). \tag{1.7}$$

Then the following assertions hold:

(a) *If C is L^p-admissible of type α for A, then W_C is bounded in $B(X,Y)$.*
(b) *If A satisfies L^p_*-estimates and W_C is bounded in $B(X,Y)$, then C is L^p-admissible of type α for A.*

Notice that for the operator $C := A^{\alpha+1/p}$, the set W_C in (1.7) is bounded, whence the assumption of L^p_*-estimates in the above characterization cannot be improved (cf. (1.6)).

THEOREM 1.8. *Let $p \in [1,\infty]$ and A be a densely defined sectorial operator of type $\omega < \pi/2$ with dense range on X. Let $B \in B(U, X_{-k})$ be a control operator for some $k \geq 1$ and let $\alpha \in (1/p' - k, 1/p')$ and consider the set*

$$W_B := \left\{ \lambda^{k+\alpha-1/p'} (\lambda + A_{-k})^{-k} B : \ \lambda > 0 \right\} \subseteq B(U, X). \tag{1.8}$$

Then the following assertions hold:

(a) *If B is L^p-admissible of type α for A, then W_B is bounded.*
(b) *If W_B is bounded and $\alpha > 0$ (for $p > 1$) or $\alpha = 0$ (for $p = 1$), then B is L^p-admissible of type α for A.*
(a)′ *If the dual condition* (1.5) *holds, then W_B is bounded.*
(b)′ *If $p < \infty$ and W_B is bounded and the adjoint operator A' satisfies $L^{p'}_*$-estimates, then* (1.5) *holds.*

As mentioned above, in case $\alpha = 0$, L^p-admissibility of type α and (1.5) are equivalent. In particular, a modified Weiss conjecture on control operators holds for $p = 1$.

Also the boundedness conditions on the sets W_C and W_B in (1.7) and (1.8) may be characterized by real interpolation methods. In fact, they are conditions on the *domain* of (a continuous extension of) the observation operator C and the *range* of the control operator B, respectively. This way of viewing the boundedness conditions on W_C and W_B, respectively, is not new for control operators (cf. [33]) under the supplementary assumption of $0 \in \varrho(A)$. The general equivalence for observation and control operators was first studied in [11]. The proof of the following is very similar to the arguments used there and is thus omitted.

THEOREM 1.9. *Let A be an injective sectorial operator on the Banach space X, and let $k \in \mathbb{N}$. Let $C \in B(X_k, Y)$ and $B \in B(U, X_{-k})$ be bounded operators, where U, Y are Banach spaces, and let $\theta \in (0,1)$. Then the following equivalences hold true:*

(a) $\sup_{0<\lambda<\infty} \|\lambda^{k\theta}(\lambda + A_{-k})^{-k} B\|_{U\to X} < \infty$ *if and only if $B : U \to X_{-k}$ is bounded in the norm $U \to (\dot{X}_{-k}, X)_{\theta,\infty}$.*
(b) $\sup_{0<\lambda<\infty} \|\lambda^{k(1-\theta)} C(\lambda + A)^{-k}\|_{X\to Y} < \infty$ *if and only if $C : X_k \to Y$ is bounded in the norm $(X, \dot{X}_k)_{\theta,1} \to Y$.*

Remark 1.10. Of course, if for some $k \in \mathbb{N}$, $C \in B(X_k, Y)$ and if the operators $\lambda^{k(1-?)} C(\lambda + A)^{-k}$, $\lambda > 0$, are uniformly bounded, then for any $n \in \mathbb{N}_0$, $C \in B(X_{k+n}, Y)$ the operators $\lambda^{n+k(1-?)} C(\lambda + A)^{-k-n}$, $\lambda > 0$, are uniformly bounded by the sectoriality of A. However, by reiteration (cf. [29, Thm. 1.10.2]),

$$\left(X, \dot{X}_k\right)_{\theta,1} = \left(X, \dot{X}_{k+n}\right)_{\sigma,1}$$

for $\sigma = \frac{n+k(1-\theta)}{k+n}$, whence the assertions of Theorem 1.9 do not depend on the question, for which $k \in \mathbb{N}$ the given boundedness conditions are satisfied. The same arguments apply to the conditions on control operators.

Wellposedness of the full system. We now study the full system (1.1) for $\mathcal{Y} = L^p_\alpha(\mathbb{R}_+, Y)$ and $\mathcal{U} = L^p_\alpha(\mathbb{R}_+, U)$.

DEFINITION 1.11. *Let X, U, Y be Banach spaces and let $k \in \mathbb{N}$, $p \in [1,\infty]$ and $\alpha \in (-1/p, 1/p')$. Let $T(\cdot)$ be a bounded analytic semigroup on X generated by $-A$. Then the system* (1.1) *is called L^p-*wellposed of type α *if $C \in B(X_k, Y)$ and $B \in B(U, X_{-k})$ are L^p-admissible of type α, and $\mathcal{F}^\alpha : L^p_\alpha(U) \to L^p_\alpha(Y)$, $\mathcal{F}^\alpha u = CT_{-k}(\cdot)B * u$ is bounded.*

If we have finite-time L^p-admissibility of type α of B and C on $I := [0,\tau]$ and if $\mathcal{F}^\alpha_\tau : L^p_\alpha(I, U) \to L^p_\alpha(I, Y)$ is bounded, we call the system *finite-time L^p-wellposed of*

type α. By resorting to $(e^{-\omega t}T(t))$ we may then assume $0 \in \varrho(A)$. The next lemma shows that the only possible singularity of the convolution kernel of $\mathcal{F}_\tau$ is at $t = 0$, whence the notion of finite-time L^p-wellposedness of type α does not depend on $\tau > 0$.

Notice that by analyticity of the semigroup $T(\cdot)$, we have $T_{-k}(t)X_{-k} \subseteq X_k$ for $t > 0$. Therefore the convolution kernel $CT_{-k}(\cdot)B$ of $\mathcal{F}^\alpha$ is a well-defined bounded operator from U to Y. (See the next lemma for norm estimates.) From Theorems 1.7, 1.8, and 1.9 we know that L^p-admissibility of type α of B and C yields $(X, \dot{X}_k)_{\theta,1} \hookrightarrow \mathcal{D}(C)$ and $\mathcal{R}(B) \subseteq (X, \dot{X}_{-k})_{\sigma,\infty}$ for $\theta = (\alpha + 1/p)/k$ and $\sigma = (1/p' - \alpha)/k$. Notice that $k(\sigma + \theta) = 1$. The next lemma is well known to specialists. We use it for $q = 1$ and $r = \infty$ and provide a proof in section 3.

LEMMA 1.12. *Let X be a Banach space and $T(\cdot)$ be a bounded analytic semigroup on X. Let $k \in \mathbb{N}$ and $\sigma, \theta \in (0,1)$ such that $k(\sigma + \theta) = 1$. Let $q, r \in [1, \infty]$ and $Z := (X, \dot{X}_k)_{\theta,q}$ and $W := (\dot{X}_{-k}, X)_{1-\sigma,r}$. Then, there exists a constant $M > 0$ such that $\|T(t)\|_{W \to Z} \le M/t$ for all $t > 0$.*

To sum up the above considerations: whenever C and B are L^p-admissible of type α, we have $\|CT(t)B\|_{U \to Y} \le M/t$ for $t > 0$ for some constant $M > 0$. A corresponding condition in the following theorem seems thus very natural.

THEOREM 1.13. *Let $p \in (1, \infty)$, $\alpha \in (-1/p, 1/p')$, and $k \in \mathbb{N}$. Let X, U, Y be Banach spaces and let $T(\cdot)$ be a bounded analytic semigroup on X, $C \in B(X_k, Y)$ and $B \in B(U, X_{-k})$ such that $\|CT(t)B\|_{U \to Y} \le M/t$. Then $\mathcal{F}^\alpha := CT(\cdot)B*$ is bounded $L^p_\alpha(U) \to L^p_\alpha(Y)$ if and only if $\mathcal{F} = \mathcal{F}^0$ is bounded $L^p(U) \to L^p(Y)$.*

The above theorem is basically a reformulation of [24, Thm. 2.4]. However, our proof also allows negative values of α. It relies on the following result.

PROPOSITION 1.14. *Let $p \in (1, \infty)$ and $\alpha < 1 - 1/p$. Let U, Y be Banach spaces and suppose that $K \in C(\mathbb{R}_+, B(U, Y))$ satisfies $\|K(t)\| \le M/t$ for some $M > 0$. Let*

$$(Tf)(t) := \int_0^t K(t-s)\left[\left(\frac{t}{s}\right)^\alpha - 1\right] f(s)\, ds, \qquad f \in L^p(\mathbb{R}_+, X).$$

Then $T \in B(L^p(\mathbb{R}_+, X), L^p(\mathbb{R}_+, Y))$ with norm bound cM where $c = c(p, \alpha)$.

We shortly discuss boundedness of $\mathcal{F}$ for $\alpha = 0$. Suppose that $C : Z \to Y$ and $B : U \to W$ are bounded, where Z and W are Banach spaces satisfying $X_1 \subseteq Z \subseteq X$ and $X \subseteq W \subseteq X_{-1}$. Suppose further that the restriction A_W of A_{-1} to W is sectorial with $\mathcal{D}(A_W) = Z$ (equivalent norms). Then $CT(\cdot)B* : L^p(U) \to L^p(Y)$ is bounded if $T(\cdot)* : L^p(W) \to L^p(Z)$ is bounded. But it is well known that the latter is equivalent to A_W having the property of *maximal L^p-regularity*. (We refer to [1, 6, 7, 19, 30] for this relation, the problem of maximal regularity, characterization results, and further references on the subject.) Thus we have obtained the next theorem.

THEOREM 1.15. *Let $p \in (1, \infty)$ and W be a Banach space $X \subseteq W \subseteq X_{-1}$ such that A_W has maximal L^p-regularity. Let $C : Z \to Y$ and $B : U \to W$ be bounded where Z denotes $\mathcal{D}(A_W)$ equipped with the graph norm. Then $\mathcal{F}^\alpha$ is bounded $L^p_\alpha(U) \to L^p_\alpha(Y)$ for any $\alpha \in (-1/p, 1/p')$.*

Our results can be used to establish existence and uniqueness of solutions to some nonlinear systems in feedback form. Let $p \in (1, \infty)$, let X, U, Y be Banach spaces, and let $x_0 \in X$. Let $F : X \to B(Y, U)$ be a Lipschitz-continuous operator-valued function, let A, B, C be linear operators such that $-A_0 := -(A - BF(x_0)C)$ generates an analytic semigroup $T(\cdot)$ in X and such that B and C are finite-time

L^p-admissible of type α for A_0. We consider the closed loop system

$$\left\{\begin{array}{lcl} x'(t)+Ax(t) &=& Bu(t),\\ x(0) &=& x_0,\\ y(t) &=& Cx(t),\\ u(t) &=& F(x(t))\,y(t) \end{array}\right. \tag{1.9}$$

on $[0,\tau]$, which we rewrite as

$$\left\{\begin{array}{lcll} x'(t)+A_0x(t) &=& B\big(F(x(t))-F(x_0)\big)\,Cx(t), & t\in[0,\tau],\\ x(0) &=& x_0. & \end{array}\right. \tag{1.10}$$

We are interested in mild solutions of (1.10), i.e., solutions of

$$x = T(\cdot)x_0 + T*B\big(F(x)-F(x_0)\big)\,Cx \quad \text{on } [0,\tau]. \tag{1.11}$$

The following result shall also be proved in section 3.

THEOREM 1.16. *Assume in addition to the preceding assumptions that $Z\hookrightarrow X$ is a Banach space such that the system corresponding to (A_0,B,Id_Z) is finite-time L^p-wellposed of type α and that $C\in B(Z,Y)$. Then there exists $\tau=\tau(x_0)>0$ such that* (1.9) *has a unique mild solution $x\in C([0,\tau],X)\cap L^p_\alpha([0,\tau],Z)$.*

Remark 1.17. The reason for introducing the space Z here is that $x\mapsto Cx$ may not induce a closed operator from $C([0,\tau],X)$ into $L^p_\alpha([0,\tau],Y)$. If C is closed as an operator from X to Y, then we can replace $C([0,\tau],X)\cap L^p_\alpha([0,\tau],Z)$ by $\{x\in C([0,\tau],X) : x(t)\in D(C) \text{ a.e., } Cx\in L^p_\alpha([0,\tau],Y)\}$ in the assertion, and the assumption on the space Z is not needed.

Remark 1.18. Our results on L^p-admissibility of type α and L^p-wellposedness of type α give more flexibility in the choice of Banach spaces X, U, and Y for the modeling of a given problem. This is important for the study of nonlinear systems via fixed point arguments, e.g., via Theorem 1.16, where an appropriate choice of α allows us to choose state spaces X as function spaces with little regularity (cf. also [24, Rem. 3.3(b)]). For the example of the controlled heat equation that we study in section 2 we refer to Remarks 2.11 and 2.12, where we describe how to obtain state spaces with negative smoothness index by suitable choices of α. In Example 2.14 we give an application of Theorem 1.16 in a nonlinear feedback situation. We mention in this context that Besov spaces of negative order have become relevant as spaces for initial values in the study of other nonlinear partial differential problems, e.g., Navier–Stokes equations (cf. [5]).

In the next section we provide examples and applications of our results. In section 3 we give proofs of the results presented so far.

2. Example: A controlled heat equation. In this section we illustrate our results with a controlled heat equation. In [4], the problem was studied in the state space $X=L^2(\Omega)$ and for $\alpha=0$. Below we discuss Lebesgue and Besov spaces.

Let $\Omega\subset\mathbb{R}^n$ be a bounded domain with boundary $\partial\Omega\in C^\infty$. Denote the outer normal unit vector on $\partial\Omega$ by $\nu:\partial\Omega\to\mathbb{R}^n$. We consider the problem

$$\left\{\begin{array}{lcl} x'(t)-\Delta x(t) &=& 0 \quad (t>0),\\ \frac{\partial x(t)}{\partial\nu}\big|_{\partial\Omega} &=& u(t) \quad (t>0),\\ x(0) &=& x_0,\\ y(t) &=& x(t)\big|_{\partial\Omega} \quad (t>0), \end{array}\right. \tag{2.1}$$

where $x(\cdot)$ takes values in some function space X. The functions $u(\cdot)$ and $y(\cdot)$ take values in function spaces on the boundary. For the modeling we follow closely [4] and set $A := -\Delta$ with Neumann boundary condition. In the state spaces we shall consider below, the operator A is sectorial of type 0, but not injective. We aim for finite-time L^c-admissibility of type γ for observation operators and finite-time L^b-admissibility of type β for control operators, where $b, c \in [1, \infty]$. Sufficient for this is $L^{b/c}$-admissibility of type β/γ on $\mathbb{R}_+$ for $\mathrm{Id} + A$. In order to use our characterizations of admissibility, we assume

$$\beta \in \left(-1/b, 1/b'\right) \quad \text{and} \quad \gamma \in \left(-1/c, 1/c'\right).$$

We start with the L^q-case.

The L^q-case. Consider $X := L^q(\Omega)$, $1 < q < \infty$. Due to the smoothness of $\partial\Omega$, we then have $\mathcal{D}(A) = \{x \in W^2_q(\Omega) : \frac{\partial x}{\partial \nu}\big|_{\partial\Omega} = 0\}$ and $-A$ generates a bounded analytic semigroup. To ensure L^c_*-estimates, we have to impose

$$\begin{aligned} X \hookrightarrow \left(\dot{X}_{-1}(\mathrm{Id}+A), \dot{X}_1(\mathrm{Id}+A)\right)_{1/2,c} &= \left(X_{-1}(A), X_1(A)\right)_{1/2,c} \\ &= \left(X_{-\delta}(A), X_\delta(A)\right)_{1/2,c}, \end{aligned} \tag{2.2}$$

where the last equality holds for any $\delta > 0$ by reiteration. In the case $X = L^q(\Omega)$, we have $X_\delta = H^{2\delta}_q(\Omega)$ for small $\delta > 0$ (cf. [26]) and $X_{-\delta} = H^{-2\delta}_q(\Omega)$ by dualization. Therefore, L^c_*-estimates for $\mathrm{Id} + A$ on $L^q(\Omega)$ are equivalent to the continuous embedding $L^q(\Omega) \hookrightarrow B^0_{q,c}(\Omega)$ (cf. [29, Thms. 2.4.1, 4.3.1]). We use the following lemma, which is proved in section 3.

LEMMA 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with boundary $\partial\Omega \in C^\infty$. If $p, q \in [1, \infty]$, we have $L^q(\Omega) \hookrightarrow B^0_{q,p}(\Omega)$ if and only if $p \geq \max(2, q)$.*

For an application of Theorem 1.7 for L^c-admissibility of C in $X = L^q(\Omega)$ we thus need $c \geq \max(2, q)$. For an application of Theorem 1.8 for L^b-admissibility of B in $X = L^q(\Omega)$ in case $\beta = 0$, we also need $L^{b'}_*$-estimates for $(I+A)'$ on X'. By the arguments above these are equivalent to $L^{q'}(\Omega) \hookrightarrow B^0_{q',b'}(\Omega)$, which means by Lemma 2.1 that for $\beta = 0$, we have to suppose $b \leq \min(q, 2)$. Obviously, if $\beta = 0$, then application of both Theorems 1.7 and 1.8 for $b = c$ would require $b = c = q = 2$ and we are back in the Hilbert space situation. We shall come back to this below.

Admissibility. Denoting the Dirichlet trace operator $\gamma_0 : x \mapsto x\big|_{\partial\Omega}$ by C we are looking for a space Y on $\partial\Omega$ such that $C : \mathcal{D}(A) \to Y$ is L^c-admissible of type γ for $\mathrm{Id} + A$. If $c \geq \max(2, q)$, then $\mathrm{Id} + A$ has L^c_*-estimates and we know by Theorem 1.7 that L^c-admissibility of type γ of $C : \mathcal{D}(A) \to Y$ is equivalent to uniform boundedness of the operators

$$\lambda^{1-\gamma-1/c} C(\lambda + \mathrm{Id} + A)^{-1}, \quad \lambda > 0.$$

By Theorem 1.9 this is equivalent to C having a continuous extension to the Banach space Z, where

$$\begin{aligned} Z &:= \left(X, \dot{X}_k(\mathrm{Id}+A)\right)_{\gamma+1/c,1} \\ &= \left(X, X_k(A)\right)_{\gamma+1/c,1} \\ &\overset{(*)}{=} \begin{cases} B^s_{q,1}(\Omega) & \text{if } s < 1 + 1/q, \\ \{x \in B^s_{q,1}(\Omega) : \ \frac{\partial x}{d\nu}\big|_{\partial\Omega} = 0\} & \text{if } s > 1 + 1/q, \end{cases} \end{aligned}$$

where $s = 2\gamma + 2/c$. For equality $(*)$ we refer to [29, Thm. 4.3.3] or [10, Thm. 3.5]. It is known that γ_0 is bounded from $B^s_{q,1}(\Omega)$ to $B^{s-1/q}_{q,1}(\partial\Omega)$ if $s > 1/q$ (cf. [29, Thm. 4.7.1]). We thus have almost proved the following proposition.

PROPOSITION 2.2. *Let $\gamma \in \mathbb{R}$, $c \in [1,\infty]$, and $q \in (1,\infty)$ satisfy $c \geq \max(q,2)$ and $2\gamma + 2/c \in (1/q, 1+1/q)$, and let Y be a Banach space and $X = L^q(\Omega)$. Then $\gamma_0 : \mathcal{D}(A) \to Y$ is L^c-admissible of type γ for $Id + A$ if and only if $B^{2\gamma+2/c-1/q}_{q,1}(\partial\Omega) \hookrightarrow Y$.*

Proof. Let $s := 2\gamma + 2/c$. The arguments above show that $B^{s-1/q}_{q,1}(\partial\Omega) \hookrightarrow Y$ is sufficient. To prove necessity we compose $\gamma_0 : B^s_{q,1}(\Omega) \to Y$ with a continuous extension operator $E_0 : B^{s-1/q}_{q,1}(\partial\Omega) \to B^s_{q,1}(\Omega)$ such that $\gamma_0 E_0 = Id$ (cf. [29, Thm. 4.7.1]). ∎

Remark 2.3. The upper bound $2\gamma + 2/c < 1 + 1/q$ appears only for simplicity of formulation. The calculation of the space Z above indicates how to proceed in other cases.

To obtain the representation of the control operator B we follow again ideas in [4] and multiply the state equation in (2.1) with a fixed function $v \in C^\infty(\overline{\Omega})$. Then, integrating by parts gives

$$\langle x'(t), v\rangle_\Omega + \langle \nabla x(t), \nabla v\rangle_\Omega = \int_{\partial\Omega} u(t) v \, d\sigma,$$

where $\langle\cdot,\cdot\rangle_\Omega$ denotes the usual duality pairing on $L^q(\Omega) \times L^{q'}(\Omega)$ and σ denotes the surface measure on $\Gamma := \partial\Omega$. Denoting extensions of the usual $L^q(\Gamma) \times L^{q'}(\Gamma)$-duality by $\langle\cdot,\cdot\rangle_\Gamma$ we thus have

$$\int_{\partial\Omega} u(t) v \, d\sigma = \langle u(t), \gamma_0 v\rangle_\Gamma,$$

which means that $B = \gamma_0'$ if we identify $X_{-1}(A)$ with the dual space of $(X')_1(A')$. Notice that $A' = -\Delta$ with Neumann boundary conditions in $X' = L^{q'}(\Omega)$.

We are interested in L^b-admissibility of type β for $\gamma_0' : U \to X_{-1}$ and $\mathrm{Id} + A$ in $X = L^q(\Omega)$. Assuming $b \leq \min(q,2)$ in case $\beta = 0$ we may use Theorem 1.8 and only have to check boundedness of $\gamma_0' : U \to W$, where the extrapolation space W is given by

$$\begin{aligned} W &= (\dot{X}_{-1}(\mathrm{Id} + A), X)_{\beta+1/b,\infty} = (X_{-1}, X)_{\beta+1/b,\infty} \\ &= \big(\big((X')_1, X'\big)_{\beta+1/b,1}\big)' = \big(\big(X', (X')_1\big)_{1-(\beta+1/b),1}\big)' \end{aligned}$$

(cf. [3, sect. 3.7]). As we have seen above, we have

$$(X', (X')_1)_{1-(\beta+1/b),1} = B^{2-2\beta-2/b}_{q',1}(\Omega)$$

for $2 - 2\beta - 2/b < 1 + 1/q' = 2 - 1/q$. Now $\gamma_0 : B^{2-2\beta-2/b}_{q',1}(\Omega) \to B^{2-2\beta-2/b-1/q'}_{q',1}(\partial\Omega)$ is continuous for $2 - 2\beta - 2/b > 1/q' = 1 - 1/q$ and $\partial\Omega$ is without boundary; hence $(B^s_{q',1}(\partial\Omega))' = B^{-s}_{q,\infty}(\partial\Omega)$. Thus we have proved one implication in the following proposition.

PROPOSITION 2.4. *Let $\beta \in \mathbb{R}$, $b \in [1,\infty]$, and $q \in (1,\infty)$ satisfy $2/b + 2\beta \in (1/q, 1+1/q)$. If $\beta = 0$, assume in addition $b \leq \min(q,2)$. Let U be a Banach space and $X = L^q(\Omega)$. Then $\gamma_0' : U \to X_{-1}$ is L^b-admissible of type β for $Id + A$ if and only if $U \hookrightarrow B^{2/b+2\beta-1-1/q}_{q,\infty}(\partial\Omega)$.*

For the remaining implication we make use of E_0', where E_0 is the extension map from the proof of Proposition 2.2.

Wellposedness of the full system. We study L^p-wellposedness of type α in finite time for the system (2.1) in the state space $X = L^q(\Omega)$, where $\alpha \in \mathbb{R}$, $p \in [1,\infty]$ and $q \in (1,\infty)$. Hence we have $c = b = p$ and $\gamma = \beta = \alpha$ in the admissibility situations above. Again, it is sufficient to study infinite-time L^p-admissibility of type α for $Id + A$. The assumptions of Propositions 2.2 and 2.4 lead, for $\alpha \neq 0$, to the restrictions $p \geq \max(q,2)$ and $2\alpha + \frac{2}{p} \in (\frac{1}{q}, 1 + \frac{1}{q})$, and for $\alpha = 0$ to the restriction $p = q = 2$.

For $p = q = 2$ we thus can state the following result extending the case $\alpha = 0$, $r = 2$, considered in [4].

PROPOSITION 2.5. *Let $|\alpha| < \frac{1}{4}$, $r \in [1,\infty]$, and $X = L^2(\Omega)$. Let Y and U be Banach spaces on $\partial\Omega$ such that $B^{2\alpha+\frac{1}{2}}_{2,r}(\partial\Omega) \hookrightarrow Y$ and $U \hookrightarrow B^{2\alpha-\frac{1}{2}}_{2,r}(\partial\Omega)$. Then the system (2.1) is L^2-wellposed of type α.*

Proof. By Propositions 2.2 and 2.4, C and B are L^p-admissible of type α for $\mathrm{Id} + A$. Notice that the semigroup associated to $\mathrm{Id} + A$ is $S(t) := e^{-t}T(t)$.

Since $B = \gamma_0 : B^{2\alpha+\frac{1}{2}}_{2,r}(\partial\Omega) \to \widetilde{W} := (X_{-1}, X)_{\beta+\frac{1}{b},r} = (X_{-1}, X)_{\alpha+\frac{1}{2},r}$ (recall $\beta = \alpha$ and $b = p = 2$) and $C = \gamma_0 : \widetilde{Z} \to B^{2\alpha+\frac{1}{2}}_{2,r}(\partial\Omega)$, where $\widetilde{Z} := B^{2\alpha+1}_{2,r}(\Omega) = (X, X_1)_{\alpha+\frac{1}{2},r}$, it suffices to prove that $S(\cdot)* : L^p_\alpha(\mathbb{R}_+, \widetilde{W}) \to L^p_\alpha(\mathbb{R}_+, \widetilde{Z})$. Theorem 1.13 applies by $|\alpha| < \frac{1}{4} < \frac{1}{2}$. Now notice that $\widetilde{Z} = \mathcal{D}(A_{\widetilde{W}})$. Hence we are left to check that $A_{\widetilde{W}}$ has maximal L^p-regularity in $\widetilde{W}$ which holds by [6, Thm. 4.7]. ∎

With essentially the same arguments we can study problem (2.1) in state spaces $X = H^\delta_q(\Omega)$ where we restrict to $\delta \in (-\frac{1}{q'}, \frac{1}{q})$ and are thus not bothered by additional boundary conditions. Here we have $X_1 = \mathcal{D}(A) = \{x \in H^{2+\delta}_q(\Omega) : \frac{\partial x}{\partial \nu}|_{\partial\Omega} = 0\}$, and a repetition of the arguments above yields for the corresponding interpolation spaces $Z = B^{\delta+2(\alpha+\frac{1}{p})}_{q,1}(\Omega)$ if $\delta + 2(\alpha + \frac{1}{p}) < 1 + \frac{1}{q}$ and $W = (B^{2-\delta-2(\alpha+\frac{1}{p})}_{q',1}(\Omega))'$ if $2 - \delta - 2(\alpha + \frac{1}{p}) < 1 + \frac{1}{q'}$, i.e., if $\frac{1}{q} < \delta + 2(\alpha + \frac{1}{p})$. Thus we obtain the next proposition.

PROPOSITION 2.6. *Let $q \in (1,\infty)$ and $p \in [1,\infty]$ satisfy $p \geq \max(2,q)$. Let $\delta \in (-\frac{1}{q'}, \frac{1}{q})$, and $\delta + 2(\alpha + \frac{1}{p}) \in (\frac{1}{q}, 1 + \frac{1}{q})$, where $\alpha \neq 0$. Let $X = H^\delta_q(\Omega)$ and let Y and U be Banach spaces on $\partial\Omega$.*

(a) *The operator $C = \gamma_0 : \mathcal{D}(A) \to Y$ is L^p-admissible of type α for $Id + A$ if and only if $B^{\delta+2(\alpha+\frac{1}{p})-\frac{1}{q}}_{q,1}(\partial\Omega) \hookrightarrow Y$.*

(b) *The operator $B = \gamma_0' : U \to X_{-1}$ is L^p-admissible of type α for $Id + A$ if and only if $U \hookrightarrow B^{\delta+2(\alpha+\frac{1}{p})-\frac{1}{q}-1}_{q,\infty}(\partial\Omega)$.*

(c) *Let $r \in [1,\infty]$ and $B^{\delta+2(\alpha+\frac{1}{p})-\frac{1}{q}}_{q,r}(\partial\Omega) \hookrightarrow Y$ and $U \hookrightarrow B^{\delta+2(\alpha+\frac{1}{p})-\frac{1}{q}-1}_{q,r}(\partial\Omega)$. Then the system (2.1) is finite-time L^p-wellposed of type α.*

Observe that the assumptions imply $\alpha + \frac{1}{p} \in (0,1)$.

We see that δ may be chosen arbitrarily close to $-\frac{1}{q'}$ by taking p large and adjusting α. Moreover, one still has free choice of $r \in [1,\infty]$.

The Besov space case. We continue the study of (2.1), but now we take as state space the Besov space $X := B^0_{q,v}(\Omega)$, where $1 < q, v < \infty$ are fixed. Again we put $A := -\Delta$ with homogeneous Neumann boundary conditions; i.e., A has domain $\mathcal{D}(A) = \{x \in B^2_{q,v}(\Omega) : \frac{\partial x}{\partial \nu}\big|_{\partial\Omega} = 0\}$. As in the L^q-case, $-A$ generates a bounded analytic semigroup in X.

Admissibility. Still denoting by γ_0 the Dirichlet trace on $\partial\Omega$, we study L^c-admissibility of type γ for $C = \gamma_0$ and L^b-admissibility of type β for $B = \gamma_0'$, where $\beta, \gamma \in \mathbb{R}$ and $b, c \in [1,\infty]$.

First we note that $\big(X_{-1}, X_1\big)_{1/2,c} = B^0_{q,c}(\Omega)$, whence $\mathrm{Id} + A$ has L^c_*-estimates on X if and only if $B^0_{q,v}(\Omega) \hookrightarrow B^0_{q,c}(\Omega)$ which is equivalent to $c \geq v$. Notice that this condition does not depend on q. If $c \geq v$, then the characterizing Theorem 1.7 applies, and L^c-admissibility of type γ of the observation operator $C = \gamma_0$ may be checked by verifying the boundedness condition on the set W_C. Again, we use Theorem 1.9 and find that, for $2\gamma + 2/c < 1 + 1/q$, W_C is bounded if and only if γ_0 has a continuous extension to the very same Banach space

$$Z = B^{2\gamma+2/c}_{q,1}(\Omega)$$

we calculated above. Hence the proof of the following can be done as in the case $X = L^q(\Omega)$.

PROPOSITION 2.7. *Let $\gamma \in \mathbb{R}$, $c \in [1,\infty]$, and $q, v \in (1,\infty)$ satisfy $c \geq v$ and $2\gamma + 2/c \in (1/q, 1+1/q)$. Let Y be a Banach space and $X = B^0_{q,v}(\Omega)$. Then $\gamma_0 : \mathcal{D}(A) \to Y$ is L^c-admissible of type γ for $Id + A$ if and only if $B^{2\gamma+2/c-1/q}_{q,1}(\partial\Omega) \hookrightarrow Y$.*

We turn to L^b-admissibility of type β for $B = \gamma_0'$. For an application of Theorem 1.8 in $X = B^0_{q,v}(\Omega)$ in case $\beta = 0$ we need $L^{b'}_*$-estimates for $(\mathrm{Id} + A)'$, which are, by the argument above, equivalent to $b \leq v$. This yields the next proposition.

PROPOSITION 2.8. *Let $\beta \in \mathbb{R}$, $b \in [1,\infty]$, and $q, v \in (1,\infty)$ satisfy $2/b + 2\beta \in (1/q, 1 + 1/q)$. Assume additionally $b \leq v$ if $\beta = 0$. Let U be a Banach space and $X = B^0_{q,v}(\Omega)$. Then $\gamma_0' : U \to X_{-1}$ is L^b-admissible of type β for $Id + A$ if and only if $U \hookrightarrow B^{2/b+2\beta-1-1/q}_{q,\infty}(\partial\Omega)$.*

Wellposedness of the full system. We study L^p-wellposedness of type α for the system (2.1) where $\alpha \in \mathbb{R}$, $p \in [1,\infty]$ and the state space $X = B^0_{q,v}(\Omega)$ with $q, v \in (1,\infty)$. Again we have $c = b = p$ and $\gamma = \beta = \alpha$ in the admissibility situations above. The assumptions of Propositions 2.7 and 2.8 now lead, for $\alpha \neq 0$, to the restrictions $p \geq v$ and $2\alpha + 2/p \in (1/q, 1 + 1/q)$. In case $\alpha = 0$ we are led to $p = v$. For example, if $p = q$, then we are led to $2\alpha \in (-1/q, 1/q')$, which, for $q = 2$, corresponds to the condition $|\alpha| < 1/4$ in Proposition 2.5 (recall $L^2(\Omega) = B^0_{2,2}(\Omega)$). By the arguments used above we hence obtain the next proposition.

PROPOSITION 2.9. *Let $\alpha \in \mathbb{R}$, $p \in (1,\infty)$, and $q \in (1,\infty)$ satisfy $2\alpha + 2/p \in (1/q, 1 + 1/q)$. Let $X = B^0_{q,p}(\Omega)$ and let Y and U be Banach spaces on $\partial\Omega$ such that $B^{2\alpha+2/p-1/q}_{q,p}(\partial\Omega) \hookrightarrow Y$ and $U \hookrightarrow B^{2/p+2\alpha-1-1/q}_{q,p}(\partial\Omega)$. Then the system (2.1) is finite-time L^p-wellposed of type α.*

Proof. The proof is similar to the case $X = L^q(\Omega)$. We check here that Theorem 1.13 applies, i.e., that $\alpha \in (-1/p, 1/p')$. The condition on α in the assumption may be rephrased as $\alpha \in (1/2q - 1/p,\ 1/2 + 1/2q - 1/p)$. In particular $\alpha > 1/2q - 1/p \geq -1/p$ and $\alpha < 1/2 + 1/2q - 1/p \leq 1 - 1/p$. ∎

We also give an analogue of Proposition 2.6 for Besov spaces, e.g., for state spaces $X = B^\delta_{q,p}(\Omega)$, where $q, p \in (1,\infty)$, and we restrict to $\delta \in (-1/q', 1/q)$ for the same reasons as before.

PROPOSITION 2.10. *Let $q, p \in (1,\infty)$, $\delta \in (-1/q', 1/q)$, $2\alpha + 2/p + \delta \in (1/q, 1 + 1/q)$, and $X = B^\delta_{q,p}(\Omega)$. Let Y and U be Banach spaces on $\partial\Omega$.*

(a) *The operator $C = \gamma_0 : \mathcal{D}(A) \to Y$ is L^p-admissible of type α for $Id + A$ if and only if $B^{2\alpha+2/p+\delta-1/q}_{q,1}(\partial\Omega) \hookrightarrow Y$.*

(b) *The operator $B = \gamma_0' : U \to X_{-1}$ is L^p-admissible of type $-\alpha$ for $Id + A$ if and only if $U \hookrightarrow B^{2\alpha+2/p+\delta-1-1/q}_{q,\infty}(\partial\Omega)$.*

(c) *Let* $r \in [1,\infty]$ *and* $B^{2\alpha+2/p+\delta-1/q}_{2,r}(\partial\Omega) \hookrightarrow Y$ *and* $U \hookrightarrow B^{2\alpha+2/p+\delta-1-1/q}_{2,r}(\partial\Omega)$. *Then the system* (2.1) *is finite-time* L^p*-wellposed of type* α.

Discussion. We discuss our results for the system (2.1) by starting from $Y := B^s_{q,r}(\partial\Omega)$ and $U := B^{s-1}_{q,r}(\partial\Omega)$, where $s \in (0,1)$, $q \in (1,\infty)$, and $r \in [1,\infty]$ are fixed. We now look for α, p, and a state space X such that (2.1) is L^p-wellposed of type α. This should be compared with the situation studied in [4], where $p = q = r = 2$, $\alpha = 0$, $s = 1/2$, and $X = L^2(\Omega)$.

Remark 2.11. In case $q = 2$, we may take $p = 2$ and $X = H^\delta_2(\Omega)$, where the restrictions may be read off Proposition 2.6: $2\alpha + \delta + 1/2 = s$ and $|\delta| < 1/2$. Thus we see that a suitable choice of α always allows us to have δ arbitrarily close to $-1/2$. In particular this applies to the "classical" case $s = 1/2$, where the restriction $\alpha = 0$ forces $\delta = 0$ and $X = L^2(\Omega)$.

Remark 2.12. For general $q \in (1,\infty)$, we may apply Proposition 2.10 and take $X = B^\delta_{q,p}(\Omega)$ under the restrictions $s = 2\alpha + 2/p + \delta - 1/q$ and $\delta \in (-1/q', 1/q)$. We see that taking p arbitrarily large and δ arbitrarily close to $-1/q'$, we still obtain finite-time L^p-wellposedness of type α for the system with state space $X = B^\delta_{q,p}(\Omega)$ by adjusting α. Observe that the state space $X = H^\delta_q(\Omega)$ would have required the additional restriction $p \geq \max(2,q)$ for $\alpha \neq 0$.

Remark 2.13. We let $s = 1/2$ and $U := H_2^{-1/2}(\partial\Omega)$, $Y := H_2^{1/2}(\partial\Omega)$; i.e., we take $q = r = 2$ in the situation discussed above. Then, for $\delta \in (-1/2, 1/2)$ and $X = B^\delta_{2,p}(\Omega)$, we have that (2.1) is finite-time L^p-wellposed of type α if $\delta = 1 - 2(\alpha + 1/p)$. The restrictions are $p \in (1,\infty)$ and $\alpha + 1/p \in (1/4, 3/4)$. This means, in other words, that for $\epsilon \in (0,1)$, $p \in [2,\infty)$, and $X = B^{-1/2+\epsilon}_{2,p}(\Omega)$, the system (2.1) is finite-time L^p-wellposed of type $\alpha = 3/4 - \epsilon/2 - 1/p$.

Now we consider a nonlinear feedback in the setting of the above remark.

Example 2.14. Since $Y = H_2^{1/2}(\partial\Omega) \hookrightarrow L^{\frac{2n-2}{n-2}}(\partial\Omega)$ and since $L^{2-2/n}(\partial\Omega) \hookrightarrow H_2^{-1/2}(\partial\Omega)$, Hölder's inequality yields that any $g \in L^{n-1}(\partial\Omega)$ induces a bounded multiplication operator $Y \to U$, $y \mapsto g \cdot y$. We take a smooth open subset Ω_0 with $\overline{\Omega_0} \subset \Omega$, e.g., a small ball, and let $\psi(x) := \int_{\Omega_0} x(\omega)\, d\omega = \langle x, 1_{\Omega_0}\rangle$ for $x \in X$. Observe that $\psi \in X'$ for all spaces X mentioned above (cf. [29]). Taking a Lipschitz-continuous function $f : \mathbb{R} \to \mathbb{R}$, we let $F(x)y := f(\psi(x))\, g \cdot y$. Then $F : X \to B(Y,U)$ is Lipschitz-continuous. We interpret $x \mapsto \psi(x)$ as a distributed measurement in Ω affecting via $f(\psi(x))$ the intensity of the linear feedback $y \mapsto g \cdot y$.

For $\epsilon \in (0,1)$, $p \in [2,\infty)$, $X = B^{-1/2+\epsilon}_{2,p}(\Omega)$, and $\alpha = 3/4 - \epsilon/2 - 1/p$ we take $x_0 \in X$ satisfying (for simplicity) $f(\psi(x_0)) = 0$ and show that Theorem 1.16 applies. We have $F(x_0) = 0$; hence $A_0 = A$. We take $Z = H^1_2(\Omega)$ whence by an application of Theorem 1.7 and Theorem 1.9, Id_Z is finite-time L^p-admissible of type α (cf. [29, Rem. 2.8.1]).

Since $B = \gamma_0' : H_2^{-1/2}(\partial\Omega) \to (H^1_2(\Omega))' =: W$ is bounded and $\mathrm{Id} + A$ has maximal L^p-regularity in W we obtain by Remark 1.4 and Theorem 1.13 that the system (A_0, B, Id_Z) is L^p-wellposed of type α in finite time. Hence we may apply Theorem 1.16 and obtain that the nonlinear system

$$\left\{\begin{array}{lcll} x'(t) - \Delta x(t) & = & 0, & t \in [0,\tau], \\ x(0) & = & x_0, & \\ \frac{\partial x(t)}{\partial \nu} & = & f\left(\int\limits_{\Omega_0} x(t)\, d\omega\right) g \cdot x(t)|_{\partial\Omega} & \end{array}\right. \tag{2.3}$$

has, for some $\tau(x_0) > 0$, a unique mild solution $x(\cdot)$ in the space $C([0,\tau], B_{2,p}^{-1/2+\epsilon}(\Omega)) \cap L^p_\alpha([0,\tau], H^1_2(\Omega))$. This means that we obtain solutions also for rather rough initial data x_0.

In the example above we made the assumption $f(\psi(x_0)) = 0$ for simplicity. For initial data x_0 with $f(\psi(x_0)) \neq 0$ one may resort to perturbation results in [11].

3. Proofs. Finite-time admissibility of type α.

Proof of Lemma 1.3. Assume C to be finite-time L^p-admissible of type α for A. For $b \geq 0$ we then have

$$\begin{aligned}\left(\int_{b+\tau/2}^{b+\tau} \|t^\alpha CT(t)x\|^p\,dt\right)^{1/p} &= \left(\int_{\tau/2}^{\tau} \|s^\alpha (1+\tfrac{b}{s})^\alpha CT(s)T(b)x\|^p\,dt\right)^{1/p}\\ &\leq \left(\int_0^\tau \|s^\alpha CT(s)T(b)x\|^p\,dt\right)^{1/p} \max((1+\tfrac{2b}{\tau})^\alpha, 1)\\ &\leq M_\tau \,\|T(b)x\|\, \max((1+\tfrac{2b}{\tau})^\alpha, 1).\end{aligned}$$

If $T(\cdot)$ is uniformly exponentially stable, i.e., if there are $c, \epsilon > 0$ such that $\|T(t)\| \leq c\,e^{-\epsilon t}$, $t \geq 0$, then we obtain for $b = (k-1)\tau/2$ and $k \in \mathbb{N}_0$,

$$\left(\int_{k\tau/2}^{(k+1)\tau/2} \|t^\alpha CT(t)x\|^p\,dt\right)^{1/p} \leq c\,M_\tau\, \max((1+\tfrac{2b}{\tau})^\alpha, 1)\, e^{-\epsilon(k-1)\tau/2}\|x\|.$$

Thus (1.4) holds.

A similar reasoning shows that the notion of finite-time L^p-admissibility of type α for control operators is independent of τ: let $a = \tau/2$ and $b = 3/2\,\tau$. Since

$$\int_0^b T(b-s)Bu(s)\,ds \leq T(a)\int_0^\tau T(\tau-s)Bu(s)\,ds + \int_a^\tau T(\tau-s)Bu(s+a)\,ds,$$

L^p-admissibility of type α on $[0,\tau]$ gives for $p < \infty$

$$\begin{aligned}\left\|\int_0^b T(b-s)Bu(s)\,ds\right\| &\leq K_\tau\|T(a)\|\,\|u\|_{L^p_\alpha([0,\tau])} + K_\tau\|\mathbb{1}_{[a,\tau]}(\cdot)u(a+\cdot)\|_{L^p_\alpha([0,\tau])}\\ &\leq K_\tau\|T(a)\|\,\|u\|_{L^p_\alpha([0,\tau])} + K_\tau\widetilde{c}_\alpha\,\|u\|_{L^p_\alpha([\tau,b])}\\ &\leq K_\tau\big(\widetilde{c}_\alpha + \|T(\tau/2)\|\big)\,\|u\|_{L^p_\alpha([0,b])},\end{aligned}$$

where $\widetilde{c}_\alpha = \max\big((\frac{1}{2})^\alpha, (\frac{2}{3})^\alpha\big)$ does not depend on τ. In case $p = \infty$, we obtain the same estimate with $\widetilde{c}_\alpha = 1$ by directly regarding the first line of the above inequalities. Thus, B is L^p-admissible of type α on $[0, 3/2\,\tau]$ as required. An iteration of the argument shows that

$$K_{(3/2)^n\tau} \leq K_\tau(\widetilde{c}_\alpha)^n \prod_{j=1}^n \big(1 + \widetilde{c}_\alpha^{-1}\|T(3\tau(\tfrac{3}{2})^j)\|\big).$$

If $T(\cdot)$ is uniformly exponentially stable and if $\alpha \geq 0$ or if $p = \infty$, the left-hand side of the above inequality is bounded since $\sum_{j\geq 1}\exp(-3\epsilon\tau(\tfrac{3}{2})^j) < \infty$ and $|\widetilde{c}_\alpha| \leq 1$. □

If $\alpha < 0$ and $p < \infty$, equivalence of finite-time and infinite-time L^p-admissibility of type α fails in general.

Example 3.1. Take $U = X = \mathbb{C}$, $-\alpha = \beta > 0$, and $T(t) = e^{-\epsilon t}$. Then, for any $\tau > 0$,

$$|T * u(\tau)| \le \|u\|_{L^p_\alpha(0,\tau)} \|e^{-\epsilon(\cdot)}\|_{L^{p'}_\beta(0,\tau)} \le c_{p'} \tau^{\beta + 1/p'} \|u\|_{L^p_\alpha(0,\tau)},$$

and the identity is finite-time L^p-admissible of type α.

On the other hand, letting $u_k := \mathbb{1}_{[k,k+1]} e^{-\epsilon(\cdot - k)}$ for $k \in \mathbb{N}$, we have $\|u_k\|_{L^p_\alpha} \le k^{-\beta} \|u_k\|_{L^p} = c k^{-\beta}$ and for $\delta \in (0,1]$,

$$(T * u_k)(k+\delta) = \int_k^{k+\delta} e^{-\epsilon(k+\delta-s)}\, e^{-\epsilon(s-k)}\, ds = \delta e^{-\epsilon\delta}.$$

Hence the identity is not infinite-time L^p-admissible of type α.

Characterization of L^p-admissibility of type α. In the proof of Theorems 1.7 and 1.8 we make use of the following lemma.

LEMMA 3.2 (see [12, Lem. 4.1]). *Let $\sigma \in (0,\pi)$, let $\varphi \in H^\infty_0(S(\sigma))$, and let $m \ge 1$ be an integer. There exist a function $f \in H^\infty_0(S(\sigma))$ and a constant $a \in \mathbb{C}$ such that*

$$\varphi(z) = z^m f^{(m)}(z) + a \frac{z^m}{(1+z)^{m+1}}, \qquad z \in S(\sigma). \tag{3.1}$$

Furthermore, if $\delta, \epsilon \in (0,1)$ are positive numbers such that

$$|\varphi(z)| = O(|z|^{-\delta}) \quad \text{at } \infty \qquad \text{and} \qquad |\varphi(z)| = O(|z|^{\epsilon}) \quad \text{at } 0,$$

then f can be chosen so that we also have $|f(z)| = O(|z|^{-\delta})$ at ∞ and $|f(z)| = O(|z|^{\epsilon})$ at 0.

Proof of Theorem 1.7. (a) Notice that for any $\lambda \in \mathbb{C}$ with positive real part and for any $x \in X$, we have

$$(\lambda + A)^{-k} x = \frac{1}{(k-1)!} \int_0^\infty t^{k-1} e^{-\lambda t} T(t) x \, dt.$$

For $x \in X_k = \mathcal{D}(A^k)$, the integrand $t \mapsto t^{k-1} e^{-\lambda k} T(t)x$ belongs to $L^1(\mathbb{R}_+, X_k)$, and so continuity of C on X_k shows

$$C(\lambda + A)^{-k} x = \frac{1}{(k-1)!} \int_0^\infty t^{k-1} e^{-\lambda t} C T(t) x \, dt.$$

By Hölder's inequality we thus have for $x \in X_k$

$$\begin{aligned}
& \big\| C(\lambda + A)^{-k} x \big\| \\
\le\ & \frac{1}{(k-1)!} \int_0^\infty \big\| t^\alpha C T(t) x \big\| t^{k-1-\alpha} e^{-Re(\lambda)\, t}\, dt \\
\le\ & \frac{1}{(k-1)!} \Big\| t \mapsto t^\alpha C T(t) x \Big\|_{L^p(\mathbb{R}_+, Y)} \left(\int_0^\infty t^{(k-1-\alpha)p'} e^{-Re(\lambda) p'\, t}\, dt \right)^{1/p'} \\
\overset{s = Re(\lambda) p'\, t}{\le}\ & \frac{M}{(k-1)!} \|x\|_X \Big((p' \mathrm{Re}(\lambda))^{-1-(k-1-\alpha)p'} \Gamma\big(1 + (k-1-\alpha)p'\big) \Big)^{1/p'} \\
=\ & K \mathrm{Re}(\lambda)^{-k+\alpha+1/p}\, \|x\|_X,
\end{aligned}$$

where Γ is the usual Gamma function and the number K depends only on k, p, and the admissibility constant M. By density of X_k in X this shows the first assertion.

(b) Without loss of generality we may assume $k \geq 2$ since by sectoriality of A, whenever W_C is bounded for some $k \in \mathbb{N}$, it is also bounded when k is replaced by $k+1$ (see Remark 1.10). We make use of the (unbounded) operator A^{-1} that is densely defined on the range of A. We set $F_k(z) := z^{k-1}e^{-z}$. Then for any $x \in X_k$ and any $t > 0$, we have

$$t^\alpha CT(t)x = t^{\alpha-k+1}CA^{-k+1}F_k(tA)x. \tag{3.2}$$

For some $\epsilon \in (0,1)$ that we will specify later, consider the decomposition $F_k(z) = \varphi(z)\psi(z)$, where

$$\varphi(z) = z^\epsilon(1+z)^{-1} \qquad \text{and} \qquad \psi(z) = z^{k-1-\epsilon}(1+z)e^{-z}. \tag{3.3}$$

Note that $\psi \in H_0^\infty(S(\theta))$ for any $\theta < \frac{\pi}{2}$, whereas $\varphi \in H_0^\infty(S(\sigma))$ for any $\sigma < \pi$. By (3.2), we have

$$\int_0^\infty \|t^\alpha CT(t)x\|_Y^p dt \leq \int_0^\infty \| \underbrace{\left[t^{\alpha+1/p-k+1}CA^{-k+1}\psi(tA)\right]}_{=:K(t)} t^{-1/p}\varphi(tA)x\|_X^p \, dt.$$

Now we will show that the operator family $K(t)$, $t > 0$, is uniformly bounded. Once this is done, the assertion of the theorem follows immediately from the assumed L^p_*-estimate for A (cf. (1.6)).

We fix $\sigma \in (\omega, \pi)$ and apply Lemma 3.2 to ψ with $m = k-1$ and $\delta = 1-\epsilon$. Let $f \in H_0^\infty(S(\sigma))$ denote the corresponding function satisfying (3.1). Note that according to that equation, $z \mapsto z^{k-1}f^{(k-1)}(z)$ belongs to $H_0^\infty(S(\sigma))$. Let $\theta = \theta_\sigma$ for some $\theta \in (\omega, \sigma)$ and let Γ denote the positively orientated boundary of $S(\theta)$. Then, as in the proof of [12, Thm. 4.2] the following representation formula for x in the dense subspace $Z := ran(A^{k-1}(I+A)^{-k})$ holds:

$$CA^{-k+1}[z^{k-1}f^{(k-1)}(z)](tA)x = \tfrac{(k-1)!}{2\pi i}\int_\Gamma f(\lambda)t^{k-1}\, CR(\lambda, tA)^k x \, d\lambda, \quad t > 0. \tag{3.4}$$

For $\lambda \in \Gamma$, by the resolvent equation we have

$$\lambda^{k-\alpha-1/p}CR(\lambda,A)^k = |\lambda|^{k-\alpha-1/p}C(|\lambda|+A)^{-k}\left[2\cosh(\pm\theta/2)\lambda R(\lambda,A) - I\right]^k, \tag{3.5}$$

and thus $\lambda^{k-\alpha-1/p}CR(z,A)^k$ is uniformly bounded by sectoriality of A. Now, by the representation (3.4),

$$K(t) = \tfrac{(k-1)!}{2\pi i}\int_\Gamma t^{\alpha+1/p}f(\lambda)\, CR(\lambda,tA)^k x\, d\lambda + at^{\alpha+1/p}C(I+tA)^{-k}$$

on Z. Next we show that for an appropriately chosen $\epsilon \in (0,1)$, $K(t) \in B(X,Y)$ and moreover the operators $K(t)$ are uniformly bounded for $t > 0$. To this end, write

$$\begin{aligned} K(t) = \tfrac{(k-1)!}{2\pi i}\int_\Gamma f(\lambda)\lambda^{1-k+\alpha+1/p}\left[\left(\tfrac{\lambda}{t}\right)^{k-\alpha-1/p} CR(\tfrac{\lambda}{t},A)^k x\right]\tfrac{d\lambda}{\lambda} \\ + a\left[\left(\tfrac{1}{t}\right)^{k-\alpha-1/p} C(\tfrac{1}{t}+A)^{-k}\right]. \end{aligned}$$

By our assumption (1.7) and scaling invariance of Γ and the measure $d\lambda/\lambda$ we obtain that $K(t)$ is uniformly bounded, provided that the integral

$$\int_\Gamma \big|f(\lambda)\big|\big|\lambda\big|^{-k+\alpha+1/p}\, d|\lambda|$$

is finite. By the estimates in Lemma 3.2 we know that $f \in O(|z|^{k-1-\epsilon})$ at zero and $f \in O(|z|^{-n})$ for any $n \in \mathbb{N}$ at infinity. Therefore the above integral is finite if

$$k - 1 - \epsilon - k + \alpha + 1/p > -1, \quad \text{i.e., if} \quad \epsilon < \alpha + 1/p.$$

This, however, due to our assumption on α, may always be satisfied by an appropriate choice of $\epsilon \in (0,1)$, and the proof is done. □

To analyze L^p-admissibility for control operators, assume that $\|T_{-k}(t)B\|_{U\to X} \le Mt^{-\gamma}$ for some $\gamma \in \mathbb{R}$. For $t > 0$ fixed we thus have

$$\begin{aligned}\left\|\int_0^t T(t-s)Bu(s)\,ds\right\|_X &\le \int_0^t \big\|T(t-s)B\big\|_{U\to X}\, s^{-\alpha}\, \|s^\alpha u(s)\|_U\, ds \\ &\le c\int_0^t (t-s)^{-\gamma} s^{-\alpha} \|s^\alpha u(s)\|_U\, ds.\end{aligned}$$

Let $k_{\alpha,\gamma}(t,s) = \mathbb{1}_{(0,t)}(s)(t-s)^{-\gamma}s^{-\alpha}$ for $s,t \in (0,\tau)$. Thus the study of the kernel $k_{\alpha,\gamma}$ which may or may not induce a bounded integral operator $K_{\alpha,\gamma} : L^p(0,\tau) \to L^\infty(0,\tau)$ gives a sufficient criterion for (in)finite-time L^p-admissibility of type α of control operators.

For $p = 1$, $K_{\alpha,\gamma}$ is bounded if and only if $k_{\alpha,\gamma}$ is uniformly bounded, which, for finite τ, is equivalent to $\gamma \le 0$ and $\alpha \le 0$. In this case we have

$$\|K_{\alpha,\gamma}\|_{L^\infty(0,\tau)\to L^\infty(0,\tau)} = \|k_{\alpha,\gamma}\|_\infty = \frac{|\alpha|^{|\alpha|}|\gamma|^{|\gamma|}}{|\alpha+\gamma|^{|\alpha+\gamma|}}\,\tau^{|\alpha+\gamma|}.$$

For $\tau = \infty$, $k_{\alpha,\gamma}$ is uniformly bounded if and only if $\alpha = \gamma = 0$.

For $p > 1$ we use Hölder's inequality and obtain

$$\begin{aligned}\int_0^t k_{\alpha,\gamma}(t,s)|f(s)|\,ds &\le \left(\int_0^t \big(k_{\alpha,\gamma}(t,s)\big)^{p'}\,ds\right)^{1/p'} \|f\|_p \\ \text{(subst. } s = t\sigma) &= \left(t^{1-(\gamma+\alpha)p'}\int_0^1 (1-\sigma)^{-\gamma p'}\sigma^{-\alpha p'}\,d\sigma\right)^{1/p'} \|f\|_p.\end{aligned}$$

Convergence of the integral in the last line is equivalent to $\gamma < 1/p'$ and $\alpha < 1/p'$. Taking the sup over $t \in (0,\tau)$ we see that, for finite τ,

$$\|K_{\alpha,\gamma}\|_{L^p(0,\tau)\to L^\infty(0,\tau)} \le c_{\alpha,\gamma,p}\tau^{1/p'-(\gamma+\alpha)}$$

provided that $\gamma < 1/p'$, $\alpha < 1/p'$, and $\gamma + \alpha \le 1/p'$, whereas, for $\tau = \infty$,

$$\|K_{\alpha,\gamma}\|_{L^p(0,\infty)\to L^\infty(0,\infty)} \le c_{\alpha,\gamma,p}$$

provided that $\gamma + 1/p < 1$, $\alpha + 1/p < 1$, and $\gamma + \alpha + 1/p = 1$. Since, in fact (cf. [16]),

$$\|K_{\alpha,\gamma}\|_{L^p(0,\tau)\to L^\infty(0,\tau)} = \sup_{t\in(0,\tau)} \|k_{\alpha,\gamma}(t,\cdot)\|_{L^{p'}(0,\tau)},$$

we have just proved the following result.

PROPOSITION 3.3. *The operator $K_{\alpha,\gamma}$ is bounded $L^p(0,\tau) \to L^\infty(0,\tau)$ if and only one of the following conditions holds:*

$$\begin{array}{lllll} \text{(i)} & p=1, & \tau<\infty, & \alpha\le 0, & \gamma\le 0, \\ \text{(ii)} & p=1, & \tau=\infty, & \alpha=0, & \gamma=0, \\ \text{(iii)} & p>1, & \tau<\infty, & \alpha+1/p<1, & \gamma+1/p<1, \quad \alpha+\gamma+1/p\le 1, \\ \text{(iv)} & p>1, & \tau=\infty, & \alpha+1/p<1, & \gamma+1/p<1, \quad \alpha+\gamma+1/p=1. \end{array} \tag{3.6}$$

Observe that condition (iv) *implies $\alpha>0$ and $\gamma>0$.*

Proof of Theorem 1.8. (a) Consider for $t>0$ the function $u(s) := \mathbb{1}_{(t/2,t)}(s)u_0$. Then

$$\left\|\int_0^t T(t-s)Bu(s)\,ds\right\|_X = \left\|\int_{t/2}^t AT(t-s)Bu(s)\,ds\right\|_{\dot X_{-1}} = \big\|\,[T(t)-T(t/2)]Bu_0\big\|_{\dot X_{-1}}.$$

We have $\|u\|_{L^p_\alpha([0,t],U)} = \frac{1}{1+\alpha p}\big(\frac{t}{2}\big)^{\alpha+1/p}$. Let $f(z) := \exp(z)-\exp(z/2)$. Notice that the function $z \mapsto z^{-\alpha-1/p} f(z) \in H_0^\infty(S(\sigma))$ for all $\sigma < \pi/2$, so Theorem 1.5 applies. We obtain

$$\begin{aligned} \mathcal{R}(B) &\subseteq \big((\dot X(A_{-1}))_{-1}, (\dot X(A_{-1}))_1\big)_{1/2(\alpha+1/p+1),\infty} \\ &= \big(\dot X_{-2}, X\big)_{1/2(\alpha+1/p+1),\infty} = \big(\dot X_{-k}, X\big)_{1-\theta,\infty} \end{aligned}$$

with $\theta = 1/k\,(1/p'-\alpha)$, $k\in\mathbb{N}$. The claim now follows from Theorem 1.9.

(b) By Theorem 1.9, boundedness of W_B is equivalent to $B: U \to (\dot X_{-k}, X)_{\theta,\infty}$ with $\theta = 1+1/k\,(\alpha-1/p')$. By analyticity of the semigroup this implies $\|T(t)B\| \le ct^{-\gamma}$ with $\gamma = 1/p'-\alpha$ (cf. the arguments in the proof of Lemma 1.12). Hence, if in addition to the assumptions of the theorem, $\alpha>0$ in case $p>1$ (or $p=1$ in case $\alpha=0$, respectively), Proposition 3.3 gives the claim.

(a)′ Let (1.5) hold. For $\mathrm{Re}(\lambda)>0$ and $u\in U$ we have

$$(\lambda + A_{-k})^{-k} Bu = \frac{1}{(k-1)!}\int_0^\infty t^{k-1} e^{-\lambda t} T_{-k}(t)Bu\,dt.$$

Then, by assumption

$$\begin{aligned} &\big\|\mathrm{Re}(\lambda)^{k+\alpha-1/p'}\,(\lambda+A)^{-k-1}Bu\big\| \\ &\le \tfrac{1}{(k-1)!}\left\|\int_0^\infty t^{k-1}\mathrm{Re}(\lambda)^{k+\alpha-1/p'}\,e^{-\lambda t}T_{-k}(t)Bu\,dt\right\| \\ &= \tfrac{1}{(k-1)!}\left\|\int_0^\infty T_{-k}(t)B\big[t^{k-1}\mathrm{Re}(\lambda)^{k+\alpha-1/p'}\,e^{-\lambda t}\otimes u\big]\,dt\right\| \\ &\le \tfrac{K}{(k-1)!}\big\|h_\lambda(t)\otimes u\big\|_{L^p_\alpha(\mathbb{R}_+,U)} \le \widetilde K\|u\|_U. \end{aligned}$$

Here the uniform boundedness of the functions $h_\lambda(t) := \mathrm{Re}(\lambda)^{k+\alpha-1/p'}\,t^{k-1}e^{-\lambda t}$, $\lambda>0$, in $L^p_\alpha(\mathbb{R}_+)$ is shown similar to the proof of Theorem 1.7.

(b)′ Without loss of generality we assume $k\ge 2$. We proceed in two steps.

Step 1. To show the existence of the integral in (1.5), we choose some $x'\in(X_{-k})'$, $x'\ne 0$. Notice that $(X_{-k})'$ may be identified with $(X')_k$, that is, the domain $\mathcal{D}((A')^k)$ with graph norm. We consider

$$\int_0^\infty \big|\langle T_{-k}(t)Bu(t), x'\rangle\big|\,dt = \int_0^\infty \big|\langle t^{-k+1}A^{-k+1}(tA)^{k-1}T_{-k}(t)Bu(t), x'\rangle\big|\,dt.$$

As we did in the proof of Theorem 1.7, we decompose $F_k(z) = z^{k-1}e^{-z}$ as $F_k(z) = \varphi(z)\psi(z)$ with φ, ψ as in (3.3). We obtain

$$\int_0^\infty \big|\langle T_{-k}(t)Bu(t), x'\rangle\big|\,dt = \int_0^\infty \big|\langle t^{-k+1}A_{-k}^{-k+1}\varphi(tA_{-k})\psi(tA_{-k})Bu(t), x'\rangle\big|\,dt.$$

Notice that by sectoriality of A, the operators $A_{-k}(\mu + A_{-k})^{-1}$, $\mu > 0$, are uniformly bounded. Moreover, $\lim_{\mu\to 0+} A_{-k}(\mu + A_{-k})^{-1}Bu = Bu$ in X_{-k} since A_{-k} has dense range in X. Therefore, applying Fatou's lemma and writing $B_\mu := A_{-k}(\mu+A_{-k})^{-1}B$, we have

$$\begin{aligned}&\int_0^\infty \big|\langle T_{-k}(t)Bu(t), x'\rangle\big|\,dt\\ &\le \liminf_{\mu\to 0+} \int_0^\infty \big|\langle t^{-k+1}A_{-k}^{-k+1}\varphi(tA_{-k})\psi(tA_{-k})B_\mu u(t), x'\rangle\big|\,dt.\end{aligned}$$

Notice that $\psi(tA_{-k})Bu(t) \in \mathcal{D}(A_{-k}^{-k+1+\epsilon})$, whereas $\psi(tA_{-k})B_\mu u(t) \in \mathcal{D}(A_{-k}^{-k+\epsilon})$. This observation allows us to interchange the operators $\varphi(tA_{-k})$ and A_{-k}^{-k+1} as follows:

$$\begin{aligned}&= \liminf_{\mu\to 0+} \int_0^\infty \big|\langle \varphi(tA_{-k})t^{-k+1}A_{-k}^{-k+1}\psi(tA_{-k})B_\mu u(t), x'\rangle\big|\,dt\\ &= \liminf_{\mu\to 0+} \int_0^\infty \big|\langle t^{1+1/p'-k}A_{-k}^{-k}\psi(tA_{-k})B_\mu u(t), t^{-1/p'}\varphi(tA_{-k})'x'\rangle\big|\,dt\\ &\le \liminf_{\mu\to 0+} \Big\| t \mapsto \underbrace{\big[t^{1/p'-\alpha-k+1}A_{-k}^{-k-1}\psi(tA_{-k})B_\mu\big]}_{=:L_\mu(t)} t^\alpha u(t)\Big\|_{L^p(\mathbb{R}_+,X)} \big\|\varphi(tA')x'\big\|_{L^{p'}(\mathbb{R}_+,\frac{dt}{t},X')}.\end{aligned}$$

Notice that by assumption on A', the $L^{p'}$-norm has an estimate against the norm of x', whence the existence of the integral is proved if we show the uniform boundedness of the operators $L_\mu(t)$ for $t > 0$ and $\mu > 0$. This step is very similar to the proof of uniform boundedness of the family $K(t)$, $t > 0$, in the proof of Theorem 1.7: we apply Lemma 3.2 with $m = k - 1$ to the function $\psi(z)$ and obtain for fixed $t > 0$

$$\begin{aligned}&A_{-k}^{-k+1}\psi(tA_{-k})B_\mu u(t)\\ =\ &\tfrac{(k-1)!}{2\pi i}\int_\Gamma f(\lambda)t^{k-1}R(\lambda, tA_{-k})^k B_\mu u(t)\,d\lambda + at^{k-1}(I + tA_{-k})^{-k}B_\mu u(t).\end{aligned}$$

Now for $u \in U$ write

$$\begin{aligned}L_\mu(t)u &= \tfrac{(k-1)!}{2\pi i}\int_\Gamma f(\lambda)t^{1/p'-\alpha}R(\lambda, tA_{-k})^k B_\mu u\,d\lambda + at^{1/p'-\alpha}(I + tA_{-k})^{-k}B_\mu u(t)\\ &= \tfrac{(k-1)!}{2\pi i}\int_\Gamma f(\lambda)\lambda^{-k-\alpha+1/p'}\left[A_{-k}(\mu+A_{-k})^{-1}\right]\left[\left(\tfrac{\lambda}{t}\right)^{k+\alpha-1/p'+1} R(\tfrac{\lambda}{t}, A_{-k})^k B\right] u\,\tfrac{d\lambda}{\lambda}\\ &\quad + a\left[A_{-k}(\mu+A_{-k})^{-1}\right]\left[\left(\tfrac{1}{t}\right)^{k+\alpha-1/p'}(\tfrac{1}{t} + A_{-k})^{-k}B\right] u.\end{aligned}$$

Therefore, by the assumption (1.8) and a similar calculation to (3.5) the set $\{L_\mu(t) : t > 0,\ \mu > 0\}$ is bounded in $B(U, X)$ provided that the integral

$$\int_\Gamma \big|f(\lambda)\big|\,|\lambda|^{-k-\alpha+1/p'}\,d|\lambda|$$

is finite. Since $f \in O(|z|^{k-1-\epsilon})$ in zero and $f \in O(|z|^{-n})$ for any $n \in \mathbb{N}$, this boils down to

$$k - 1 - \epsilon - k - \alpha + 1/p' > -1, \quad \text{i.e., to} \quad \alpha < 1/p' - \epsilon,$$

which, due to our assumption on α, always may be satisfied by some $\epsilon \in (0,1)$.

Step 2. Now let $x' \in X'$. We show that $t \mapsto \langle T(t)Bu(t), x'\rangle \in L^1(\mathbb{R}_+)$ with a norm estimate against $K'\|u\|_{L^p_\alpha(R_+,U)}\|x'\|_{X'}$. To this end we first notice that by analyticity of the semigroup, $T_{-k}(t)Bu(t) \in X$ for $t > 0$. Moreover, for t positive, $T_{-k}(t)Bu(t) = \lim_{\epsilon\to 0} T_{-k}(t+\epsilon)Bu(t)$ in X. Therefore, Fatou's lemma yields

$$\int_0^\infty \big|\langle T_{-k}(t)Bu(t), x'\rangle\big|\,dt \le \liminf_{\epsilon\to 0} \int_0^\infty \big|\langle T_{-k}(t)Bu(t), T_{-k}(\epsilon)'x'\rangle\big|\,dt.$$

Notice that $y'_\epsilon := T_{-k}(\epsilon)'x' \in \mathcal{D}((A')^k)$ and by Step 1,

$$\int_0^\infty \big|\langle T(t)Bu(t), y'_\epsilon\rangle\big|\,dt \le K\|u\|_{L^p_\alpha}\|y'_\epsilon\|_{X'}.$$

Since $\|T_{-k}(\epsilon)'x'\| \le K_0\|x'\|$ the integral $\int_0^\infty T(t)Bu(t)\,dt$ exists as a Pettis integral in X. The above argumentation shows that we have a bounded linear mapping $\Phi : L^p(\mathbb{R}_+, U) \to X''$, $\Phi(u) = \int_0^\infty T_{-k}(t)Bu(t)\,dt$. If u is a step function with compact support that does not contain zero, the integral in question even exists as a Bochner integral. In this case, it takes values in X by analyticity of the semigroup $T(\cdot)$. Since such step functions are dense in $L^p(\mathbb{R}_+, U)$ (recall $p < \infty$) we obtain $\mathcal{R}\Phi \subseteq X$ and thus Φ is necessarily bounded from $L^p(\mathbb{R}_+, U)$ to X. This finishes our proof. ∎

Regularity and wellposedness.

Proof of Lemma 1.12. It is not hard to see that for bounded analytic semigroups and $k \in \mathbb{N}$, $\|t^k A^k T(t)\| \le c_k < \infty$. Indeed, by the elementary functional calculus for sectorial operators (cf. [21]) and substituting $tz = \lambda$ one has

$$\begin{aligned}\big\|(tA)^k T(t)\big\| &= \left\| \tfrac{1}{2\pi i} \int_\Gamma (tz)^k e^{-tz} R(z, A)\,dz \right\| \\ &\le M \int_\Gamma |tz|^k e^{-t\,\mathrm{Re}(z)}\, \tfrac{|dz|}{|z|} = M \int_\Gamma |\lambda|^k e^{-\ \mathrm{Re}\ \lambda}\, \tfrac{|d\lambda|}{|\lambda|} =: c_k < \infty.\end{aligned}$$

This shows $\|T(t)\|_{X\to \dot X_k} \le c_k t^{-k}$, $t > 0$. On the other hand, clearly $\|T(t)\|_{X\to X} \le c_0$, $t \ge 0$. By real interpolation we obtain immediately $\|T(t)\|_{X\to Z} \le k_1 t^{-\theta k}$. Similarly, considering $\dot X_{-k}$ and X in place of X and $\dot X_k$, one obtains $\|T(t)\|_{W\to X} \le k_2 t^{-\sigma k}$. Both estimates together give the claim by the semigroup property. ∎

Proof of Proposition 1.14. Let $\phi(s) := \big|(1+s)^\alpha - 1\big|$. Then

$$\|(Tf)(t)\|_Y \le M \int_0^t \tfrac{1}{t-s}\phi(\tfrac{t-s}{s})\|f(s)\|_U\,ds,$$

whence T is pointwise bounded by a multiple of the scalar integral operator

$$(\widetilde{T}u)(t) := \int_0^t \tfrac{1}{t-s}\phi(\tfrac{t-s}{s})u(s)\,ds,$$

which has the kernel $k(t,s) = \mathbb{1}_{[0,t]}(s)\frac{1}{t-s}\phi(\frac{t-s}{s})$. Notice that, substituting $s = t\sigma$,

$$\big\|k(t,\cdot)\big\|_{p'}^{p'} = \int_0^t \left|\frac{(1+\frac{t-s}{s})^\alpha - 1}{t-s}\right|^{p'} ds = t^{1-p'} \int_0^1 \left|\frac{\sigma^{-\alpha}-1}{\sigma-1}\right|^{p'} d\sigma \le t^{1-p'}\widetilde{c},$$

where $\widetilde{c} = \widetilde{c}(p,\alpha) < \infty$ since letting $g(\sigma) := \sigma^{-\alpha}$, the limit for $\sigma \to 1$ equals $g'(1) = -\alpha$. It follows by Hölder's inequality that $|\widetilde{T}u(t)| \le cM\|u\|_p t^{-1/p}$ (notice $(1-p')/p' = -1/p$). Therefore,

$$\lambda\,\mu\big(\{t>0:\ |\widetilde{T}u(t)|\ge\lambda\}\big)^{1/p} \le \lambda\,\mu\big(\{t>0:\ M\|u\|_p t^{-1/p}\ge\lambda\}\big)^{1/p} = M\|u\|_p,$$

showing that $\widetilde{T}$ is of weak type (p,p) for every $p\in(1,\infty)$. By Marcinkiewicz interpolation, $\widetilde{T}$ is bounded on $L^p(\mathbb{R}_+)$ for $p\in(1,\infty)$, which implies the assertion. □

Proof of Theorem 1.13. Let $\mathcal{F}$ denote the convolution operator acting $L^p(\mathbb{R}_+,U) \to L^p(\mathbb{R}_+,Y)$ and $\mathcal{F}^\alpha$ denote the same operator acting $L^p_\alpha(\mathbb{R}_+,U) \to L^p_\alpha(\mathbb{R}_+,Y)$. Further let $\Phi_\alpha : L^p_\alpha(\mathbb{R}_+,\cdot) \to L^p(\mathbb{R}_+,\cdot)$ be the canonical isometric isomorphism given by $(\Phi_\alpha f)(t) := t^\alpha f(t)$. Then $T := \Phi_\alpha \mathcal{F}^\alpha \Phi_\alpha^{-1} - \mathcal{F}$ satisfies

$$(Tu)(t) = \int_0^t \big[\big(\tfrac{t}{s}\big)^\alpha - 1\big] K(t-s)\,u(s)\,ds$$

with $K(t-s) = CT(t-s)B$. By analyticity of the semigroup, $K\in C(\mathbb{R}_+, B(U,Y))$ and by hypothesis $\|K(t)\| \le M/t$. Therefore, Proposition 1.14 applies and yields boundedness of $T : L^p(\mathbb{R}_+,U) \to L^p(\mathbb{R}_+,Y)$. Hence $\mathcal{F}$ is bounded if and only if $\mathcal{F}^\alpha$ is. □

Proof of Theorem 1.16. We let $v := T(\cdot)x_0$ and denote, for $\rho,\tau>0$ to be fixed later,

$$\Sigma_{\rho,\tau} = \{x\in C([0,\tau],X)\cap L^p_\alpha([0,\tau],Z) : x(0)=x_0, \|x-w\|_\Sigma \le \rho\},$$

where $\|x\|_\Sigma := \max\{\|x\|_{C([0,\tau],X)}, \|x\|_{L^p_\alpha([0,\tau],Z)}\}$. Observe that we dropped τ in notation of the norm and that $\Sigma_{\rho,\tau}$ is complete for the metric induced by $\|x\|_\Sigma$. We let $\Gamma x := v + T(\cdot) * B(F(x)-F(x_0))Cx$ for $x\in\Sigma_{\rho,\tau}$ and shall choose ρ and τ such that Γ is a contraction on $\Sigma_{\rho,\tau}$. Then Banach's fixed point theorem ends the proofs.

If f is a simple function with values in U with compact support in $(0,\tau)$, then by analyticity of the semigroup $T(\cdot)B * f$ is a continuous, X-valued function. Since such functions are dense in $L^p_\alpha(U)$ (recall $p<\infty$), $T(\cdot)B*$ maps $L^p_\alpha(U)$ boundedly into $C([0,\tau],X)$. We let $c_v(\tau) := \|v-x_0\|_{C([0,\tau],X)}$ and $l_v(\tau) := \|v\|_{L^p_\alpha([0,\tau],Z)}$. In the following we shall drop the time interval in the norms. The assumptions imply that $\|T(\cdot)B*u\|_\Sigma \le K\|u\|_{L^p_\alpha(U)}$ for some $K>0$. We write L for the Lipschitz-constant of the function F. Then we have, for $x\in\Sigma_{\rho,\tau}$,

$$\begin{aligned}
\|\Gamma x - v\|_\Sigma &\le K\|(F(x)-F(x_0))Cx\|_{L^p_\alpha(U)}\\
&\le KL\|C\|\|x-x_0\|_{L^\infty(X)}\|x\|_{L^p_\alpha(Z)}\\
&\le KL\|C\|(\|x-v\|_{C(X)} + \|v-x_0\|_{C(X)})(\|x-v\|_{L^p_\alpha(Z)} + \|v\|_{L^p_\alpha(Z)})\\
&\le KL\|C\|(\rho + c_v(\tau))(\rho + l_v(\tau)).
\end{aligned}$$

Similarly, we obtain, for $x,\tilde{x}\in\Sigma_{\rho,\tau}$,

$$\begin{aligned}
\|\Gamma x - \Gamma\tilde{x}\|_\Sigma &\le K\|(F(x)-F(x_0))Cx - (F(\tilde{x})-F(x_0))C\tilde{x}\|_{L^p_\alpha(U)}\\
&\le K(\|(F(x)-F(x_0))C(x-\tilde{x})\|_{L^p_\alpha(U)} + \|(F(x)-F(\tilde{x}))C\tilde{x}\|_{L^p_\alpha(U)})\\
&\le KL\|C\|(\|x-x_0\|_{L^\infty(X)}\|x-\tilde{x}\|_{L^p_\alpha(Z)} + \|x-\tilde{x}\|_{C(X)}\|\tilde{x}\|_{L^p_\alpha(Z)})\\
&\le KL\|C\|(\rho + c_v(\tau) + \rho + l_v(\tau))\|x-\tilde{x}\|_\Sigma.
\end{aligned}$$

Now we choose $\rho > 0$ such that $\eta := 4KL\|C\|\rho < 1$ and then $\tau > 0$ such that $\max\{c_v(\tau), l_v(\tau)\} \le \rho$. Thus we obtain

$$\|\Gamma x - v\|_\Sigma \le \eta\rho < \rho \quad \text{and} \quad \|\Gamma x - \Gamma\tilde{x}\|_\Sigma \le \eta\|x - \tilde{x}\|_\Sigma$$

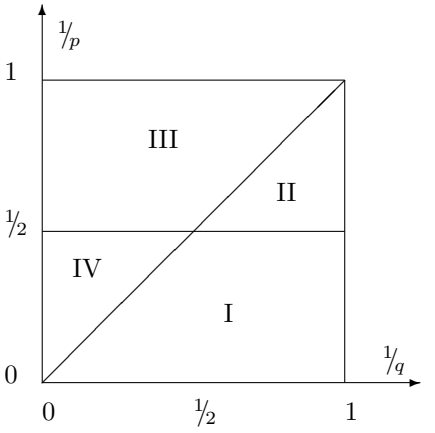
for $x, \tilde{x} \in \Sigma_{\rho,\tau}$, as desired. □

L^q-spaces and Besov spaces.

Proof of Lemma 2.1. The definition of Besov and Triebel–Lizorkin spaces together with Minkowski's inequality yield, for any $s \in \mathbb{R}$,

$$\begin{aligned} &B^s_{q,p}(\Omega) \hookrightarrow F^s_{q,p}(\Omega) \quad \text{provided that } q \ge p \text{ and} \\ &F^s_{q,p}(\Omega) \hookrightarrow B^s_{q,p}(\Omega) \quad \text{provided that } q \le p. \end{aligned}$$

This will be used in what follows. First we show the "if" part; that is, we show $L^q(\Omega) \hookrightarrow B^0_{q,p}(\Omega)$ in the case $p \ge \max(2, q)$, that is, $(1/q, 1/p) \in$ I, where area I is as depicted below. By the above embeddings of Besov and Triebel–Lizorkin spaces, $L^q(\Omega) = F^0_{q,2}(\Omega) \hookrightarrow F^0_{q,p} \hookrightarrow B^0_{q,p}$.



Next, we consider the case $(1/q, 1/p) \in$ III, that is, $p \le \min(q, 2)$ and $(p, q) \ne (2, 2)$. If $p \le 2 \le q$ and if $p < q$, we have

$$B^0_{q,p}(\Omega) \hookrightarrow B^0_{q,2} \underset{\neq}{\hookrightarrow} F^0_{q,2}(\Omega) = L^q(\Omega),$$

and if $p, q < 2$ and $p \le q$, we have

$$B^0_{q,p}(\Omega) \hookrightarrow F^0_{q,p}(\Omega) \underset{\neq}{\hookrightarrow} F^0_{q,2}(\Omega) = L^q(\Omega).$$

Therefore, obviously $L^q(\Omega) \not\hookleftarrow B^0_{q,p}(\Omega)$.

For counterexamples in areas II and IV we construct specific functions $f \in L^q$ by wavelet decompositions (cf. [23]) that show why the Besov norm cannot be estimated by the L^q-norm.

Let Λ be the set of all points $\lambda = 2^{-j}k + 2^{-j-1}\epsilon$, where $j \in \mathbb{Z}$, $k \in \mathbb{Z}^n$, and $0 \ne \epsilon \in \{0, 1\}^n$. Then every $\lambda \in \Lambda$ corresponds to unique j, k, and ϵ. Let Q_λ be the dyadic cube defined by $Q_\lambda := \{x \in \mathbb{R}^n : \ 2^j x - k \in [0, 1)^n\}$. Finally, by [23, Thm. III.8.1] choose some 1-regular wavelet basis (ψ_λ) with compact support. Then supp $\psi_\lambda \subset cQ_\lambda$ for some $c > 0$. We let $\Lambda' := \{\lambda \in \Lambda : \text{ supp } \psi_\lambda \subset \Omega\}$.

First we treat $(1/q, 1/p)$ in area II, that is, $q < p < 2$. By [23, Thm. VI.2.1], for $f = \sum_\lambda \alpha(\lambda)\psi_\lambda(x)$ in $L^q(\mathbb{R}^n)$, we have equivalence

$$\|f\|_{L^q} \sim \left\| \left(\sum_{\lambda\in\Lambda} |\alpha(\lambda)|^2 |Q_\lambda|^{-1} \mathbb{1}_{Q_\lambda} \right)^{1/2} \right\|_{L^q}. \tag{3.7}$$

In the following, it will be sufficient to consider only functions f that decompose in a finite sum. If $Q \subset \Omega$ for some $Q = Q_{\lambda_0}$, $\lambda_0 \in \Lambda'$, set $\alpha(\cdot)$ such that only dyadic subcubes of Q are considered in the above summation: if Q_λ belongs to the jth dyadic subdivision of Q, then let $\alpha(\lambda) := \alpha_j$; otherwise let $\alpha(\alpha) := 0$. Then the expression on the right-hand side of (3.7) L^q-norm of f simplifies to

$$\left\| \left(\sum_{j=0}^{N} \sum_{\Lambda'_j} \left|\alpha(\lambda)\right|^2 \mathbb{1}_{Q_\lambda} \right)^{1/2} \right\|_{L^q} = \left|Q\right|^{1/q} \left(\sum_{j=0}^{N} |\alpha_j|^2 \right)^{1/2}.$$

On the left-hand side, an equivalent $B^0_{q,p}$-norm of $f = \sum_\lambda \alpha(\lambda)\psi_\lambda(x)$ is given by

$$\|f\|_{B^0_{q,p}} \sim \left(\sum_{j=0}^{N} \left(\left(\sum_{\lambda\in\Lambda'_j} \left|\alpha(\lambda)\right|^q \right)^{1/q} 2^{-nj(1/q-1/2)} \right)^p \right)^{1/p}; \tag{3.8}$$

see [23, VI.10.5]. But, for $\lambda \in \Lambda'_j$ such that $Q_\lambda \subset Q$, $|\alpha(\lambda)| = |\alpha_j| 2^{jn/q}$, whence

$$\|f\|_{B^0_{q,p}} \sim \left(\sum_{j=0}^{N} |\alpha_j| 2^{njp/2} \right)^{1/p}.$$

Therefore, setting $\alpha_j := 2^{njp/2}$ for $j = 0, \dots, N$ and letting $N \to \infty$ reveals that $L^q(\Omega) \hookrightarrow B^0_{q,p}$ implies $p \geq 2$.

Finally, consider the case IV, that is, $q > 2$ and $2 < p < q$. If we set the wavelet coefficients $\alpha(\lambda)$ of f in (3.7) such that the cubes in $\{Q(\lambda) : \alpha(\lambda) \neq 0\}$ are piecewise disjoint, then

$$\begin{aligned}
\|f\|_q \sim \left\| \left(\sum_{\lambda\in\Lambda} |\alpha(\lambda)|^2 |Q_\lambda|^{-1} \mathbb{1}_{Q_\lambda} \right)^{1/2} \right\|_{L^q} &= \left\| \sum_{\lambda\in\Lambda} |\alpha(\lambda)| |Q_\lambda|^{-1/2} \mathbb{1}_{Q_\lambda} \right\|_{L^q} \\
&= \left(\sum_{\lambda\in\Lambda} |\alpha(\lambda)|^q |Q_\lambda|^{-q/2} |Q_\lambda| \right)^{1/q} \\
&= \left(\sum_j \left(\left(\sum_{\lambda\in\Lambda_j} |\alpha(\lambda)|^q \right)^{1/q} |Q_\lambda|^{1/q-1/2} \right)^q \right)^{1/q}.
\end{aligned}$$

On the left-hand side, notice that $|Q_\lambda| = 2^{-nj}$. Thus, comparing the L^q-norm of f with the equivalent $B^0_{q,p}$-norm given by (3.8), we find $L^q(\Omega) \hookrightarrow B^0_{q,p}(\Omega)$ requires $q \geq p$, contradicting the assumption $p < q$. □

REFERENCES

[1] H. Amann, *Linear and Quasilinear Parabolic Problems, Abstract Linear Theory*, Vol. I, Monogr. Math. 89, Birkhäuser, Boston, 1995.

[2] P. Auscher, A. McIntosh, and A. Nahmod, *Holomorphic functional calculi of operators, quadratic estimates and interpolation*, Indiana Univ. Math. J., 46 (1997), pp. 375–403.

[3] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss. 223, Springer, Berlin, 1976.

[4] C. I. Byrnes, D. S. Gilliam, V. I. Shubov, and G. Weiss, *Regular linear systems governed by a boundary controlled heat equation*, J. Dynam. Control Systems, 8 (2002), pp. 341–370.

[5] M. Cannone, *Ondelettes, paraproduits et Navier-Stokes*, with a preface by Y. Meyer, Diderot Editeur, Paris, 1995.

[6] G. Da Prato and P. Grisvard, *Sommes d'opérateurs linéaires et équations différentielles opérationnelles*, J. Math. Pures Appl. (9), 54 (1975), pp. 305–387.

[7] G. Dore, *L^p regularity for abstract differential equations*, in Functional Analysis and Related Topics, 1991 (Kyoto), Lecture Notes in Math. 1540, Springer, Berlin, 1993, pp. 25–38.

[8] K.-J. Engel, *On the characterization of admissible control- and observation operators*, Systems Control Lett., 34 (1998), pp. 225–227.

[9] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Grad. Texts in Math. 194, Springer, New York, 2000.

[10] D. Guidetti, *On interpolation with boundary conditions*, Math. Z., 207 (1991), pp. 439–460.

[11] B. H. Haak, M. Haase, and P. C. Kunstmann, *Perturbation, interpolation, and maximal regularity*, Adv. Differential Equations, 11 (2006), pp. 201–240.

[12] B. H. Haak and C. Le Merdy, *α-admissibility of observation and control operators*, Houston J. Math., 31 (2005), pp. 1153–1167.

[13] M. Haase, *The Functional Calculus for Sectorial Operators*, Oper. Theory Adv. Appl. 169, Birkhäuser, Basel, 2006.

[14] M. Haase, *Operator-valued H^∞-calculus in inter- and extrapolation spaces*, Integral Equations Operator Theory, 56 (2006), pp. 197–228.

[15] B. Jacob and J. R. Partington, *Admissibility of control and observation operators for semigroups: A survey*, in Current Trends in Operator Theory and Its Applications, Oper. Theory Adv. Appl. 149, Birkhäuser, Basel, 2004, pp. 199–221.

[16] K. Jörgens, *Lineare Integraloperatoren*, B. G. Teubner, Stuttgart, 1970.

[17] N. Kalton, P. Kunstmann, and L. Weis, *Perturbation and interpolation theorems for the H^∞-calculus with applications to differential operators*, Math. Ann., 336 (2006), pp. 747–801.

[18] N. Kalton and L. Weis, *Euclidian Structures*, in preparation.

[19] P. C. Kunstmann and L. Weis, *Maximal L_p-regularity for parabolic equations, Fourier multiplier theorems and H^∞-functional calculus*, in Functional Analytic Methods for Evolution Equations, Lecture Notes in Math. 1855, Springer, Berlin, 2004, pp. 65–311.

[20] C. Le Merdy, *The Weiss conjecture for bounded analytic semigroups*, J. London Math. Soc. (2), 67 (2003), pp. 715–738.

[21] A. McIntosh, *Operators which have an H_∞ functional calculus*, in Miniconference on Operator Theory and Partial Differential Equations (North Ryde, 1986), Proc. Centre Math. Anal. Austral. Nat. Univ. 14, Australian National University, Canberra, 1986, pp. 210–231.

[22] A. McIntosh and A. Yagi, *Operators of type ω without a bounded H_∞ functional calculus*, in Miniconference on Operators in Analysis (Sydney, 1989), Proc. Centre Math. Anal. Austral. Nat. Univ. 24, Australian National University, Canberra, 1990, pp. 159–172.

[23] Y. Meyer, *Ondelettes et opérateurs*. I, Actualités Mathématiques, Hermann, Paris, 1990.

[24] J. Prüss and G. Simonett, *Maximal regularity for evolution equations in weighted L_p-spaces*, Arch. Math. (Basel), 82 (2004), pp. 415–431.

[25] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.

[26] R. Seeley, *Interpolation in L^p with boundary conditions*, Stud. Math., 44 (1972), pp. 47–60.

[27] O. Staffans, *Well-posed linear systems, Lax-Phillips scattering, and L^p-multipliers*, in Systems, Approximation, Singular Integral Operators, and Related Topics (Bordeaux, 2000), Oper. Theory Adv. Appl. 129, Birkhäuser, Basel, 2001, pp. 445–464.

[28] O. Staffans, *Well-Posed Linear Systems*, Encyclopedia Math. Appl. 103, Cambridge University Press, Cambridge, UK, 2005.

[29] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth, Heidelberg, 1995.

[30] L. Weis, *Operator-valued Fourier multiplier theorems and maximal L_p-regularity*, Math. Ann., 319 (2001), pp. 735–758.
[31] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[32] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
[33] G. Weiss, *Two conjectures on the admissibility of control operators*, in Estimation and Control of Distributed Parameter Systems (Vorau, 1990), Internat. Ser. Numer. Math. 100, Birkhäuser, Basel, 1991, pp. 367–378.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2119–2137

CONJUGATE POINTS IN FORMATION CONSTRAINED OPTIMAL MULTI-AGENT COORDINATION: A CASE STUDY*

JIANGHAI HU†, MARIA PRANDINI‡, AND CLAIRE TOMLIN§

Abstract. In this paper, an optimal coordinated motion planning problem for multiple agents subject to constraints on the admissible formation patterns is formulated. Solutions to the problem are reinterpreted as distance minimizing geodesics on a certain manifold with boundary. A geodesic on this manifold may fail to be a solution for different reasons. In particular, if a geodesic possesses conjugate points, then it will no longer be distance minimizing beyond its first conjugate point. We study a particular instance of the formation constrained optimal coordinated motion problem, where a number of initially aligned agents tries to switch positions by rotating around their common centroid. The complete set of conjugate points of a geodesic naturally associated with this problem is characterized analytically. This allows us to prove that the geodesic will not correspond to an optimal coordinated motion when the angle of rotation exceeds a threshold that decreases to zero as the number of agents increases. Moreover, infinitesimal perturbations that improve the performance of the geodesic after it passes the conjugate points are also determined, which, interestingly, are characterized by a certain family of orthogonal polynomials.

Key words. conjugate point, multi-agent coordination, geodesics, orthogonal polynomials

AMS subject classifications. 93C85, 53C22, 58E25, 05E35

DOI. 10.1137/040616334

1. Introduction.

Multi-agent coordinated motion planning problems arise in various contexts, such as air traffic management (ATM) [12, 21], robotics [3], and unmanned aerial vehicles (UAVs) [17, 20, 22]. In most cases, certain separation constraints between the agents have to be guaranteed due to physical, safety, or efficiency reasons. For example, in ATM systems, aircraft flying at the same altitude are required to maintain a minimal horizontal separation of 5 nautical miles in en-route airspace and 3 nautical miles when close to airports. When multiple mobile robots are performing a coordinated task such as lifting a common object, specific formations have to be kept by the robots throughout the operation. For UAVs, flying in formation may reduce the fuel expenditure and the communication power needed for information exchange.

In this paper, we formulate an optimal coordinated motion planning problem for multiple agents under formation constraints. We consider all the coordinated motions that can lead a group of agents from given initial positions to given destination positions within a certain time horizon, while satisfying the additional constraint that the formation patterns of the agents belong to a prescribed subset. Among the coordinated motions in this restricted set, we try to find the ones that minimize a weighted sum of the energy functions of each individual agent's motions, with the weights

*Received by the editors October 5, 2004; accepted for publication (in revised form) July 13, 2006; published electronically January 12, 2007. This research was partially supported by the ONR MURI "CoMotion" under grant N00014-02-1-0720. A preliminary version of the results in this paper has appeared in [8].

http://www.siam.org/journals/sicon/45-6/61633.html

†School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 (jianghai@purdue.edu).

‡Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy (prandini@elet.polimi.it).

§Department of Aeronautics and Astronautics, Stanford University, Stanford, CA 94305 (tomlin@stanford.edu).

representing agent priorities. In our problem formulation, we use simple kinetic models for the agent dynamics and consider only holonomic constraints, as opposed to the numerous works dealing with nonholonomic constraints (e.g. [2, 11, 19]). See, for example, [5, 13, 15] for relevant work on the problems of stable and optimal coordinated control of vehicle formations.

A geometric interpretation of the considered optimal coordinated motion planning problem is given in this paper. According to this interpretation, a solution to the problem is a shortest curve with constant speed between two fixed points in a certain manifold with boundary, where the boundary is determined by the feasible formation patterns. Being a shortest curve between two points, which can be far away from each other, such a *globally* distance minimizing curve is obviously also a *locally* distance minimizing curve, i.e., a curve whose sufficiently short segments are distance-minimizing between their respective end points. Locally distance minimizing curves parameterized with constant speed are called *geodesics*. Thus a solution to the problem necessarily corresponds to a geodesic of the manifold.

Conversely, however, for various reasons a geodesic of the manifold may fail to be globally distance minimizing, thus failing to solve the problem. One of the reasons is the occurrence of conjugate points. Traveling along a geodesic from a fixed starting point, a conjugate point occurs at a point where there exists a nontrivial Jacobi field along the geodesic vanishing at both the starting point and that particular point [4], or less rigorously, where there exists infinitesimally more than one geodesic connecting the starting point to that point. For a simple example, consider the sphere. Geodesics on the sphere are great circles; and conjugate points along a great circle occur at the antipodal point of its starting point. It is a well-known fact in Riemannian geometry that a geodesic will not be distance minimizing once it passes its first conjugate point [4], as one can then infinitesimally perturb it to obtain a shorter curve with the same end points. In the sphere example, when a great circle extends beyond two antipodal points, a shorter curve between its starting and ending points can be found by following the intersection of the sphere with any codimensional one plane passing through the two points. The aim of this paper is to study, through a concrete example, the loss of optimality due to the existence of conjugate points in the multiagent coordination problem, which so far has been largely ignored in the literature.

It is in general difficult, if not impossible, to characterize the conjugate points of a geodesic analytically. In this paper, we shall focus on a special instance of the formation constrained optimal coordination problem, namely, a group of initially aligned agents switching positions by rotating around their common centroid. We shall show that the conjugate points of a geodesic that arises naturally as a candidate solution to this particular problem admit nice analytic formulae. We shall also determine the infinitesimal perturbations that can shorten the geodesic with various efficiencies, once it passes its conjugate points, and characterize them using a certain family of orthogonal polynomials. A geometric interpretation of our results is that they characterize how far a curve traveling along the outer edge of a (high-dimensional) donut-shaped state space can be extended before it is no longer distance minimizing. Another interesting interpretation of our results can be given with reference to a mechanical model of a multisegment snake-like robot that needs to rotate, moving between two completely stretched configurations. If the angle of rotation exceeds a certain threshold, then the motion corresponding to the robot rotating around its center of mass at constant angular velocity, with all of its segments constantly aligned in a straight line, will not be energy minimizing. More efficient motions can be found by "folding" the snake in various ways.

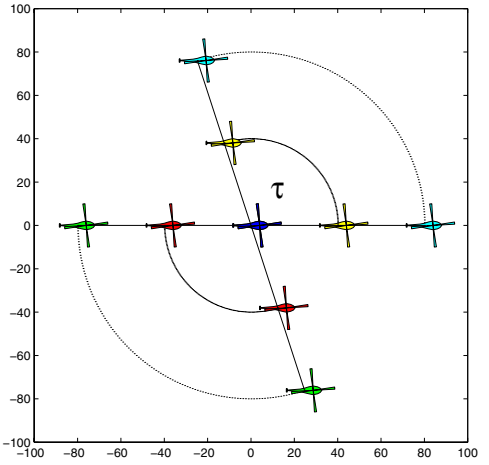


FIG. 1.1. *A five-agent example.*

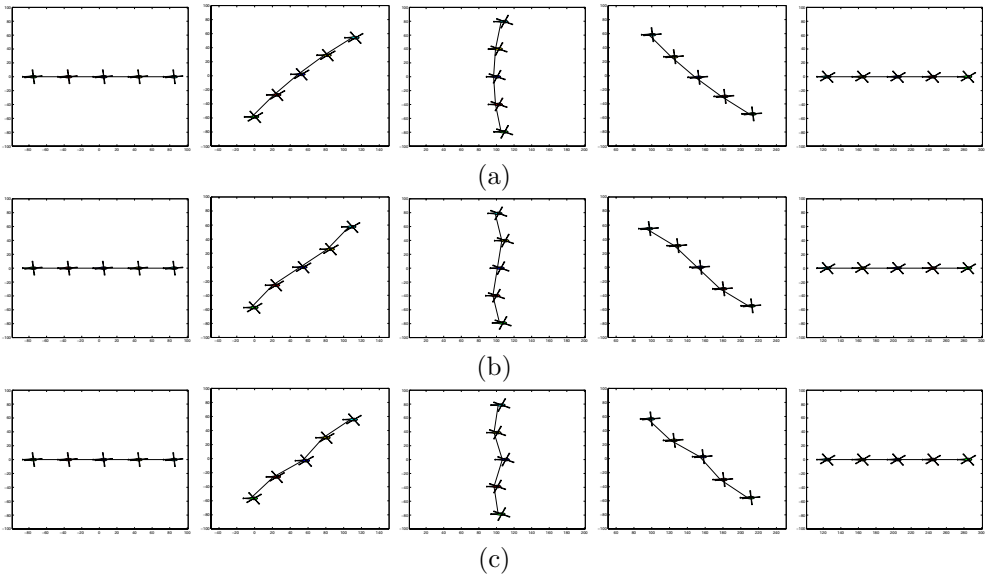


FIG. 1.2. *Three joint maneuvers with increasingly lower cost than the one shown in Figure* 1.1 *when* $\tau = \pi$. *Top row: maneuver* (a)*; middle row: maneuver* (b)*; bottom row: maneuver* (c). *From left to right: snapshots at times* $t = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$, *respectively.*

As a preview of the results and a specific example, see Figure 1.1, where five helicopters initially flying in a straight line try to reach new positions by rotating counterclockwise around their centroid, i.e., the middle helicopter, at the same angular velocity during the time period $[0, 1]$ (thus the five helicopters form a straight line at all times). The results in this paper will show that this joint maneuver is optimal in terms of minimizing a cost function defined as the sum of energy of an individual helicopter's maneuvers only if the angle τ of rotation is small, and that it is not optimal if $\tau > \frac{\pi}{3}$. Indeed, if for example $\tau = \pi$, then we can find better maneuvers than the one in Figure 1.1. We plot in the three rows of Figure 1.2 three such maneuvers, where in each row the five figures from left to right represent the snapshots of the

maneuver at times $t = 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$, respectively. In terms of performance, we will show that maneuver (c) is the best; maneuver (a) is the worst of the three, but still better than the original one in Figure 1.1.

A related work [7] studies the simpler problem of optimal collision avoidance—with no formation constraints. Avoidance maneuvers are formally classified into homotopy types, and the optimal ones are characterized by local and global conditions obtained through a variational analysis approach based on a geometric interpretation of a joint maneuver as a braid. In [7, sect. 3.6], a special instance of the case study investigated in this paper, with only three agents involved in the coordinated maneuver, was presented as an example to show that optimal solutions within a homotopy type are not unique.

The results proved in the present paper represent a highly nontrivial extension of those in the example of [7] to the general case of k (possibly larger than 3) agents and contain contributions that go far beyond the scope of [7]. First of all, formation constraints play a central role in the optimal coordination problem studied here, whereas in [7] they are only marginally treated in the example of section 3.6. Second, as the number k of agents increases beyond 3, computing the conjugate points becomes a markedly more difficult task because (i) many intermediate results, such as the inverse Riemannian metric, the Christoffel symbols, and the curvature tensors, are no longer directly computable; (ii) the Jacobi equation that one needs to solve is of the form $\ddot{x} = -Bx$ with $B \in \mathbb{R}^{(k-1)\times(k-1)}$, and for $k > 3$ it cannot be reduced to a scalar ODE as in [7, sect. 3.6], thus necessitating the spectrum decomposition of B; and (iii) the perturbed solutions with better performance after passing the conjugate points are now multidimensional instead of one-dimensional. Lastly, some completely new results are obtained in this paper, such as that the first conjugate point along the candidate solution occurs at a distance that decreases to zero at the rate of $\frac{1}{k}$ as the number k of agents grows to infinity, and that the infinitesimal perturbation that can improve the performance of the candidate solution after it passes the conjugate points can be described by a family of orthogonal polynomials of order $k-1$, called the ultraspherical polynomials.

This paper is organized as follows. In section 2 the formation patterns of a group of agents are defined and the problem of formation constrained optimal multi-agent coordination is formulated. In section 3 the conjugate points of the geodesic naturally arising in a particular instance of the problem are characterized analytically. Infinitesimal perturbations that improve the performance of the geodesic beyond its conjugate points are also determined. Finally, some conclusions are drawn in section 4.

2. Formation constrained optimal multi-agent coordination. In this section we formulate the problem of formation constrained optimal multi-agent coordination and describe some properties of its solutions that will be useful in the next section. For simplicity, the problem is formulated in Euclidean spaces. See [9] for an extension of the results to general Riemannian manifolds with a group of symmetries.

2.1. Problem formulation. We first introduce some notions. Let $\mathbb{R}^n$ be the Euclidean space with the standard Euclidean metric. Denote by $\langle q_i \rangle_{i=1}^k = (q_1, \ldots, q_k)$ an (ordered) k-tuple of points in $\mathbb{R}^n$ for some positive integer k. We say that $\langle q_i \rangle_{i=1}^k$ satisfies the *r-separation condition* for some $r > 0$ if $d(q_i, q_j) = \|q_i - q_j\| \geq r$ for all $i \neq j$, i.e., if the minimum pairwise distance among the k points is at least r.

With each k-tuple $\langle q_i \rangle_{i=1}^k$ satisfying the r-separation constraint we associate an undirected graph $(\mathcal{V}, \mathcal{E})$ with a set of vertices $\mathcal{V} = \{1, \ldots, k\}$ and a set of edges $\mathcal{E} = \{(i,j) : \|q_i - q_j\| = r\}$. We call $(\mathcal{V}, \mathcal{E})$ the *formation pattern* of the k-tuple

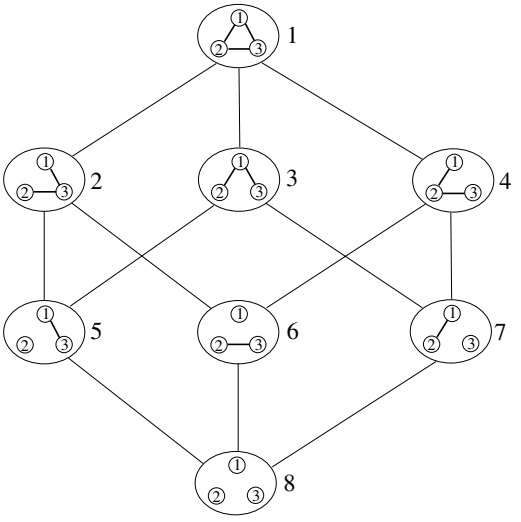


FIG. 2.1. *A graphical representation of $\mathcal{F}$ in the cases $n = 2$ and $k = 3$. Eight feasible formation patterns are represented by eight nodes. There is a line segment upward from node $(\mathcal{V}, \mathcal{E}_1)$ to node $(\mathcal{V}, \mathcal{E}_2)$ if and only if $(\mathcal{V}, \mathcal{E}_1) \prec (\mathcal{V}, \mathcal{E}_2)$ and no other $(\mathcal{V}, \mathcal{E}) \in \mathcal{F}$ satisfies $(\mathcal{V}, \mathcal{E}_1) \prec (\mathcal{V}, \mathcal{E}) \prec (\mathcal{V}, \mathcal{E}_2)$.*

$\langle q_i \rangle_{i=1}^k$. Denote by $\mathcal{F}$ the set of (*feasible*) formation patterns, i.e., the set of all undirected graphs $(\mathcal{V}, \mathcal{E})$ associated with k-tuples of points satisfying the r-separation condition. A partial order $\prec$ can be defined on $\mathcal{F}$ so that two formation patterns satisfy $(\mathcal{V}, \mathcal{E}_1) \prec (\mathcal{V}, \mathcal{E}_2)$ if and only if $(\mathcal{V}, \mathcal{E}_1)$ is a subgraph of $(\mathcal{V}, \mathcal{E}_2)$. A graphical representation of $\mathcal{F}$ with this partial order relation is shown in Figure 2.1 in the cases $n = 2$ and $k = 3$.

REMARK 1. *For given n, r, and k, all graphs with k vertices are not feasible formation patterns, i.e., are not the formation patterns of a k-tuple of points in $\mathbb{R}^n$ satisfying the r-separation condition. For example, if $n = 2$ and $k = 4$, the complete graph with four vertices and edges between each pair of them is not the formation pattern of any $\langle q_i \rangle_{i=1}^4$ satisfying the r-separation condition, regardless of r. Characterizing the set $\mathcal{F}$ of all feasible formation patterns for general n, r, and k remains a challenging problem.*

Consider k agents moving in $\mathbb{R}^n$. Suppose that they start at time 0 from the initial positions $a_1, \ldots, a_k \in \mathbb{R}^n$ and must reach the destination positions $b_1, \ldots, b_k \in \mathbb{R}^n$ at time t_f, respectively. We assume that both $\langle a_i \rangle_{i=1}^k$ and $\langle b_i \rangle_{i=1}^k$ satisfy the r-separation condition. Denote the joint trajectory of the agents by a k-tuple of curves $\gamma = \langle \gamma_i \rangle_{i=1}^k$, where the trajectory of agent i during the time interval $[0, t_f]$ is modeled as a continuous and piecewise C^1 curve $\gamma_i : [0, t_f] \to \mathbb{R}^n$ such that $\gamma_i(0) = a_i$ and $\gamma_i(t_f) = b_i$. The joint trajectory γ is said to be *collision-free* if for all $t \in [0, t_f]$ the k-tuple $\langle \gamma_i(t) \rangle_{i=1}^k$ satisfies the r-separation condition, or equivalently, if the distance between any two agents is at least r at all times during $[0, t_f]$. For a collision-free joint trajectory γ, its formation pattern at time $t \in [0, t_f]$ is the formation pattern of the k-tuple of points $\langle \gamma_i(t) \rangle_{i=1}^k$, which is time varying.

Define the cost of the joint trajectory γ as

$$J(\gamma) = \sum_{i=1}^{k} \mu_i E(\gamma_i), \tag{2.1}$$

where $\mu_1, \dots, \mu_k$ are positive numbers representing the priorities of the agents, and

$$E(\gamma_i) = \frac{1}{2}\int_0^{t_f} \|\dot{\gamma}_i(t)\|^2\, dt \tag{2.2}$$

is the standard energy of the trajectory γ_i as a curve in $\mathbb{R}^n$, for $i = 1, \dots, k$. If we denote by $L(\gamma_i) = \int_0^{t_f} \|\dot{\gamma}_i(t)\|\, dt$ the arc length of γ_i, then by the Cauchy–Schwarz inequality [14], $E(\gamma_i) \geq \frac{1}{2}L^2(\gamma_i)/t_f$, where the equality holds if and only if $\|\dot{\gamma}_i(t)\|$ is constant for $t \in [0, t_f]$.

Based on the introduced notions, the problem of optimal formation constrained coordinated motions for k agents moving in $\mathbb{R}^n$ can be formulated as follows.

PROBLEM 1 (optimal formation constrained coordination (OFC)). *Among all the collision-free joint trajectories $\gamma = \langle \gamma_i \rangle_{i=1}^k$ that start from $\langle a_i \rangle_{i=1}^k$ at time 0 and end at $\langle b_i \rangle_{i=1}^k$ at time t_f, find the ones that minimize the cost $J(\gamma)$ and satisfy the constraint that the formation pattern of γ at any time $t \in [0, t_f]$ belongs to some prescribed subset $\tilde{\mathcal{F}}$ of $\mathcal{F}$. Here we assume that $\tilde{\mathcal{F}}$ contains the formation patterns of both $\langle a_i \rangle_{i=1}^k$ and $\langle b_i \rangle_{i=1}^k$.*

If $\tilde{\mathcal{F}} = \mathcal{F}$, i.e., all formation patterns are allowed, the resulting OFC problem is also called the optimal collision avoidance (OCA) problem, which has important implication in the ATM application [21].

As an example, in Figure 2.1, one can choose $\tilde{\mathcal{F}}$ to consist of formation patterns 1, 2, 3, and 4, thus requiring that every two agents "contact" each other either directly or indirectly via the third agent at all time instants. This makes sense in practical situations, where the three agents have to share some common data and where information exchange is possible only at the minimum allowed distance. As another example, $\tilde{\mathcal{F}}$ can be chosen to consist of formation patterns 1, 3, 4, and 7. In this case, agent 1 and agent 2 are required to be bound together during the whole time interval $[0, t_f]$; thus the OFC problem can be viewed as the OCA problem between agent 3 and this two-agent group.

2.2. Geometric interpretation. The OFC problem can be interpreted in the following geometric way.

Each k-tuple $\langle q_i \rangle_{i=1}^k$ of points in $\mathbb{R}^n$ corresponds to a single point $q = (q_1, \dots, q_k)$ in $\mathbb{R}^{nk} = \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ (k times). Thus each joint trajectory $\gamma = \langle \gamma_i \rangle_{i=1}^k$ of the k agents corresponds to a curve γ in $\mathbb{R}^{nk}$ starting from the point $a = (a_1, \dots, a_k)$ at time 0 and ending at the point $b = (b_1, \dots, b_k)$ at time t_f. The collision-free condition is equivalent to γ avoiding the obstacle

$$W \triangleq \cup_{i \neq j} \{(q_1, \dots, q_k) \in \mathbb{R}^{nk} : \|q_i - q_j\| < r\}, \tag{2.3}$$

or equivalently, γ being a curve in the set $\mathbb{R}^{nk} \setminus W$. With the formation constraint $\tilde{\mathcal{F}}$, a solution γ to the OFC problem is further restricted to lie in a subset $X_{\tilde{\mathcal{F}}}$ of $\mathbb{R}^{nk} \setminus W$. Indeed, corresponding to each allowed formation pattern $(\mathcal{V}, \mathcal{E})$ in $\tilde{\mathcal{F}}$, the feasible set of γ is a submanifold (cell) $X_{(\mathcal{V},\mathcal{E})}$ of $\mathbb{R}^{nk} \setminus W$ of proper dimension: the less edges (i.e., distance constraints) in $(\mathcal{V}, \mathcal{E})$, the higher the dimension of the cell $X_{(\mathcal{V},\mathcal{E})}$. For example, the cell corresponding to the "free" formation pattern that consists of no edges is the interior of $\mathbb{R}^{nk} \setminus W$ and has the highest possible dimension nk. Thus the feasible set of the solution γ to the OFC problem is the union of all such cells: $X_{\tilde{\mathcal{F}}} = \cup_{(\mathcal{V},\mathcal{E}) \in \tilde{\mathcal{F}}} X_{(\mathcal{V},\mathcal{E})}$. In order for γ to be a solution to the OFC problem, γ as a curve in $\mathbb{R}^{nk}$ must lie in $X_{\tilde{\mathcal{F}}}$.

As a subset of $\mathbb{R}^{nk}$, $X_{\tilde{\mathcal{F}}}$ inherits a Riemannian metric from the standard one on $\mathbb{R}^{nk}$ after properly scaling its coordinate axes according to $\langle \mu_i \rangle_{i=1}^k$. For a solution γ to the OFC problem, its cost $J(\gamma)$ defined in (2.1) is nothing but the energy of γ as a curve in $X_{\tilde{\mathcal{F}}}$ under this inherited metric. In particular, if $\mu_1 = \cdots = \mu_k = 1$, $J(\gamma)$ is the standard energy of γ as a curve in $\mathbb{R}^{nk}$. Thus the OFC problem is equivalent to finding the energy minimizing curves connecting a and b in $X_{\tilde{\mathcal{F}}}$. It is well known in Riemannian geometry [16] (see also the discussion after (2.2)) that such curves are necessarily the shortest curves in $X_{\tilde{\mathcal{F}}}$ between a and b parameterized with constant speed. Therefore, solving the OFC problem is equivalent to finding the geodesics from a to b in $X_{\tilde{\mathcal{F}}}$ that are also globally distance minimizing.

REMARK 2. *Solutions to the OFC problem may not exist. The OCA problem of two agents on a line trying to switch positions is one such example. The geometric interpretation of the OFC problem allows us to easily formulate conditions for its feasibility. In general, to ensure the existence of solutions, it is sufficient (though not necessary) to require that the set $X_{\tilde{\mathcal{F}}}$ be closed and that a and b lie in the same connected component of $X_{\tilde{\mathcal{F}}}$. The first requirement translates into the following property of $\tilde{\mathcal{F}}$: For each $(\mathcal{V}, \mathcal{E}) \in \tilde{\mathcal{F}}$, any formation pattern $(\mathcal{V}, \mathcal{E}_1)$ such that $(\mathcal{V}, \mathcal{E}) \prec (\mathcal{V}, \mathcal{E}_1)$ is also an element of $\tilde{\mathcal{F}}$. The second requirement is satisfied if there exists at least one collision-free joint trajectory $\langle \gamma_i \rangle_{i=1}^k$ from a to b whose formation pattern is in $\tilde{\mathcal{F}}$ at all times.*

2.3. Conservation law for the solutions. We now describe some properties of the solutions to the OFC problem that will be used in the next sections.

PROPOSITION 1. *Suppose that the joint trajectory $\gamma = \langle \gamma_i \rangle_{i=1}^k$ is a solution to the OFC problem. Then the quantities*

$$\sum_{i=1}^k \mu_i \dot{\gamma}_i(t), \quad \sum_{i=1}^k \mu_i \big(\gamma_i(t)\dot{\gamma}_i^T(t) - \dot{\gamma}_i(t)\gamma_i^T(t)\big) \tag{2.4}$$

are constant for all $t \in [0, t_f]$.

Note that in this paper an element in $\mathbb{R}^n$ is regarded by default as a column vector; thus the second quantity in (2.4) is an n-by-n matrix.

If one thinks of each agent i as a point in $\mathbb{R}^n$ with mass μ_i, then Proposition 1 implies that the linear and (generalized) angular momenta of the k-point mass system are conserved along the solutions to the OFC problem. The proof of this proposition can be found in, e.g., [6], and, for the case of general Riemannian manifolds with a group of symmetries, in [9]. Thus we omit the proof here.

One implication of Proposition 1 is that, if both a and b are μ-aligned in the sense that $\sum_{i=1}^k \mu_i a_i = \sum_{i=1}^k \mu_i b_i = 0$, then so is any k-tuple $\langle \gamma_i(t) \rangle_{i=1}^k$, $t \in [0, t_f]$, for a solution $\gamma = \langle \gamma_i \rangle_{i=1}^k$ to the OFC problem, namely,

$$\sum_{i=1}^k \mu_i \gamma_i \equiv 0.$$

Hence in the μ-aligned case the solution γ as a curve in $\mathbb{R}^{nk}$ must lie in a subspace

$$V = \left\{ (q_1, \ldots, q_k) \in \mathbb{R}^{nk} : \sum_{i=1}^k \mu_i q_i = 0 \right\}$$

of $\mathbb{R}^{nk}$. This reduces the dimension of the state space by n.

3. An interesting example. Consider the OFC problem on $\mathbb{R}^2$ with the k agents having the same priority $\mu_1 = \cdots = \mu_k = 1$ and the minimal allowed separation $r = 1$. Suppose that the starting positions $\langle a_i \rangle_{i=1}^k$ of the agents are given by $\langle (\frac{2i-k-1}{2}, 0) \rangle_{i=1}^k$. In other words, at time $t = 0$, the k agents are aligned on the x-axis with a common centroid at the origin and with consecutive agents at the minimal allowed separation. For each $t \geq 0$, denote by $R_t : \mathbb{R}^2 \to \mathbb{R}^2$ the counterclockwise rotation of $\mathbb{R}^2$ by an angle t in radians. Suppose that the destination positions are $\langle b_i \rangle_{i=1}^k = \langle R_{t_f}(a_i) \rangle_{i=1}^k$. Both the initial and destination positions have the same formation pattern $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, k\}$ and $\mathcal{E} = \{(i, i+1) : i = 1, \ldots, k-1\}$. We choose the admissible formation pattern set $\tilde{\mathcal{F}}$ to consist of this formation pattern only. Therefore, in considering the OFC problem, we require that agents i and $i+1$ be kept at constant distance r throughout $[0, t_f]$ for $i = 1, \ldots, k-1$, and all other pairs of agents maintain a distance greater than r.

Since $\langle a_i \rangle_{i=1}^k$ and $\langle b_i \rangle_{i=1}^k$ are μ-aligned, by the discussion after Proposition 1 in section 2.3, a solution $\gamma = \langle \gamma_i \rangle_{i=1}^k$ to the above OFC problem satisfies $\sum_{i=1}^k \gamma_i \equiv 0$. Thus γ as a curve in $\mathbb{R}^{2k} \setminus W$ lies in the subspace $V = \{(q_1, \ldots, q_k) \in \mathbb{R}^{2k} : \sum_{i=1}^k q_i = 0\}$. Furthermore, due to the admissible formation pattern set $\tilde{\mathcal{F}}$, γ belongs to the subset $X_{\tilde{\mathcal{F}}}$ of $(\mathbb{R}^{2k} \setminus W) \cap V$ given by

$$X_{\tilde{\mathcal{F}}} = \Bigg\{ (q_1, \ldots, q_k) \in \mathbb{R}^{2k} : \|q_i - q_{i+1}\| = 1, i = 1, 2, \ldots, k-1, \\ \|q_i - q_j\| > 1 \ \forall j > i+1, \text{ and } \sum_{i=1}^k q_i = 0 \Bigg\}. \tag{3.1}$$

Therefore, according to section 2.2, the solutions to the OFC problem are the distance minimizing geodesics in $X_{\tilde{\mathcal{F}}}$ connecting $a = (a_1, \ldots, a_k)$ to $b = (b_1, \ldots, b_k)$.

In this section, we shall consider $\gamma^* = \langle R_t(a_i) \rangle_{i=1}^k$, $t \in [0, t_f]$, as a natural candidate solution to the above OFC problem.[1] Under the motions specified by γ^*, the k agents rotate counterclockwise at a constant unit angular velocity around the origin from their starting to their destination positions. We shall show that $\langle R_t(a_i) \rangle_{i=1}^k$, $t \geq 0$, is a geodesic in $X_{\tilde{\mathcal{F}}}$; thus γ^* is optimal for t_f small enough. We shall also show that the first conjugate point along this geodesic occurs at $\langle R_{\tau_k}(a_i) \rangle_{i=1}^k$ for some time τ_k, implying that γ^* is no longer optimal if $t_f > \tau_k$. We shall derive the analytical expression of τ_k and show that $\tau_k \sim \frac{1}{k} \to 0$ as $k \to \infty$.

3.1. Geometry of the manifold $X_{\tilde{\mathcal{F}}}$. We start by constructing a convenient coordinate system on $X_{\tilde{\mathcal{F}}}$ and then proceed to derive the geometry of $X_{\tilde{\mathcal{F}}}$ as a submanifold of $\mathbb{R}^{2k}$ in this coordinate system, such as its Riemannian metric, its covariant derivatives, and its curvature tensors. A general reference on Riemannian geometry can be found in [4].

First of all, $X_{\tilde{\mathcal{F}}}$ is a $(k-1)$-dimensional smooth submanifold of $\mathbb{R}^{2k}$ and admits global coordinates $(\theta_1, \ldots, \theta_{k-1})$, where θ_i is the angle $q_{i+1} - q_i \in \mathbb{R}^2$ makes with respect to the positive x-axis (see Figure 3.1). The coordinate map $f : (\theta_1, \ldots, \theta_{k-1}) \mapsto (q_1, \ldots, q_k) \in X_{\tilde{\mathcal{F}}}$ is defined by

$$q_i = q_1 + \sum_{j=1}^{i-1} \begin{bmatrix} \cos\theta_j \\ \sin\theta_j \end{bmatrix}, \quad i = 2, \ldots, k, \tag{3.2}$$

[1] In the following, γ^* is also sometimes used to denote the whole curve $\langle R_t(a_i) \rangle_{i=1}^k$, $t \geq 0$. In such cases, its meaning should be clear from the context.

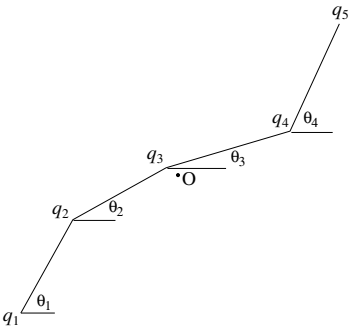


FIG. 3.1. *Coordinates of the manifold $X_{\tilde{\mathcal{F}}}$ when $k = 5$.*

where q_1 is chosen such that $\sum_{i=1}^{k} q_i = 0$, namely,

$$q_1 = -\frac{1}{k}\sum_{j=1}^{k-1}(k-j)\begin{bmatrix}\cos\theta_j\\ \sin\theta_j\end{bmatrix}. \tag{3.3}$$

In these coordinates, γ^* corresponds to $\theta_i(t) = t$, $t \in [0, t_f]$, $i = 1, \ldots, k-1$.

At any $q \in X_{\tilde{\mathcal{F}}}$, $\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_{k-1}}$ form a basis of $T_q X_{\tilde{\mathcal{F}}}$. In this basis, the Riemannian metric $\langle\cdot,\cdot\rangle$ that $X_{\tilde{\mathcal{F}}}$ inherits from $\mathbb{R}^{2k}$ as a submanifold can be computed as

$$g_{ij} \triangleq \left\langle \frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j} \right\rangle = \left\langle \frac{\partial f}{\partial\theta_i}, \frac{\partial f}{\partial\theta_j} \right\rangle_{\mathbb{R}^{2k}}, \quad 1 \le i, j \le k-1. \tag{3.4}$$

Here f is the map defined in (3.2) and (3.3), and each $\frac{\partial f}{\partial\theta_i}$ is a vector in $\mathbb{R}^{2k}$. $\langle\cdot,\cdot\rangle_{\mathbb{R}^{2k}}$ is the standard inner product on $\mathbb{R}^{2k}$. With this definition of metric, the map f becomes an isometry, and the cost of a joint trajectory γ of the k-agent system given by (2.1) can be expressed in two equivalent ways: In $(q_1, \ldots, q_k)$ coordinates it is

$$J(\gamma) = \frac{1}{2}\int_0^{t_f} \sum_{i=1}^{k} \|\dot{q}_i\|^2\, dt,$$

and in $(\theta_1, \ldots, \theta_{k-1})$ coordinates it is

$$J(\gamma) = \frac{1}{2}\int_0^{t_f} \left\langle \sum_{i=1}^{k-1}\dot{\theta}_i\frac{\partial}{\partial\theta_i}, \sum_{i=1}^{k-1}\dot{\theta}_i\frac{\partial}{\partial\theta_i} \right\rangle dt = \frac{1}{2}\int_0^{t_f} \sum_{i,j=1}^{k-1} g_{ij}\dot{\theta}_i\dot{\theta}_j\, dt.$$

After some careful computation, (3.4) in our case yields

$$g_{ij} = \Delta_{ij}\cos(\theta_i - \theta_j),$$

where Δ_{ij} are constants given by

$$\Delta_{ij} = \begin{cases} \frac{i(k-j)}{k} & \text{if } i \le j, \\ \frac{(k-i)j}{k} & \text{if } i > j. \end{cases} \tag{3.5}$$

Denote by $(g^{ij})_{1\le i,j\le k-1}$ the inverse matrix of $(g_{ij})_{1\le i,j\le k-1}$.

REMARK 3. *For a general configuration $(\theta_1, \dots, \theta_{k-1})$, the metric g defined above may become degenerate; i.e., the inverse matrix of $(g_{ij})_{1\le i,j\le k-1}$ may not exist. This is the case, for example, when $\theta_i - \theta_{i+1} = \pi$ for some $i = 1, \dots, k-2$. However, due to the condition that $\|q_i - q_j\| > 1$ for $j > i+1$ in the definition (3.1) of $X_{\tilde{\mathcal{F}}}$, this degeneracy will not be encountered by configurations in $X_{\tilde{\mathcal{F}}}$.*

The following lemma can be verified directly.

LEMMA 1. *Let $\Delta = (\Delta_{ij})_{1\le i,j\le k-1} \in \mathbb{R}^{(k-1)\times(k-1)}$ be the symmetric matrix with components Δ_{ij} defined in (3.5). Then*

$$\Delta^{-1} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

With the Riemannian metric g defined above, the covariant derivative with respect to the Levi–Civita connection on $X_{\tilde{\mathcal{F}}}$ is given by [4]:

$$\nabla_{\frac{\partial}{\partial\theta_i}} \frac{\partial}{\partial\theta_j} = \sum_{m=1}^{k-1} \Gamma_{ij}^m \frac{\partial}{\partial\theta_m}, \tag{3.6}$$

where Γ_{ij}^m are the Christoffel symbols defined as

$$\Gamma_{ij}^m = \frac{1}{2} \sum_{l=1}^{k-1} \left\{ \frac{\partial g_{jl}}{\partial\theta_i} + \frac{\partial g_{li}}{\partial\theta_j} - \frac{\partial g_{ij}}{\partial\theta_l} \right\} g^{lm}, \quad 1 \le i, j, m \le k-1.$$

A curve γ in $X_{\tilde{\mathcal{F}}}$ is a geodesic if and only if $\nabla_{\dot\gamma}\dot\gamma \equiv 0$. By definition (3.6), this equation, also called the geodesic equation, can be written in the $(\theta_1, \dots, \theta_{k-1})$ coordinates as a group of second order differential equations:

$$\ddot\theta_m = -\sum_{i,j=1}^{k-1} \Gamma_{ij}^m \dot\theta_i \dot\theta_j, \quad m = 1, \dots, k-1. \tag{3.7}$$

In our case, we can compute that, for $1 \le i, j, m \le k-1$,

$$\Gamma_{ij}^m = \begin{cases} 0 & \text{if } i \ne j, \\ \sum_{l=1}^{k-1} \Delta_{il} \sin(\theta_l - \theta_i) g^{lm} & \text{if } i = j. \end{cases} \tag{3.8}$$

Notice that along γ^* we have $\theta_1 = \cdots = \theta_{k-1}$. Therefore, we have the following.

LEMMA 2. *Along γ^* we have $\Gamma_{ij}^m = 0$ for all i, j, m; hence $\nabla_{\frac{\partial}{\partial\theta_i}} \frac{\partial}{\partial\theta_j} = 0$ for all i, j.*

Since $\dot\gamma^* = \frac{\partial}{\partial\theta_1} + \cdots + \frac{\partial}{\partial\theta_{k-1}}$, by Lemma 2 and the linearity of covariant derivatives,

$$\nabla_{\dot\gamma^*} \frac{\partial}{\partial\theta_j} = \nabla_{\sum_{i=1}^{k-1} \frac{\partial}{\partial\theta_i}} \frac{\partial}{\partial\theta_j} = \sum_{i=1}^{k-1} \nabla_{\frac{\partial}{\partial\theta_i}} \frac{\partial}{\partial\theta_j} = 0, \quad j = 1, \dots, k-1. \tag{3.9}$$

Thus

$$\nabla_{\dot\gamma^*}\dot\gamma^* = \nabla_{\dot\gamma^*} \sum_{j=1}^{k-1} \frac{\partial}{\partial\theta_j} = \sum_{j=1}^{k-1} \nabla_{\dot\gamma^*} \frac{\partial}{\partial\theta_j} = 0,$$

which is exactly the condition for γ^* to be a geodesic in $X_{\tilde{\mathcal{F}}}$. Alternatively, since $\Gamma_{ij}^m = 0$ along γ^*, the geodesic equation (3.7) reduces to $\ddot{\theta}_m = 0$, $m = 1, \ldots, k-1$, which are trivially satisfied by $\gamma^*(t)$ with coordinates $\theta_i(t) = t$, $i = 1, \ldots, k-1$, $t \in [0, t_f]$.

COROLLARY 1. *γ^* is a geodesic in $X_{\tilde{\mathcal{F}}}$.*

Since by definition a geodesic is locally distance minimizing, we conclude that γ^* is a solution to the OFC problem for t_f small enough.

3.2. Conjugate points along γ^*. To characterize the conjugate points of γ^*, we need to calculate the curvature of $X_{\tilde{\mathcal{F}}}$. The curvature tensor of $X_{\tilde{\mathcal{F}}}$ is given by [4]

$$R\left(\frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j}\right)\frac{\partial}{\partial\theta_l} = \sum_{m=1}^{k-1} R_{ijl}^m \frac{\partial}{\partial\theta_m}, \tag{3.10}$$

where R_{ijl}^m are defined from the Christoffel symbols as

$$R_{ijl}^m = \sum_{\beta=1}^{k-1} \Gamma_{il}^\beta \Gamma_{j\beta}^m - \sum_{\beta=1}^{k-1} \Gamma_{jl}^\beta \Gamma_{i\beta}^m + \frac{\partial \Gamma_{il}^m}{\partial\theta_j} - \frac{\partial \Gamma_{jl}^m}{\partial\theta_i}, \quad 1 \le i, j, l,\ m \le k-1. \tag{3.11}$$

A Jacobi field X along γ^* is a vector field along γ^* satisfying the Jacobi equation

$$\nabla_{\dot{\gamma}^*}\nabla_{\dot{\gamma}^*} X + R(\dot{\gamma}^*, X)\dot{\gamma}^* = 0. \tag{3.12}$$

A conjugate point $\gamma^*(\tau)$ along γ^* occurs at time $t = \tau$ if there is a nontrivial Jacobi field X along γ^* that vanishes at both times 0 and τ, i.e., $X(0) = X(\tau) = 0$. In the following, we shall characterize the conjugate points of γ^* starting from $\gamma^*(0) = a$.

Write an arbitrary vector field X along γ^* in coordinates as

$$X = \sum_{i=1}^{k-1} x_i \frac{\partial}{\partial\theta_i} \tag{3.13}$$

for some C^2 functions $x_i : [0, t_f] \to \mathbb{R}$. Note that here for simplicity we drop the time t in the above expression, following the convention in [4]. The explicit form for the vector field X should be $X(t) = \sum_{i=1}^{k-1} x_i(t) \frac{\partial}{\partial\theta_i}\big|_{\gamma^*(t)}$, $t \in [0, t_f]$. Equation (3.13) thus represents X as a vector $x = (x_1, \ldots, x_{k-1})$ in $\mathbb{R}^{k-1}$ that varies with time $t \in [0, t_f]$.

By (3.9) and the property of covariant derivatives, we have

$$\nabla_{\dot{\gamma}^*} X = \nabla_{\dot{\gamma}^*} \sum_{m=1}^{k-1} x_m \frac{\partial}{\partial\theta_m} = \sum_{m=1}^{k-1} \left(x_m \nabla_{\dot{\gamma}^*} \frac{\partial}{\partial\theta_m} + \dot{x}_m \frac{\partial}{\partial\theta_m} \right) = \sum_{m=1}^{k-1} \dot{x}_m \frac{\partial}{\partial\theta_m},$$

and similarly,

$$\nabla_{\dot{\gamma}^*}\nabla_{\dot{\gamma}^*} X = \sum_{m=1}^{k-1} \ddot{x}_m \frac{\partial}{\partial\theta_m}.$$

On the other hand, since R defined in (3.10) is a trilinear tensor, by expansion we have

$$R(\dot{\gamma}^*, X)\dot{\gamma}^* = \sum_{i,j,l,m=1}^{k-1} R_{ijl}^m x_j \frac{\partial}{\partial\theta_m}.$$

So the Jacobi equation (3.12) along γ^* is reduced to

$$\sum_{m=1}^{k-1}\left(\ddot{x}_m+\sum_{i,j,l=1}^{k-1}R^m_{ijl}x_j\right)\frac{\partial}{\partial\theta_m}=0,$$

or equivalently, to the following group of second order differential equations on $\mathbb{R}^{k-1}$:

$$\ddot{x}_m+\sum_{i,j,l=1}^{k-1}R^m_{ijl}x_j=0,\quad m=1,\dots,k-1. \tag{3.14}$$

Equations (3.14) can be written in matrix form as

$$\ddot{x}+B_kx=0, \tag{3.15}$$

where $B_k=(b_{mj})_{1\le m,j\le k-1}\in\mathbb{R}^{(k-1)\times(k-1)}$ is a constant matrix whose component on the mth row and jth column is defined by

$$b_{mj}=\sum_{i,l=1}^{k-1}R^m_{ijl}.$$

In Appendix A we will prove the following simple expression for B_k.

LEMMA 3. *Define* $\Lambda\triangleq\operatorname{diag}(\frac{k-1}{2},\dots,\frac{i(k-i)}{2},\dots,\frac{k-1}{2})_{1\le i\le k-1}\in\mathbb{R}^{(k-1)\times(k-1)}$. *Then*

$$B_k=\Delta^{-1}\Lambda-I_k, \tag{3.16}$$

where I_k *is the* $(k-1)$*-by-*$(k-1)$ *identity matrix.*

REMARK 4. B_k *is a constant matrix independent of* t *since the metric of* $X_{\tilde{\mathcal{F}}}$ *is homogeneous along* γ^*, *or more precisely, for each* $\tau>0$, *the map* $(q_1,\dots,q_k)\in X_{\tilde{\mathcal{F}}}\mapsto(R_\tau(q_1),\dots,R_\tau(q_k))\in X_{\tilde{\mathcal{F}}}$ *is an isometry of* $X_{\tilde{\mathcal{F}}}$ *mapping* $\gamma^*(t)$ *to* $\gamma^*(t+\tau)$, *whose differential map takes* $\frac{\partial}{\partial\theta_i}\big|_{\gamma^*(t)}$ *to* $\frac{\partial}{\partial\theta_i}\big|_{\gamma^*(t+\tau)}$ *for each* $i=1,\dots,k-1$.

The solutions of the Jacobi equation (3.15) are closely related to the spectral decomposition of B_k. To compute the eigenvalues of B_k, define a matrix

$$U\triangleq\begin{bmatrix}u_1&\cdots&u_{k-1}\end{bmatrix}$$

whose column vectors u_j for $j=1,\dots,k-1$ are given by

$$u_j=\left[\left(\frac{2-k}{k}\right)^{j-1},\dots,\left(\frac{2i-k}{k}\right)^{j-1},\dots,\left(\frac{k-2}{k}\right)^{j-1}\right]^T\in\mathbb{R}^{k-1}. \tag{3.17}$$

U is a Vandermonde matrix and hence nonsingular.

The following lemma can be verified directly.

LEMMA 4. *For each* $j=1,\dots,k-1$,

$$\Delta^{-1}\Lambda u_j=\frac{j(j+1)}{2}u_j+\textit{linear combination of } u_1,\dots,u_{j-1}. \tag{3.18}$$

In matrix form, (3.18) is equivalent to $\Delta^{-1}\Lambda U=U\Sigma$, i.e., $U^{-1}\Delta^{-1}\Lambda U=\Sigma$, where Σ is an upper triangular matrix whose elements on the main diagonal are

$\frac{j(j+1)}{2}$, $j = 1, \ldots, k-1$. Thus $\Delta^{-1}\Lambda$ and Σ have the same set of eigenvalues. As a result, we have the following.

COROLLARY 2. *B_k has $k-1$ distinctive eigenvalues $\lambda_j = \frac{j(j+1)}{2} - 1$ for $j = 1, \ldots, k-1$.*

Denote by v_j an eigenvector of $\Delta^{-1}\Lambda$ corresponding to the eigenvalue $\lambda_j + 1$, i.e.,

$$\Delta^{-1}\Lambda v_j = (\lambda_j + 1)v_j = \frac{j(j+1)}{2}\, v_j, \tag{3.19}$$

for $j = 1, \ldots, k-1$. Then by (3.16) v_j is also an eigenvector of B_k corresponding to the eigenvalue λ_j. Lemma 4 and (3.17) imply that v_j is of the form

$$v_j = \left[P_j\left(\frac{2-k}{k}\right), \ldots, P_j\left(\frac{2i-k}{k}\right), \ldots, P_j\left(\frac{k-2}{k}\right)\right]^T \tag{3.20}$$

for some nontrivial polynomial P_j of degree $j-1$.

In Appendix B, we shall derive from (3.19) and (3.20) the general condition (B.3) on the polynomials P_j and prove that P_j are the discrete version of some orthogonal polynomials called the *ultraspherical* (or *Gegenbauer*) polynomials $C_{j-1}^{(\alpha)}$ with parameter $\alpha = \frac{3}{2}$. In particular, P_j is an even polynomial when j is odd, and an odd polynomial when j is even [1].

Using condition (B.3), one can determine the first few P_j (up to a scaling factor):

$$P_1(x) = 1, \quad P_2(x) = 3x, \quad P_3(x) = 5x^2 - \left(1 - \frac{4}{k^2}\right), \quad P_4(x) = 7x^3 + \left(\frac{20}{k^2} - 3\right)x.$$

The first few v_j can then be obtained by (3.20). Therefore, we have the next lemma.

LEMMA 5. *B_k has the following eigenvectors:*

- $v_1 = [1, \ldots, 1]^T$ *for* $\lambda_1 = 0$;
- $v_2 = [2-k, \ldots, 2i-k, \ldots, k-2]^T$ *for* $\lambda_2 = 2$;
- $v_3 = [4k^2 - 20k + 24, \ldots, 5(2i-k)^2 - k^2 + 4, \ldots, 4k^2 - 20k + 24]^T$ *for* $\lambda_3 = 5$.

REMARK 5. *B_k has an eigenvalue 0 with the corresponding eigenvector $[1, \ldots, 1]^T$ as a consequence of the fact that $\dot{\gamma}^* = \frac{\partial}{\partial\theta_1} + \cdots + \frac{\partial}{\partial\theta_{k-1}}$ is parallel along γ^*.*

For large values of j, the expression of the polynomial P_j can be quite complicated. Fortunately, one can verify directly the following lemma.

LEMMA 6. *B_k has the following eigenvectors:*

- $v_{k-2} = [(2-k)\binom{k}{1}, \ldots, (-1)^{i+1}(2i-k)\binom{k}{i}, \ldots, (-1)^k(k-2)\binom{k}{k-1}]^T$ *for* $\lambda_{k-2} = \frac{(k-1)(k-2)}{2} - 1$;
- $v_{k-1} = [\binom{k}{1}, \ldots, (-1)^{i+1}\binom{k}{i}, \ldots, (-1)^k\binom{k}{k-1}]^T$ *for* $\lambda_{k-1} = \frac{k(k-1)}{2} - 1$.

Lemmas 5 and 6 together give the eigenvectors of B_k corresponding to a few of its smallest and largest eigenvalues.

The eigenvectors $v_1, \ldots, v_{k-1}$ of B_k form a basis of $\mathbb{R}^{k-1}$. We can then express x in this basis as $x = \sum_{j=1}^{k-1} y_j v_j$, so that the Jacobi equation (3.15) becomes

$$\ddot{y}_j + \lambda_j y_j = 0, \quad 1 \le j \le k-1.$$

Assume that X, hence x, vanishes at $t = 0$. Then $y_1(0) = \cdots = y_{k-1}(0) = 0$. Nontrivial solutions to the above equations with $y_j(0) = 0$, $1 \le j \le k-1$, are

$$\begin{aligned} y_1(t) &= c_1 t, \\ y_j(t) &= c_j \sin(t\sqrt{\lambda_j}), \quad j = 2, \ldots, k-1, \end{aligned}$$

for some constants $c_1, \dots, c_{k-1}$ not identically zero. Conjugate points are encountered at those time epochs $\tau > 0$, where $y_j(\tau) = 0$ for all j. This is possible only if $c_1 = 0$ and τ is an integer multiple of $\pi/\sqrt{\lambda_j}$ for some $j \in \{2, \dots, k-1\}$. Therefore, we have the next theorem.

THEOREM 1. *The set of conjugate points along γ^* is*

$$\{\gamma^*(\tau) : \tau = m\pi/\sqrt{\lambda_j} \text{ for some } m \in \mathbb{N} \text{ and some } j \in \{2, \dots, k-1\}, \tau \le t_f\}.$$

The first conjugate point along γ^* occurs at time

$$\tau_k \triangleq \frac{\pi}{\sqrt{\lambda_{k-1}}} = \frac{\pi\sqrt{2}}{\sqrt{(k-2)(k+1)}}. \tag{3.21}$$

A geodesic is no longer distance minimizing beyond its first conjugate point. Thus we have the following.

COROLLARY 3. *γ^* is not a solution to the OFC problem if $t_f > \tau_k$.*

Note that $\tau_k \sim \frac{1}{k}$ as $k \to \infty$. The result for the case $k = 3$ was first proved in [7].

REMARK 6. *Traveling along a geodesic emitting from a fixed starting point, the last point for which the corresponding geodesic segment remains distance minimizing is called a* cut point [4]. *Along the geodesic a cut point is either the first conjugate point, or the first point that is connected by at least two distinct (possibly far way from each other) geodesics to the starting point. In the latter case, the cut point is encountered before the first conjugate point. In our case we conjecture that the cut point along γ^* coincides with the first conjugate point, or in other words, γ^* is an optimal solution to the OFC problem for all t_f up to τ_k.*

3.3. Infinitesimal perturbations beyond the conjugate points. We now show how shorter curves than γ^* with the same end points look, at least infinitesimally, once γ^* surpasses its conjugate points.

Let $\{\gamma_s^*\}_{-\epsilon<s<\epsilon}$ be a C^∞ proper variation of γ^* in $X_{\tilde{\mathcal{F}}}$ with variation field

$$X \triangleq \frac{\partial \gamma_s^*}{\partial s}\Big|_{s=0},$$

which is a vector field along γ^*. For each s, define $E(s)$ as the energy of γ_s^* in $X_{\tilde{\mathcal{F}}}$, which coincides with the cost function of the joint trajectory corresponding to γ_s^*. By the variation of energy formulas [10], $E'(0) = 0$ since γ^* is a geodesic, and

$$E''(0) = -\int_0^{t_f} \langle X, \nabla_{\dot\gamma^*}\nabla_{\dot\gamma^*}X + R(\dot\gamma^*, X)\dot\gamma^*\rangle \, dt.$$

Write $X = \sum_{i=1}^{k-1} x_i \frac{\partial}{\partial\theta_i}$ in the basis $\frac{\partial}{\partial\theta_1}, \dots, \frac{\partial}{\partial\theta_{k-1}}$ along γ^*. Since $\{\gamma_s^*\}_{-\epsilon<s<\epsilon}$ is a proper variation, vector $x = (x_1, \dots, x_{k-1})$ vanishes at times 0 and t_f. In this coordinate system, the above equation reduces to

$$E''(0) = -\int_0^{t_f} x^T \Delta(\ddot{x} + B_k x) \, dt.$$

Suppose now that $t_f > \pi/\sqrt{\lambda_j}$ for some $j \in \{2, \dots, k-1\}$. Then, by choosing $\{\gamma_s^*\}_{-\epsilon<s<\epsilon}$ such that $x(t) = v_j \sin(\pi t/t_f)$, where we recall that v_j is an eigenvector of B_k associated with the eigenvalue λ_j, we have

$$E''(0) = -\left(\lambda_j - \frac{\pi^2}{t_f^2}\right)(v_j^T \Delta v_j)\int_0^{t_f} \sin^2\left(\frac{\pi t}{t_f}\right) dt < 0, \tag{3.22}$$

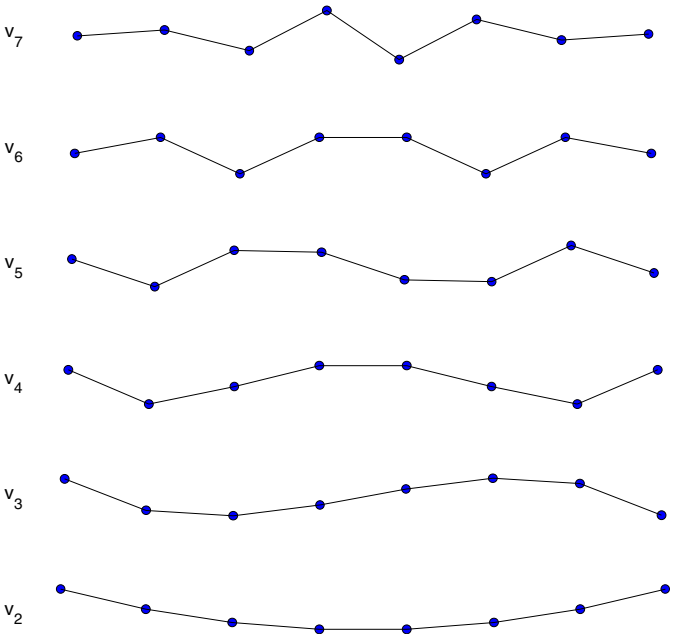


FIG. 3.2. *Perturbed shapes corresponding to* v_l, $l = 2, \ldots, k-1$, *when* $k = 8$.

since $v_j^T \Delta v_j > 0$ and $\lambda_j - \pi^2/t_f^2 > 0$. Therefore, γ_s^* is shorter than γ^* for sufficiently small s.

To sum up, the above analysis shows that if $t_f > \pi/\sqrt{\lambda_j}$ for some $j \in \{2, \ldots, k-1\}$, a solution better than γ^* can be obtained by infinitesimally perturbing γ^* in such a way that, at each $t \in [0, t_f]$, $(\theta_1, \ldots, \theta_{k-1})$ is incremented by an amount of $v_j \sin(\pi t/t_f) ds$. The linked-rod system (snake) formed by connecting successive agents will then assume a shape determined by the signs of the components of v_j. For example, the alternating signs of the components of v_{k-1} indicate a perturbation, where the $k-1$ rods are first folded into a saw-like shape during the first half of the time interval $[0, t_f]$, with the degree of folding of each rod depending on its position (in fact, proportional to $\binom{k}{l}$ for the lth rod from the edge, $l = 1, \ldots, k-1$), and then straightened up during the latter half of the time interval. In contrast, v_2 indicates that the $k-1$ rods bend into a bow-like shape, whereas the shape specified by v_{k-2} is a mixing (product) of the bending specified by v_2 and the folding specified by v_{k-1}. The maximal perturbation occurs at $t = t_f/2$. Figure 3.2 plots the various shapes of the linked-rod system at time $t = t_f/2$ caused by the perturbations $v_j \sin(\pi t/t_f)$ when $k = 8$. Note that these shapes have been rotated to align with the x-axis.

The efficiency of the perturbations specified by different v_j, provided that $t_f > \pi/\sqrt{\lambda_j}$, can be studied by comparing the respective $E''(0)$ under the requirement that $\int_0^{t_f} \|X\|^2 \, dt$ be constant. Since $\int_0^{t_f} \|X\|^2 \, dt = (v_j^T \Delta v_j) \int_0^{t_f} \sin^2(\pi t/t_f) \, dt$, we can conclude by (3.22) that the larger the eigenvalue λ_j, the more efficient the perturbation specified by its corresponding eigenvector v_j. This implies that the most efficient perturbation is the one given by v_{k-1}.

4. Conclusions. In this paper, we formulate the optimal formation constrained multi-agent coordination problem and give a geometric interpretation of its solutions

as geodesics in a certain manifold. We study an instance of the problem and characterize analytically the conjugate points of a geodesic proposed as a candidate solution. We conclude that this geodesic is optimal for sufficiently close starting and destination positions, but no longer optimal after surpassing its first conjugate point, which occurs from the starting position at a distance that decreases to zero at the same rate as $1/k$ as the number k of agents increases. In this case, infinitesimally better coordinated motions are determined. We show that the analytical forms of these motions can be derived from a certain family of orthogonal polynomials.

Appendix A. Proof of Lemma 3. In this proof, all terms in the equations, such as R^m_{ijl}, the Christoffel symbols, and their derivatives, are evaluated along γ^*, for which $\theta_1 = \cdots = \theta_{k-1}$ and the conclusion of Lemma 2 holds. Thus (3.11) becomes

$$R^m_{ijl} = \frac{\partial \Gamma^m_{il}}{\partial \theta_j} - \frac{\partial \Gamma^m_{jl}}{\partial \theta_i}, \quad i,j,l,m \in \{1,\dots,k-1\},$$

and for each $1 \le m, j \le k-1$,

$$b_{mj} = \sum_{i,l} R^m_{ijl} = \sum_{i,l} \frac{\partial \Gamma^m_{il}}{\partial \theta_j} - \sum_{i,l} \frac{\partial \Gamma^m_{jl}}{\partial \theta_i} = \sum_i \frac{\partial \Gamma^m_{ii}}{\partial \theta_j} - \sum_i \frac{\partial \Gamma^m_{jj}}{\partial \theta_i}, \tag{A.1}$$

where the summations are all from 1 to $k-1$. The first term can be simplified to

$$\begin{aligned}\sum_i \frac{\partial \Gamma^m_{ii}}{\partial \theta_j} &= \sum_{i\neq j} \frac{\partial \Gamma^m_{ii}}{\partial \theta_j} + \frac{\partial \Gamma^m_{jj}}{\partial \theta_j} \\ &= \sum_{i\neq j} \frac{\partial}{\partial \theta_j}\left[\sum_l \Delta_{il}\sin(\theta_l - \theta_i) g^{lm}\right] + \frac{\partial}{\partial \theta_j}\left[\sum_{l\neq j} \Delta_{jl}\sin(\theta_l - \theta_j) g^{lm}\right] \\ &= \sum_{i\neq j} \Delta_{ij} g^{jm} - \sum_{l\neq j} \Delta_{jl} g^{lm},\end{aligned}$$

where we have used the fact that $\theta_1 = \cdots = \theta_{k-1}$ on γ^*. Similarly,

$$\begin{aligned}\sum_i \frac{\partial \Gamma^m_{jj}}{\partial \theta_i} &= \sum_{i\neq j} \frac{\partial}{\partial \theta_i}\left[\sum_l \Delta_{jl}\sin(\theta_l - \theta_j) g^{lm}\right] + \frac{\partial}{\partial \theta_j}\left[\sum_{l\neq j} \Delta_{jl}\sin(\theta_l - \theta_j) g^{lm}\right] \\ &= \sum_{i\neq j} \Delta_{ji} g^{im} - \sum_{l\neq j} \Delta_{jl} g^{lm}.\end{aligned}$$

Hence (A.1) can be rewritten as

$$b_{mj} = \sum_{i\neq j} \Delta_{ij} g^{jm} - \sum_{i\neq j} \Delta_{ji} g^{im} = \Big(\sum_i \Delta_{ij}\Big) g^{jm} - \sum_i \Delta_{ji} g^{im}.$$

Since $g_{ji} = \Delta_{ji}$ on γ^*, $\sum_i \Delta_{ji} g^{im} = \sum_i g_{ji} g^{im} = \delta_{mj}$ by the definition of g^{im}. Moreover, $\sum_i \Delta_{ij} = \sum_{i\le j} \frac{i(k-j)}{2} + \sum_{i>j} \frac{(k-i)j}{2} = \frac{j(k-j)}{2}$. Therefore,

$$b_{mj} = \frac{j(k-j)}{2} g^{jm} - \delta_{mj} = \frac{j(k-j)}{2} g^{mj} - \delta_{mj} \tag{A.2}$$

by the symmetry of g^{mj}. Note that, on γ^*, $(g^{mj})_{1\le m,j\le k-1} = [(g_{mj})_{1\le m,j\le k-1}]^{-1} = \Delta^{-1}$. So (A.2) is exactly the desired conclusion.

Appendix B. Computation of polynomials P_j. In this appendix we shall derive the expressions of the polynomials P_j in (3.20) and prove that they belong to a certain family of orthogonal polynomials.

Fix a $j \in \{1, \dots, k-1\}$. Recall that the polynomial P_j is defined in (3.20), where v_j is given in (3.19). By substituting (3.20) into (3.19) and using Lemma 1, we have

$$-\frac{1}{2}(i-1)(k-i+1)P_j\left(\frac{2i-2-k}{k}\right) + i(k-i)P_j\left(\frac{2i-k}{k}\right)$$

$$-\frac{1}{2}(i+1)(k-i-1)P_j\left(\frac{2i+2-k}{k}\right) = \frac{j(j+1)}{2}P_j\left(\frac{2i-k}{k}\right)$$

for $i = 1, \dots, k-1$. This is equivalent to

$$\nabla^2_{2/k}[(1-x^2)P_j(x)] = -j(j+1)P_j(x) \tag{B.1}$$

for $x = \frac{2i-k}{k}$, $i = 1, \dots, k-1$. Here, for $h > 0$, ∇^2_h denotes the operator

$$\nabla^2_h[P(x)] = \frac{1}{h^2}[-2P(x) + P(x-h) + P(x+h)], \tag{B.2}$$

which is a difference approximate of $P''(x)$ with step size h. ∇^2_h maps a polynomial $P(x)$ of degree $m \geq 2$ to a polynomial of degree $m-2$. Thus in (B.1) both sides are polynomials of degree $j-1 \leq k-2$. Since the equality (B.1) holds for $k-1$ distinct values of x, it holds for all x. To sum up, P_j satisfies the following condition:

$$\nabla^2_{2/k}[(1-x^2)P_j(x)] = -j(j+1)P_j(x) \quad \forall x \in \mathbb{R}. \tag{B.3}$$

Note that $\nabla^2_h[P(x)] \to P''(x)$ as $h \to 0$. Therefore, for large k, the above condition can be approximated by

$$(1-x^2)P_j''(x) - 4xP_j'(x) + (j-1)(j+2)P_j(x) = 0.$$

The polynomial solution to this differential equation is called the $(j-1)$th *ultraspherical* (or *Gegenbauer*) *polynomial* $C^{(\alpha)}_{j-1}(x)$ with parameter $\alpha = \frac{3}{2}$ [1, p. 781]. The polynomials $C^{(3/2)}_{j-1}(x)$ obtained for different $j \geq 1$ are orthogonal on $[-1, 1]$ with respect to the weight function $1-x^2$, i.e., $\int_{-1}^{1}(1-x^2)C^{(3/2)}_l(x)C^{(3/2)}_j(x)\,dx = 0$ for $l \neq j$ (see [18]). Since $P_j(x)$ tends to $C^{(3/2)}_{j-1}(x)$ as $k \to \infty$, we expect that the polynomials $P_j(x)$, $j = 1, \dots, k-1$, are "approximately" orthogonal with respect to $1-x^2$ on $[-1, 1]$ as well. Indeed, a discrete version of this orthogonality condition holds:

$$\sum_{x=\frac{2i-k}{k},\, i=1,\dots,k-1} (1-x^2)P_l(x)P_j(x) = 0 \quad \text{for } l, j \in \{1, \dots, k-1\},\ l \neq j. \tag{B.4}$$

Condition (B.4) can be easily proved as follows. Observe that

$$\sum_{x=\frac{2i-k}{k},\, i=1,\dots,k-1} (1-x^2)P_l(x)P_j(x) = \frac{8}{k^2}v_l^T \Lambda\, v_j.$$

Thus we need only to show that $v_l^T \Lambda\, v_j = 0$ for $l \neq j$, or equivalently, that $V^T \Lambda V$ is diagonal, where V is the matrix defined as $V = \begin{bmatrix} v_1 & \cdots & v_{k-1} \end{bmatrix}$. Equation (3.19) implies that $\Lambda V = \Delta V \Omega$, where $\Omega = \mathrm{diag}(\lambda_1 + 1, \ldots, \lambda_i + 1, \ldots, \lambda_{k-1} + 1)$. Thus $V^T \Lambda V = (V^T \Delta V)\Omega$. Note that the left-hand side $V^T \Lambda V$ is a symmetric matrix, while the right-hand side is the product of a symmetric matrix $V^T \Delta V$ and a diagonal matrix Ω with distinct diagonal elements. In order for them to be equal, we must have that both $V^T \Lambda V$ and $V^T \Delta V$ are diagonal matrices. The diagonality of $V^T \Lambda V$ implies condition (B.4), and the diagonality of $V^T \Delta V$ implies that the tangent vectors represented by v_j, $j = 1, \ldots, k-1$, at any point on γ^* are orthogonal.

Acknowledgment. The first author would like to thank Alan Weinstein for his helpful comments and discussions on the results of this paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1992 (reprint of the 1972 edition).

[2] A. Bicchi and L. Pallottino, *Optimal planning for coordinated vehicles with bounded curvature*, in Algorithmic and Computational Robotics: New Directions (Hanover, NH, 2000), A K Peters, Natick, MA, 2001, pp. 181–189.

[3] J. P. Desai and V. Kumar, *Nonholonomic motion planning for multiple mobile manipulators*, in Proc. IEEE International Conference on Robotics and Automation, Vol. 4, Albuquerque, NM, 1997, pp. 3409–3414.

[4] M. P. do Carmo, *Riemannian Geometry*, Birkhäuser Boston, Boston, MA, 1992.

[5] J. A. Fax, *Optimal and Cooperative Control of Vehicle Formations*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 2002.

[6] J. Hu, M. Prandini, K. H. Johansson, and S. Sastry, *Hybrid geodesics as optimal solutions to the collision-free motion planning problem*, in Hybrid Systems: Computation and Control, 4th International Workshop (Rome, Italy, 2001), Lecture Notes in Comput. Sci. 2034, M. Domenica, D. Benedetto, and A. Sangiovanni-Vincentelli, eds., Springer-Verlag, Berlin, pp. 305–318.

[7] J. Hu, M. Prandini, and S. Sastry, *Optimal coordinated motions of multiple agents moving on a plane*, SIAM J. Control Optim., 42 (2003), pp. 637–668.

[8] J. Hu, M. Prandini, and C. Tomlin, *Interesting conjugate points in formation constrained optimal multi-agent coordination*, in Proc. American Control Conference (Portland, OR), Vol. 3, IEEE, Piscataway, NJ, 2005, pp. 1871–1876.

[9] J. Hu and S. Sastry, *Optimal collision avoidance and formation switching on Riemannian manifolds*, in Proc. 40th IEEE Conference on Decision and Control, Vol. 2, Orlando, FL, 2001, pp. 1071–1076.

[10] J. Jost, *Riemannian Geometry and Geometric Analysis*, 3rd ed., Springer-Verlag, Berlin, 2002.

[11] P. S. Krishnaprasad, *Eulerian many-body problems*, in Dynamics and Control of Multibody Systems (Brunswick, ME, 1988), AMS, Providence, RI, 1989, pp. 187–208.

[12] J. Kuchar and L. C. Yang, *Survey of conflict detection and resolution modeling methods*, IEEE Trans. Intell. Transportation Systems, 1 (2000), pp. 179–189.

[13] N. E. Leonard and E. Fiorelli, *Virtual leaders, artificial potentials and coordinated control of groups*, in Proc. 40th IEEE Conference on Decision and Control, Vol. 3, Orlando, FL, 2001, pp. 2968–2973.

[14] J. W. Milnor, *Morse Theory*, Princeton University Press, Princeton, NJ, 1963.

[15] R. O. Saber, W. B. Dunbar, and R. M. Murray, *Cooperative control of multi-vehicle systems using cost graphs and optimization*, in Proc. American Control Conference (Denver, CO), Vol. 3, IEEE, Piscataway, NJ, 2003, pp. 2217–2222.

[16] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, Vol. I, 2nd ed., Publish or Perish, Wilmington, DE, 1979.

[17] D. Stipanovic, G. Inalhan, R. Teo, and C. J. Tomlin, *Decentralized overlapping control of a formation of unmanned aerial vehicles*, Automatica J. IFAC, 40 (2004), pp. 1285–1296.

[18] G. Szegő, *Orthogonal Polynomials*, 4th ed., AMS, Providence, RI, 1975.

[19] P. Tabuada, G. J. Pappas, and P. Lima, *Feasible formations of multi-agent systems*, in Proc. American Control Conference (Arlington, VA), Vol. 1, IEEE, Piscataway, NJ, 2001, pp. 56–61.

[20] M. Tillerson, L. Breger, and J. P. How, *Distributed coordination and control of formation flying spacecraft*, in Proc. American Control Conference (Denver, CO), Vol. 2, IEEE, Piscataway, NJ, 2003, pp. 1740–1743.

[21] C. Tomlin, G. Pappas, and S. Sastry, *Conflict resolution for air traffic management: A study in multi-agent hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 509–521.

[22] J. D. Wolfe, D. F. Chichka, and J. L. Speyer, *Decentralized controllers for unmanned aerial vehicle formation flight*, in Proc. AIAA Guidance, Navigation and Control Conference, San Diego, CA, 1996, pp. 29–31.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2138–2155

CHARACTERIZATION OF DELAY-INDEPENDENT STABILITY AND DELAY INTERFERENCE PHENOMENA*

WIM MICHIELS† AND SILVIU-IULIAN NICULESCU‡

Abstract. The problem of the asymptotic stability independent of delays for a class of linear systems including multiple delays is addressed. Both cases where the delays are allowed to vary independently of each other and where they are restricted to a one-dimensional subspace of the delay-parameter space are considered. It the latter case it turns out that the resulting dependency between the delays (rationally independent, rationally dependent, commensurate) plays an important role. The stability conditions are expressed in terms of the spectral properties of some appropriate complex matrices. As a consequence of the stability study, a complete characterization of the delay interference phenomenon is given. Furthermore, a connection is established with the stability theory for continuous-time delay-difference equations, subjected to delay perturbations. Various illustrative examples complete the paper.

Key words. time-delay systems, stability, delay interference

AMS subject classifications. 34K06, 34K20

DOI. 10.1137/050641223

1. Introduction.

We address the asymptotic stability in the delay-parameter space of the linear system

$$\dot{x}(t) = A_0 x(t) + \sum_{i=1}^{m} A_i x(t-\tau_i), \tag{1.1}$$

where $A_i \in \mathbb{R}^{n\times n},\ i = 0,\dots,m$, are given system matrices and $\tau_i \geq 0,\ i = 1,\dots,m$, are delay constants.

The subject of delay-independent stability of the zero solution of (1.1), that is, asymptotic stability guaranteed for all delay values, received a lot of attention in the literature starting in the 1980s. Without being exhaustive, we cite some of the approaches proposed to handle the problem: two-variable criteria [12, 13], matrix pencil techniques [3, 20], frequency-sweeping tests [5, 4], and finite-dimensional LMI conditions derived by using some appropriate quadratic Lyapunov–Krasovskii functionals [2]. For further discussions and references, see, for instance, [24, 8, 21, 6]. Essentially, in most such existing criteria for delay-independent stability and, more generally, in work on characterizing and computing stability regions in the delay-parameter space, the delays are *either* allowed to vary completely independently of each other *or* they are restricted to be commensurate (multiples of the same number).

*Received by the editors September 26, 2005; accepted for publication (in revised form) August 1, 2006; published electronically January 12, 2007. This article presents results of the research project IAP P5, funded by the program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, and of the bilateral research project PAI Tournesol (France, Flemish Community) Distributed Delays in Dynamical Systems: Analysis and Applications (2006–2007).

http://www.siam.org/journals/sicon/45-6/64122.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium (Wim.Michiels@cs.kuleuven.be). This author is a postdoctoral fellow of the Fund for Scientific Research, Flanders, Belgium.

‡Laboratoire des Signaux et Systèmes (L2S, UMR CNRS 8506), CNRS- SUPELEC 3, rue Joliot-Curie, 91190 Gif-sur-Yvette, France (Silviu.Niculescu@lss.supelec.fr).

The latter implies a relation of the form

$$(\tau_1, \dots, \tau_m) = \tau_0\ (n_1, \dots, n_m), \quad n_i \in \mathbb{N},\ i = 1, \dots, m, \tag{1.2}$$

where $n_1, \dots, n_m$ are fixed and $\tau_0 \in \mathbb{R}_+$ is the free parameter. Geometrically, the parameterization (1.2) corresponds to a particular *ray* or direction in the complete delay-parameter space.

In this paper we study the eigenvalue distribution of a collection of complex matrices, derived from the delay system (1.1). Among others, this leads us to new necessary and sufficient conditions for delay-independent stability. The criteria cover both the case where all delays vary independently of each other and the case where they are restricted to *an arbitrary* ray in the delay-parameter space, more precisely,

$$(\tau_1, \dots, \tau_m) = \tau_0\ (r_1, \dots, r_m), \quad r_i \in \mathbb{R}_+,\ i = 1, \dots, m, \tag{1.3}$$

with $\tau_0 \in \mathbb{R}_+$ being once again the free parameter. The relaxation from (1.2) to (1.3) has the following consequences:

1. It allows us to make assertions about *delay-independent stability* when the ray under consideration consists of delay values with *any type of interdependence* (commensurate, rationally (in)dependent delays). This interdependency turns out to be important.
2. A study of the *sensitivity* of delay-independent stability *along a ray* w.r.t. changes of the direction (determined by the r_i in (1.3)) becomes possible. Among others, it leads us to a complete characterization of the so-called *delay interference phenomenon*, that is, the presence of delay-independent stability along a particular ray, which is not robust against arbitrarily small perturbations of the direction of the ray.

Due to the fragility of delay-independent stability properties (note that they involve a *noncompact* set of delay values), these issues are nontrivial and deserve special attention, as we shall see. They have, however, barely been treated in the literature, excepting some contributions on the interference phenomenon and related topics [16, 17, 5, 15, 22].

To the best of the authors' knowledge, the notion of delay interference[1] was first mentioned by MacDonald [17] (see also [16]), where a second-order system including two delays is shown to be subjected to delay interference if it has the following property: delay-independent stability if the delays are equal, and delay-dependent stability with respect to each delay if the other is equal to zero. A further example (scalar system including two delays) of delay interference can be found in [5]. This is also discussed in [15] and [22], where some characterizations regarding interference are given (a frequency-sweeping test combined with a matrix pencil condition in [22], a sector characterization in [15]). As we shall discuss at the end of section 5, these results appear as corollaries of the general theory developed throughout the paper.

The paper is organized as follows. Some definitions, notations, and assumptions are presented in the next section. Section 3 contains preliminary results, in support of section 4, which is devoted to the main results (delay-independent stability and interference characterizations). In section 5 some illustrative examples are presented and discussed. In section 6 the main results are connected with the stability theory

[1] In physics, interference represents the combination of two or more wave motions to form a resultant wave in which the displacement is reinforced or canceled.

of continuous-time delay-difference equations. Finally, some concluding remarks end the paper. The appendices contain some background material.

2. Notation, definitions, and assumptions. The following notation is used:

$\mathbb{C}$	set of complex numbers
$\mathbb{C}_-,\ \mathbb{C}_+$	open left half-plane, open right half-plane
$d(\rho, E),\quad \rho \in \mathbb{C},\ E \subseteq \mathbb{C}$	$\inf_{t\in E} \lvert\rho - t\rvert$
$D(E, F),\quad E, F \subseteq \mathbb{C}$	$\sup_{\rho\in E} d(\rho, F)$
$D_h(E, F),\quad E, F \subseteq \mathbb{C}$	$\max\{D(E, F), D(F, E)\}$, Hausdorff distance
ϕ	empty set
j	imaginary unit, $j = \sqrt{-1}$
$\mathbb{N}$	set of natural numbers, includes zero
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	$\{r \in \mathbb{R} :\ r \geq 0\}$.
$\Re(\lambda), \Im(\lambda),\quad \lambda \in \mathbb{C}$	real and imaginary part of λ
$\vec{r} \in \mathbb{R}^m, \vec{n} \in \mathbb{N}^m, \ldots$	short notation for $(r_1, \ldots, r_m), (n_1, \ldots, n_m), \ldots$
$\Sigma(A)$	spectrum of the matrix A
$\mathbb{Z}$	set of integer numbers
$\bar{E}$	closure of the set E
$\mathcal{P}(E)$	power set of E, set of all subsets
$H(\lambda;\ \vec{\tau})$	characteristic function of (1.1), $\det(\lambda I - A_0 - \sum_{i=1}^m A_i e^{-\lambda\tau_i})$
$\lVert x\rVert, x \in \mathbb{R}^n$	Euclidean norm of x
$\mathcal{B}_+^m$	$\{\vec{r} \in \mathbb{R}_+^m :\ \lVert\vec{r}\rVert = 1\}$

Given a direction $(r_1, \ldots, r_m) \in \mathcal{B}_+^m$ in the delay-parameter space, we define the associated *ray*, $\mathcal{T}(\vec{r})$, as follows.

DEFINITION 2.1. *For $\vec{r} \in \mathcal{B}_+^m$, let $\mathcal{T}(\vec{r}) := \{\tau_0\ \vec{r} :\ \tau_0 \in \mathbb{R}_+\}$.*

As we address delay-independent stability properties for both cases where all delays vary independently of each other, and where they are restricted to a particular ray, we will from now on consequently use the following terminology, to avoid confusion.

DEFINITION 2.2. *The system* (1.1) *is delay-independent stable if and only if its zero solution is asymptotically stable for all $\vec{\tau} \in \mathbb{R}_+^m$.*

DEFINITION 2.3. *The ray $\mathcal{T}(\vec{r})$ is stable if and only if the zero solution of* (1.1) *is asymptotically stable for all $\vec{\tau} \in \mathcal{T}(\vec{r})$.*

Regarding delay interference we have the next definition.

DEFINITION 2.4. *A stable ray $\mathcal{T}(\vec{r})$ is subjected to the delay interference phenomenon if and only if for all $\epsilon > 0$ there exists a $\vec{s} \in \mathcal{B}_+^m$ with $\lVert\vec{r} - \vec{s}\rVert < \epsilon$ such that the ray $\mathcal{T}(\vec{s})$ is not stable.*

Recall that the zero solution of (1.1) is asymptotically stable if and only if all its *characteristic roots*, the zeros of

$$H(\lambda;\ \vec{\tau}) = \det\left(\lambda I - A_0 - \sum_{i=1}^{m} A_i e^{-\lambda\tau_i}\right),$$

are in the open left half-plane. See, for instance, [14] for stability definitions and their relation with the location of the characteristic roots.

The following technical assumption will be made throughout the paper.

Assumption 2.5.

$$\det\left(A_0 + \sum_{i=1}^{m} A_i\right) \neq 0.$$

Note that if this assumption is not satisfied, we have

$$\forall \vec{\tau} \in \mathbb{R}_+^m \quad H(0;\ \vec{\tau}) = 0,$$

which implies that the zero solution is not asymptotically stable, whatever the values of the delays. As the article is devoted to characterizations of delay-independent stability, the assumption can be made without losing generality.

3. Preliminaries. The following sets and quantities will play a major role in the characterization of delay-independent stability and the interference phenomenon.

DEFINITION 3.1. *Let*

$$\mathbb{W} = \bigcup_{\vec{\theta}\in[0,\ 2\pi]^m} \Sigma\left(A_0 + \textstyle\sum_{i=1}^{m} A_i e^{-j\theta_i}\right),$$

$$\alpha_0 = \sup\left\{\Re(\lambda) :\ \lambda \in \mathbb{W}\right\},$$

and for $\vec{r} \in \mathcal{B}_+^m$, *let*

$$\mathbb{V}(\vec{r}) = \bigcup_{\theta\geq 0} \Sigma\left(A_0 + \textstyle\sum_{i=1}^{m} A_i e^{-j\theta r_i}\right),$$

$$\alpha(\vec{r}) = \sup\left\{\Re(\lambda) :\ \lambda \in \mathbb{V}(\vec{r})\right\}.$$

It directly follows that

$$\begin{aligned} \forall \vec{r} \in \mathbb{B}_+^m \quad & \overline{\mathbb{V}(\vec{r})} \subseteq \mathbb{W}, \\ & \alpha(\vec{r}) \leq \alpha_0. \end{aligned} \tag{3.1}$$

As we shall see the relations (3.1) can be strict ($\subset$, resp., $<$). However, if the components of $\vec{r}$ are rationally independent,[2] then the next proposition applies.

PROPOSITION 3.2. *If the components of* $\vec{r}$ *are rationally independent, then*

$$\overline{\mathbb{V}(\vec{r})} = \mathbb{W}, \quad \alpha(\vec{r}) = \alpha_0.$$

Proof. Since $\mathbb{V}(\vec{r}) \subseteq \mathbb{W}$ and $\mathbb{W}$ is closed, it remains to prove for the first statement that every element of W can be approximated arbitrarily well with elements from $\mathbb{V}(\vec{r})$.

For any $\tilde{\lambda} \in \mathbb{W}$ there exists a $\vec{\theta} \in [0,\ 2\pi]^m$ such that $\tilde{\lambda}$ is a zero of the polynomial

$$p(\lambda) := \det\left(\lambda I - A_0 - \sum_{i=1}^{m} A_i e^{-j\theta_i}\right).$$

Since the delays $\vec{r}$ are rationally independent, we can apply Kronecker's theorem [11, Theorem 444], which leads to the following: for all $n \geq 1$, there exists a $k_n > 0$ such that

$$\left|e^{-j\theta_i} - e^{-jk_n r_i}\right| < \frac{1}{n}, \quad i = 1, \ldots, m.$$

[2]See Appendix B for definitions regarding the interdependency of numbers.

It follows that the set of polynomials $\{p_n(\lambda)\}_{n\geq 1}$, where

$$p_n(\lambda) := \det\left(\lambda I - A_0 - \sum_{i=1}^{m} A_i e^{-jk_n r_i}\right),$$

uniformly converges to $p(\lambda)$ on compact sets. Applying Theorem A1 of [18], a variant of Rouché's theorem, yields the following: for all $\epsilon > 0$, there exists a $\bar{n} \geq 0$ such that for all $n \geq \bar{n}$, p_n has a zero λ_n satisfying $|\lambda_n - \tilde{\lambda}| < \epsilon$. Since for any n the zeros of p_n are contained in $\mathbb{V}(\vec{r})$ the proof the first statement is complete. The second statement is a direct corollary of the first. □

Next, with $\mathcal{B}_+^m$ equipped with the Euclidean norm and $\mathcal{P}(\mathbb{C})$ with the Hausdorff metric, we address the continuity of the functions $\mathbb{V}(\cdot)$ and $\alpha(\cdot)$; see Appendix A for precise definitions. As we illustrate at the end of the section, these functions in general are not continuous at each point; however, the following weaker property holds.

PROPOSITION 3.3. *The function* $\mathbb{V} : \mathcal{B}_+^m \to \mathcal{P}(\mathbb{C}),\ \vec{r} \mapsto \mathbb{V}(\vec{r})$ *is lower semicontinuous at each* $\vec{r} \in \mathcal{B}_+^m$.

Proof. We prove lower semicontinuity of $\mathbb{V}(\cdot)$ at an arbitrary point $\vec{s} \in \mathcal{B}_+^m$.

Fix $\epsilon > 0$. Put a grid on the complex plane, whose lines are parallel to the real or imaginary axis and equally spaced with $\epsilon/4$ in both directions. This partitions the complex plane into disjunct squares. If the closure of such a square has a nonempty intersection with $\mathbb{V}(\vec{s})$, we assign a complex number to it, obtained by sampling exactly one point from the intersection. Let the set $\mathcal{G}_\epsilon$ be defined as a collection of all these samples. Since $\mathbb{V}(\vec{r})$ is bounded we have $\mathrm{card}(\mathcal{G}_\epsilon) < \infty$, and by construction,

$$\forall v \in \mathbb{V}(\vec{s}) \quad d(v, \mathcal{G}_\epsilon) < \epsilon/2. \tag{3.2}$$

Next, take any point $g \in \mathcal{G}_\epsilon$. By definition, there exists a $\tilde{\theta} > 0$ such that the matrix $A_0 + \sum_{i=1}^m A_i e^{-j\tilde{\theta}s_i}$ has a characteristic root equal to g. From continuity properties of the characteristic roots, there exists a $\delta > 0$ such that for all $\vec{r} \in \mathcal{B}_+^m$ with $\|\vec{r} - \vec{s}\| < \delta$, the matrix $A_0 + \sum_{i=1}^m A_i e^{-j\tilde{\theta}r_i}$ has an eigenvalue, say, λ, which satisfies $|\lambda - g| < \epsilon/2$. This can be repeated for all points of $\mathcal{G}_\epsilon$, and, since the cardinality of this set is finite, we can choose the threshold δ *independently* of $g \in \mathcal{G}_\epsilon$. Hence, we obtain

$$\exists \delta > 0\ \forall g \in \mathcal{G}_\epsilon\ \forall \vec{r} \in \mathcal{B}_+^m \quad \|\vec{r} - \vec{s}\| < \delta \Rightarrow d(g, \mathbb{V}(\vec{r})) < \epsilon/2. \tag{3.3}$$

The combination of (3.2) and (3.3) results in

$$\exists \delta > 0\ \forall v \in \mathbb{V}(\vec{s})\ \forall \vec{r} \in \mathcal{B}_+^m \quad \|\vec{r} - \vec{s}\| < \delta \Rightarrow d(v, \mathbb{V}(\vec{r})) < \epsilon. \tag{3.4}$$

Since the above analysis can be repeated for any $\epsilon > 0$, we end up with

$$\forall \epsilon > 0\ \exists \delta > 0 \quad \forall \vec{r} \in \mathcal{B}_+^m \quad \|\vec{r} - \vec{s}\| < \delta \Rightarrow D(\mathbb{V}(\vec{s}), \mathbb{V}(\vec{r})) \leq \epsilon, \tag{3.5}$$

that is, lower semicontinuity of $\mathbb{V}$ at $\vec{s}$. □

COROLLARY 3.4. *The function* $\alpha : \mathcal{B}_+^m \to \mathbb{R},\ \vec{r} \mapsto \alpha(\vec{r})$ *is lower semicontinuous at each* $\vec{r} \in \mathcal{B}_+^m$.

By combining Propositions 3.2 and 3.3 with relations (3.1) we arrive at a stronger result at rationally independent $\vec{r}$, as follows.

PROPOSITION 3.5. *The function* $\mathbb{V} : \mathcal{B}_+^m \to \mathcal{P}(\mathbb{C}),\ \vec{r} \mapsto \mathbb{V}(\vec{r})$ *is continuous at rationally independent* $\vec{r}$. *The function* $\alpha : \mathcal{B}_+^m \to \mathbb{R},\ \vec{r} \mapsto \alpha(\vec{r})$ *is continuous at rationally independent* $\vec{r}$.

We illustrate the above concepts and properties with the equation

$$\dot{x}(t) = -1.3x(t) - x(t-\tau_1) - \frac{1}{2}x(t-\tau_2), \tag{3.6}$$

for which we have

$$\mathbb{W} = \bigcup_{(\theta_1,\theta_2)\in[0,\ 2\pi]^2} \left(-1.3 - e^{-j\theta_1} - \tfrac{1}{2}e^{-j\theta_2}\right) = \left\{\lambda \in \mathbb{C} :\ \tfrac{1}{2} \le |\lambda + 1.3| \le \tfrac{3}{2}\right\},$$
$$\mathbb{V}(\vec{r}) = \bigcup_{\theta\ge 0} \left(-1.3 - e^{-jr_1\theta} - \tfrac{1}{2}e^{-jr_2\theta}\right).$$

In Figure 1 we show the set $\mathbb{V}(\vec{r})$ for some values of $\vec{r}$. We also indicate the boundaries of $\mathbb{W}$. When the components of $\vec{r}$ are commensurate, $\mathbb{V}(\vec{r})$ forms a closed curve. In all the cases displayed we have $\alpha(\vec{r}) < \alpha_0$ and $\overline{\mathbb{V}(\vec{r})} \subset \mathbb{W}$ (strict relations). *Discontinuities* of the function $\vec{r} \mapsto \mathbb{V}(\vec{r})$ occur at all rationally dependent $\vec{r}$. To illustrate the mechanism, let us first compare the cases (a) and (f). Ratio $r_2/r_1 = 1.9$ can be seen as a perturbation of $r_2/r_1 = 2$ and, although the components of $\vec{r}$ are still commensurate, they are more independent, since the coprime numbers 19 and 10 (having ratio 1.9) are larger than 2 and 1 (having ratio 2). As a consequence, the curve

$$\theta \ge 0 \mapsto -1.3 - e^{-j\theta} - \frac{1}{2}e^{-j\frac{r_2}{r_1}\theta} \tag{3.7}$$

with $r_2/r_1 = 1.9$ only closes at $\theta = 20\pi$ when θ is increased from zero (instead of $\theta = 2\pi$ for $r_2/r_1 = 2$), and a larger portion of $\mathbb{W}$ is covered by $\mathbb{V}(\vec{r})$. This is clearly visible in subplot (f), where we have also depicted the values of (3.7) for $\theta \in [0,\ 2\pi]$ (bold part of curve). Next, if $r_2/r_1 = 2$ would be perturbed instead to an irrational value, such as $2 - \pi/n, n \ge 2$, then the corresponding curve $\mathbb{V}(\vec{r})$ would never close and its points would densely fill $\mathbb{W}$, as follows from Proposition 3.2. Since n can be taken arbitrarily large, the perturbation can be taken arbitrarily small.

4. Main results. Step by step we characterize the stability of the zero solution of (1.1) in the delay parameters, in terms of the sets $\mathbb{V}(\cdot)$ and $\mathbb{W}$. The main results will be stated in Theorems 4.1, 4.5, and 4.7, and they will be summarized in Table 2.

We first recall some properties of the characteristic roots of (1.1), which are needed in the subsequent analysis. The number of characteristic roots in any right half-plane is finite. Furthermore, the function

$$g :\ \mathbb{R}_+^m \to \mathbb{R},\ \vec{\tau} \mapsto g(\vec{\tau}) = \max_{\lambda\in\mathbb{C}} \{\Re(\lambda) :\ H(\lambda;\ \vec{\tau}) = 0\} \tag{4.1}$$

is continuous [21]. As a consequence, if the zero solution of (1.1) is asymptotically stable for some delay values but the system is not delay-independent stable, then there exist delay values for which the rightmost characteristic roots are *on the imaginary axis*.

Necessary and sufficient conditions for delay-independent stability are expressed by the following theorem.

THEOREM 4.1. *The system* (1.1) *is delay-independent stable if and only if*

$$\mathbb{W} \subset (\mathbb{C}_- \cup \{0\}). \tag{4.2}$$

Proof. Both implications are proved by *contradiction*.

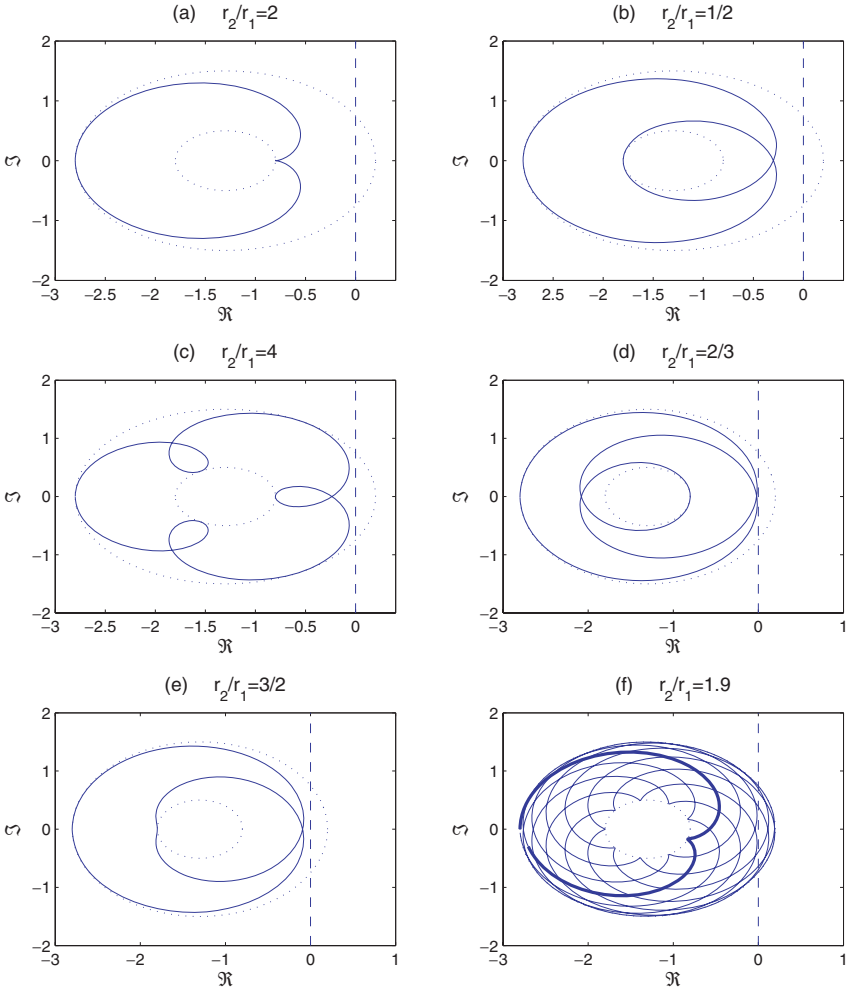


FIG. 1. *The set* $\mathbb{V}(\vec{r})$ *for the system* (3.6)*, for different values of* $\vec{r}$ *(solid curves). The dotted curves are the boundaries of* $\mathbb{W}$*.*

$\Rightarrow$ Assume that $\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$. We distinguish between two cases, as follows.

Case 1. $\mathbb{W}$ contains an element $j\omega,\ \omega \neq 0$, i.e., there exists a $\vec{\theta} \in [0,\ 2\pi]^m$ such that

$$\det\left(j\omega I - A_0 - \sum_{i=1}^{m} A_i e^{-j\theta_i}\right) = 0.$$

It follows that $H(j\omega;\ \vec{\tau}_0) = 0$, where

$$\vec{\tau}_0 = \left(\frac{\theta_1}{\omega}, \ldots, \frac{\theta_m}{\omega}\right).$$

This contradicts the delay-independent stability.

Case 2. $\mathbb{W}$ contains an element $\tilde{\lambda} \in \mathbb{C}_+$. Hence, there exists a $\vec{\theta_0} \in [0,\ 2\pi]^m$ such that

$$A_0 + \sum_{i=1}^{m} A_i e^{-j\theta_i} \tag{4.3}$$

has a characteristic root $\tilde{\lambda}$ for $\vec{\theta} = \vec{\theta_0}$. If $\vec{\theta} \in \mathbb{R}_+^m$ is varied continuously (w.r.t. the Euclidean norm) from $\vec{\theta_0}$ to $(k_1 2\pi, \ldots, k_m 2\pi)$, for some $\vec{k} \in \mathbb{N}^m$, the spectrum of the matrix (4.3) continuously changes (w.r.t. the Hausdorff metric) to the spectrum of $A_0 + \sum_{i=1}^m A_i$. If the latter matrix is not Hurwitz, we have a contradiction (instability for all delays equal to zero). If not, it implies that $\mathbb{W} \cap \mathbb{C}_+$ is not separated from the imaginary axis, and we are the situation described with Case 1.

$\Leftarrow$ Assume that (1.1) is not delay-independent stable. In the case where $A_0 + \sum_{i=1}^m A_i$ is not Hurwitz, this matrix has a characteristic root in the closed right half-plane, different from zero (Assumption 2.5), and we already have a contradiction since $\Sigma(A_0 + \sum_{i=1}^m A_i) \subseteq \mathbb{W}$. In the other case, the system (1.1) is asymptotically stable for all delays equal to zero and, by the continuity of (4.1), there must be characteristic roots on the imaginary axis for some delay values, i.e.,

$$\exists \omega > 0\ \exists \vec{\tau}_0 \in \mathbb{R}_+^m \quad H(j\omega;\ \vec{\tau}_0) = 0.$$

It follows that $j\omega \in \mathbb{W}$ and we have a contradiction. ∎

Remark 4.2. For a delay-independent stable system, the condition $0 \notin \mathbb{W}$, respectively, $0 \in \mathbb{W}$, is referred to in the literature as *strong*, respectively, *weak*, delay-independent stability (see [2, 21] and the references therein), notions which are related to stability properties of some associated 2D system.

The stability of a ray $\mathcal{T}(\vec{r})$ can be characterized in an analogous way, as follows.

PROPOSITION 4.3. *The ray $\mathcal{T}(\vec{r})$ is stable if and only if*

$$\mathbb{V}(\vec{r}) \subset (\mathbb{C}_- \cup \{0\})\,. \tag{4.4}$$

Proof. The proof is similar to the proof of Theorem 4.1. ∎

The next proposition states that if the system (1.1) is not delay-independent stable, then it is always prone to the delay interference phenomenon.

PROPOSITION 4.4. *Assume that $\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$. If a ray $\mathcal{T}(\vec{r})$ is stable, then it is subjected to the delay interference phenomenon.*

Proof. From the assumption $\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$ it follows that (1.1) is not delay-independent stable. So there exist delay values $\vec{\tau} \in \mathbb{R}_+^m$ and a frequency $\omega > 0$ such that $H(j\omega; \vec{\tau}) = 0$. For any given $\vec{k} \in \mathbb{N}^m,\ \vec{k} \neq 0$, this implies that $H(j\omega; \vec{\tau}(n)) = 0$ for all $n \in \mathbb{N}$, where

$$\vec{\tau}(n) = \left(\tau_1 + k_1 n\ \frac{2\pi}{\omega}, \ldots, \tau_m + k_m n\ \frac{2\pi}{\omega}\right). \tag{4.5}$$

Therefore, the ray

$$\mathcal{T}\left(\frac{\vec{\tau}(n)}{\|\vec{\tau}(n)\|}\right)$$

is unstable for all $n \in \mathbb{N}$. From

$$\lim_{n\to\infty} \frac{\vec{\tau}(n)}{\|\vec{\tau}(n)\|} = \frac{\vec{k}}{\|\vec{k}\|}$$

and the fact that this analysis can be repeated for every nonzero $\vec{k} \in \mathbb{N}^m$, it follows that every ray $\mathcal{T}(\vec{r})$ where $\vec{r}$ has *commensurate* components becomes unstable when applying certain infinitesimal perturbations to $\vec{r}$.

If the components of $\vec{r}$ are noncommensurate, then there exist infinitesimal $\delta\vec{r}$ such that $\vec{r} + \delta\vec{r}$ has commensurate components ($\mathbb{Q}^m$ is dense in $\mathbb{R}^m$), and the above arguments can be repeated. This completes the proof. ∎

Combining the previous results for the case where $\mathbb{W}$ has a nonempty intersection with the open right half-plane yields the following.

THEOREM 4.5. *Assume that* $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$. *Then the following hold:*

1. *If the components of* $\vec{r}$ *are rationally independent, then the ray* $\mathcal{T}(\vec{r})$ *is unstable.*
2. *If the ray* $\mathcal{T}(\vec{r})$ *is stable, then it is subjected to the delay interference phenomenon.*
3. *The set* $\left\{\vec{r} \in \mathcal{B}_+^m : \ \mathcal{T}(\vec{r}) \text{ stable}\right\}$ *is nowhere dense in* $\mathcal{B}_+^m$.

Proof. If the components of $\vec{r}$ are rationally independent, then $\overline{\mathbb{V}(\vec{r})} = \mathbb{W}$ by Proposition 3.2. A combination with the assumption of the theorem leads to $\mathbb{V}(\vec{r}) \cap \mathbb{C}_+ \neq \phi$. Applying Proposition 4.3 then yields the first assertion.

The second assertion can be proven in two different ways. It directly follows from Proposition 4.4, whose proof mainly relies on the invariance property (4.5). Alternatively, let $\vec{r} \in \mathbb{R}_+^m$ be such that $\mathcal{T}(\vec{r})$ is stable. Following from the density of the rationally independent numbers in $\mathbb{C}^m$, there exist arbitrarily small perturbations $\delta\vec{r} \in \mathbb{R}_+^m$ such that $\vec{r}+\delta\vec{r}$ has rationally independent components. By the first assertion of the theorem, the ray $\mathcal{T}((\vec{r} + \delta\vec{r})/(\|\vec{r} + \delta\vec{r}\|))$ is unstable. This implies that $\mathcal{T}(\vec{r})$ is subjected to the delay interference phenomenon.

The third assertion follows from Propositions 3.2 and 3.3 and Proposition 4.3. ∎

Under the assumption of the previous theorem, we comment on the detection of *stable* rays $\mathcal{T}(\vec{r})$ (if any), *without* the explicit computation of stability/instability regions of (1.1) in the delay-parameter space. From Theorem 4.5, the components of $\vec{r}$ cannot be rationally independent. By Proposition 4.3, we have to search for values for which $\mathbb{V}(\vec{r}) \subset (\mathbb{C}_- \cup \{0\})$, whereas we have $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$ and $\mathbb{V}(\vec{r}) \subseteq \mathbb{W}$. Thus, values of $\vec{r}$, for which $\mathbb{V}(\vec{r})$ doesn't cover $\mathbb{W}$ very well, are good candidates. Following from Propositions 3.2 and 3.3, such values must be characterized by a large dependence, for instance, $\vec{r} = \vec{n}/\|\vec{n}\|$, with $\vec{n} \in \mathbb{N}^m$ and $\|\vec{n}\|$ small. To fix the ideas, we pick up the example (3.6), which satisfies $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$ since $\alpha_0 = 0.2$. Table 1 displays the corresponding values of $\alpha(\vec{r})$, for rationally dependent $\vec{r}$. It becomes apparent that stable rays correspond to

$$(n_1, n_2) \in \{(1,0),\ (0,1),\ (1,2),\ (1,4),\ (2,1),\ (2,3),\ (3,2)\}\,. \tag{4.6}$$

Notice that for five of these cases the set $\mathbb{V}(\vec{r})$ is depicted in Figure 1, subplots (a) to (e).

Finally, we look at the special case, where $\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$ but still lies in the *closed* left half-plane. But first we prove the following lemma.

LEMMA 4.6. *Assume that* (1.1) *has a characteristic root in* $\mathbb{C}_+$ *for some delay values. Then* $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$.

Proof. Inspired by the proof of Lemma 2.2 in [10], we use an approximation and a continuation argument.

TABLE 1
Values of $\alpha(\vec{r})$, where $\vec{r} = (n_1,\ n_2)/\sqrt{n_1^2 + n_2^2}$, computed for the system (3.6).

$\alpha(\vec{r})$	0	1	2	3	4	5	6	7	8	$\cdots$ n_1
0		−0.8								
1	−1.8	0.20	−0.27	0.20	0.06	0.20	0.13	0.20	0.16	
2		−0.55		−0.01		0.11		0.15		
3		0.20	−0.08		0.08	0.20		0.20	0.16	
4		−0.06		0.06		0.13		0.16		
5		0.20	0.05	0.20	0.11		0.15	0.20	0.17	
6		0.07				0.14		0.16		
7		0.20	0.11	0.20	0.14	0.20	0.16		0.17	
8		0.13		0.14		0.16		0.17		
$\vdots$										
n_2										

Let $\vec{\tau}$ be delay values, for which (1.1) has a characteristic root $\tilde{\lambda} \in \mathbb{C}_+$, i.e.,

$$\det\left\{\tilde{\lambda}I - A_0 - \sum_{i=1}^{m} A_i e^{-\tilde{\lambda}\tau_i}\right\} = 0. \tag{4.7}$$

Without losing generality we may assume that the components of $\vec{\tau}$ are commensurate,[3] and we can set

$$\vec{\tau} = \tau_0\left(\frac{p_1}{q}, \ldots, \frac{p_m}{q}\right), \quad p_i \in \mathbb{N},\ i = 1, \ldots, m, \quad q \in \mathbb{N}.$$

Then (4.7) can be written as

$$\det\left\{\tilde{\lambda}I - A_0 - \sum_{i=1}^{m} A_i \mu_0^{p_i}\right\} = 0, \tag{4.8}$$

where

$$\mu_0 = e^{-\frac{\tilde{\lambda}\tau_0}{q}}.$$

From (4.8) it follows that the polynomial

$$p(\mu; k) := \det\left\{k\tilde{\lambda}I - A_0 - \sum_{i=1}^{m} A_i\ \mu^{p_i}\right\}, \tag{4.9}$$

parameterized by $k \in [1, \infty)$, satisfies $p(\mu_0; 1) = 0$. By definition we have $|\mu_0| < 1$. Obviously, as $k \to \infty$, *all* zeros of (4.9) must become unbounded. From this and the continuity of the zeros w.r.t. k, there must be a pair $(\tilde{k}, \tilde{\mu})$ with $\tilde{k} > 1$ and $|\tilde{\mu}| = 1$ such that

$$\det\left\{\tilde{k}\tilde{\lambda}I - A_0 - \sum_{i=1}^{m} A_i\ \tilde{\mu}^{p_i}\right\} = 0.$$

[3]If not, they can be made commensurate by adding perturbations, which can be chosen sufficiently small, such that characteristic roots in $\mathbb{C}_+$ are maintained.

TABLE 2
Characterization of stability of the steady-state solution of (1.1) *along rays in the delay parameter space, as a function of* $\mathbb{W}$ *and* $\mathbb{V}(\cdot)$.

<table>
<tr><th>$\mathbb{W} \subset (\mathbb{C}_- \cup \{0\})$</th><th>$\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$ and
$\mathbb{W} \cap \mathbb{C}_+ = \phi$</th><th>$\mathbb{W} \cap \mathbb{C}_+ \neq \phi$</th></tr>
<tr><td rowspan="3">delay-independent stable:

$0 \notin \mathbb{W}$: strong

$0 \in \mathbb{W}$: weak</td><td colspan="2">not delay-independent stable
ray $\mathcal{T}(\vec{r})$ stable if and only if $\mathbb{V}(\vec{r}) \subset (\mathbb{C}_- \cup \{0\})$</td></tr>
<tr><td colspan="2">stable ray (if any) subjected to the
delay interference phenomenon</td></tr>
<tr><td>characteristic roots in $\mathbb{C}_+$
not possible</td><td>ray $\mathcal{T}(\vec{r})$ unstable
for $\vec{r}$ rationally independent</td></tr>
</table>

This means that $\tilde{k}\tilde{\lambda} \in \mathbb{W}$ and, consequently, $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$. ∎

The combination of Theorem 4.1, Proposition 4.4, and Lemma 4.6 results in the next theorem.

THEOREM 4.7. *Assume that* $\mathbb{W} \cap \mathbb{C}_+ = \phi$ *and* $\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$. *Then there exist delay values for which* (1.1) *has characteristic roots on the imaginary axis, while there are no delay values for which* (1.1) *has characteristic roots in* $\mathbb{C}_+$. *Every stable ray is subjected to the delay interference phenomenon.*

The main results of this section are summarized in Table 2. The cases $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$ and $\mathbb{W} \subset (\mathbb{C}_- \cup \{0\})$ with $0 \notin \mathbb{W}$ are generic. The other cases, where the rightmost elements of $\mathbb{W}$ are on the imaginary axis, characterize situations where a system is on the edge of loosing or acquiring delay-independent stability. Recall that this analysis has been performed under Assumption 2.5, which can be checked a priori. (If the assumption is not satisfied, then there exist no values of $\vec{\tau}$ for which the zero solution of (1.1) is asymptotically stable.)

To conclude, we briefly discuss some existing results related to the interference phenomenon. In [15] it is shown that delay-independent stability is equivalent to asymptotic stability for all delay values lying in a nontrivial sector in the delay-parameter space, and to the robustness of stability of a ray, consisting of commensurate delay values, w.r.t. small perturbations of the direction (see [15] for precise formulations). Note that the latter two statements imply the existence of a stable ray, which is not subjected to the interference phenomenon. Given a stable ray, the additional condition to have interference of [17] (delay-dependent stability when one of the (two) delays is set to zero) and of [22] (frequency-sweeping test) are in fact conditions for the presence of characteristic roots on the imaginary axis for some delay values, thus conditions for *not* having delay-independent stability. In this way, the above cited results are a direct corollary of the fact that a system is *either* delay-independent stable *or* every stable ray is subjected to the delay interference phenomenon; see Table 2.

5. Illustrations. We present two examples, which together illustrate all results and phenomena described in the previous sections. Besides the delays, the examples will exhibit another parameter. This allows us to illustrate, in addition, two scenarios for the transition between delay-independent stability and delay-dependent stability: one via weak delay-independent stability, the other via the case $\mathbb{W} \not\subset (\mathbb{C}_- \cup \{0\})$, $\mathbb{W} \cap \mathbb{C}_+ = \phi$.

As a first example we consider

$$\dot{x}(t) = -ax(t) - x(t-\tau_1) - \frac{1}{2}x(t-\tau_2), \tag{5.1}$$

where $a \in \mathbb{R}$ is a parameter. We have

$$\mathbb{W} = \left\{ \lambda \in \mathbb{C} : \ \frac{1}{2} \leq |\lambda + a| \leq \frac{3}{2} \right\}.$$

According to Table 2 the zero solution of (5.1) is delay-independent stable if and only if $a \geq \frac{3}{2}$. For $a > \frac{3}{2}$ we have strong delay-independent stability, and for $a = \frac{3}{2}$ we have weak delay-independent stability. For $a < \frac{3}{2}$ we are directly in the case $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$.

For $a = 1.3$ the system reduces to (3.6), for which the analysis in the previous section and Table 1 shows that there are seven stable rays, characterized by $\vec{r} = \vec{n}/\|\vec{n}\|$, with $\vec{n}$ given by (4.6). Such an analysis can easily be repeated for other values of a, since a change of this parameter, say, δa, affects only the sets $\mathbb{W}$ and $\mathbb{V}(\vec{r})$ by a shift along the real axis of $-\delta a$, while the numerical values of Table 1 change accordingly with $-\delta a$. For $a = 1$, there are only three stable rays left, characterized by

$$(n_1, n_2) \in \{(1,0),\ (0,1),\ (1,2)\}, \tag{5.2}$$

and for $a < -0.5$ all rays are unstable.

In Figure 2 we show the complete stability/instability regions of (5.1) in the delay-parameter space,[4] for $a = 1$ and $a = 1.3$. The solid lines correspond to delay values for which there are characteristic roots on the imaginary axis. The dashed lines indicate the stable rays. Notice that small perturbations of their slope lead to intersections with the solid curves, a consequence of the delay interference phenomenon. As $a \to 1.5$, we have in fact a scenario toward delay-independent stability, characterized by an increase of the number of stable rays. That number becomes arbitrarily large when getting arbitrarily closed to the bifurcation value $a = 1.5$, for which we have weak delay-independent stability.

Next, we analyze the system

$$\dot{x}(t) = A_0 x(t) + A_1 x(t - \tau_1) + A_2 x(t - \tau_2), \tag{5.3}$$

where

$$A_0 = \begin{bmatrix} 0 & 1 \\ a - a^2 - \frac{5}{4} & 2a - 1 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ \frac{1}{5} & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 \\ -\frac{4}{5} & 0 \end{bmatrix} \tag{5.4}$$

and $a \in \mathbb{R}$ is a parameter. The corresponding set $\mathbb{W}$ is depicted in Figure 3. A change of parameter a results once again in a shift of this set along the real axis.

For $a < 0$ the system is strongly delay-independent stable and for $a > 0$ we have $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$; see Table 2. For the intermediate value, $a = 0$, we are in the special case where the assumptions of Theorem 4.7 are satisfied.

In Figure 4, we plot the stability region in the (τ_1, τ_2)-space for $a = 1/16$ and $a = 0$. In the first case there are three stable rays, determined by $\vec{r} = \vec{n}/\|\vec{n}\|$, with

$$\vec{n} \in \{(1,0),\ (1,1),\ (4,1)\}.$$

These directions can also be obtained directly by constructing a table similar to Table 1. As $a \to 0+$, all closed curves in the (τ_1, τ_2)-space, corresponding to charac-

[4]The stability crossing curves were computed as Hopf bifurcation curves with the package DDE-BIFTOOL [7], thereby exploiting (frequency-dependent) invariance properties w.r.t. delay shifts. Other approaches are described in [9] and [23].

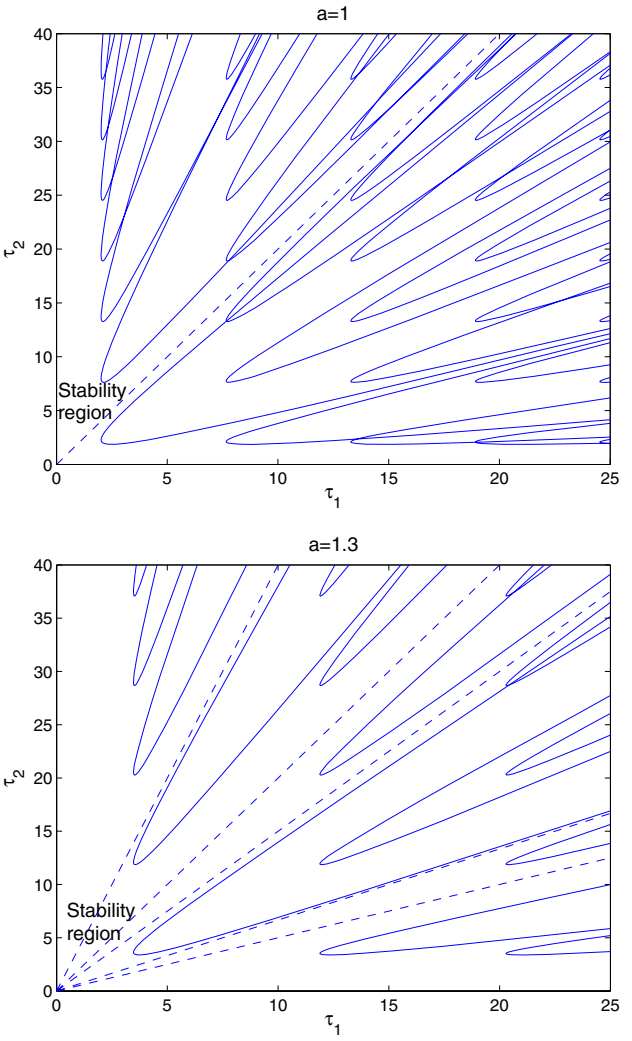


FIG. 2. *Stability/instability regions of the zero solution of* (5.1) *in the* (τ_1, τ_2)*-space, for* $a = 1$ *(top) and* $a = 1.3$ *(bottom). For* $a = 1$*, the directions of the stable rays (dashed lines) are given by* (5.2)*, for* $a = 1.3$ *by* (4.6)*. For* $a \geq 1.5$ *the systems is delay-independent stable.*

teristic roots on the imaginary axis, shrink, and at the limit $a = 0$, where Theorem 4.7 applies, they have collapsed to equally spaced points. For such delay values, which can be computed analytically as

$$(\tau_1, \tau_2) = \left(\frac{10\pi}{3\sqrt{3}} + k\frac{4\pi}{\sqrt{3}}, \frac{4\pi}{3\sqrt{3}} + l\frac{4\pi}{\sqrt{3}} \right), \quad k, l \in \mathbb{N}, \tag{5.5}$$

the system has characteristic roots on the imaginary axis (more precisely, $\pm j\sqrt{3}/2$), but for all other delay values its zero solution is asymptotically stable. This is illustrated in Figure 5, where the real parts of the rightmost characteristic roots are displayed along a particular ray, containing some of the points (5.5).

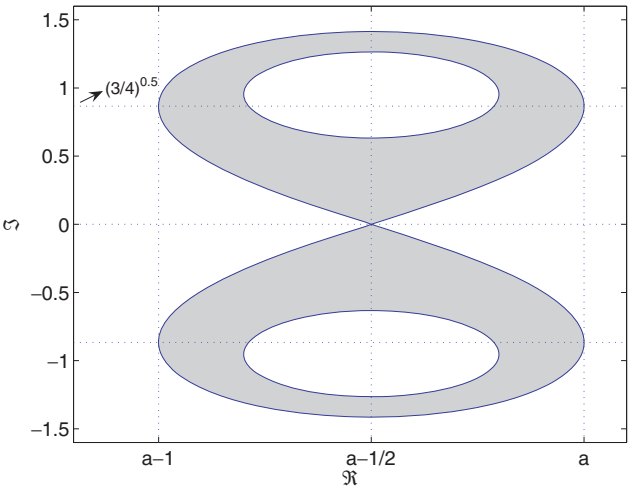


FIG. 3. *The set* $\mathbb{W}$*, corresponding to the system* (5.3)–(5.4)*, is shaded.*

Notice that for the limit case, $a = 0$, the number of unstable rays is infinite but *countable*, since the unstable rays have to contain at least one of the points (5.5). Hence, they are characterized by a slope

$$\frac{r_2}{r_1} = \frac{2+6l}{5+6k},\ k, l \in \mathbb{N},$$

being rational numbers. This leads to a paradox: for $a = 0$ all rays $\mathcal{T}(\vec{r})$, with r_2/r_1 irrational, are stable, whereas for any $a > 0$, they are unstable (since $\mathbb{W} \cap \mathbb{C}_+ \neq \phi$ and Theorem 4.5 applies). The explanation is as follows: for all rationally independent $\vec{r}$ and every value of a, we have for the example

$$\begin{aligned} &\alpha(\vec{r}) = \alpha_0, \\ &\Re(\lambda) < \alpha_0 \quad \forall \lambda \in \mathbb{V}(\vec{r}). \end{aligned}$$

Consequently, $\mathbb{V}(\vec{r})$ intersects the closed right half-plane if $\alpha_0 > 0$ but not if $\alpha_0 = 0$.

6. Relation with the stability of continuous-time delay-difference equations. We relate the results of the previous sections with stability conditions for continuous-time delay-difference equations of the form

$$x(t) - \sum_{i=1}^{m} A_i x(t - r_i) = 0, \tag{6.1}$$

which are discussed in, e.g., [10, 19, 18] and the references therein.

We assume that $\det(I - \sum_{i=1}^{m} A_i) \neq 0$, which prevents a characteristic root at zero for all $\vec{r} \in \mathbb{R}_+^m$. Then the stability conditions for the zero solution of (6.1) from [10, Theorem 2.2 and Corollary 2.2] can be rephrased as displayed in Table 3.

A comparison between Tables 2 and Table 3 reveals a strong correspondence between the stability of the zero solution of (1.1) along rays in the delay space

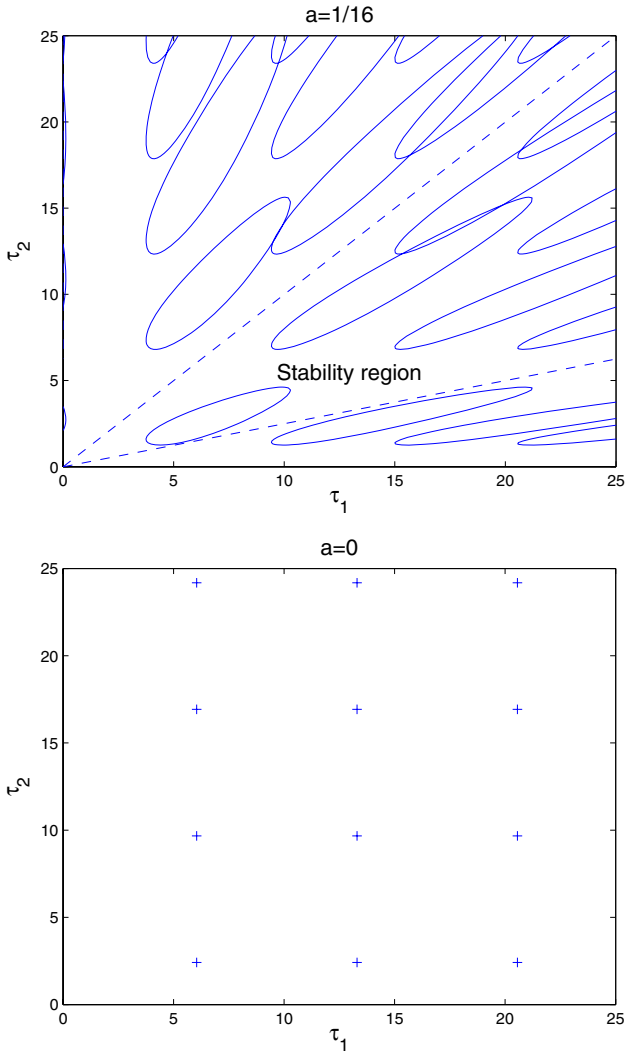


FIG. 4. *Stability/instability regions of the zero solution of* (5.3)–(5.4) *in the* (τ_1, τ_2)*-space, for* $a = 1/16$ *(top) and* $a = 0$ *(bottom).*

and the stability of the zero solution of (6.1). This correspondence is described in Table 4.

7. Concluding remarks. A complete characterization of delay-independent stability, stability of rays in the delay-parameter space, and the delay interference phenomenon was made. The results were illustrated with numerical examples. In addition, scenarios for the transition between delay-dependent stability and delay-independent stability were shown, and the correspondence with the stability theory for delay-difference equations was outlined.

It is worthwhile to mention that the stability results for rays in the delay-parameter space can easily be extended toward higher dimensional subspaces. For instance, asymptotic stability for all delay values lying in a subspace spanned by $\vec{r}^{(1)} \in \mathcal{B}_+^m$ and

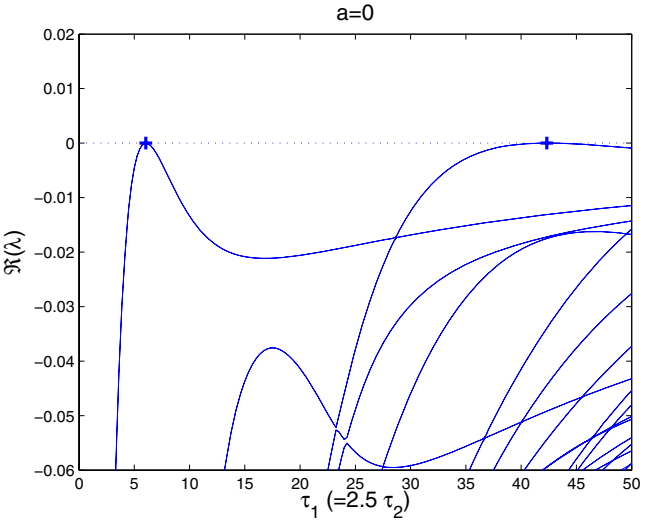


FIG. 5. *Real parts of the rightmost characteristic roots of the system* (5.3)–(5.4) *with* $a = 0$*, for delay values along the ray* $\mathcal{T}(5/\sqrt{29}, 2/\sqrt{29})$.

TABLE 3

Conditions for the stability of the zero solution of (6.1). *The set* $\mathbb{W}$ *is defined as in Definition* 3.1*, with* A_0 *set to zero.*

$\mathbb{W} \subset \{\lambda \in \mathbb{C} : \ \lvert\lambda\rvert < 1\}$	$\sup\{\lvert\lambda\rvert : \ \lambda \in \mathbb{W}\} > 1$
asympt. stable for all values of $\vec{r}$	not asympt. stable for all values of $\vec{r}$
	if $\vec{r}$ is stabilizing, then there exist infinitesimal perturbations on $\vec{r}$ which destroy stability $\equiv$ not strongly stable
	unstable for rationally independent $\vec{r}$

$\vec{r}^{(2)} \in \mathcal{B}_+^m$ will be determined by the position of the set

$$\bigcup_{\theta_1,\theta_2 \geq 0} \Sigma \left(A_0 + \sum_{i=1}^m A_i e^{-j\left(\theta_1 r_i^{(1)} + \theta_2 r_i^{(2)}\right)} \right) \subseteq \mathbb{W}.$$

Furthermore, the study of the sets $\mathbb{W}$ and $\mathbb{V}(\cdot)$ does not only lead to a powerful tool for characterizing delay-independent stability properties. Note, for instance, that the intersection of $\mathbb{W}$ with the imaginary axis contains all possible values where characteristic roots can cross the imaginary axis when the delays are varied. Actually, from these values and/or the corresponding parameters, that is, $\vec{\theta}$ in Definition 3.1, detailed information can be retrieved about the geometry of the stability crossing curves/(hyper)surfaces in the delay-parameter space, thereby generalizing the results from [9].

Appendix A. Continuity properties. Throughout the paper we encounter functions from $\mathcal{B}_+^m \subset \mathbb{R}^m$ to $\mathbb{R}$ and from $\mathcal{B}_+^m$ to $\mathcal{P}(\mathbb{C})$. We assume that $\mathbb{R}^m$ is equipped with the Euclidean norm and $\mathcal{P}(\mathbb{C})$ with the Hausdorff metric.

A function $f : \mathcal{B}_+^m \to \mathbb{R}$ is lower semicontinuous, respectively, upper semicontinuous, at $\vec{s}$ if and only if (see, for instance, [1])

$$\forall \epsilon > 0 \ \exists \delta > 0 \ \ \forall \vec{r} \in \mathcal{B}_+^m \ \ \|\vec{r} - \vec{s}\| < \delta \Rightarrow f(\vec{r}) - f(\vec{s}) > -\epsilon,$$

TABLE 4
The similarity between the stability of the zero solution of (1.1) *along rays in the delay space, and the stability of the zero solution of* (6.1).

Equation (1.1)	$\rightsquigarrow$	Equation (6.1)
relation of $\mathbb{W}$ w.r.t. imaginary axis	$\rightsquigarrow$	relation of $\mathbb{W}$ w.r.t. unit circle
direction $\vec{r}$ in delay space	$\rightsquigarrow$	nominal delay value $\vec{r}$
(un)stable ray $\mathcal{T}(\vec{r})$	$\rightsquigarrow$	(de)stabilizing delay value $\vec{r}$
stable ray subjected to the delay interference phenomenon	$\rightsquigarrow$	stability for nominal $\vec{r}$ destroyed by infinitesimal perturbations on $\vec{r}$

respectively,

$$\forall \epsilon > 0 \ \exists \delta > 0 \ \ \forall \vec{r} \in \mathcal{B}_+^m \ \ \|\vec{r} - \vec{s}\| < \delta \Rightarrow f(\vec{r}) - f(\vec{s}) < \epsilon.$$

It is continuous at $\vec{s}$ when it is both upper and lower semicontinuous at $\vec{s}$.

A function $f : \mathcal{B}_+^m \to \mathcal{P}(\mathbb{C})$ is lower semicontinuous at $\vec{s}$ if and only if

$$\forall \epsilon > 0 \ \exists \delta > 0 \ \ \forall \vec{r} \in \mathcal{B}_+^m \ \ \|\vec{r} - \vec{s}\| < \delta \Rightarrow D(f(\vec{s}), f(\vec{r})) < \epsilon.$$

When replacing $D(f(\vec{s}), f(\vec{r}))$ with $D(f(\vec{r}), f(\vec{s}))$, respectively, $D_h(f(\vec{s}), f(\vec{r}))$, we have upper semicontinuity, respectively, continuity, at $\vec{s}$.

Appendix B. Interdependency of numbers. The real numbers $r_1, r_2, \ldots, r_m$ are *rationally independent* if and only if

$$\sum_{i=1}^{m} z_i r_i = 0, \quad z_i \in \mathbb{Z},$$

implies $z_i = 0, \ \ i = 1, \ldots, m$. For example, two numbers are rationally independent if and only if their ratio is an irrational number.

If the real numbers $r_1, \ldots r_m$ are *rationally dependent* (that is, not rationally independent), then there always exists an integer $p < m$ and a matrix $\Gamma \in \mathbb{Z}^{m \times p}$ of full column rank such that

$$\begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix} = \Gamma \begin{bmatrix} s_1 \\ \vdots \\ s_p \end{bmatrix},$$

with the numbers $s_1, \ldots, s_p$ rationally independent. Thus, rationally dependent numbers depend on a smaller number of rationally independent numbers. In the special case where $p = 1$, the numbers $r_1, \ldots, r_m$ are called *commensurate*, as they are all multiples of the same number.

For example, the numbers 1, π and $\exp(1)$ are rationally independent, the numbers $1, 2$, and $5/3$ commensurate. The numbers 1, π and $1 + \pi$ are rationally dependent, yet not commensurate, as

$$\begin{bmatrix} 1 \\ \pi \\ 1+\pi \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \pi \end{bmatrix},$$

with 1 and π rationally independent.

Acknowledgments. The authors would like to thank the associate editor and the anonymous reviewers for the careful reading of the manuscript and the careful comments.

REFERENCES

[1] C. Berge, *Espaces Topologiques. Fonctions Multivoques*, Dunod, Paris, 1966.
[2] P. A. Bliman, *LMI characterization of strong delay-independent stability of linear delay systems via quadratic Lyapunov–Krasovskii functionals*, Systems Control Lett., 43 (2001), pp. 263–274.
[3] J. Chen, G. Gu, and C. A. Nett, *A new method for computing delay margins for stability of linear delay systems*, Systems Control Lett., 26 (1995), pp. 107–117.
[4] J. Chen and H. A. Latchman, *Frequency sweeping tests for stability independent of delay*, J. Math. Anal. Appl., 40 (1995), pp. 1640–1645.
[5] R. Datko, *Remarks concerning the asymptotic stability and stabilization of linear delay differential equations*, J. Math. Anal. Appl., 111 (1985), pp. 571–584.
[6] L. Dugard and E. I. Verriest, *Stability and Control of Time-Delay Systems*, Lecture Notes in Control and Inform. Sci. 228, Springer-Verlag, 1998.
[7] K. Engelborghs, T. Luzyanina, and G. Samaey, *DDE-BIFTOOL v.* 2.00: *A Matlab Package for Bifurcation Analysis of Delay Differential Equations*, TW Report 330, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, 2001.
[8] K. Gu, V. L. Kharitonov, and J. Chen, *Stability of Time-Delay Systems*, Control Eng., Birkhäuser, Boston, 2003.
[9] K. Gu, S. I. Niculescu, and J. Chen, *On stability of crossing curves for general systems with two delays*, J. Math. Anal. Appl., 311 (2005), pp. 231–253.
[10] J. K. Hale and S. M. Verduyn Lunel, *Strong stabilization of neutral functional differential equations*, IMA J. Math. Control Inform., 19 (2002), pp. 5–23.
[11] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, Oxford, UK, 1968.
[12] E. W. Kamen, *On the relationship between zero criteria for two-variable polynomials and asymptotic stability of delay differential equations*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 983–984.
[13] E. W. Kamen, *Linear systems with commensurate time delays: Stability and stabilization independent of delay*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 367–375.
[14] V. B. Kolmanovskii and V. R. Nosov, *Stability of Functional Differential Equations*, Mathematics in Science and Engineering 180, Academic Press, New York, 1986.
[15] J. Louisell, *Absolute stability in linear delay-differential systems: Ill-posedness and robustness*, IEEE Trans. Automat. Control, 40 (1995), pp. 1288–1291.
[16] N. MacDonald, *Two delays may not destabilize although either delay can*, Math Biosci., 82 (1986), pp. 127–140.
[17] N. MacDonald, *An interference effect of independent delays*, IEE Proc. Control Theory Appl., D, 134 (1987), pp. 38–42.
[18] W. Michiels, K. Engelborghs, D. Roose, and D. Dochain, *Sensitivity to infinitesimal delays in neutral equations*, SIAM J. Control Optim., 40 (2002), pp. 1134–1158.
[19] W. Michiels and T. Vyhlídal, *An eigenvalue based approach to the stabilization of linear time-delay systems of neutral type*, Automatica, 41 (2005), pp. 991–998.
[20] S.-I. Niculescu, *Stability and hyperbolicity of linear systems with delayed state: A matrix pencil approach*, IMA J. Math. Control Inform., 15 (1998), pp. 331–347.
[21] S.-I. Niculescu, *Delay Effects on Stability: A Robust Control Approach*, Lecture Notes in Comput. Sci., Springer-Verlag, Heidelberg, 2001.
[22] S.-I. Niculescu, *On interference phenomena in stability of linear systems with multiple delays*, in Proceedings of the Fourth Annual Asian Control Conference, Singapore, 2002.
[23] R. Sipahi and N. Olgac, *Complete stability robustness of third-order LTI multiple time-delay systems*, Automatica, 41 (2005), pp. 1413–1422.
[24] O. Toker and H. Özbay, *Complexity issues in robust stability of linear delay-differential systems*, Math. Control Signals Syst., 9 (1996), pp. 386–400.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2156–2168

ON MARKOV GAMES WITH AVERAGE REWARD CRITERION AND WEAKLY CONTINUOUS TRANSITION PROBABILITIES*

HEINZ-UWE KÜENLE†

Abstract. In this paper we consider two-person zero-sum stochastic games with the average reward criterion and Borel state and action spaces. A geometric drift condition is assumed. We show that the optimality (Shapley) equation has a unique solution if the transition probability function is weakly continuous, the stage reward is lower semicontinuous, and the set-valued mappings of admissible actions satisfy some semicontinuity assumptions. Furthermore, the minimizing player has an optimal stationary strategy and the maximizing player has an ε-optimal stationary strategy for every $\varepsilon > 0$.

Key words. zero-sum Markov games, average reward criterion, weakly continuous transition probabilities, optimality equation

AMS subject classifications. 91A15, 91A05, 91A25

DOI. 10.1137/040617303

1. Introduction.

This paper deals with two-person zero-sum stochastic games with average payoff criterion and unbounded reward function. As in related papers (see, for instance, [1], [2], [5], [7], [11], [13]), we make some stochastic stability assumptions. For the existence of the value and of optimal or ε-optimal strategies further assumptions are necessary. In the literature especially (semi-)continuity and compactness conditions are known. Concerning discounted Markov games or Markov games with finite horizon, two classes of such conditions are considered. In the one class the transition probability function is strongly continuous, in the other class it is weakly continuous. But in all the papers cited above only the case of strongly continuous transition probabilities is investigated. The reason may be that there are some difficulties in considering stochastic games with weakly continuous transition probabilities under the average reward criterion. On the other hand, in many applications the transition probabilities are weakly, not strongly, continuous. Jaśkiewicz and Nowak [3] were the first to treat stochastic games with average payoff criterion, unbounded rewards, stochastic stability conditions, and weakly continuous transition probabilities. They assume continuity of the reward function and of the set-valued mappings of admissible actions and prove the existence of the value of the game and of optimal stationary strategies using a vanishing discount factor approach. But the question of the existence of a solution of the optimality (Shapley) equation is not completely answered in their paper.

In this paper we use the same stochastic stability conditions. However, we assume only that the reward function is lower semicontinuous, the set-valued mapping of admissible actions of the maximizing player is lower semicontinuous (and complete-valued), and that of the minimizing player is upper semicontinuous (and compact-valued). We show that the optimality equation has a unique solution, the game has a value, the minimizing player has an optimal stationary strategy, and the maximizing player has an ε-optimal stationary strategy for every $\varepsilon > 0$.

*Received by the editors October 20, 2004; accepted for publication (in revised form) August 3, 2006; published electronically January 22, 2007.
http://www.siam.org/journals/sicon/45-6/61730.html
†Department of Mathematics, Brandenburg University of Technology Cottbus, PF 10 13 44, D-03013 Cottbus, Germany (kuenle@math.tu-cottbus.de).

The main ideas of our proof are the following. First we substitute a certain Borel measurable function δ in the stochastic stability conditions (which is the generalization of the indicator function of a small set (in the sense of [8])) by an upper semicontinuous function. Then we introduce a parametrized family of operators and show that there are parameters for which these operators have lower semicontinuous fixed points. Finally, we prove that there is one such parameter for which the solution of the fixed point equation defines a solution of the optimality equation.

This proof can be considered as a modification of the proofs in [5] and [7]. We remark that in the case of strongly continuous transition function another fixed point approach is given by Vega-Amaya [13]. But if δ is not continuous, then Vega-Amaya's approach as well as the original approach in [5] and [7] does not work in the case of weakly continuous transition probability function, since the operators used do not map the corresponding sets of lower semicontinuous functions into itself.

The paper is organized as follows. In section 2 we introduce the Markov game model. Section 3 contains our assumptions and some preliminary results. In section 4 we present the main result and its proof.

2. The Markov game model. For a metric space Y we denote by $\mathfrak{B}_Y$ the σ-algebra of all Borel subsets of Y and by $P(Y)$ the family of all probability measures on $\mathfrak{B}_Y$. If Y is a Borel subset of a complete separable metric space, endowed with $\mathfrak{B}_Y$, then Y is called a Borel space. Measurability of a function on Y stands for $\mathfrak{B}_Y$-measurability.

We study two-person zero-sum stochastic games

$$\mathcal{M} = (X, A, B, K_A, K_B, p, r),$$

whereby

1. X is the *state space* and A and B are the *action spaces* for players 1 and 2. All three sets are assumed to be Borel spaces.
2. K_A and K_B are the *constraint sets*, which are Borel subsets of $X \times A$ and $X \times B$. For each state $x \in X$, the set

$$A(x) := \{a \in A : (x, a) \in K_A\}$$

represents the *set of admissible actions* for player 1 in state x and is assumed to be nonempty. Analogously, we define $B(x)$ as the (nonempty) *set of admissible actions* for player 2 in state x. The set

$$K := \{(x, a, b) : x \in X, a \in A(x), b \in B(x)\}$$

is a Borel subset of $X \times A \times B$ (see [9]).
3. p is a Borel measurable transition probability from K to X, the *law of motion.*
4. $r : K \to \mathbb{R}$ is a Borel measurable function from K to $\mathbb{R}$, the *reward function* for player 1.

We can describe the situation of a zero-sum Markov game as follows. At any discrete time point $n = 0, 1, 2, \ldots$, the game is in state $x \in X$. Both players simultaneously decide for actions $a \in A(x)$ and $b \in B(x)$. Player 1 receives a reward $r(x, a, b)$, while player 2 has to pay an amount of $r(x, a, b)$. At the beginning of the next period $n + 1$ the game is in a state x', chosen according to the probability distribution $p(\cdot|x, a, b)$. The goal of player 1 is to maximize his reward, while player 2 wants to minimize his costs.

Let $H_n = K^n \times X$. An element $h_n \in H_n$ is called the *history* of the game up to time n.

A sequence $\pi = (\pi_0, \pi_1, \dots)$ is a *strategy* for player 1 if each π_n is a transition probability from H_n to A, and $\pi_n(A(x_n)|x_0, a_0, b_0, \dots, x_n) = 1$ for all n. The class of all strategies for player 1 will be denoted by Π. Analogously, we define $\mathbf{P}$ as the class of all strategies for player 2. We assume $\Pi \neq \emptyset$ and $\mathbf{P} \neq \emptyset$, which follows from the assumptions to be given later.

A Borel measurable transition probability f from X to A with $f(A(x)|x) = 1$ for all $x \in X$ is called a *Markovian decision rule (of the first player)*. The set of all Markovian decision rules for the first player is denoted $\mathbf{F}$. For $f \in \mathbf{F}$, we write $f(x)$ for the probability measure on $A(x)$ with $f(x)(D) = f(D|x)$ for all Borel sets $D \subseteq A(x)$. A sequence $\pi = (f_0, f_1, \dots)$ with $f_n \in \mathbf{F}$ for all n is a *Markovian strategy* for player 1. If additionally $f_n = f$ for some $f \in \mathbf{F}$ and all n, $\pi = (f, f, \dots) =: f^\infty$ is a *stationary* strategy for player 1. Hence, the set of all stationary strategies for player 1 can be identified by $\mathbf{F}$. Similarly, we define $\mathbf{G}$ as the set of all Markovian decision rules for the second player (transition probabilities g from X to B with $g(B(x)|x) = 1$ for all $x \in X$) which can be identified by the set of all stationary strategies of this player.

Let us consider the space $\Omega := (X \times A \times B)^\infty$, endowed with the product σ-algebra $\mathcal{F}$. It follows from the Ionescu–Tulcea theorem that there exists a suitable probability measure $P_x^{\pi\rho}$ and a stochastic process $\{X_n, A_n, B_n\}$ on $(\Omega, \mathcal{F})$ for every pair $(\pi, \rho) \in \Pi \times \mathbf{P}$ and each initial state $x = x_0 \in X$. Here, the random variables X_n, A_n, B_n represent the state and the actions of players 1 and 2 in stage n. Let $E_x^{\pi\rho}$ be the expectation operator with respect to $P_x^{\pi\rho}$. Then, for a pair $\pi \in \Pi, \rho \in \mathbf{P}$ of strategies, the *total expected n-stage reward* for player 1 is

$$J_n(x, \pi, \rho) := E_x^{\pi\rho}\left[\sum_{m=0}^{n-1} r(X_m, A_m, B_m)\right]$$

if the expectation exists. The *expected average reward* for player 1 is

$$J(x, \pi, \rho) := \liminf_{n\to\infty} J_n(x, \pi, \rho)/n.$$

We define the *lower value* and the *upper value* of the game by

$$v_l(x) := \sup_{\pi\in\Pi} \inf_{\rho\in\mathbf{P}} J(x, \pi, \rho) \qquad \text{and} \qquad v_u(x) := \inf_{\rho\in\mathbf{P}} \sup_{\pi\in\Pi} J(x, \pi, \rho).$$

Obviously, there is always $v_l(x) \leq v_u(x)$ for all $x \in X$. If the equality $v_l(x) = v_u(x)$ holds for all $x \in X$, the common function is called the *value* v of the game.

Assume that v exists and let $\varepsilon \geq 0$. Then a strategy $\pi^* \in \Pi$ is called *ε-optimal* for player 1 if

$$\inf_{\rho\in\mathbf{P}} J(x, \pi^*, \rho) \geq v(x) - \varepsilon$$

for all $x \in X$ and similarly a strategy $\rho^* \in \mathbf{P}$ is *ε-optimal* for player 2 if

$$\sup_{\pi\in\Pi} J(x, \pi, \rho^*) \leq v(x) + \varepsilon$$

for all $x \in X$. As is usual, 0-optimal strategies are called *optimal*.

3. Assumptions and preliminary results. Let Y and Z be metric spaces and $\Gamma : Y \to \mathfrak{B}_Z$ a set-valued mapping. Γ is called lower (upper) semicontinuous if $\{y \in Y : \Gamma(y) \cap S \neq \emptyset\}$ is open (closed) for every open (closed) subset of $S \subseteq Z$. Γ is called continuous if Γ is lower (l.s.c) and upper semicontinuous (u.s.c).

A real function $v : Y \to \mathbb{R}$ is called l.s.c. if for every $r \in \mathbb{R}$ the set $\{y \in Y : v(y) > r\}$ is open. v is called u.s.c. if $-v$ is l.s.c. Clearly, a semicontinuous function is Borel measurable.

A transition probability q from Y to Z is called weakly continuous if $\int_Z u(z)q(dz|\cdot)$ is continuous on Y for every bounded continuous function $u : Z \to \mathbb{R}$.

We now formulate our main assumptions.

Assumption 3.1.

A1. $A(\cdot)$ is l.s.c. $A(x)$ is complete for every $x \in X$.

A2. $B(\cdot)$ is u.s.c. $B(x)$ is compact for every $x \in X$.

A3. r is l.s.c.

A4. p is weakly continuous.

A5. There is a continuous function $W : X \to [1, \infty)$, a Borel measurable function $\delta : K \to [0, 1]$ with $\delta \not\equiv 0$, a probability measure $\mu \in P(X)$, and constants $\alpha \in (0, 1)$, $\beta > 0$, such that

(a) $\sup_{(x,a,b)\in K} \frac{|r(x,a,b)|}{W(x)} < \infty$,

(b) $\int_X W(\xi)p(d\xi|\cdot)$ is continuous on K,

(c) $\int_X W(\xi)p(d\xi|x, a, b) \leq \alpha W(x) + \beta\delta(x, a, b)$ for each $(x, a, b) \in K$,

(d) $\delta(\cdot)\mu(D) \leq p(D|\cdot)$ for each Borel set $D \subseteq X$.

We remark that under A1 and A2 the strategy sets $\mathbf{F}$ and $\mathbf{G}$ are not empty. (See Lemma 2.1. and Lemma 2.2. in [10], for instance.)

Furthermore, A5(c) and A5(d) imply $\int_X W(\xi)\mu(d\xi) < \infty$, because there is $(x_0, a_0, b_0) \in K$ with $\delta(x_0, a_0, b_0) > 0$. Hence,

$$\begin{aligned}\int_X W(\xi)\mu(d\xi) &\leq \frac{1}{\delta(x_0, a_0, b_0)} \int_X W(\xi)p(d\xi|x_0, a_0, b_0) \\ &\leq \frac{1}{\delta(x_0, a_0, b_0)}(\alpha W(x_0) + \beta\delta(x_0, a_0, b_0)) < \infty.\end{aligned}$$

Let $w : K \to \mathbb{R}$ be a measurable function for which $w(x, \cdot)$ is bounded from below for all $x \in X$. For $\nu_1 \in P(A(x))$, $\nu_2 \in P(B(x))$ we set

$$w(x, \nu_1, \nu_2) := \int_{A(x)} \int_{B(x)} w(x, a, b)\nu_2(db)\nu_1(da) \tag{3.1}$$

and $w_{fg}(x) := w(x, f(x), g(x))$ for each pair $f \in \mathbf{F}$, $g \in \mathbf{G}$. Especially,

$$p(D|x, \nu_1, \nu_2) := \int_{A(x)} \int_{B(x)} p(D|x, a, b)\nu_2(db)\nu_1(da)$$

and $p_{fg}(D|x) := p(D|x, f(x), g(x))$ for any $D \in \mathfrak{B}_X$, $\nu_1 \in P(A(x))$, $\nu_2 \in P(B(x))$, $f \in \mathbf{F}$, $g \in \mathbf{G}$.

We denote by V the function $V : X \to \mathbb{R}$ with $V(x) := W(x) + \beta$ for each $x \in X$. For a function $u : X \to \mathbb{R}$ we define $\|u\|_V := \sup_{x\in X} \frac{|u(x)|}{V(x)}$.

Let $\mathcal{V}$ be the set of all real Borel measurable functions u with finite $\|u\|_V$. $\mathcal{V}$ is a Banach space.

Let Y and Z be metric spaces, q a substochastic kernel from Y to Z, and u a measurable function on $Y \times Z$. We denote by $qu(y)$ the integral

$$qu(y) := \int_Z u(y,z)q(dz|y) \tag{3.2}$$

(if it exists for all $y \in Y$). Then we can consider q also as an operator which maps corresponding sets of measurable functions on $Y \times Z$ into the set of measurable functions on Y.

We then have

$$pv(x,a,b) = \int_X v(\xi)p(d\xi|x,a,b)$$

for each $v \in \mathcal{V}$ and $(x,a,b) \in K$ and

$$\mu v = \int_X v(\xi)\mu(d\xi).$$

The operators p and μ are well defined on $\mathcal{V}$ since the assumptions A5(a) and A5(c) hold.

Let $\mathcal{C}$ be the set of all nonnegative bounded continuous functions u on X with $\mu u > 0$. We put

$$\zeta(x,a,b) := \inf_{u\in\mathcal{C}} \frac{pu(x,a,b)}{\mu u}.$$

Obviously, $\delta \le \zeta \le 1$. The following lemma shows that we can replace δ by the u.s.c. function ζ.

LEMMA 3.2. *Let* A4 *and* A5 *hold.*

(a) ζ *is u.s.c.*

(b)

$$\int_X W(\xi)p(d\xi|x,a,b) \le \alpha W(x) + \beta\zeta(x,a,b) \text{ for each } (x,a,b) \in K. \tag{3.3}$$

(c)

$$\zeta(\cdot)\mu(D) \le p(D|\cdot) \text{ for each } D \in \mathfrak{B}_X. \tag{3.4}$$

Proof.

(a) Assumption A4 implies the continuity of pu for $u \in \mathcal{C}$. From the properties of semicontinuous functions it follows that the infimum of an arbitrary nonempty set of continuous functions is u.s.c.

(b) This inequality follows from $\delta \le \zeta$.

(c) Let U be an open subset of X and I_U the indicator function of U. Since I_U is l.s.c., there is a sequence of bounded continuous functions $(w_n)_{n\in\mathbb{N}}$ with $w_n \nearrow I_U$. (See, for instance, Lemma 5.1. in [10]). We can assume $w_n \ge 0$, because for a continuous function w the function $\max\{w,0\}$ is also continuous. From

$$\zeta(x,a,b)\mu w_n \le pw_n(x,a,b)$$

it follows then by the monotone convergence theorem that

$$\zeta(x,a,b)\mu(U) \le p(U|x,a,b)$$

for all $(x,a,b) \in K$.

Because X is a Borel subset of a metric space, every probability measure on $\mathfrak{B}_X$ is regular. (See, for instance, Lemma 1.16 in [4].) That means especially

$$\mu(D) = \inf\{\mu(U) : \ D \subseteq U, \ U \text{ open}\}$$

and

$$p(D|x,a,b) = \inf\{p(U|x,a,b) : \ D \subseteq U, \ U \text{ open}\}$$

for all $D \in \mathfrak{B}_X$. It follows (3.4). □

We now put

$$P(D|x,a,b) := p(D|x,a,b) - \zeta(x,a,b)\mu(D)$$

for all $D \in \mathfrak{B}_X$, $(x,a,b) \in K$. It follows from (3.4) that P is a substochastic kernel. The operator P according to (3.2) is then well defined on $\mathcal{V}$. It holds that $Pv = pv - \zeta \cdot \mu v$ for $v \in \mathcal{V}$.

We set

$$\lambda := \frac{\alpha + \beta}{1 + \beta}.$$

Obviously, $\alpha < \lambda < 1$. The following properties play an essential role in our approach.

LEMMA 3.3. *Let* A4 *and* A5 *hold.*

(a) *P is isotonic.*

(b) *$0 \le PV(x,a,b) \le \lambda V(x)$ for all $(x,a,b) \in K$.*

Proof.

(a) $Pv_1 \le Pv_2$ for $v_1 \le v_2$ for $v_1, v_2 \in \mathcal{V}$ follows immediately from Lemma 3.2(c).

(b) It holds that

$$\beta = \lambda - \alpha + \lambda\beta \le (\lambda - \alpha)W + \lambda\beta \tag{3.5}$$

and

$$\beta \le \mu V. \tag{3.6}$$

It follows that

$$\begin{aligned} pV &= pW + \beta \\ &\le \alpha W + \beta\zeta + \beta \quad \text{(see (3.3))} \\ &\le \alpha W + \zeta\mu V + (\lambda - \alpha)W + \lambda\beta \quad \text{(see (3.5) and (3.6))} \\ &= \lambda V + \zeta\mu V. \end{aligned}$$

This yields $PV \le \lambda V$. $PV \ge 0$ follows from (a). □

Let $\tilde{\mathcal{V}}$ be the set of all l.s.c. functions from $\mathcal{V}$. $\tilde{\mathcal{V}}$ is complete. To establish this fact, we consider a Cauchy sequence (v_n) with $v_n \in \tilde{\mathcal{V}}$ and put $v^* := \lim_{n\to\infty} v_n \in \mathcal{V}$. The sequence (u_n) with $u_n := v_n - \|v^* - v_n\|_V V$ has the properties $u_n \in \tilde{\mathcal{V}}$, $u_n \le v^*$, and $u_n \to v^*$. Hence, $v^* = \sup_{n\in\mathbb{N}} u_n \in \tilde{\mathcal{V}}$.

The following minimax selection theorem is a slight modification of Nowak's Theorem 5.1. in [10]. Similar results can be found also in [6] and [12].

LEMMA 3.4. *Let assumptions* A1 *and* A2 *be satisfied. Let $u : K \to \mathbb{R}$ be an l.s.c. function with $\sup_{(x,a,b)\in K} \frac{|u(x,a,b)|}{V(x)} < \infty$ and $v(x) := \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} u(x, f(x), g(x))$*

for all $x \in X$. Then it holds that

(a) $v \in \tilde{\mathcal{V}}$,

(b) $v(x) = \inf_{g\in\mathbf{G}} \sup_{f\in\mathbf{F}} u(x, f(x), g(x))$ *for all* $x \in X$,

(c) *there is* $g^* \in \mathbf{G}$ *with* $v(x) = \sup_{f\in\mathbf{F}} u(x, f(x), g^*(x))$ *for all* $x \in X$,

(d) *for every* $\varepsilon > 0$ *there is* $f_\varepsilon \in \mathbf{F}$ *with* $v(x) - \varepsilon \le \inf_{g\in\mathbf{G}} u(x, f_\varepsilon(x), g(x))$ *for all* $x \in X$.

Proof. There is $c \in \mathbb{R}$ with $u' := u + cV > 0$. Hence, u' is l.s.c. and bounded from below. Then (a)–(c) follow directly from Theorem 5.1 in [10]. To show (d), Nowak [10] assumes that u is also bounded from above. But it is easy to see that his proof works also under our assumption $\sup_{(x,a,b)\in K} \frac{|u(x,a,b)|}{V(x)} < \infty$. ∎

4. Main result. In this section we will prove our main result, namely, that under Assumption 3.1 the optimality equation has a solution $(h^*, v^*) \in \mathbb{R} \times \tilde{\mathcal{V}}$, where h^* is unique and v^* is unique up to an additive constant. Furthermore, we show that h^* is the value of the game, the first player has an ε-optimal stationary strategy for every $\varepsilon > 0$, and the second player has an optimal stationary strategy.

THEOREM 4.1. *Let Assumption* 3.1 *be satisfied. Then the following hold.*

(a) *The optimality equation*

$$v(x) + h = \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} \left\{ r_{fg}(x) + \int_X v(\xi) p_{fg}(d\xi|x) \right\} \quad \text{for all } x \in X \tag{4.1}$$

has a solution $(h^*, v^*) \in \mathbb{R} \times \tilde{\mathcal{V}}$.

(b) *There is a Markovian decision rule* $g^* \in \mathbf{G}$ *and, for every* $\varepsilon > 0$, *a Markovian decision rule* $f_\varepsilon \in \mathbf{F}$, *with*

$$\begin{aligned} v^*(x) + h^* &= \sup_{f\in\mathbf{F}} \left\{ r_{fg^*}(x) + \int_X v^*(\xi) p_{fg^*}(d\xi|x) \right\} \\ &\le \inf_{g\in\mathbf{G}} \left\{ r_{f_\varepsilon g}(x) + \int_X v^*(\xi) p_{f_\varepsilon g}(d\xi|x) \right\} + \varepsilon \end{aligned} \tag{4.2}$$

for all $x \in X$.

(c) f_ε^∞ *is an* ε*-optimal strategy of the first player and* $g^{*\infty}$ *is an optimal strategy of the second player.*

(d) h^* *is the value of the game for every solution* $(h^*, v^*) \in \mathbb{R} \times \tilde{\mathcal{V}}$ *of* (4.1).

(e) v^* *is unique in* $\tilde{\mathcal{V}}$ *up to additive constants.*

Before we prove Theorem 4.1 we will show some auxiliary results. But let us introduce first two further operators. In this section it is assumed that Assumption 3.1 is satisfied, which implies that the following operators are well defined. We set

$$T_{fg}^\eta v := r_{fg} - (1 - \zeta_{fg})\eta + P_{fg} v$$

and

$$LT^\eta v(x) := \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} T_{fg}^\eta v(x)$$

for all $f \in \mathbf{F}$, $g \in \mathbf{G}$, $v \in \mathcal{V}$, $x \in X$, and $\eta \in \mathbb{R}$.

We remark that it holds that $T_{fg}^\eta v = [T^\eta v]_{fg}$ in the sense of (3.1).

The proof of the next lemma can be omitted since it follows immediately from Lemma 3.3 by standard methods.

LEMMA 4.2. *Fix an arbitrary* $\eta \in \mathbb{R}$.

(a) T^η_{fg} *is contractive in* $\mathcal{V}$, *which means* $T^\eta_{fg}\mathcal{V} \subseteq \mathcal{V}$, *and for all* $u, v \in \mathcal{V}$, $f \in \mathbf{F}$, *and* $g \in \mathbf{G}$ *it holds*

$$\|T^\eta_{fg}u - T^\eta_{fg}v\|_V \le \lambda\|u - v\|_V.$$

(b) *For all* $u, v \in \mathcal{V}$ *it holds*

$$\|LT^\eta u - LT^\eta v\|_V \le \lambda\|u - v\|_V.$$

(c) T^η_{fg} *and* LT^η *are isotonic.*

We remark that we have $LT^\eta\tilde{\mathcal{V}} \subseteq \tilde{\mathcal{V}}$ for continuous ζ, which means that in this case LT^η is contractive in $\tilde{\mathcal{V}}$ and the idea of the proof in [5] works also under the above assumptions. But the assumption of continuous ζ seems very strong. (See also the corresponding remarks in [3].) Therefore, we have to modify the proof in [5]. We will first find a set $\check{H} \subseteq \mathbb{R}$ and for every $\eta \in \check{H}$ a complete set $\check{\mathcal{V}}_\eta \subseteq \mathcal{V}$ with $LT^\eta\check{\mathcal{V}}_\eta \subseteq \check{\mathcal{V}}_\eta$. Hence, LT^η is contractive in $\check{\mathcal{V}}_\eta$ and has a unique fixed point $v^*_\eta \in \check{\mathcal{V}}_\eta$. Then we will show that there is an $\eta^* \in \check{H}$ such that $(\eta^*, v^*_{\eta^*})$ is a solution of the optimality equation.

The next lemma gives a condition under which $LT^\eta v \in \tilde{\mathcal{V}}$ for $v \in \tilde{\mathcal{V}}$.

LEMMA 4.3. $LT^\eta v \in \tilde{\mathcal{V}}$ *for every* $v \in \tilde{\mathcal{V}}$ *and* $\eta \in \mathbb{R}$ *with* $\mu v \ge \eta$.

Proof. Let

$$u(x,a,b) := r(x,a,b) + \int_X v(\xi)p(d\xi|x,a,b) - \eta + (\eta - \mu v)\zeta(x,a,b)$$

for all $(x,a,b) \in K$. It is well known that for weakly continuous p the integral $\int_X v'(\xi)p(d\xi|\cdot)$ is l.s.c. if v' is l.s.c. and bounded from below. (See, e.g., Lemma 5.3. in [10].) Setting $v' := v + \|v\|_V V$ we see also that u is l.s.c. The statement follows then by Lemma 3.4. □

Let $v \in \tilde{\mathcal{V}}$. We set

$$\check{L}T^\eta v := \begin{cases} LT^\eta v & \text{if } \eta \le \mu v, \\ -\infty & \text{if } \eta > \mu v, \end{cases}$$

where $-\infty$ means a numerical function identical with $-\infty$. Especially, $\check{L}T^\eta(-\infty) := -\infty$. From Lemma 4.3 it follows that we can consider $\check{L}T^\eta$ as an operator $\check{L}T^\eta : \tilde{\mathcal{V}} \cup \{-\infty\} \to \tilde{\mathcal{V}} \cup \{-\infty\}$.

LEMMA 4.4. *Let* $c := \sup_{(x,a,b)\in K} \frac{|r(x,a,b)|}{V(x)}$, $C := -c(1-\lambda)^{-1}$, *and* $\eta := C\mu V$. *If* $u_0 \equiv 0$ *and* $u_n = \check{L}T^\eta u_{n-1}$ *for* $n \ge 1$, *then* $u_n \ge CV$ *and* $\mu u_n \ge \eta$ *for all* $n \in \mathbb{N}$.

Proof. Assume that $u_{n-1} \in \tilde{\mathcal{V}}$ and $u_{n-1} \ge CV$. Then $u_n \in \tilde{\mathcal{V}}$, and

$$\begin{aligned} u_n = LT^\eta u_{n-1} &\ge \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} \{-cV - \eta(1-\zeta_{fg}) + P_{fg}(-c(1-\lambda)^{-1}V)\} \\ &\ge \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} \{-cV - c(1-\lambda)^{-1}P_{fg}V\} \\ &\ge -c(1+\lambda(1-\lambda)^{-1})V = CV. \end{aligned}$$

Hence, $\mu u_n \ge C\mu V = \eta$.

We now get the statement by mathematical induction. □

This lemma allows us now to define the aforementioned sets $\check{H}$ and $\check{\mathcal{V}}_\eta$. We put

$$\check{\mathcal{V}}_\eta := \{v \in \tilde{\mathcal{V}} : [\check{L}T^\eta]^n v > -\infty \text{ for every } n \in \mathbb{N}\} \text{ and } \check{H} := \{\eta \in \mathbb{R} : \check{\mathcal{V}}_\eta \neq \emptyset\}.$$

From Lemma 4.4 it follows that $\check{H} \neq \emptyset$.

We now show an inequality which will be helpful to prove the next lemmas.

LEMMA 4.5. *Let* $u_0, u_0' \in \tilde{\mathcal{V}}$, $\eta, \eta' \in \mathbb{R}$. *If* $u_n := [\check{L}T^\eta]^n u_0 > -\infty$, $u_n' := [\check{L}T^{\eta'}]^n u_0' > -\infty$, *then*

$$\|u_n - u_n'\|_V \leq \lambda^n \|u_0 - u_0'\|_V + \frac{1}{1-\lambda}|\eta - \eta'|.$$

Proof. It follows from the definition of $\check{L}T^\eta$ and from Lemma 4.3 $\eta \leq \mu u_{k-1}$, $\eta' \leq \mu u_{k-1}'$, and $u_k, u_k' \in \tilde{\mathcal{V}}$ for $1 \leq k \leq n$. Hence, $u_k = LT^\eta u_{k-1}$ and $u_k' = LT^{\eta'} u_{k-1}'$ for $1 \leq k \leq n$.

Assume that

$$u_{k-1} \leq u_{k-1}' + \lambda^{k-1}\|u_0 - u_0'\|_V V + \frac{1}{1-\lambda}|\eta - \eta'|V.$$

Then we get by means of Lemma 3.3 for all $f \in \mathbf{F}$, $g \in \mathbf{G}$,

$$\begin{aligned} T_{fg}^\eta u_{k-1} &= r_{fg} - \eta(1 - \zeta_{fg}) + P_{fg} u_{k-1} \\ &= r_{fg} - \eta'(1 - \zeta_{fg}) + P_{fg} u_{k-1}' + (\eta' - \eta)(1 - \zeta_{fg}) + P_{fg}(u_{k-1} - u_{k-1}') \\ &\leq T_{fg}^{\eta'} u_{k-1}' + |\eta' - \eta|V + \lambda^{k-1}\|u_0 - u_0'\|_V \lambda V + \frac{\lambda}{1-\lambda}|\eta' - \eta|V \\ &= T_{fg}^{\eta'} u_{k-1}' + \lambda^k \|u_0 - u_0'\|_V V + \frac{1}{1-\lambda}|\eta' - \eta|V. \end{aligned}$$

Hence,

$$\begin{aligned} u_k = LT^\eta u_{k-1} &\leq LT^{\eta'} u_{k-1}' + \lambda^k \|u_0 - u_0'\|_V V + \frac{1}{1-\lambda}|\eta' - \eta|V \\ &= u_k' + \lambda^k \|u_0 - u_0'\|_V V + \frac{1}{1-\lambda}|\eta' - \eta|V. \end{aligned}$$

By mathematical induction it follows

$$u_n \leq u_n' + \lambda^n \|u_0 - u_0'\|_V V + \frac{1}{1-\lambda}|\eta' - \eta|V.$$

Because we can exchange u_n and u_n', we obtain the statement from this inequality. ∎

LEMMA 4.6. $\check{\mathcal{V}}_\eta$ *is complete for every* $\eta \in \check{H}$.

Proof. Let (u_n) be a Cauchy sequence in $\check{\mathcal{V}}_\eta$. Since $\tilde{\mathcal{V}}$ is complete, there is $u^* \in \tilde{\mathcal{V}}$ with $\lim_{n\to\infty} u_n = u^*$.

From $\mu u_n \geq \eta$ it follows $\mu u^* \geq \eta$ by the dominated convergence theorem. Let $v_k := [\check{L}T^\eta]^k u^*$. Assume that there is $N \in \mathbb{N}$ with $\mu v_{N-1} \geq \eta$, $\gamma := \eta - \mu v_N > 0$. We consider $n_0 \in \mathbb{N}$ with $\|u_{n_0} - u^*\|_V \leq \frac{\gamma}{2\mu V}$. For $w_k := [LT^\eta]^k u_{n_0}$ we get, from Lemma 4.5,

$$\|v_N - w_N\|_V \leq \lambda^N \|u^* - u_{n_0}\|_V \leq \frac{\gamma}{2\mu V}.$$

Hence,

$$\eta - \mu w_N \geq \eta - \mu v_N - \frac{\gamma}{2\mu V}\mu V = \gamma - \frac{\gamma}{2} = \frac{\gamma}{2} > 0,$$

which contradicts $u_{n_0} \in \check{\mathcal{V}}_\eta$.

Therefore, $\mu v_n \geq \eta$ for all $n \in \mathbb{N}$. This implies $u^* \in \check{\mathcal{V}}_\eta$, and the completeness of $\check{\mathcal{V}}_\eta$. □

LEMMA 4.7. *For $\eta \in \check{H}$, the operator LT^η has in $\check{\mathcal{V}}_\eta$ a unique fixed point.*

Proof. $\check{\mathcal{V}}_\eta \neq \emptyset$, $\check{\mathcal{V}}_\eta$ is complete (see Lemma 4.6), and $LT^\eta \check{\mathcal{V}}_\eta \subseteq \check{\mathcal{V}}_\eta$. Furthermore, LT^η is contractive in $\check{\mathcal{V}}_\eta$ (see Lemma 4.2). Then the statement follows by Banach's fixed point theorem. □

By means of the following two lemmas we will construct suitable $\eta^* \in \mathbb{R}$ and $v^* \in \check{\mathcal{V}}$, and finally we will show that (η^*, v^*) is a solution of the optimality equation. We denote the unique fixed point of LT^η in $\check{\mathcal{V}}_\eta$ by v^*_η.

LEMMA 4.8. *Let $\eta \in \check{H}$. Then it holds for all $\eta' \in \check{H}$ with $\eta \leq \eta'$*

(a) $v^*_\eta \geq v^*_{\eta'}$,

(b) $\eta' \leq \mu v^*_\eta$.

Proof.

(a) Let $\eta, \eta' \in \check{H}$. For $u \geq u'$ with $\mu u' \geq \eta'$ we have $\mu u \geq \mu u' \geq \eta' \geq \eta$. Then we get by Lemma 4.2

$$v := LT^\eta u \geq LT^\eta u' \geq LT^{\eta'} u' =: v'.$$

Hence, if $\mu v' \geq \eta'$, then $\mu v \geq \eta$. Therefore, if $u_0 \in \check{\mathcal{V}}_{\eta'}$, then $u_0 \in \check{\mathcal{V}}_\eta$. For $u_0 \in \check{\mathcal{V}}_{\eta'}$ we then have

$$v^*_{\eta'} = \lim_{n\to\infty} [LT^{\eta'}]^n u_0 \leq \lim_{n\to\infty} [LT^\eta]^n u_0 = v^*_\eta.$$

(b) For an arbitrary $\eta \in \check{H}$ we have for all $\eta' \in \check{H}$ with $\eta \leq \eta'$

$$\eta' \leq \mu v^*_{\eta'} \leq \mu v^*_\eta.$$

Therefore, μv^*_η is an upper bound for $\check{H}$. □

LEMMA 4.9. *For $\eta^* := \sup \check{H}$, $v^* := \inf_{\eta \in \check{H}} v^*_\eta$, it holds that $\mu v^* = \eta^*$. Furthermore, $\lim_{\eta \nearrow \eta^*} \|v^*_\eta - v^*\|_V = 0$ and $v^* \in \tilde{\mathcal{V}}$.*

Proof. From Lemma 4.8(b) it follows that $\mu v^* \geq \eta^*$ by the monotone convergence theorem. Let $\varepsilon := \mu v^* - \eta^*$.

Assume $\varepsilon > 0$. Let $\gamma := \varepsilon \frac{1-\lambda}{4\mu V} > 0$. It holds that $\gamma \leq \frac{\varepsilon}{4}$. Let $\eta := \eta^* - \gamma$, $\eta' := \eta^* + \gamma$. Then $\eta \in \check{H}$ and $\eta' = \mu v^* - \varepsilon + \gamma \leq \mu v^* \leq \mu v^*_\eta$. Let $u_n := [\check{L}T^{\eta'}]^n v^*_\eta$. If $\mu u_k \geq \eta'$ for all $k < n$, then it follows from Lemma 4.5 that

$$\|u_n - v^*_\eta\|_V \leq \frac{1}{1-\lambda}|\eta - \eta'| = \frac{2}{1-\lambda}\gamma.$$

Hence,

$$u_n \geq v^*_\eta - \frac{2}{1-\lambda}\gamma V.$$

Then, it follows

$$\mu u_n \geq \mu v^*_\eta - \frac{2}{1-\lambda}\gamma\mu V = \mu v^*_\eta - \frac{\varepsilon}{2} \geq \mu v^* - \frac{\varepsilon}{2} = \eta^* + \varepsilon - \frac{\varepsilon}{2} \geq \eta^* + \gamma = \eta'.$$

By mathematical induction we get $\mu u_n \geq \eta'$ for all $n \in \mathbb{N}$. That means $\eta' \in \check{H}$, which contradicts $\eta^* := \sup \check{H}$. Hence, $\mu v^* = \eta^*$.

For $\eta_n \nearrow \eta^*$ we have $v^*_{\eta_n} \searrow v^*$ (see Lemma 4.8). From Lemma 4.5 it follows that $(v^*_{\eta_n})$ is a Cauchy sequence. Then $v^* \in \tilde{\mathcal{V}}$, since $\tilde{\mathcal{V}}$ is complete. □

Now we are ready to prove Theorem 4.1. Especially, we will show that (η^*, v^*) is a solution of the optimality equation. We remark that we will not use $\eta^* \in \check{H}$ and $v^* = v^*_{\eta^*}$ in this proof. We will, however, derive these facts later (see Corollary 4.10).

Proof of Theorem 4.1. Let v^*, η^* as in Lemma 4.9 and (η_n) a sequence with $\eta_n \in \check{H}$ and $\lim_{n\to\infty} \eta_n = \eta^*$. Then it follows from Lemma 4.9 $\lim_{n\to\infty} \|v^*_{\eta_n} - v^*\|_V = 0$. We get for all $f \in \mathbf{F}$, $g \in \mathbf{G}$, by Lemma 3.3,

$$\begin{aligned} r_{fg} - \eta^* + p_{fg}v^* &= r_{fg} - (1-\zeta_{fg})\eta^* + p_{fg}v^* - \zeta_{fg}\mu v^* \\ &= r_{fg} - (1-\zeta_{fg})\eta^* + P_{fg}v^* \\ &= r_{fg} - (1-\zeta_{fg})\eta_n + P_{fg}v^*_{\eta_n} + (1-\zeta_{fg})(\eta_n - \eta^*) + P_{fg}(v^* - v^*_{\eta_n}) \\ &\leq r_{fg} - (1-\zeta_{fg})\eta_n + P_{fg}v^*_{\eta_n} + |\eta_n - \eta^*| + \lambda\|v^*_{\eta_n} - v^*\|_V V \\ &= T^{\eta_n}_{fg} v^*_{\eta_n} + |\eta_n - \eta^*| + \lambda\|v^*_{\eta_n} - v^*\|_V V, \end{aligned}$$

and analogously

$$r_{fg} - \eta^* + p_{fg}v^* \geq T^{\eta_n}_{fg} v^*_{\eta_n} - |\eta_n - \eta^*| - \lambda\|v^*_{\eta_n} - v^*\|_V V.$$

Therefore,

$$\begin{aligned} & v^*_{\eta_n}(x) - |\eta_n - \eta^*| - \lambda\|v^*_{\eta_n} - v^*\|_V V(x) \\ &= LT^{\eta_n} v^*_{\eta_n}(x) - |\eta_n - \eta^*| - \lambda\|v^*_{\eta_n} - v^*\|_V V(x) \\ &\leq \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} \left\{ r_{fg}(x) + \int_X v^*(\xi) p_{fg}(d\xi|x) \right\} - \eta^* \\ &\leq LT^{\eta_n} v^*_{\eta_n}(x) + |\eta_n - \eta^*| + \lambda\|v^*_{\eta_n} - v^*\|_V V(x) \\ &= v^*_{\eta_n}(x) + |\eta_n - \eta^*| + \lambda\|v^*_{\eta_n} - v^*\|_V V(x) \end{aligned}$$

for all $x \in X$. For $n \to \infty$, it follows that

$$v^*(x) = \sup_{f\in\mathbf{F}} \inf_{g\in\mathbf{G}} \left\{ r_{fg}(x) + \int_X v^*(\xi) p_{fg}(d\xi|x) \right\} - \eta^*$$

for all $x \in X$. Hence, (η^*, v^*) is a solution of the optimality equation.

For every solution $(h^*, v^*) \in \mathbb{R} \times \tilde{\mathcal{V}}$ of the optimality equation (4.1) we obtain the existence of $g^* \in \mathbf{G}$ and $f_\varepsilon \in \mathbf{F}$ with (4.2) by Lemma 3.4. The proof of (d) uses standard arguments (see, for instance, [5]). The proof of (e) is also similar to the proof in [5]. Let $(\eta^*, \tilde{v})) \in \mathbb{R} \times \tilde{\mathcal{V}}$ an arbitrary solution of (4.1). We set $\tilde{u}(x) := \tilde{v}(x) + \eta^* - \int_X \tilde{v}(\xi)\mu(d\xi)$ for all $x \in X$. Then $(\eta^*, \tilde{u}) \in \mathbb{R} \times \tilde{\mathcal{V}}$ and $(\eta^*, \tilde{u})$ is

also a solution of (4.1). It follows that

$$\begin{aligned}\tilde{u}(x) &= \sup_{f\in\mathbf{F}}\inf_{g\in\mathbf{G}}\left\{r_{fg}(x)-(1-\zeta_{fg}(x))\eta^*+\int_X \tilde{u}(\xi)p_{fg}(d\xi|x)-\zeta_{fg}(x)\eta^*\right\}\\ &= \sup_{f\in\mathbf{F}}\inf_{g\in\mathbf{G}}\left\{r_{fg}(x)-(1-\zeta_{fg}(x))\eta^*+\int_X \tilde{u}(\xi)P_{fg}(d\xi|x)\right.\\ &\qquad \left.+\zeta_{fg}(x)\int_X \tilde{u}(\xi)\mu(d\xi)-\zeta_{fg}(x)\eta^*\right\}\\ &= \sup_{f\in\mathbf{F}}\inf_{g\in\mathbf{G}}\left\{r_{fg}(x)-(1-\zeta_{fg}(x))\eta^*+\int_X \tilde{u}(\xi)P_{fg}(d\xi|x)\right\}\end{aligned}$$

for all $x\in X$. Hence, $\tilde{u}$ is a solution of the fixed point equation

$$u = LT^{\eta^*}u.$$

Since this fixed point equation has only one solution (see Lemma 4.2(b)), we obtain the statement.

Finally, we present some properties of the set $\check{H}$ and the fixed points of the operators LT^η for $\eta\in\check{H}$.

COROLLARY 4.10. *Let v^*, η^* as in Lemma 4.9.*

(a) *v^* is the unique solution of the fixed point equation $v = LT^{\eta^*}v$. Hence, $v^* = v^*_{\eta^*}$.*

(b) *$\check{H} = (-\infty, \eta^*]$*

(c) *$(v^*_\eta)_{\eta\in\check{H}}$ is continuous (left-continuous at η^*) in the V-norm $\|\cdot\|_V$.*

(d) *For $\eta\in\check{H}$ it holds that $\mu v^*_\eta = \eta$ if and only if $\eta = \eta^*$.*

Proof.

(a) If we choose $\tilde{v} = v^*$ in the proof of Theorem 4.1, we then get $\tilde{u} = v^*$.

(b) $\eta^*\in\check{H}$ follows from (a). $\eta\le\eta^*$ implies $v^* = LT^{\eta^*}v^* \le LT^\eta v^*$. By mathematical induction we then get $[\check{L}T^\eta v^*]^n = [LT^\eta v^*]^n \ge v^* > -\infty$ for all $n\in\mathbb{N}$. Therefore, $v^*\in\check{\mathcal{V}}_\eta$ and $\eta\in\check{H}$.

(c) See Lemma 4.5.

(d) See Lemmas 4.8 and 4.9. □

5. Conclusions. This paper deals with two-person zero-sum stochastic games with the average reward criterion and Borel state and action spaces under a geometric drift condition. We assume that the law of motion p is weakly continuous, the reward function r is l.s.c., the set-valued mapping A of admissible actions of the maximizing player is l.s.c (and complete-valued), and the set-valued mapping B of admissible actions of the minimizing player is u.s.c. (and compact-valued). It is shown that the optimality equation has a unique solution, the game has a value, the minimizing player has an optimal stationary strategy, and the maximizing player has an ε-optimal stationary strategy for every $\varepsilon > 0$.

In contrast to this, the assumptions in the paper of Jaśkiewicz and Nowak [3] are more restrictive. It is required that r, A, and B are continuous and that the probability measure μ and the function δ in the geometric drift condition satisfy the additional property $\int_X \inf_{a\in A(x)}\inf_{b\in B(x)}\delta(x,a,b)\mu(dx) > 0$. Under these conditions it is proved that both players have optimal stationary strategies. But it is only shown that the optimality equation holds almost everywhere in respect of a maximal irreducibility measure (whose existence must be assumed).

We remark that it is not difficult to derive the existence of an optimal stationary strategy of the maximizing player from our results if r, A, and B are continuous. It is sufficient to consider a Markov game with $-r$ instead of r.

Acknowledgments. The author wishes to thank an associate editor and the anonymous referees for many helpful comments. One of the referees provided a very detailed review and proposed various modifications to improve the paper. He also suggested the inclusion of Corollary 4.10.

REFERENCES

[1] O. Hernández-Lerma and J. B. Lasserre, *Zero-sum stochastic games in Borel spaces: Average payoff criteria*, SIAM J. Control Optim., 39 (2001), pp. 1520–1539.
[2] A. Jaśkiewicz and A. S. Nowak, *On the optimality equation for zero-sum ergodic stochastic games*, Math. Methods Oper. Res., 54 (2001), pp. 291–301.
[3] A. Jaśkiewicz and A. S. Nowak, *Zero-sum ergodic stochastic games with Feller transition probabilities*, SIAM J. Control Optim., 45 (2006), pp. 773–789.
[4] O. Kallenberg, *Foundations of Modern Probability*, Springer-Verlag, New York, 1997.
[5] H.-U. Küenle and R. Schurath, *The optimality equation and ε-optimal strategies in Markov games with average reward criterion*, Math. Methods Oper. Res., 56 (2002), pp. 451–471.
[6] H.-U. Küenle, *Stochastische Spiele und Entscheidungsmodelle*, Teubner-Verlag, Leipzig, 1986.
[7] H.-U. Küenle, *Markov Games Under a Geometric Drift Condition*, in Advances in Dynamic Games, Ann. Internat. Soc. Dynam. Games 7, Birkhäuser, Boston, 2005, pp. 21–38.
[8] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1994.
[9] A. S. Nowak, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466–476.
[10] A. S. Nowak, *Semicontinuous nonstationary stochastic games*, J. Math. Anal. Appl., 117 (1986), pp. 84–99.
[11] A. S. Nowak, *Optimal strategies in a class of zero-sum ergodic games*, Math. Methods Oper. Res., 50 (1999), pp. 399-420.
[12] U. Rieder, *On Semicontinuous Dynamic Games*, Technical report, University of Karlsruhe, Karlsruhe, Germany, 1978.
[13] O. Vega-Amaya, *Zero-sum average semi-Markov games: Fixed point solutions of the Shapley equation*, SIAM J. Control Optim., 42 (2003), pp. 1876–1894.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2169–2206

CONVERGENT NUMERICAL SCHEME FOR SINGULAR STOCHASTIC CONTROL WITH STATE CONSTRAINTS IN A PORTFOLIO SELECTION PROBLEM*

AMARJIT BUDHIRAJA† AND KEVIN ROSS†

Abstract. We consider a singular stochastic control problem with state constraints that arises in problems of optimal consumption and investment under transaction costs. Numerical approximations for the value function using the Markov chain approximation method of Kushner and Dupuis are studied. The main result of the paper shows that the value function of the Markov decision problem (MDP) corresponding to the approximating controlled Markov chain converges to that of the original stochastic control problem as various parameters in the approximation approach suitable limits. All our convergence arguments are probabilistic; the main assumption that we make is that the value function be finite and continuous. In particular, uniqueness of the solutions of the associated HJB equations is neither needed nor available (in the generality under which the problem is considered). Specific features of the problem that make the convergence analysis nontrivial include unboundedness of the state and control space and the cost function; degeneracies in the dynamics; mixed boundary (Dirichlet–Neumann) conditions; and presence of both singular and absolutely continuous controls in the dynamics. Finally, schemes for computing the value function and optimal control policies for the MDP are presented and illustrated with a numerical study.

Key words. singular control, Hamilton–Jacobi–Bellman equations, portfolio selection, stochastic control, free boundary problem, Skorohod problem

AMS subject classifications. 60K25, 68M20, 90B22, 90B35, 60J70

DOI. 10.1137/050640515

1. Introduction.

Singular control is an important and challenging class of problems in stochastic control theory. Roughly speaking, by singular control we mean that the control terms in the dynamics of the state process need not be absolutely continuous with respect to the Lebesgue measure and are only required to have paths of bounded variation. The Hamilton–Jacobi–Bellman (HJB) equations for such problems, which are variational inequalities with gradient constraints, are typically quite hard to work with. Despite the fact that over the past 20 years there has been a significant development in the theory of weak and viscosity solutions of HJB equations for such diffusion control problems (cf. [5, 26, 4, 15, 22, 29]), the existence/uniqueness and regularity theory for this class of PDEs is not well understood. In view of the various applications in mathematical finance (cf. [11]) and stochastic networks (cf. [12]) that lead to singular control, it is particularly important to develop methods for numerical approximations for such control problems.

Over the last 30 years, Kushner, Dupuis, and coworkers (cf. [19] and references therein) have developed a powerful machinery, the so-called Markov chain approximation method, for a wide spectrum of computational problems in stochastic control theory. This probabilistic approach has two main advantages. First, approximation with a Markov chain allows one to use physical insights derived from the dynamics

*Received by the editors September 17, 2005; accepted for publication (in revised form) August 20, 2006; published electronically January 22, 2007. This research was supported in part by ARO grant W911NF-04-1-0230. This work was undertaken while the second author was at the University of North Carolina at Chapel Hill.

http://www.siam.org/journals/sicon/45-6/64051.html

†Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, 210 Smith Building, CB# 3260, Chapel Hill, NC 25799-3260 (budhiraj@email.unc.edu, kjross@email.unc.edu).

of the controlled diffusion in obtaining a stable approximation scheme. Second, the Markov chain approximation method does not require the smoothness of the cost or value function, nor does it rely on the uniqueness properties of the associated HJB equations. This is a particularly significant advantage in problems where the PDE theory for the associated HJB equations is hard to tackle.

In this work we study a problem of optimal consumption and portfolio selection with proportional transaction costs that has been studied by several authors [24, 6, 29, 23, 31]. The basic problem can be described as follows. Consider a single investor who has two instruments available for investment: a risk-free asset such as a bank account, which pays a fixed interest rate $r > 0$, and a risky asset, such as a stock, whose price evolution is modeled via a geometric Brownian motion with a mean value of return $b > r$ and constant volatility $\sigma > 0$. We assume that the investor may buy or sell stock continuously over time in not necessarily integer valued quantities. The investor is assumed to consume wealth at some time-dependent rate $C(t)$, and without loss of generality we assume that the consumption is deducted from the bank account. The investor may instantaneously transfer money from the bank account to stock, and vice versa, by paying a proportional transaction cost; namely, there are $\lambda \in (0, \infty)$ and $\mu \in (0, 1)$ such that the investor pays λ times the amount moved from the bank account to stock as a transaction fee, and similarly, he pays μ times the amount moved from stock to the bank account as a transaction fee. All transaction fees are charged from the bank account. The basic constraint on the consumption control C and the portfolio selection control, denoted (M, N), is that the investor must be solvent at all times. More precisely, if $X(t)$ and $Y(t)$ represent the amount of investment in the bank account and the stock, respectively, at time t, then we require $(X(t), Y(t)) \in \mathbb{S}$ for all $t \geq 0$, where

$$\mathbb{S} \doteq \{(x, y) \in \mathbb{R}^2 : x + (1 + \lambda)y \geq 0 \text{ and } x + (1 - \mu)y \geq 0\}.$$

These solvency constraints ensure that at all times the investor has sufficient wealth to settle any obligations due to selling stock short (the first inequality) or borrowing from the bank account (the second inequality). The goal of the investor is to maximize the expected total discounted utility of consumption, $\mathbb{E}\int_0^\infty e^{-\beta t} f(C(t))dt$, where $\beta \in (0, \infty)$ is the discount factor and the utility function $f : [0, \infty) \to [0, \infty)$ is a continuous function satisfying $f(0) = 0$. The condition $f(0) = 0$ can be relaxed if f is nondecreasing and $f(0) > -\infty$ by replacing f with $f - f(0)$.

In the absence of transaction costs, Merton proved in the classical paper [25] that when the utility function is $f(c) = c^p/p, p < 1, p \neq 0$ or $f(c) = \log c$ (note that the latter utility function does not satisfy the conditions of the current paper) the investor's optimal policy is to keep a constant proportion of total wealth in the risky asset and to consume at a rate proportional to total wealth. (For a simple and self-contained treatment, see [6]). This "Merton line" target can always be achieved since transactions can be made continuously and instantaneously without affecting wealth. However, when transaction costs apply, such a policy results in immediate bankruptcy. Magill and Constantinides first conjectured in [24] that there must exist a "no-transaction region" taking the form of a wedge in the wealth space. When wealth is inside this region, consumption is the only control that can be exercised. Purchase or sale of stock occurs only when the wealth attempts to exit the no-transaction region. The formal arguments of [24] were put on a rigorous footing by Davis and Norman in [6] for the cases $f(c) = c^p/p$ and $f(c) = \log c$. In their work, under suitable conditions on model parameters, the free boundary problem associated with optimal

consumption in the presence of proportional transaction costs is solved explicitly, and C^2 regularity of the value function is established. The authors show that the (optimal) no-transaction region is a wedge; in particular, the optimal policy is to exercise the minimal amount of trading necessary to keep wealth inside the no-transaction region. Inside the region, consumption occurs at a finite rate. In [29] Shreve and Soner consider the same problem as in [6] but with conditions on the model parameters that are weaker and much more explicit. Once more, regularity properties of the value function and the associated free boundary are proved. A more general utility function, which satisfies suitable smoothness, concavity and growth properties, was considered in [31]. Using viscosity solution methods, the authors sketch a proof for unique solvability of the associated HJB equation by the value function. A finite difference approximation scheme for approximating the value function is introduced; however, convergence of the proposed scheme for the portfolio selection problem is not proved. The authors do provide results from several numerical studies which identify near optimal control policies and the (numerical) free boundary.

In the current work, we do not impose any concavity, smoothness, or growth conditions on the utility function; the key condition (Condition 2.1) that we require is that the value function is finite and continuous. In particular, we do not claim or use the fact that the value function is the unique solution of the associated HJB equation. The main goal of the study is to obtain convergent numerical approximations for the value function. The basic approach, as in [19], is to introduce a Markov decision problem (MDP) for an approximating, finite state, discrete time, controlled Markov chain. The main result of the paper (Theorem 5.12) shows that the value function of the MDP converges to the value function of the original singular control problem as various parameters in the approximation approach their limits suitably. In section 6 we use the approximating MDP to obtain computational schemes for obtaining near optimal control policies. The key result of this section is Lemma 6.1, which allows us to characterize the value function and optimal control policies via the solution of suitable dynamic programming equations (see Theorem 6.2). Finally, in section 7 results from a numerical study using the algorithm of section 6 are described. In particular, Figure 2 shows the numerical no-transaction region and the associated (numerical) free boundary obtained by an implementation of the algorithm.

The only paper (to the best of our knowledge) that carries out a complete convergence analysis for a numerical scheme for a singular control problem is [20]. Although several ideas developed in [20] are crucial to the ideas in the current paper, there are key differences in the model that make our analysis substantially delicate. First, the above paper considers a queuing problem with "finite buffers" which essentially means that the state space and control space are bounded. In the current study we first have to suitably approximate the original unbounded model by one in which the consumption control and the state space are bounded. This two stage approximation procedure is carried out in Propositions 2.2 and 2.4. This is the only place where the assumption on the continuity of the value function (Condition 2.1) is used. Next, in contrast to [20], in addition to singular control terms, we also have an absolutely continuous control term (consumption control) that appears in a nonlinear fashion in the cost (reward) criterion through the utility function f. This requires us to introduce the relaxed formulation for the stochastic control problem in order to carry out the convergence analysis. Lemma 5.1 ensures that the relaxed formulation does not change the value function of the control problem. The next substantial difficulty in our analysis is the state constraint feature of the dynamics. Although in [20] also the state is constrained to be in a bounded polyhedral region, that can be easily handled

by introducing the so-called "Skorohod map." In the current model the directions of control, which do not point inward into the state space (see Figure 1), do not allow for a similar reduction. However, one useful feature of the dynamics (see (1)) is that once the state of the system reaches the boundary of $\mathbb{S}$, the only admissible control corresponds to moving the state process instantaneously to the origin and keeping it there at all times. This observation allows us to convert an infinite horizon cost to an exit time criterion (see (2)–(4)). This reformulation makes some aspects of the convergence analysis simpler; however, the degeneracies in the state dynamics make the treatment of convergence properties of exit times quite subtle. To see the basic difficulty, consider the following simple example. Let ξ_n be a sequence of positive reals such that $\xi_n \to 0$ as $n \to \infty$. Let x_n be the solution of the ODE $\dot{x} = x$ with initial condition ξ_n and x be the solution of the same ODE with 0 initial condition. Clearly $x_n \to x$ uniformly on compacts; however, if $\tau_n \doteq \inf\{t|x_n(t) = 0\}$ and $\tau \doteq \inf\{t|x(t) = 0\}$, then clearly $\tau_n \not\to \tau$. In other words, convergence of processes in general need not imply the convergence of the corresponding exit times. The issue is especially problematic when, as is the case for the controlled dynamics considered in this paper, the diffusion coefficients in the state dynamics are not uniformly nondegenerate. This is another key difference between the current model and the problem studied in [20].

One of the major obstacles in proving the convergence of the value function of a sequence of approximating discrete MDPs to the value of the original singular control problem is proving the tightness of the sequence of singular control terms in the Skorohod $D[0, \infty)$ space. A powerful technique for bypassing this tightness issue, based on suitable stretching of a time scale, was introduced in [20]. Although such time transformation ideas go back to the work of Meyer and Zheng [27] (see also Kurtz [18]), the papers [20, 21] were the first to use such ideas in stochastic control problems. A similar technique was also recently used in [3]. A key ingredient to this technique is the uniform moment estimate obtained in Lemma 4.4. In [20] such a moment estimate follows easily from the form of the cost function where a strictly positive proportional cost is incurred for exercising the singular control. In the current problem there is no direct contribution to the (cost) reward function from the singular control term and, as a result, the proof of this uniform estimate becomes more involved. Roughly speaking, the main idea of the proof is that a controller cannot make too much use of a singular control without pushing the process to the boundary of the domain.

The paper is organized as follows. In section 2 we give a precise formulation of the control problem of interest. We also present here two propositions (Propositions 2.2 and 2.4) which allow approximation of the original control problem by one with a bounded state space and bounded consumption actions. Section 3 introduces the discrete MDP that approximates the original singular control problem, and section 4 defines the continuous time interpolations and the time transformation that are key to the convergence analysis. In section 5 we present the main convergence result that establishes the convergence of the value function of the MDP to that of the original singular control problem. Section 6 is devoted to computational methods for the MDP. A key result here is Lemma 6.1 which allows, via Theorems 6.2 and 6.3, iterative methods for computation of the value function and optimal control policies for the MDP. In problems with only absolutely continuous controls, estimates of the form in Lemma 6.1 are straightforward consequences of a contraction property that follows from the strictly positive discount factor in the cost (cf. Chapter 6 of [19]). However, for singular control problems, due to the instantaneous nature of the control, such contraction estimates are typically unavailable. Here, once again, we

use the special feature of the dynamics, which says that excessive use of the singular control will rapidly bring the process to the boundary, in obtaining such an estimate. In section 7 we present results from a numerical study of the algorithm. Finally, the appendix contains proofs of the more technical results.

The following notation will be used in the paper. Given a Polish space E, $D_E \equiv D([0,\infty) : E)$ will denote the space of paths that are right-continuous with left limits (RCLL) on $[0,\infty)$ taking values in E, endowed with the usual Skorohod topology. For an RCLL path $\{\xi(t)\}$, the jump at $t > 0$ will be denoted by $\delta\xi(t)$. As a convention we take $\delta\xi(0) \doteq \xi(0)$. For a sequence of random variables $\{\xi_n\}_{n\geq 0}$, we will use the notation $\delta\xi_n$ for the increment $\xi_{n+1} - \xi_n$. For a point $x \in \mathbb{R}^k$ and a set $G \in \mathbb{R}^k$, $\mathrm{dist}(x, G)$ will denote the distance of x from $\overline{G}$. The Borel sigma field for a metric space E will be denoted by $\mathcal{B}(E)$.

2. Optimal consumption and portfolio selection with transaction costs. We begin with a precise mathematical formulation of the optimal consumption and investment problem described in the previous section. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which is given a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ satisfying the usual hypothesis. Let W be a real-valued $\{\mathcal{F}_t\}$-Brownian motion. We will denote the probability system $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}, W)$ by Φ. The Wiener process represents the source of uncertainty of the risky asset. The state process, which represents the wealth of the investor, is a controlled Markov process $Z \equiv (X, Y)$ given on the above probability system via the equations

$$\begin{aligned} dX(t) &= (rX(t) - C(t))dt - (1+\lambda)dM(t) + (1-\mu)dN(t), \\ dY(t) &= bY(t)dt + \sigma Y(t)dW(t) + dM(t) - dN(t), \end{aligned} \tag{1}$$

with initial condition $X(0-) = x, Y(0-) = y$, where $z \doteq (x, y) \in \mathbb{S}$. Here C is an $\{\mathcal{F}_t\}$-progressively measurable process such that for all $t \in [0,\infty)$, $C(t) \geq 0$ a.s. and $\mathbb{E}\int_0^t e^{-rs}C(s)ds < \infty$. Also, M and N are $\{\mathcal{F}_t\}$-adapted, nondecreasing, RCLL processes satisfying $M(0) \geq 0$ and $N(0) \geq 0$ a.s. The processes X and Y represent the amounts invested in the bond and the stock, respectively; $M(t), N(t)$ denote the cumulative purchases and sales of stock, respectively, over $[0, t]$. The process C represents the consumption of the investor. The processes C, M, and N are the control processes. Since M and N are not required to be absolutely continuous (with respect to the Lebesgue measure), they are referred to as singular controls. Denote by $\mathcal{A}(\Phi, z) \equiv \mathcal{A}(z)$ the set of "admissible controls," i.e., all $U \equiv (C, M, N)$ of the form described above. Let $\partial\mathbb{S}$ denote the boundary of $\mathbb{S}$. From the dynamical description of Z it follows that if $z \in \partial\mathbb{S}$, then the only control that keeps the investor solvent takes Z to the origin instantly and keeps it there at all times (see Figure 1).

Recall the utility function f in the introduction. Since $f(0) = 0$, one can reformulate the state constraint control problem on an infinite time horizon described in the introduction to an exit time control problem, as follows. For $z \in \mathbb{S}$ and $U \in \mathcal{A}(z)$, let $\tau \equiv \tau(z, U)$ be defined as

$$\tau \doteq \inf\{t \in [0,\infty) : Z(t) \notin \mathbb{S}^o\}, \tag{2}$$

where Z is the controlled process corresponding to initial condition z and control U. Define the cost, $J(z, U)$, for using the control U by

$$J(z, U) \doteq \mathbb{E}\int_{[0,\tau)} e^{-\beta t} f(C(t))dt. \tag{3}$$

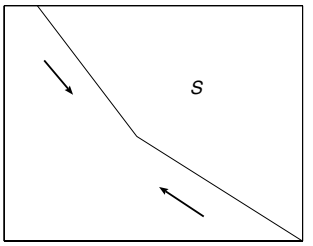


FIG. 1. *State space and singular control directions.*

The value function of the control problem is then given by

$$V(z) = \sup_{\Phi} \sup_{U \in \mathcal{A}(z)} J(z, U), \tag{4}$$

where the outside supremum is over all probability systems Φ. The following will be a standing assumption in this work.

CONDITION 2.1. *For all $z \in \mathbb{S}$, $V(z) < \infty$ and $V : \mathbb{S} \to \mathbb{R}_+$ is a continuous map.*

We refer the reader to [16, 29, 31] for some sufficient conditions needed for the above assumption to hold.

State and control space truncation. In order to develop numerical methods for computing $V(z)$, we will need to first approximate the control problem by an analogous control problem with a bounded state space and control set. We now present the convergence result, which says that the value function of the "truncated control problem" converges to V as the truncation parameters approach their limits. We begin by considering the control space truncation.

For $p \in (0, \infty)$, let $\mathcal{A}_p(\Phi, z) \equiv \mathcal{A}_p(z)$ be the subset of $\mathcal{A}(z)$ consisting of $U = (C, M, N)$, which satisfy $0 \le C(t) \le p$, for all $t \ge 0$ a.s. Define $V_p(z)$ by replacing $\mathcal{A}(z)$ with $\mathcal{A}_p(z)$ in (4). The following is the first convergence result.

PROPOSITION 2.2. *V_p converges to V, uniformly on compact subsets of $\mathbb{S}$, as $p \to \infty$.*

Proof. We first establish pointwise convergence, i.e., $V_p(z) \to V(z)$ as $p \to \infty$. Since $V_p(z) \le V(z)$, it suffices to show that, for all $z \in \mathbb{S}$,

$$\liminf_{p \to \infty} V_p(z) \ge V(z).$$

Fix $\epsilon > 0$ and choose an "ϵ-optimal control," i.e., $U_\epsilon \in \mathcal{A}(z)$ such that $V(z) - \epsilon < J(z, U_\epsilon)$. Suppose τ_ϵ is the associated exit time from $\mathbb{S}^o$. Define a control $\tilde{U}_p \equiv (\tilde{C}_p, \tilde{M}_p, \tilde{N}_p)$ by $\tilde{C}_p(t) \doteq C_\epsilon(t) \wedge p$, $\tilde{M}_p(t) \doteq M_\epsilon(t)$, $\tilde{N}_p(t) \doteq N_\epsilon(t)$, $t \ge 0$. It follows from the fact that $\tilde{C}_p \le C_\epsilon$ and standard comparison results for solutions of stochastic differential equations (cf. Proposition 5.2.18 of [17]) that the wealth process under control $\tilde{U}_p$ is never less than the wealth process under control U_ϵ. In particular, denoting by τ_p the exit time from $\mathbb{S}^o$ by the controlled process corresponding to the control $\tilde{U}_p$, we have $\tau_p \ge \tau_\epsilon$. Combining this with the observations that $\tilde{C}_p(t) \uparrow C_\epsilon(t)$ as $p \to \infty$ a.s. for all $t \ge 0$ and f is continuous, we have from Fatou's lemma

$$\liminf_{p \to \infty} J(z, \tilde{U}_p) \ge \liminf_{p \to \infty} \mathbb{E} \int_{[0, \tau_\epsilon)} e^{-\beta t} f(\tilde{C}_p(t)) dt \ge \mathbb{E} \int_{[0, \tau_\epsilon)} e^{-\beta t} f(C_\epsilon(t)) dt \ge V(z) - \epsilon.$$

Since $\epsilon > 0$ is arbitrary, the pointwise convergence of V_p to V follows. Next we show that for each p, V_p is lower semicontinuous (l.s.c.). Fix $z \in \mathbb{S}$ and let $\mathbb{S} \ni z_n \to z$ as $n \to \infty$. To prove that V_p is l.s.c. it suffices to show that

$$\liminf_{n\to\infty} V_p(z_n) \geq V_p(z). \tag{5}$$

Fix $\epsilon > 0$ and let $U_\epsilon = (C_\epsilon, M_\epsilon, N_\epsilon) \in \mathcal{A}_p(z)$ be an ϵ-optimal control, i.e., $V_p(z) - \epsilon < J(z, U_\epsilon)$. Let Z_ϵ be the controlled process according to U_ϵ and define τ_ϵ via (2) with Z replaced by Z_ϵ. Define $U_n \equiv (C_n, M_n, N_n)$ as $C_n \doteq C_\epsilon$, $M_n(t) \doteq M_\epsilon(t)1_{t<\tau_\epsilon} + M_n^* 1_{t\geq\tau_\epsilon}$, $N_n(t) \doteq N_\epsilon(t)1_{t<\tau_\epsilon} + N_n^* 1_{t\geq\tau_\epsilon}$, where $M_n^*, N_n^* \geq 0$ are chosen so that the controlled process Z_n corresponding to U_n and initial condition z_n satisfies $Z_n(\tau_\epsilon) \notin \mathbb{S}^o$. (Note that, clearly, $U_n \in \mathcal{A}_p(z_n)$.) This ensures that $\tau_n \doteq \inf\{t : Z_n(t) \notin \mathbb{S}^o\}$ is at most τ_ϵ. Note that on the set $\{\tau_\epsilon = \infty\}$, we have $U_n(t) = U_\epsilon(t)$ for all $t \geq 0$. We claim that on the set $\{\tau_\epsilon < \infty\}$ we have $\liminf_{n\to\infty} \tau_n \geq \tau_\epsilon$ a.s., which implies $\tau_n \to \tau_\epsilon$ a.s. as $n \to \infty$ on the set $\{\tau_\epsilon < \infty\}$. To see the claim, suppose that $\liminf \tau_n < \tau_\epsilon - \delta$ for some $\delta > 0$. Then there exists $N_0 \geq 1$ such that $\tau_n < \tau_\epsilon - \delta/2$ for all $n \geq N_0$. Also, from the choice of the control U_n we see that, for all $\delta > 0$ and $L \in (0, \infty)$, $\sup_{0\leq t\leq(\tau_\epsilon-\delta/2)\wedge L} |Z_n(t) - Z(t)| \to 0$ in probability as $n \to \infty$. Combining this with the fact that $Z_n(\tau_n) \notin \mathbb{S}^o$, we have that $Z(t) \notin \mathbb{S}^o$ for some $t \leq \tau_\epsilon - \delta/2$. However, this contradicts the definition of τ_ϵ. Thus we have shown $\tau_n \to \tau_\epsilon$ a.s. on the set $\{\tau_\epsilon < \infty\}$.

Next, recalling the choice of U_ϵ and that $C_\epsilon(t) = C_n(t)$ for all $t \geq 0$ on the set $\{\tau_\epsilon = \infty\}$, we have

$$\begin{aligned} V_p(z) - V_p(z_n) &\leq J(z, U_\epsilon) - J(z_n, U_n) + \epsilon \\ &= \mathbb{E}\Big[1_{\{\tau_\epsilon<\infty\}} \int_{[\tau_n,\tau_\epsilon)} e^{-\beta t}(f(C_\epsilon(t)) - f(C_n(t)))dt\Big] + \epsilon \\ &\leq f_*(p)\mathbb{E}\Big[1_{\{\tau_\epsilon<\infty\}} \int_{[\tau_n,\tau_\epsilon)} e^{-\beta t}dt\Big] + \epsilon, \end{aligned}$$

where $f_*(p) \doteq \sup_{0\leq c\leq p} f(c) < \infty$. Since $\tau_n \to \tau_\epsilon$ a.s. on the set $\{\tau_\epsilon < \infty\}$, the first term on the right in the last line above approaches 0 as $n \to \infty$. Inequality (5) now follows from the above display on taking $n \to \infty$ and then $\epsilon \to 0$. Finally, note that for each z, $V(z) - V_p(z) \downarrow 0$. The result now follows from Dini's theorem (cf. Theorem M8 in [1]). □

Next, we consider the truncation of the state space. The reduction will be achieved by replacing the original dynamical system given by (1) with one which evolves exactly as before in the interior of some compact domain but is instantaneously reflected back when the controlled process is about to exit the domain. The reflection mechanism is made precise via the notion of a Skorohod map. We begin with the following definition. Fix $\ell \in (0, \infty)$.

DEFINITION 2.3. *Let $\phi \in \mathcal{D} \doteq D([0, \infty) : \mathbb{R}^2)$ be such that $\phi(0) \in (-\infty, \ell] \times (-\infty, \ell]$. We will denote the space of all such ϕ by $\mathcal{D}_0$. We say a pair $(\psi, \eta) \in \mathcal{D} \times \mathcal{D}$ solves the Skorohod problem (SP) for ϕ in $(-\infty, \ell] \times (-\infty, \ell]$, with normal reflection, if the following hold:*

(i) $\psi(0) = \phi(0)$.

(ii) $\psi(t) = \phi(t) - \eta(t)$, $t \in (0, \infty)$.

(iii) $\psi(t) \in (-\infty, \ell] \times (-\infty, \ell]$ *for all* $t \geq 0$.

(iv) $\eta(\cdot)$ *is componentwise nondecreasing.*

(v) $\eta_i(t) = \int_{(0,t]} 1_{\{\psi_i(t)=\ell\}} d\eta_i(t)$, $i = 1, 2$, *where* $\eta(t) = (\eta_1(t), \eta_2(t))'$, $\psi(t) = (\psi_1(t), \psi_2(t))'$.

It is well known (cf. [7, 13]) that for every $\phi \in \mathcal{D}_0$, there is a unique solution (ψ, η) to the above SP. We will write $\psi = \Gamma(\phi)$ and refer to the map $\Gamma : \mathcal{D}_0 \to \mathcal{D}_0$ as the Skorohod map. The following Lipschitz property (cf. [7]) is quite useful in various estimates. There exists $\kappa \in (0, \infty)$, independent of ℓ, such that, for all $\phi_1, \phi_2 \in \mathcal{D}_0$,

$$|\Gamma(\phi_1) - \Gamma(\phi_2)|_T^* \le \kappa |\phi_1 - \phi_2|_T^*, \ T \in (0, \infty). \tag{6}$$

We will now introduce the modified constrained dynamics of the controlled Markov process. Set $\mathbb{S}_\ell \doteq \mathbb{S} \cap (-\infty, \ell] \times (-\infty, \ell]$. Let $Z_\ell \equiv (X_\ell, Y_\ell)$ solve the following system of equations:

$$\begin{aligned} dX_\ell(t) &= (rX_\ell(t) - C(t))dt - (1+\lambda)dM(t) + (1-\mu)dN(t) - dR_1(t), \\ dY_\ell(t) &= bY_\ell(t)dt + \sigma Y_\ell(t)dW(t) + dM(t) - dN(t) - dR_2(t), \end{aligned} \tag{7}$$

where $Z_\ell(0-) = z$, $U \equiv (C, M, N) \in \mathcal{A}_p(z)$, $z = (x, y) \in \mathbb{S}_\ell$, and $R = (R_1, R_2)'$ is a componentwise nondecreasing, RCLL, $\{\mathcal{F}_t\}$-adapted process satisfying

$$\int_0^\infty 1_{\{X_\ell(t)<\ell\}} dR_1(t) = 0, \quad \int_0^\infty 1_{\{Y_\ell(t)<\ell\}} dR_2(t) = 0. \tag{8}$$

The unique solvability of (7) and (8) follows from the Lipschitz continuity property (6) of the Skorohod map and the usual Picard iteration method. Define τ_ℓ and $J_\ell(z, U)$ as in (2) and (3) with Z replaced by Z_ℓ in (2) and τ replaced by τ_ℓ in (3). Define $V_{\ell,p}$ as

$$V_{\ell,p}(z) = \sup_{\Phi} \sup_{U \in \mathcal{A}_p(\Phi, z)} J_\ell(z, U). \tag{9}$$

The following is the second convergence result of this section.

PROPOSITION 2.4. *For all* $p \in (0, \infty)$, $V_{\ell,p}$ *converges to* V_p, *uniformly on compact subsets of* $\mathbb{S}$, *as* $\ell \to \infty$.

Proof. Let $Z \equiv (X, Y)$ be as in (1) and τ as in (2), with $C \equiv 0$. It is easy to check that for each $T \in (0, \infty)$ and compact subset $\mathbb{S}_0 \subset \mathbb{S}$, there exists $\Lambda \equiv \Lambda(T) \in (0, \infty)$ such that

$$\sup_{\Phi} \sup_{(M,N)} \sup_{z \in \mathbb{S}_0} \mathbb{E} \sup_{0 \le t \le T \wedge \tau} (X^+(t) + Y^+(t)) \le \Lambda,$$

where the supremum is taken over all $\{\mathcal{F}(t)\}$-adapted, nondecreasing, RCLL processes M and N such that $M(0) \ge 0$, $N(0) \ge 0$, and over all systems Φ. Thus in particular we have that

$$\sup_{\ell} \sup_{\Phi} \sup_{U \in \mathcal{A}_p(\Phi, z)} \sup_{z \in \mathbb{S}_0} \mathbb{E} \sup_{0 \le t \le T \wedge \tau_\ell} (X_\ell^+(t) + Y_\ell^+(t)) \le \Lambda, \tag{10}$$

where $Z_\ell \equiv (X_\ell, Y_\ell)$ are as defined in (7), and τ_ℓ is as introduced below (8).

Fix $\delta > 0$. Let $z \in \mathbb{S}_0$ and $\epsilon > 0$ be arbitrary. Let Φ and $U \in \mathcal{A}_p(z, \Phi)$ be such that $V_{\ell,p}(z) \le J_\ell(z, U) + \epsilon$. Choose $T \in (0, \infty)$ such that $f_*(p)e^{-\beta T}/T < \epsilon$. Then $V_{\ell,p}(z) \le \mathbb{E} \int_0^{T \wedge \tau_\ell} e^{-\beta t} f(C(t))dt + 2\epsilon$.

Choose $\ell_0 \equiv \ell_0(\delta)$ such that $\ell_0 > (\Lambda f_*(p))/(\delta\beta)$. Define

$$A_{\ell_0} \doteq \left\{ \omega : \sup_{0 \le t \le T \wedge \tau_\ell} (X_\ell^+(t) + Y_\ell^+(t)) > \ell_0 \right\}.$$

Then

$$\mathbb{E}\int_0^{T\wedge\tau_\ell} e^{-\beta t} f(C(t))dt$$
$$= \mathbb{E}\left[1_{A_{\ell_0}} \int_0^{T\wedge\tau_\ell} e^{-\beta t} f(C(t))dt\right] + \mathbb{E}\left[1_{A^c_{\ell_0}} \int_0^{T\wedge\tau_\ell} e^{-\beta t} f(C(t))dt\right]. \tag{11}$$

It follows from Markov's inequality and (10) that $\mathbb{P}[A_{\ell_0}] \le \Lambda/\ell_0$. Thus the first integral on the right side of (11) is bounded by $(f_*(p)/\beta)\mathbb{P}[A_{\ell_0}] \le \delta$. Next, for $\ell \ge \ell_0$, on the set $A^c_{\ell_0}$, $Z_\ell(\cdot \wedge T \wedge \tau_\ell) = Z(\cdot \wedge T \wedge \tau_\ell)$. In particular, $T \wedge \tau \ge T \wedge \tau_\ell$. Thus

$$\mathbb{E}\left[1_{A^c_{\ell_0}} \int_0^{T\wedge\tau_\ell} e^{-\beta t} f(C(t))dt\right] \le \mathbb{E}\left[\int_0^{T\wedge\tau} e^{-\beta t} f(C(t))dt\right] \le V_p(z).$$

Combining the above bounds, we have $V_{\ell,p}(z) \le V_p(z)+\delta+2\epsilon$. Since $\epsilon > 0$ is arbitrary, we have that, for all $\ell \ge \ell_0$ and $z \in \mathbb{S}_0$, $V_{\ell,p}(z) \le V_p(z) + \delta$. It is easily seen that the roles of V_p and $V_{\ell,p}$ can be interchanged in the above argument. Thus we have that, given $\delta > 0$, there exists an ℓ_0 such that $|V_{\ell,p}(z) - V_p(z)| \le \delta$ if $\ell > \ell_0$, for all $z \in \mathbb{S}_0$. Since $\mathbb{S}_0$ is an arbitrary compact subset of $\mathbb{S}$, the result follows. □

COROLLARY 2.5. *For all* $z \in \mathbb{S}$, $\lim_{p\to\infty} \lim_{\ell\to\infty} V_{\ell,p}(z) = V(z)$.

3. An approximating Markov decision problem. In this section we will present the Markov decision problem whose value function approximates $V_{\ell,p}$. Since throughout this section ℓ, p will be fixed, we will drop them from the notation: $V_{\ell,p}, \tau_\ell, J_\ell, \mathcal{A}_p(z)$ and $Z_\ell \equiv (X_\ell, Y_\ell)$. We will introduce a discrete time, discrete state controlled Markov chain to approximate the continuous time process given by (7).

Fix $h > 0$ and define the two-dimensional h-grid, $L^h \doteq \{(jh, kh) : -\infty < j, k < +\infty\}$. The symbol h denotes the approximation parameter, and as h approaches 0, a suitable interpolation of the controlled Markov chain, to be introduced below, "approaches" a controlled diffusion process of the form in (7). We will assume for simplicity that ℓ is an integer multiple of h.

A natural definition of the state space for the approximating chain is $\mathbb{S}^h_\ell \doteq \mathbb{S}_\ell \cap L^h$. However, due to reflection terms in the dynamics of the controlled process, it is convenient to consider a slightly "enlarged" state space, namely, $\mathbb{S}^{h+}_\ell \doteq \mathbb{S}_{\ell+h} \cap L^h$. The "solvency boundary" and reflecting boundary of the space $\mathbb{S}^{h+}_\ell$ are defined, respectively, as

$$\partial^h \doteq \{(x, y) \in \mathbb{S}^{h+}_\ell : x + (1 + \lambda)y \le h(1 + \lambda), \text{ or } x + (1 - \mu)y \le h\}$$
$$\partial^h_{\mathrm{R}} \doteq \{(x, y) \in \mathbb{S}^{h+}_\ell : x = \ell + h, \text{ or } y = \ell + h\}.$$

Let $\{Z^h_n, n = 0, 1, 2, \dots\}$ be a discrete time controlled Markov chain with state space $\mathbb{S}^{h+}_\ell$, with $Z^h_n = (X^h_n, Y^h_n)$. The transition probabilities will be defined so that the chain's evolution law well approximates the local behavior of the controlled diffusion (7). For each n, the increments of the chain δZ^h_n will approximate exactly one of the following dynamical descriptions:

- "Controlled diffusion step": $(rX_t - C_t, bY_t)'dt + (0, \sigma)'dW_t$.
- "Purchase control step": $(-(1 + \lambda), 1)'dM_t$.
- "Sales control step": $(1 - \mu, -1)'dN_t$.
- "Reflection step": dR_t.

Each of these steps is described precisely in what follows. We also introduce a family of "interpolation intervals" $\{\Delta^h, h > 0\}$ used in defining the approximating cost function and in the convergence arguments. For each pair $(z, c) \in \mathbb{S}_\ell^{h+} \times [0, p]$ we first define a family $\tilde{\Delta}^h(z, c)$. For the controlled diffusion steps, if the state of the chain is z and the exercised consumption control is c, Δ^h will be taken to be $\tilde{\Delta}^h(z, c)$; whereas for singular control steps and reflection steps, Δ^h will be taken to be 0. This reflects the fact that for the controlled diffusion (7), reflection and singular control terms can change the state instantaneously. Suitable conditions on $\tilde{\Delta}^h(z, c)$ needed to obtain convergence of the continuous time interpolated processes to corresponding controlled diffusions are introduced below.

Controlled diffusion steps and local consistency. By a controlled diffusion step we mean that the Markov chain evolves according to a transition law which is "locally consistent" in the sense of [19], with a (controlled) diffusion given as

$$d\tilde{X}(t) = (r\tilde{X}(t) - C(t))dt,\ d\tilde{Y}(t) = b\tilde{Y}(t)dt + \sigma\tilde{Y}(t)dW(t).$$

Formally, given $h > 0$, we choose for each $c \in [0, p]$ and $z \in \mathbb{S}_\ell^{h+} \setminus \partial^h$ a probability measure $q_h^{(0)}(z, c, d\tilde{z})$ on L^h, along with an interpolation interval $\tilde{\Delta}^h(z, c)$, which satisfies the following local consistency conditions for some $\rho > 0$:

$$m_0(z, c) \doteq \int_{L^h} (\tilde{z} - z) q_h^{(0)}(z, c, d\tilde{z}) = \begin{pmatrix} rx - c \\ by \end{pmatrix} \tilde{\Delta}^h(z, c) + O(h^\rho \tilde{\Delta}^h(z, c)), \tag{12}$$

$$\begin{aligned} \sigma_0(z, c) &\doteq \int_{L^h} (\tilde{z} - z - m_0(z, c))(\tilde{z} - z - m_0(z, c))' q_h^{(0)}(z, c, d\tilde{z}) \\ &= \begin{pmatrix} 0 & 0 \\ 0 & |\sigma y|^2 \end{pmatrix} \tilde{\Delta}^h(z, c) + O(h^\rho \tilde{\Delta}^h(z, c)). \end{aligned} \tag{13}$$

In the above displays, $\tilde{z} = (\tilde{x}, \tilde{y})$, and throughout, by the symbol $O(k)$ we will mean an expression which is bounded above by $\alpha|k|$, where α is a constant depending only on the coefficients of the model and the truncation parameters ℓ, p. In addition we assume that there exists $\zeta \in (0, \infty)$ such that $q_h^{(0)}(z, c, B_{\zeta h}(z)) = 1$ for all $c \in [0, p]$ and $h > 0$, where $B_{\zeta h}(z)$ is a ball of radius ζh centered at z. The interpolation intervals are required to satisfy

$$\tilde{\Delta}_*^h \doteq \sup_{z,c} \tilde{\Delta}^h(z, c) \to 0 \text{ as } h \to 0,\ \inf_{z,c} \tilde{\Delta}^h(z, c) > 0 \text{ for each } h > 0, \tag{14}$$

where the sup and inf in the above displays are taken over all $(z, c) \in \mathbb{S}_\ell^{h+} \times [0, p]$. For the sake of specificity we make the following choice for $q_h^{(0)}$. Let $Q(x, y) \equiv Q^h(x, y) \doteq hr|x| + hp + hb|y| + \sigma^2 y^2$. Define for all $(x, y) \in \mathbb{S}_\ell^{h+} \setminus \partial^h$

$$\begin{aligned} q_h^{(0)}((x, y), c, (x + h, y)) &\doteq \frac{hrx^+}{Q(x, y)}, \quad q_h^{(0)}((x, y), c, (x - h, y)) \doteq \frac{hrx^- + hc}{Q(x, y)}, \\ q_h^{(0)}((x, y), c, (x, y + h)) &\doteq \frac{hby^+ + \frac{1}{2}\sigma^2 y^2}{Q(x, y)}, \quad q_h^{(0)}((x, y), c, (x, y - h)) \doteq \frac{hby^- + \frac{1}{2}\sigma^2 y^2}{Q(x, y)}, \\ q_h^{(0)}((x, y), c, (x, y)) &\doteq \frac{h(p - c)}{Q(x, y)}, \\ \tilde{\Delta}^h(z, c) &\doteq \frac{h^2}{Q(x, y)}. \end{aligned} \tag{15}$$

It is easy to check that $q_h^{(0)}, \tilde{\Delta}^h$ defined above satisfy (12), (13), and (14).

Singular control steps. The singular control terms in the controlled diffusion are the nondecreasing RCLL processes M and N. The process M pushes the state process in the direction $v_1 = (-(1+\lambda), 1)'$, whereas N pushes the state process in the direction $v_2 = ((1-\mu), -1)'$. For the approximating chain we will assume that at most one among the sales control and purchase control are exercised at any given time instant and that the magnitude of the corresponding displacement is $O(h)$. In order to capture the "singular" behavior of the limit diffusion—namely, the feature that the state process can instantaneously be displaced by large amounts—we will take the interpolation interval for all singular control steps in the approximating chain to be 0.

In order to obtain weak convergence of the interpolated chain to the controlled diffusion, we need to ensure that the control directions match asymptotically those for the physical problem. More precisely, given $h > 0$ we define for each $z \in \mathbb{S}_\ell^{h+}$ two probability measures $q_h^{(i)}(z, d\tilde{z}), i = 1, 2$, on L^h as follows. For states $(x, y) \in \mathbb{S}_\ell^{h+} \setminus \partial^h$,

$$q_h^{(1)}((x,y),(x-h,y)) = \lambda/(\lambda+1), q_h^{(1)}((x,y),(x-h,y+h)) = 1/(\lambda+1); \tag{16}$$

$$q_h^{(2)}((x,y),(x,y-h)) = \mu, q_h^{(2)}((x,y),(x+h,y-h)) = 1-\mu. \tag{17}$$

It is easy to check that $q_h^{(1)}$ and $q_h^{(2)}$ introduced above satisfy the following consistency conditions:

$$m_i(z) \doteq \int_{L^h} (\tilde{z} - z) q_h^{(i)}(z, d\tilde{z}) = h v_i, \tag{18}$$

$$\sigma_i(z) \doteq \int_{L^h} (\tilde{z} - z - m_i(z))(\tilde{z} - z - m_i(z))' q_h^{(i)}(z, d\tilde{z}) = O(h^2). \tag{19}$$

Reflection steps. We will define a transition kernel that with probability 1 moves a state in ∂_{R}^h to some state in $\mathbb{S}_\ell^h$. Once more, since reflection in the diffusion control problem occurs instantaneously, we take the interpolation interval at reflection steps to be 0. Since the directions of reflection in the diffusion control problem are normal, a natural choice of the transition kernel for reflection step is as follows for $z \in \partial_{\mathrm{R}}^h$:

$$q_h^{(3)}((\ell+h,y),(\ell,y)) = 1, \quad q_h^{(3)}((x,\ell+h),(x,\ell)) = 1, \quad q_h^{(3)}((\ell+h,\ell+h),(\ell,\ell)) = 1. \tag{20}$$

For $z \notin \partial_{\mathrm{R}}^h$, $q_h^{(3)}(z, \cdot)$ can be defined arbitrarily. It will be seen from the definition of admissible controls given below that for such states, the definition of $q_h^{(3)}$ is immaterial.

The controlled Markov chain. As described above, the control at each step is first specified by the choice of an action: controlled diffusion, singular control, or reflection. Therefore, we define a sequence of control actions $\{I_n^h, n = 0, 1, 2, \dots\}$ with $I_n^h = 0, 1, 2, 3$ if the nth step in the chain is a controlled diffusion step, purchase control step, sales control step, or reflection step, respectively. In the case of a controlled diffusion step, the magnitude of the consumption control must also be specified. Consequently, the space of controls is given by $\mathcal{U} \doteq \{0, 1, 2, 3\} \times [0, p]$.

The probability measures associated with each of the control actions will now be combined into a single probability measure for use in defining the controlled Markov chain. For each $z \in \mathbb{S}_\ell^{h+} \setminus \partial^h, u \in \mathcal{U}$ $(u = (i, c))$, we define a probability measure

$p_h(z,u,d\tilde{z})$ on L^h by

$$p_h(z,u,d\tilde{z}) = q_h^{(0)}(z,c,d\tilde{z})1_{\{i=0\}} + q_h^{(i)}(z,d\tilde{z})1_{\{i\in\{1,2,3\}\}}. \tag{21}$$

The definition of the transition function for $z \in \partial^h$ is not important since in the analysis of the control problem the chain will be stopped the first time it hits ∂^h. For the sake of specificity we set $p_h(z,u,z) = 1$ for all $z \in \partial^h$ and $u \in \mathcal{U}$.

We are now ready to specify the controlled Markov chains. Given a sequence $U^h = \{U_n^h, n = 0,1,2,\dots\}$ (where $U_n^h = (I_n^h, C_n^h)$) of $\mathcal{U}$-valued random variables we construct a controlled Markov chain $\{Z_n^h, n = 0,1,2,\dots\}$ with initial condition $z_h = (x_h, y_h) \in \mathbb{S}_\ell^{h+}$ and state space $\mathbb{S}_\ell^{h+}$, as follows:

$$Z_0^h = z_h, \;\; \mathbb{P}[Z_{n+1}^h \in E|\mathcal{F}_n^h] = p_h(Z_n^h, U_n^h, E), \;\; n \geq 0, \;\; E \in \mathcal{B}(\mathbb{S}_\ell^{h+}), \tag{22}$$

where $\mathcal{F}_n^h = \sigma\{Z_0^h,\dots,Z_n^h,U_0^h,\dots,U_n^h\}$. The following definition of admissible controls ensures that $Z_n^h \in \mathbb{S}_\ell^{h+}$ for all n, and so the definition in (22) is meaningful.

DEFINITION 3.1. *The control sequence $U^h = \{U_n^h, n = 0,1,2,\dots\}$ is said to be admissible for the initial condition z_h, and $\{Z_n^h\}$ ($\{Z_n^h, U_n^h\}$) is called the corresponding controlled Markov chain (respectively, controlled pair) if*

1. *U_n^h is $\sigma\{Z_0^h,\dots,Z_n^h,U_0^h,\dots,U_{n-1}^h\}$-adapted.*
2. *$\mathbb{P}[I_n^h = 3|Z_n^h \in \mathbb{S}_\ell^h] = 0$ and $\mathbb{P}[I_n^h = 3|Z_n^h \in \partial_R^h \setminus \partial^h] = 1$ for all n.*
3. *Condition (22) holds.*

The class of all admissible control sequences for initial state z_h will be denoted by $\mathcal{A}^h(z_h)$.

We also define for each $z \in \mathbb{S}_\ell^{h+}$ and $u = (i,c) \in \mathcal{U}$ the interpolation intervals

$$\Delta^h(z,u) = \tilde{\Delta}^h(z,c)1_{\{i=0\}}. \tag{23}$$

For an admissible pair $\{Z_n^h, U_n^h\}$, we denote the associated sequence of interpolation intervals $\Delta^h(Z_n^h, U_n^h)$ by $\{\Delta_n^h, n = 0,1,2,\dots\}$. Define, $t_0^h \doteq 0$ and $t_n^h \doteq \sum_{i=0}^{n-1} \Delta_i^h$ for $n \geq 1$.

MDP for the chain. Given an admissible pair $\{Z_n^h, U_n^h\}$ let $\eta_h \doteq \inf\{n : Z_n^h \in \partial^h\}$. The cost function for the controlled Markov chain is defined as

$$J^h(z_h, U^h) = \mathbb{E}\sum_{n=0}^{\eta_h - 1} e^{-\beta t_n^h} f(C_n^h)\left(\frac{1 - e^{-\beta\Delta_n^h}}{\beta}\right). \tag{24}$$

Note that we have used the factor $(1 - e^{-\beta\Delta_n^h})/\beta$ rather than the more intuitive (and asymptotically equivalent) Δ_n^h. This somewhat simplifies the convergence proofs without affecting the limiting results. The value function of the MDP is defined as

$$V^h(z_h) = \sup_{U^h \in \mathcal{A}^h(z_h)} J^h(z_h, U^h). \tag{25}$$

4. Continuous time interpolation and time rescaling. One of the main goals of the study is to show that the value function of the MDP defined in (25) converges, as $h \to 0$, to the value function of the limit diffusion control problem. This convergence result allows for the computation of near optimal policies for the diffusion control problem introduced below (6) by numerically solving the above MDP. We next introduce the continuous time interpolation and time rescaling techniques that will be used in the proof of our main convergence result.

The continuous time interpolations of various processes will be constructed to be piecewise constant on the time intervals $[t_n^h, t_{n+1}^h)$, $n \geq 0$. For use in this construction we define $n^h(t) \doteq \max\{n : t_n^h \leq t\}$, $t \geq 0$. Note that $n^h(t)$ is an $\{\mathcal{F}_n^h\}$-stopping time. Setting $\mathcal{F}^h(t) \doteq \mathcal{F}_{n^h(t)}^h$ we obtain a continuous time filtration $\{\mathcal{F}^h(t), t \geq 0\}$. Define $U^h(t) \doteq U_{n^h(t)}^h, t \geq 0$. Also, define the continuous time processes associated with the controlled diffusion steps as follows. First, let $B_0^h = 0$ and $S_0^h = 0$ and define for $n \geq 1$,

$$B_n^h \doteq \sum_{k=0}^{n\wedge\eta_h - 1} \mathbb{E}[\delta Z_k^h | \mathcal{F}_k^h] 1_{\{I_k^h = 0\}}, \quad S_n^h \doteq \sum_{k=0}^{n\wedge\eta_h - 1} \Big(\delta Z_k^h - \mathbb{E}[\delta Z_k^h | \mathcal{F}_k^h]\Big) 1_{\{I_k^h = 0\}}. \tag{26}$$

Define the continuous time process B^h by setting $B^h(0) \doteq 0$ and $B^h(t) \doteq B_{n^h(t)}^h$ for $t > 0$. The process S^h is defined in a similar manner. We define the interpolations associated with the purchase control and sales control as follows. Let $M_0^h = 0, N_0^h = 0$, $E_{i,0}^h = 0$, $i = 1, 2$, and define for $n \geq 1$,

$$M_n^h \doteq \sum_{k=0}^{n\wedge\eta_h - 1} h 1_{\{I_k^h = 1\}}, \quad N_n^h \doteq \sum_{k=0}^{n\wedge\eta_h - 1} h 1_{\{I_k^h = 2\}}, \quad E_{i,n}^h \doteq \sum_{k=0}^{n\wedge\eta_h - 1} (\delta Z_k^h - h v_i) 1_{\{I_k^h = i\}}.$$

The continuous time processes M^h and N^h are defined as $M^h(0) \doteq 0, N^h(0) \doteq 0$ and $M^h(t) \doteq M_{n^h(t)}^h, N^h(t) \doteq N_{n^h(t)}^h$ for $t \geq 0$. The processes E_1^h and E_2^h are defined analogously. The continuous time process associated with reflection is defined as follows. If $n^h(t) = 0$, define $R^h(t) = 0$; otherwise let

$$R^h(t) \doteq - \sum_{k=0}^{n^h(t)-1} \delta Z_k^h 1_{\{I_k^h = 3\}}. \tag{27}$$

We define the continuous time interpolation Z^h of the controlled Markov chain Z_n^h introduced in Definition 3.1 by $Z^h(t) \doteq Z_{n^h(t)}^h$, $t \geq 0$. The following representation for $Z^h(t)$ is easily verified:

$$Z^h(t) = z_h + B^h(t) + S^h(t) + v_1 M^h(t) + v_2 N^h(t) + E_1^h(t) + E_2^h(t) - R^h(t), \ t \geq 0. \tag{28}$$

Also, it follows from condition (12) that on the set $\{I_n^h = 0, \eta_h > n\}$,

$$\mathbb{E}[\delta Z_n^h | \mathcal{F}_n^h] = \begin{pmatrix} r X_n^h - C_n^h \\ b Y_n^h \end{pmatrix} \Delta^h(Z_n^h, 0, C_n^h) + O(h^\rho \Delta^h(Z_n^h, 0, C_n^h)) \text{ a.s.}$$

This fact, together with the piecewise constant nature of the processes, yields

$$B^h(t) = \int_0^{t\wedge\tau^h} \begin{pmatrix} r X^h(s) - C^h(s) \\ b Y^h(s) \end{pmatrix} ds + \delta_1^h(t), \tag{29}$$

where $\tau^h \doteq t_{\eta_h}^h$ and δ_1^h is an $\{\mathcal{F}^h(t)\}$-adapted process which, in view of (14), satisfies for all $t \geq 0$ and $m \geq 1$,

$$\sup_{0 \leq s \leq t} \mathbb{E}|\delta_1^h(s)|^m \to 0 \text{ as } h \to 0.$$

A similar calculation gives the following representation of the cost function (24):

$$J^h(z_h, U^h) = \mathbb{E}\int_{[0,\tau^h]} e^{-\beta t} f(C^h(t))dt. \tag{30}$$

Time rescaling. A common approach for proving the convergence of V^h to V as $h \to 0$ is to begin by showing that the collection $\{(Z^h(\cdot), \tau^h), h \geq 0\}$ is tight and then characterize the subsequential weak limits suitably. However, for problems with singular controls, showing the tightness of the above family becomes problematic since, in general, the processes $\{(M^h(\cdot), N^h(\cdot)), h \geq 0\}$ may fail to be tight. A powerful method for handling this tightness issue was introduced by Martins and Kushner [21]. The basic idea is to suitably stretch out the time scale so that the various processes involved in the convergence analysis, in the new time scale, are tight; carry out the weak convergence analysis with the rescaled processes; and finally, revert back to the original time scale to argue the convergence of V^h to V.

We now introduce the time rescaling that will be used in our study. The rescaled time increments, $\{\hat{\Delta}_n^h, n = 0, 1, 2, \dots\}$, are defined as $\hat{\Delta}_n^h \doteq \Delta_n^h 1_{\{I_n^h=0\}} + h1_{\{I_n^h \in \{1,2\}\}}$. Define $\hat{t}_0^h \doteq 0$ and $\hat{t}_n^h \doteq \sum_{i=0}^{n-1} \hat{\Delta}_i^h$ for $n \geq 1$.

DEFINITION 4.1. *The rescaled time process $\hat{T}^h(t)$ is the unique continuous nondecreasing process satisfying the following:* (1) $\hat{T}^h(0) = 0$*;* (2) *the derivative of $\hat{T}^h(t)$ is* 1 *for $t \in (\hat{t}_n^h, \hat{t}_{n+1}^h)$ if $I_n^h = 0$;* (3) *the derivative of $\hat{T}^h(t)$ for $t \in (\hat{t}_n^h, \hat{t}_{n+1}^h)$ is* 0 *if $I_n^h = 1, 2, 3$.*

It is easy to check that $\hat{T}^h(\hat{t}_n^h) = t_n^h$ and that $\hat{T}^h(\hat{t}_{n+1}^h) - \hat{T}^h(\hat{t}_n^h) = \Delta_n^h$. Let $\hat{n}^h(t) \doteq \max\{n : \hat{t}_n^h \leq t\}$, $t \geq 0$. Using the observation that every reflection step must be followed by either a singular control step or a diffusion control step, it follows that $\hat{n}^h(t)$ is a bounded $\{\mathcal{F}_n^h\}$-stopping time, with bound

$$\hat{n}^h(t) \leq 2\left(\frac{t}{h} + \frac{t}{\inf_{z,c} \hat{\Delta}^h(z, 0, c)}\right) < \infty. \tag{31}$$

Define the continuous time filtration $\{\hat{\mathcal{F}}^h(t), t \geq 0\}$ by setting $\hat{\mathcal{F}}^h(t) \doteq \mathcal{F}_{\hat{n}^h(t)}$.

The rescaled processes (denoted with a $\hat{}$) are defined in a manner similar to the processes defined below (26) with appropriate adjustments to the time variable. For example, we define $\hat{B}^h(0) = 0$ and $\hat{B}^h(t) \doteq B^h_{\hat{n}^h(t)}$ if $\hat{n}^h(t) > 0$. We define the processes $\hat{U}^h(t)$, $\hat{S}^h(t)$, $\hat{M}^h(t)$, $\hat{N}^h(t)$, $\hat{E}_1^h(t)$, $\hat{E}_2^h$, $\hat{R}^h(t)$, $\hat{Z}^h(t)$ analogously (that is, by replacing $n^h(t)$ with $\hat{n}^h(t)$ in the definitions below (26)). Then we have the following rescaled version of (28):

$$\hat{Z}^h(t) = z_h + \hat{B}^h(t) + \hat{S}^h(t) + v_1\hat{M}^h(t) + v_2\hat{N}^h(t) + \hat{E}_1^h(t) + \hat{E}_2^h - \hat{R}^h(t). \tag{32}$$

Remark 4.2. From the definition of $\hat{T}^h(t)$ if follows that $\hat{n}^h(t) = n^h(\hat{T}^h(t))$. This equality yields a straightforward relationship between the original interpolated processes and the rescaled processes. For example, $\hat{B}^h(t) = B^h(\hat{T}^h(t))$. Similar equations hold between $U^h(t)$, $S^h(t)$, $M^h(t)$, $N^h(t)$, $E_1^h(t)$, E_2^h, $R^h(t)$, $Z^h(t)$, and their corresponding rescaled versions.

Using the fact that $\hat{T}^h(\hat{t}_{n+1}^h) - \hat{T}^h(\hat{t}_n^h) = \Delta_n^h$, which is 0 for singular control and reflection steps, a calculation similar to that which produced (29) yields

$$\hat{B}^h(t) = \int_0^{t\wedge\hat{\tau}_1^h} \begin{pmatrix} r\hat{X}^h(s) - \hat{C}^h(s) \\ b\hat{Y}^h(s) \end{pmatrix} d\hat{T}^h(s) + \hat{\delta}_1^h(t), \tag{33}$$

where $\hat{\tau}_1^h \doteq \inf\{t : \hat{Z}^h(t) \in \partial^h\}$ and $\hat{\delta}_1^h$ is an $\{\hat{\mathcal{F}}^h(t)\}$-adapted process satisfying, for all $m \geq 1$,

$$\mathbb{E} \sup_{0\leq s\leq t} |\hat{\delta}_1^h(s)|^m \to 0 \text{ as } h \to 0. \tag{34}$$

We now state several lemmas related to the time rescaling. The following "change of variables" formula (cf. Theorem IV.3.45 in [28]) will be used several times in our analysis.

LEMMA 4.3. *Let $\hat{G} : [0,\infty) \to [0,\infty)$ be a continuous and nondecreasing function. Suppose that $\hat{G}(t) \to \infty$ as $t \to \infty$. Define the inverse $G : [0,\infty) \to [0,\infty)$ as $G(t) = \inf\{s : \hat{G}(s) > t\}$. Then for all bounded and measurable functions $g : [0,\infty) \to [0,\infty)$,*

$$\int_{[0,G(t)]} g(s)d\hat{G}(s) = \int_{[0,t]} g(G(s))ds. \tag{35}$$

The following lemma is at the heart of the time transformation idea. It ensures that the weak limits of $\hat{T}^h(t)$ increase to ∞ as $t \to \infty$ and thus makes reverting back to the original time scale, in the limit, possible (see Theorem 5.6). The proof of the lemma is contained in the appendix.

LEMMA 4.4. *Let $\{U_n^h, n = 0, 1, 2, \dots\}_{h>0}$ be a family of admissible control sequences. Then for all $t \geq 0$,*

$$\sup_h \mathbb{E}|M^h(t) + N^h(t)| < \infty. \tag{36}$$

An important consequence of the above lemma is the following.

LEMMA 4.5. *There exists an $h_0 \in (0,\infty)$ such that for all $h < h_0$, $\hat{T}^h(t) \to \infty$ with probability 1 as $t \to \infty$.*

Proof. Since $\tilde{\Delta}_*^h \to 0$ as $h \to 0$, we can find an h_0 such that $\tilde{\Delta}_*^h < 1$ for all $h < h_0$. We will argue via contradiction. Suppose $h < h_0$ and $\mathbb{P}[\sup_{t\geq 0} \hat{T}^h(t) < \infty] > 0$. Then there exist $\epsilon > 0$ and $T_0 > 0$ such that

$$\mathbb{P}\left[\sup_{t\geq 0} \hat{T}^h(t) < T_0 - 1\right] > \epsilon. \tag{37}$$

Using Lemma 4.4 we can find a K large enough so that

$$\mathbb{P}[M^h(T_0) \geq K] \leq \frac{\mathbb{E}M^h(T_0)}{K} < \frac{\epsilon}{4}, \;\; \mathbb{P}[N^h(T_0) \geq K] \leq \frac{\mathbb{E}N^h(T_0)}{K} < \frac{\epsilon}{4}.$$

We will now show that

$$\mathbb{P}[\hat{T}^h(T_0 + 2K) < T_0 - 1] \leq \frac{\epsilon}{2}. \tag{38}$$

This will lead to a contradiction in view of (37) and hence prove the lemma. Note that

$$\begin{aligned}
&\mathbb{P}[\hat{T}^h(T_0 + 2K) < T_0 - 1] \\
&\quad \leq \mathbb{P}[\hat{T}^h(T_0 + M^h(T_0) + N^h(T_0)) < T_0 - 1, M^h(T_0) < K, N^h(T_0) < K] \\
&\quad +\mathbb{P}[M^h(T_0) \geq K] + \mathbb{P}[N^h(T_0) \geq K] \\
&\quad \leq \mathbb{P}[\hat{T}^h(T_0 + M^h(T_0) + N^h(T_0)) < T_0 - 1] + \frac{\epsilon}{4} + \frac{\epsilon}{4}. \qquad (39)
\end{aligned}$$

Furthermore, for each fixed t, $t+M^h(t)+N^h(t) \geq \sum_{k=0}^{n^h(t)-1}(\Delta_k^h 1_{\{I_k^h=0\}}+h1_{\{I_k^h=1,2\}})$. Since $\hat{T}^h$ is nondecreasing and $\hat{T}^h(\hat{t}_n^h)=t_n^h$,

$$\hat{T}^h\left(t+M^h(t)+N^h(t)\right) \geq \hat{T}^h\left(\sum_{k=0}^{n^h(t)-1}\left(\Delta_k^h 1_{\{I_k^h=0\}}+h1_{\{I_k^h=1,2\}}\right)\right)$$
$$= \hat{T}^h(\hat{t}^h_{n^h(t)}) = t^h_{n^h(t)} = \sum_{k=0}^{n^h(t)-1}\Delta_k^h 1_{\{I_k^h=0\}} \geq t-\tilde{\Delta}_*^h.$$

The last inequality above is a consequence of the inequalities $\sum_{k=0}^{n^h(t)-1}\Delta_k^h 1_{\{I_k^h=0\}} \leq t \leq \sum_{k=0}^{n^h(t)}\Delta_k^h 1_{\{I_k^h=0\}}$. Recalling that $\tilde{\Delta}_*^h<1$, we see that $\hat{T}^h(t+M^h(t)+N^h(t)) \geq t-1$ for all $t\geq 0$. Using this inequality in (39) proves (38) and hence we have the result. □

Let $T^h(t) \doteq \inf\{s : \hat{T}^h(s) > t\}$. Observe that $\hat{T}^h(T^h(t))=t$ and that, due to Lemma 4.5, $T^h(t)<\infty$ a.s. for all $t\geq 0$. Define $\hat{\tau}^h \doteq T^h(\tau^h)$.

LEMMA 4.6. *For $z_h \in \mathbb{S}_\ell^{h+}$ and $\{U_n^h\} \in \mathcal{A}^h(z_h)$,*

$$J^h(z_h,U^h) = \mathbb{E}\int_{[0,\hat{\tau}_1^h]} e^{-\beta\hat{T}^h(t)} f(\hat{C}^h(t))d\hat{T}^h(t).$$

Proof. Note that

$$\hat{\tau}_1^h = \inf\{t : \hat{Z}^h(t)\in\partial^h\} \;=\; \inf\{t : Z^h(\hat{T}^h(t))\in\partial^h\} \;=\; \inf\{t : \hat{T}^h(t)\geq\tau^h\}.$$

If $\tau^h=\infty$, then clearly $\hat{\tau}_1^h=\infty$. Suppose $\tau^h<\infty$. Then the above display shows that $\hat{\tau}_1^h = T^h(\tau^h-)$. Also, clearly $\hat{T}^h$ is constant over the interval $(T^h(\tau^h-), T^h(\tau^h)]$. The result now follows from (30) and Lemma 4.3. □

5. Main convergence result. In this section we show that $V^h(z_h)$ converges to $V(z)$ whenever $z_h \to z$. The basic approach will be as follows. First, we establish tightness of the continuous time (rescaled) processes defined in the previous section and characterize their subsequential limits. Then we define a time transformation for the limit processes to revert back to the original scale. We will show that the time transformed versions of the limit processes have the same laws as those of the various processes in the diffusion control problem. Using this characterization result we will show that, given a sequence of admissible controls $\{U^h, h>0\}$, the lim sup of the corresponding cost functions is bounded above by the cost for an admissible control for the diffusion control problem. This will establish that $\limsup_{h\to 0} V^h(z_h) \leq V(z)$ whenever $z_h \to z$. Finally, we prove convergence of the value functions by proving the reverse inequality. The main idea of this proof is to select a near optimal control for the limit diffusion control problem and to construct from this an admissible control for the controlled Markov chain which is asymptotically near optimal.

We begin by introducing the following "relaxed control" formulation which arises naturally in the weak convergence arguments for convergence of the cost functions.

Relaxed control formulation. Let $\tilde{\mathcal{M}}$ denote the space of all Borel measures ϑ on $[0,p]\times[0,\infty)$ such that if $\vartheta(d\alpha,dt)=\vartheta_t(d\alpha)\nu(dt)$, then (i) ϑ_t is a probability measure on $[0,p]$ for ν-almost every t, and (ii) $\nu(a,b]\leq b-a$ for all $0\leq a\leq b<\infty$. Let $\mathcal{M}$ be the subset of $\tilde{\mathcal{M}}$ consisting of ϑ that satisfy, for all $t\geq 0$, $\vartheta([0,p]\times[0,t])=t$. Given a probability system Φ and initial condition $z\in\mathbb{S}_\ell$, let $\bar{\mathcal{A}}_p(\Phi,z)$ be the set of

all processes $\bar{U} \equiv (m, M, N)$, where M and N are as introduced below (1), $m \in \mathcal{M}$ a.s., and $m(A \times [0,t])$ is $\{\mathcal{F}_t\}$-adapted for all $t \in [0,\infty)$, $A \in \mathcal{B}[0,p]$. Set $C(t) \doteq \int_{[0,p]} \alpha m_t(d\alpha)$, where m_t, a probability measure on $[0,p]$, is defined by the relation $m(d\alpha, dt) = m_t(d\alpha)dt$.

Let Z be defined via (7) with (C, M, N) as above and τ be given by (2). Define for $\bar{U} \in \bar{\mathcal{A}}_p(\Phi, z)$,

$$\bar{J}(z, \bar{U}) \doteq \mathbb{E} \int_{[0,p]\times[0,\tau)} e^{-\beta t} f(\alpha) m(d\alpha, dt),$$

and let

$$\bar{V}(z) \doteq \sup_{\Phi} \sup_{\bar{U} \in \bar{\mathcal{A}}(\Phi, z)} \bar{J}(z, \bar{U}).$$

The following lemma establishes the equivalence between the relaxed control formulation and the precise control formulation. The proof is contained in the appendix.

LEMMA 5.1. *For all* $z \in \mathbb{S}$, $\bar{V}(z) = V(z)$.

The space $\tilde{\mathcal{M}}$ can be metrized using the Prohorov metric in the usual way (see pages 263–264 of [19]). Furthermore, with this metric, $\tilde{\mathcal{M}}$ is a compact space, and a sequence $\vartheta_n \in \tilde{\mathcal{M}}$ converges to ϑ if and only if for all continuous functions ψ on $[0,p] \times [0,\infty)$ with compact support,

$$\int_{[0,p]\times[0,\infty)} \psi(\alpha, t) m_n(d\alpha, dt) \to \int_{[0,p]\times[0,\infty)} \psi(\alpha, t) m(d\alpha, dt). \tag{40}$$

We now define $\tilde{\mathcal{M}}$-valued random variables $\hat{m}^h$ by the relation

$$\hat{m}^h(A \times [0,t]) \doteq \int_{[0,t]} 1_A(\hat{C}^h(s)) d\hat{T}^h(s), \quad A \in \mathcal{B}([0,p]), t \in [0,\infty).$$

Noting that the right side above is equal to $\int_{[0,t]} (\int_A \delta_{\hat{C}^h(s)}(d\alpha)) d\hat{T}^h(s)$, where δ_x is the probability measure concentrated at x, we can write $\hat{m}^h(d\alpha, dt)$ as $\hat{m}^h_t(d\alpha)\hat{\nu}^h(dt)$, where $\hat{m}^h_t$ and $\hat{\nu}^h$ are given by $\hat{m}^h_t(A) = \delta_{\hat{C}^h(t)}(A)$, $\hat{\nu}^h(a,b] = \hat{T}^h(b) - \hat{T}^h(a)$ for $A \in \mathcal{B}([0,p])$ and $0 \le a \le b < \infty$.

Convergence of the time rescaled processes. Recall the definitions and notation, found in section 4, relating to the continuous time interpolated processes and the corresponding rescaled versions. We begin by showing that the processes $\hat{E}^h_1$ and $\hat{E}^h_2$ converge weakly to the 0 process as $h \to 0$.

LEMMA 5.2. *Let* $\hat{E}^h_i, i = 1, 2$, *be as defined above* (27). *Then* $\hat{E}^h_i$ *converges in probability to* 0 *in* $D([0,\infty) : \mathbb{R}^2)$.

Proof. The local consistency condition (18) and property (22) imply that $E^h_{i,n}$ is an $\{\mathcal{F}^h_n\}$-martingale. As $\hat{n}^h(t)$ is a bounded stopping time (cf. (31)) and the increments of $E^h_{i,n}$ are bounded, it follows from the optional sampling theorem that the continuous time process $\hat{E}^h_i(t)$ is an $\{\hat{\mathcal{F}}^h(t)\}$-martingale, the trace of the quadratic variation of which is given by $\mathrm{Tr}\langle \hat{E}^h_i \rangle(t) = \sum_{k=0}^{\hat{n}^h(t)\wedge\eta_h - 1} \mathbb{E}[|\delta z^h_k - h v_i|^2 1_{\{I^h_k = i\}} | \mathcal{F}^h_{k-1}]$. Finally, applying Doob's inequality, (19), and the observation that the maximum number of steps of either singular control in the first $\hat{n}^h(t)$ steps is t/h, we have for $i = 1, 2$, $\mathbb{E}[\sup_{s\le t} |\hat{E}^h_i(s)|]^2 \le 4\mathbb{E}\mathrm{Tr}\langle \hat{E}^h_1 \rangle(t) \le O(h^2)(t/h) = O(h)$. The result now follows. ∎

Define the process $\hat{A}^h$ by $\hat{A}^h(t) \doteq \int_{[0,t)} \hat{C}^h(s) d\hat{T}^h(s)$. Let $\bar{\mathbb{R}}$ denote the one point compactification of $\mathbb{R}$. The following proposition gives the tightness of the various time rescaled processes. The proof is similar to that of Theorem 5.3 of [20] and is therefore omitted.

PROPOSITION 5.3. *Let $\hat{H}^h \doteq (\hat{Z}^h, \hat{T}^h, \hat{A}^h, \hat{M}^h, \hat{N}^h, \hat{R}^h, \hat{B}^h, \hat{S}^h)$. Then the family $\{(\hat{H}^h, \hat{\tau}_1^h, \hat{m}^h), h > 0\}$ is tight in $D([0,\infty) : E) \times \bar{\mathbb{R}} \times \mathcal{M}$, where $E = \mathbb{S}_\ell^{h+} \times \mathbb{R}_+^6 \times \mathbb{R}^4$.*

We now turn our attention to characterizing subsequential limit points of the family $\{(\hat{H}^h, \hat{\tau}_1^h, \hat{m}^h), h > 0\}$. Suppose that the initial condition sequence $\{z_h\}$ converges to some $z \in \mathbb{S}_\ell$. Slightly abusing notation, let h index a weakly convergent subsequence of $(\hat{H}^h, \hat{\tau}_1^h, \hat{m}^h)$ with weak limit, $(\hat{H}, \hat{\tau}_1, \hat{m})$, where $\hat{H} \doteq (\hat{Z}, \hat{T}, \hat{A}, \hat{M}, \hat{N}, \hat{R}, \hat{B}, \hat{S})$, given on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\hat{\mathcal{F}}^*(t) \doteq \sigma(\hat{H}(s), \hat{m}(A \times [0,s)) | A \in \mathcal{B}([0,p]), 0 \le s \le t)$, and let $\hat{\mathcal{F}}(t) \doteq \hat{\mathcal{F}}^*(t+) \vee \mathcal{N}$, where $\mathcal{N}$ denotes the collection of all $\mathbb{P}$-null sets.

THEOREM 5.4. *The limit point $(\hat{H}, \hat{\tau}_1, \hat{m})$ has the following properties.*

1. *$\hat{T}$ is nondecreasing and Lipschitz continuous with Lipschitz coefficient 1.*
2. *There exists an $\{\hat{\mathcal{F}}(t)\}$-progressively measurable process $\hat{C}$ with $\hat{C}(t) \in [0,p]$ for all $t \ge 0$, such that*

$$\hat{B}(t) = \int_0^{t \wedge \hat{\tau}_1} \begin{pmatrix} r\hat{X}(s) - \hat{C}(s) \\ b\hat{Y}(s) \end{pmatrix} d\hat{T}(s). \tag{41}$$

3. *$\hat{S}_1(t) = 0$, for all $t \ge 0$, and $\hat{S}_2$ is a continuous $\{\hat{\mathcal{F}}(t)\}$-martingale with quadratic variation $\langle \hat{S}_2 \rangle_t = \int_0^t |\sigma \hat{Y}(s)|^2 d\hat{T}(s), t \ge 0$.*
4. *$\hat{M}$ and $\hat{N}$ are nondecreasing and continuous.*
5. *$\hat{R}$ is a vector of nondecreasing continuous processes which satisfy*

$$\int_0^\infty 1_{\{\hat{X}(t) < \ell\}} d\hat{R}_1(t) = 0, \quad \int_0^\infty 1_{\{\hat{Y}(t) < \ell\}} d\hat{R}_2(t) = 0. \tag{42}$$

6. *$\hat{Z}$ is a continuous process satisfying $\mathbb{P}[\hat{Z}(t) \in S_\ell] = 1$ for all $t \ge 0$ and*

$$\hat{Z}(t) = z + \hat{B}(t) + \hat{S}(t) + v_1 \hat{M}(t) + v_2 \hat{N}(t) - \hat{R}(t). \tag{43}$$

7. *Writing $\hat{m}(d\alpha, dt)$ as $\hat{m}_t(d\alpha)\hat{\nu}(dt)$ we have $\hat{\nu}(a,b] = \hat{T}(b) - \hat{T}(a)$, $0 \le a \le b < \infty$.*
8. *$\hat{C}(t) = \int_{[0,p]} \alpha \hat{m}_t(d\alpha)$ for $\hat{\nu}$-almost every $t \in [0,\infty)$.*

Proof. By appealing to the Skorohod representation theorem and by relabeling the convergent subsequence, we can assume without loss of generality that $\hat{H}^h \to \hat{H}$ a.s. The fact that the process $\hat{T}$ is nondecreasing and Lipschitz continuous with Lipschitz coefficient 1 follows easily from similar properties for $\hat{T}^h$. Since $|\hat{A}^h(t) - \hat{A}^h(s)| \le p|\hat{T}^h(t) - \hat{T}^h(s)|$ it follows that $\hat{A}$ is absolutely continuous with respect to $\hat{T}$. Therefore there exists a $[0,p]$-valued process $\hat{C}$, progressively measurable with respect to $\{\hat{\mathcal{F}}^*(t)\}$ such that $\hat{A}(t) = \int_0^t \hat{C}(s) d\hat{T}(s)$. This fact, together with $(\hat{Z}^h, \hat{T}^h) \to (\hat{Z}, \hat{T})$ a.s. and an application of the dominated convergence theorem, yields 2. We next show that $\hat{S}$ has continuous paths. First, note that by local consistency ((12), (13)) there exists $\zeta \in (0,\infty)$ such that for all $u \ge 0, h \ge 0$ $j(\hat{S}^h, u) \doteq \sup_{t \le u} |\hat{S}^h(t) - \hat{S}^h(t-)| \le 2\zeta h$. Thus for h small enough, $j(\hat{S}^h) \doteq \int_0^\infty e^{-u}(j(\hat{S}^h, u) \wedge 1) du \le 2\zeta h$. Therefore, by Theorem 3.10.2 in [9] the limiting process $\hat{S}$ has continuous paths. One can check

that the quadratic variation of $\hat{S}^h$, which is an $\{\hat{\mathcal{F}}^h(t)\}$-martingale, is given by

$$\langle \hat{S}^h \rangle(t) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \int_0^{t \wedge \hat{\tau}_1^h} |\sigma \hat{Y}^h(s)|^2 d\hat{T}^h(s) + \hat{\delta}_2^h(t), \tag{44}$$

where due to (14) and using (31) the process $\hat{\delta}_2^h$ satisfies for all $m \geq 1$,

$$\mathbb{E} \sup_{s \leq t} |\hat{\delta}_2^h(s)|^m \to 0 \text{ as } h \to 0.$$

From (44) it follows that $\hat{S}_1(t) = 0$ for all $t \geq 0$. From (44) and the Burkholder–Gundy inequalities we also have

$$\mathbb{E}|\hat{S}_2^h(t)|^4 \leq \alpha \left[T^2 + \mathbb{E} \sup_{0 \leq u \leq t} |\hat{\delta}_2^h(u)|^2 \right].$$

Thus the family $\{(\hat{S}_2^h(t))^2, h > 0\}$ is uniformly integrable. A standard argument (cf. pages 1457–1458 in [20]) shows that $\hat{S}_2$ is an $\{\hat{\mathcal{F}}_t\}$-martingale with quadratic variation as given in 3. Part 4 is immediate on noting that $\hat{M}^h, \hat{N}^h$ are nondecreasing, and since the maximum number of purchase or sales steps over $(\hat{n}^h(t), \hat{n}^h(t+s))$ is $s/h + 1$,

$$|\hat{M}^h(t+s) - \hat{M}^h(t)| \leq s + h, \;\; |\hat{N}^h(t+s) - \hat{N}^h(t)| \leq s + h.$$

From Definition 3.1(3) it follows that (42) holds with $(\hat{X}, \hat{Y}, \hat{R})$ replaced by $(\hat{X}^h, \hat{Y}^h, \hat{R}^h)$. Also clearly $\hat{Z}^h \in (-\infty, \ell] \times (-\infty, \ell]$. Parts 5 and 6 are now immediate consequences of (32) and continuity properties of the Skorohod map (see (6)). Next, write $\hat{m}(d\alpha, dt)$ as $\hat{m}_t(d\alpha)\hat{\nu}(dt)$. Since $\hat{m}^h([0,p],(a,b]) = \hat{T}^h(b) - \hat{T}^h(a)$ for $0 \leq a \leq b < \infty$, taking limits yields $\hat{\nu}(a,b] = \hat{T}(b) - \hat{T}(a)$. This proves part 7. Part 8 is immediate from the representation $\int_{(a,b]} \hat{C}^h(s) d\hat{T}^h(s) = \int_{(a,b] \times [0,p]} \alpha \hat{m}^h(d\alpha, ds)$, $0 \leq a \leq b < \infty$ and the fact that (see the proof of part 2) $\int_{(a,b]} \hat{C}^h(s) d\hat{T}^h(s)$ converges to $\int_{(a,b]} \hat{C}(s) d\hat{T}(s)$. □

Time inversion. We now define an inverse time transformation that will revert the limit processes back to the original time scale. We will see that the time inverted processes lead to an admissible control pair for the diffusion control problem in (7)–(9). The key step in returning to the original time scale is the following result analogous to Lemma 4.5.

LEMMA 5.5. *$\hat{T}(t) \to \infty$ with probability 1 as $t \to \infty$.*

Proof. We will argue via contradiction. Suppose $\mathbb{P}[\sup_{t \geq 0} \hat{T}(t) < \infty] > 0$. Then there exist $\epsilon > 0$ and $T_0 > 0$ such that

$$\mathbb{P}\left[\sup_{t \geq 0} \hat{T}(t) < T_0 - 1 \right] > \epsilon. \tag{45}$$

As in the proof of Lemma 4.5 we can find a $K \in (0, \infty)$ such that $\liminf_{h \to 0} \mathbb{P}[\hat{T}^h(T_0 + 2K) < T_0 - 1] \leq \epsilon/2$. The weak convergence $\hat{T}^h \Rightarrow \hat{T}$ now implies $\mathbb{P}[\hat{T}(T_0 + 2K) < T_0 - 1] \leq \epsilon/2$. This contradicts (45), and hence the result follows. □

Let T be the inverse of $\hat{T}$, defined as $T(t) \doteq \inf\{s : \hat{T}(s) > t\}$. From Lemma 5.5 it follows that $T(t) < \infty$ a.s. for all $t \geq 0$. Since $\hat{T}(t)$ is nondecreasing and continuous,

it follows that $T(t)$ is nondecreasing and right-continuous. Also note the following properties for all $t \geq 0$:

$$T(t) \geq t, \quad \hat{T}(T(t)) = t, \quad T(\hat{T}(t)) \geq t,$$
$$T(t) \uparrow \infty \text{ as } t \uparrow \infty, \quad T(t) < \infty \text{ a.s.}, \quad \hat{T}(s) \in [0,t] \Leftrightarrow s \in [0, T(t)]. \tag{46}$$

Let $\hat{H}$ be as in Theorem 5.4. Define $H(t) \doteq \hat{H}(T(t))$. We will use similar notation for the various components of H; for example, $Z(t) \doteq \hat{Z}(T(t))$, etc. Let $\tau_1 \doteq \hat{T}(\hat{\tau}_1)$. Then by (43), for $t \geq 0$,

$$Z(t) = z + B(t) + S(t) + v_1 M(t) + v_2 N(t) - R(t). \tag{47}$$

Before characterizing the various terms in (47) we note that for $t \geq 0$, $\{T(s) < t\} = \{\hat{T}(t) > s\} \in \hat{\mathcal{F}}(t)$ since $\hat{T}(t)$ is $\hat{\mathcal{F}}(t)$-measurable. Therefore, since $\hat{\mathcal{F}}(t)$ is right-continuous, $T(s)$ is an $\{\hat{\mathcal{F}}(t)\}$-stopping time for each $s \geq 0$. Let $\mathcal{F}_0(t) \doteq \hat{\mathcal{F}}(T(t))$ and note that $H(t) \doteq \hat{H}(T(t))$ and $m(A \times [0,t]) \doteq \hat{m}(A \times [0, T(t)])$ are $\mathcal{F}_0(t)$-measurable. Define $\mathcal{F}(t) \doteq \sigma(H(s), m(A \times [0,s]) : A \in \mathcal{B}([0,p]), 0 \leq s \leq t)$. Then $\mathcal{F}(t) \subseteq \mathcal{F}_0(t)$.

THEOREM 5.6. *The processes in* (47) *have the following properties.*

1. $B(t) = \int_0^{t\wedge\tau_1} \binom{rX(s) - C(s)}{bY(s)} ds$.
2. $S_1 \equiv 0$ *and* S_2 *is a continuous* $\{\mathcal{F}_0(t)\}$*-martingale with quadratic variation*

$$\langle S_2 \rangle(t) = \int_0^{t\wedge\tau_1} |\sigma Y(s)|^2 ds. \tag{48}$$

There exists an enlargement of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ *and of the filtration* $\{\mathcal{F}_0(t)\}$ *that supports a Wiener process* W*, which is a martingale with respect to the enlarged filtration and such that*

$$S_2(t) = \int_0^{t\wedge\tau_1} \sigma Y(s) dW(s). \tag{49}$$

3. *The process* C *is* $\{\mathcal{F}_0(t)\}$*-progressively measurable with* $C(t) \in [0,p]$ *a.s. for all* $t \geq 0$*. In addition,* $M(0), N(0) \geq 0$ *and the processes* M *and* N *are right-continuous, nondecreasing, and* $\{\mathcal{F}_0(t)\}$*-adapted. Hence defining* $\Phi \doteq (\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_0(t)\}, W)$ *we have* $U \equiv (C, M, N) \in \mathcal{A}_p(\Phi, z)$.
4. *For all* $t \geq 0$*,* $Z(t) \in (-\infty, \ell] \times (-\infty, \ell]$ *a.s., R is a vector of nondecreasing right-continuous processes, and the representation* (7) *holds with* (X_ℓ, Y_ℓ, Z_ℓ) *there replaced by* (X, Y, Z).

Remark 5.7. Note that Theorem 5.6 does not prove that (Z, R) is a solution to the Skorohod problem introduced in Definition 2.3, since in general (8) may fail to hold for the process R. However, note that if (Z^*, R^*) is the solution of (7) and (8) with $U = (C, M, N)$ as in part 3 of Theorem 5.6, then by classical comparison results for (reflected) stochastic differential equations one can show that $Z^*(t) \geq Z(t)$ for all t a.s., and so $\tau^* \geq \tau$ where τ is as in (2) and τ^* is defined by the right side of (2) with Z replaced by Z^*. This in particular shows that

$$\int_{[0,p]\times[0,\tau]} e^{-\beta t} f(\alpha) dm_t(d\alpha) dt \leq \int_{[0,p]\times[0,\tau^*]} e^{-\beta t} f(\alpha) dm_t(d\alpha) dt. \tag{50}$$

Proof of Theorem 5.6. Part 1 is an immediate consequence of Theorem 5.4(2) and Lemma 4.3 on noting that

$$\int 1_{[T(\tau_1-),T(\tau_1)]} d\hat{T}(s) = 0 \text{ a.s.} \tag{51}$$

Clearly, $S_1(t) \doteq \hat{S}_1(T(t)) \equiv 0$ a.s. From Theorem 5.4 we have that $\{\hat{S}_2(t), \hat{\mathcal{F}}(t)\}$ is a continuous martingale. Thus for all $n \geq 1$, $\mathbb{E}[\hat{S}_2(T(t) \wedge n)|\hat{\mathcal{F}}(T(s))] = \hat{S}_2(T(s) \wedge n)$ a.s. Also as $\hat{S}_2$ has continuous paths and $T(t) < \infty$ a.s., we have as $n \to \infty$ for all $t \geq 0$, $\hat{S}_2(T(t) \wedge n) \to \hat{S}_2(T(t)) = S_2(t)$ a.s. Furthermore, from Theorem 5.4, part 3, there exists $\alpha \in (0, \infty)$ such that $\mathbb{E}|\hat{S}_2(T(t) \wedge n)|^2 \leq \alpha t$ for all $t \geq 0, n \in \mathbb{N}$. Hence, for each fixed t, the family $\{\hat{S}_2(T(t) \wedge n), n \geq 1\}$ is uniformly integrable and therefore $\hat{S}_2(T(t) \wedge n) \to \hat{S}_2(T(t))$ in L^1. Taking limits as $n \to \infty$, it follows that $\mathbb{E}[\hat{S}_2(T(t))|\hat{\mathcal{F}}(T(s))] = \hat{S}_2(T(s))$, that is, $\mathbb{E}[S_2(t)|\mathcal{F}_0(s)] = S_2(s)$. This proves that $\{S_2(t), \mathcal{F}_0(t)\}$ is a martingale. Although T in general may fail to be continuous, $S_2(\cdot) \doteq \hat{S}_2(T(\cdot))$ has continuous paths a.s. This is a consequence of the fact that $\{S_2^h(\cdot)\}_{h>0}$ is tight in $D([0, \infty))$, and an argument similar to that for $\{\hat{S}_2^h\}$ in the proof of Theorem 5.4 shows that any weak limit point, $\tilde{S}_2$, of $\{S_2^h\}$ must have continuous paths a.s. Also, since $\hat{S}_2^h(\cdot) = S_2^h(\hat{T}^h(\cdot))$, we must have that if $(\tilde{S}_2, \hat{S}_2, \hat{T})$ is a limit point of the tight sequence $(S_2^h, \hat{S}_2^h, \hat{T}^h)$, then $\hat{S}_2(t) = \tilde{S}_2(\hat{T}(t))$, and thus from (46), $S(t) \doteq \hat{S}_2(T(t)) = \tilde{S}_2(t)$. Thus we have shown that S_2 is a continuous $\mathcal{F}_0(t)$-martingale. We next consider its quadratic variation. By the Burkholder–Davis–Gundy inequalities (cf. Theorem 3.3.28 in [17]) there exists a constant α independent of n such that

$$\mathbb{E}[|\hat{S}_2^2(T(t) \wedge n)|^2] \leq \mathbb{E}\left[\left(\sup_{0 \leq s \leq T(t)} |\hat{S}_2(s \wedge n)|\right)^4\right] \leq \alpha(\alpha_1^2 t^2).$$

Thus the families $\{\hat{S}_2^2(T(t) \wedge n), n \geq 1\}$ and $\{\langle \hat{S}_2 \rangle(T(t) \wedge n), n \geq 1\}$ are uniformly integrable for each fixed $t \geq 0$. Since $\hat{S}_2^2$ (respectively, $\langle \hat{S}_2 \rangle$) has continuous paths and $T(t) < \infty$ a.s. $\hat{S}^2(T(t) \wedge n) \to \hat{S}_2^2(T(t))$ (respectively, $\langle \hat{S}_2 \rangle(T(t) \wedge n) \to \langle \hat{S}_2 \rangle(T(t))$) a.s. as $n \to \infty$. By the uniform integrability, this convergence also holds in the L^1 sense. Thus

$$\mathbb{E}[\hat{S}_2^2(T(t) \wedge n) - \langle \hat{S}_2 \rangle(T(t) \wedge n)|\hat{\mathcal{F}}(T(s))] \to \mathbb{E}[\hat{S}_2^2(T(t)) - \langle \hat{S}_2 \rangle(T(t))|\hat{\mathcal{F}}(T(s))]. \tag{52}$$

The above relation and the fact that $\hat{S}_2^2 - \langle \hat{S}_2 \rangle$ is an $\hat{\mathcal{F}}_t$-martingale now show that $\mathbb{E}[S_2^2(t) - \langle \hat{S} \rangle(T(t))|\mathcal{F}_0(s)] = S_2^2(s) - \langle \hat{S} \rangle(T(s))$. Thus the quadratic variation of S_2 is given by $\langle S_2 \rangle(t) = \langle \hat{S}_2 \rangle(T(t))$. The representation (48) now follows on using Theorem 5.4, Lemma 4.3, and (51). By the martingale representation theorem (e.g., Theorem 3.4.2 in [17]) it now follows that there exists a one-dimensional Brownian motion W, possibly defined on an enlarged probability space, that is, a martingale with respect to an enlargement of the filtration $\{\mathcal{F}_0(t)\}$ and is such that (49) holds.

The $\{\mathcal{F}_0(t)\}$-progressive measurability (respectively, adaptedness) of C (respectively, M and N) follows from the $\{\hat{\mathcal{F}}(t)\}$-progressive measurability of $\hat{C}$ (respectively, adaptedness of $\hat{M}$ and $\hat{N}$). Also, since $\hat{C}$ takes values in $[0, p]$, the same is true of C. Right continuity of M and N is a consequence of the fact that $\hat{M}$ and $\hat{N}$ are continuous and T is right continuous. This proves 3. Part 4 is once more an immediate consequence of Theorem 5.4 (part (5)) and Lemma 4.3. □

Until now the parameters ℓ, p have been fixed and thus excluded from the notation. It is convenient to include these parameters in the notation for the remainder of this section.

Convergence of the value functions. Let $z \in \mathbb{S}_\ell$ and let $\{z_h, h > 0\}$ be a sequence with $z_h \in \mathbb{S}_\ell^h$ such that $z_h \to z$ as $h \to 0$. Recall the definitions of $V_{\ell,p}(z)$ in (9) and $V^h(z_h)$ in (25). Our main goal in this section is to show that $V^h(z_h) \to V_{\ell,p}(z)$ as $h \to 0$. We begin with the following proposition.

PROPOSITION 5.8. *Let $\{z_h\}, z$ be as above. Then* $\limsup_{h\to 0} V^h(z_h) \le V_{\ell,p}(z)$.

Proof. Fix for each $h > 0$ an admissible control sequence for the initial condition z_h, $U^h \equiv \{U_n^h, n \ge 1\} \in \mathcal{A}^h(z_h)$. Recall the definition of $J^h(z_h, U^h)$ in (24). In order to prove the proposition it suffices to show that

$$\limsup_{h\to 0} J^h(z_h, U^h) \le V_{\ell,p}(z). \tag{53}$$

Using Lemma 4.6 and boundedness of f, we can find, for each $\epsilon \in (0,\infty)$, a $c \equiv c(\epsilon) \in (0,\infty)$ such that

$$J^h(z_h, U^h) \le \mathbb{E}\int_{[0,\hat\tau_1^h \wedge c]\times[0,p]} e^{-\beta \hat T^h(t)} f(\alpha)\hat m^h(d\alpha, dt) + \frac{\epsilon}{2}. \tag{54}$$

Let $(\hat H^h, \hat\tau_1^h, \hat m^h)$ be as in Proposition 5.3 and $(\hat H, \hat\tau_1, \hat m)$ be one of its weak limit points. Once more, as in the proof of Theorem 5.4 we can assume, by relabeling and appealing to the Skorohod representation theorem, that $(\hat H^h, \hat\tau_1^h, \hat m^h) \to (\hat H, \hat\tau_1, \hat m)$ a.s. Taking limits as $h \to 0$ in (54), we have

$$\limsup_{h\to 0} J^h(z_h, U^h) \le \mathbb{E}\int_{[0,p]\times[0,\hat\tau_1\wedge c]} e^{-\beta \hat T(t)} f(\alpha)\hat m(d\alpha, dt) + \frac{\epsilon}{2}.$$

As $\epsilon > 0$ and $c = c(\epsilon)$ are arbitrary,

$$\limsup_{h\to 0} J^h(z_h, U^h) \le \mathbb{E}\int_{[0,p]\times[0,\hat\tau_1]} e^{-\beta \hat T(t)} f(\alpha)\hat m(d\alpha, dt). \tag{55}$$

Let, as before, $\hat\tau \doteq \inf\{t \ge 0 : \hat Z(t) \notin \mathbb{S}^o\}$. Recall that $\hat\tau_1 \ge \hat\tau$. Then clearly

$$\begin{aligned}&\mathbb{E}\left[1_{\{\hat\tau=\infty\}}\int_{[0,p]\times[0,\hat\tau_1]} e^{-\beta \hat T(t)} f(\alpha)\hat m(d\alpha, dt)\right]\\ &\quad= \mathbb{E}\left[1_{\{\hat\tau=\infty\}}\int_{[0,p]\times[0,\hat\tau]} e^{-\beta \hat T(t)} f(\alpha)\hat m(d\alpha, dt)\right].\end{aligned} \tag{56}$$

Now suppose that $\hat\tau < \infty$. Let τ^* denote the first point of increase of $\hat T$ in $[\hat\tau, \hat\tau_1]$. More precisely, let $\tau^* \doteq \inf\{t \in [\hat\tau, \infty) : \hat T(t+\delta) > \hat T(t) \text{ for all } \delta > 0\} \wedge \hat\tau_1$. Note that

$$\begin{aligned}&\mathbb{E}\left[1_{\{\hat\tau<\infty\}}\int_{(\hat\tau,\hat\tau_1]} e^{-\beta \hat T(t)}\left(\int_{[0,p]} f(\alpha) d\hat m_t(d\alpha)\right) d\hat T(t)\right]\\ &\quad= \mathbb{E}\left[1_{\{\hat\tau<\infty\}}\int_{[\tau^*,\hat\tau_1]} e^{-\beta \hat T(t)}\left(\int_{[0,p]} f(\alpha) d\hat m_t(d\alpha)\right) d\hat T(t)\right].\end{aligned} \tag{57}$$

We now show that the above quantity is equal to 0. Consider the evolution of the process $\hat Z$ over the interval $[\hat\tau, \tau^*]$. By definition, $\hat Z(\hat\tau) \notin \mathbb{S}^o$. Since $\hat T$ is constant over

this time interval, we see from Theorem 5.4(2),(3) that $\hat{B}_1$ and $\hat{S}$ are both constant over this interval, and since neither v_1 nor v_2 can push the process into the interior of $\mathbb{S}$, we see that $\hat{Z}(\tau^*) \notin \mathbb{S}^o$. Define $s^* \doteq \hat{T}(\tau^*)$. Since τ^* is a point of increase of $\hat{T}$ we have $T(s^*) = T(\hat{T}(\tau^*)) = \tau^*$. Thus $Z(s^*) = \hat{Z}(\tau^*) \notin \mathbb{S}^o$, where Z is defined by (47). Consider first the case $Z(s^*) \neq 0$; then from (48), $\langle S_2 \rangle$ is strictly increasing at s^*. From this it follows that for all $\delta > 0$ there exists $s_\delta \in [s^*, s^* + \delta]$ such that $\text{dist}(Z(s_\delta), \mathbb{S}) > 0$, i.e., $\text{dist}(\hat{Z}(T(s_\delta)), \mathbb{S}) > 0$. Now since $\mathbb{S}^h \to \mathbb{S}$ and $\hat{Z}^h \to \hat{Z}$ as $h \to 0$ we have $\text{dist}(\hat{Z}^h(T(s_\delta)), \mathbb{S}^h) > 0$ for all h small enough. Therefore, by definition of $\hat{\tau}_1^h$ we must have $\hat{\tau}_1^h \leq T(s_\delta)$ for all h small enough. This implies $\hat{\tau}_1 \leq T(s_\delta)$. Now, taking $\delta \to 0$ and using the right continuity of T at s^* it follows that $\hat{\tau}_1 \leq T(s^*) = \tau^*$. Hence the quantity in (57) is equal to 0.

Finally, consider the case when $Z(s^*) = 0$ (and $\hat{\tau} < \infty$). Let $s^{**} \doteq \inf\{s > s^* | Z(s^*) \neq 0\}$. From the dynamics of Z (see (47)) it follows that for every $\delta > 0$, there exists $s_\delta \in [s^{**}, s^{**} + \delta]$ such that $\text{dist}(Z(s_\delta), \mathbb{S}) > 0$. Arguing as before, we have $\hat{\tau}_1 \leq T(s^{**})$. Define $m_t(d\alpha) \doteq \hat{m}_{T(t)}(d\alpha)$ for $t \geq 0$. Since $C(t) = 0$ for $t \in [s^*, s^{**}]$ we get that $m_t = \delta_0$ for t in this interval. Thus since $f(0) = 0$, we have $\int_{[0,p]} f(\alpha) dm_t(d\alpha) = 0$ for all $t \in [s^*, s^{**}]$. Combining this with the fact that $[\tau^*, \hat{\tau}_1] \subset [T(s^*), T(s^{**})]$ we now see that the expression in (57) is 0. Thus

$$\begin{aligned} &\mathbb{E}\left[1_{\{\hat{\tau}<\infty\}} \int_{[0,p]\times[0,\hat{\tau}_1]} e^{-\beta \hat{T}(t)} f(\alpha) \hat{m}(d\alpha, dt)\right] \\ &= \mathbb{E}\left[1_{\{\hat{\tau}<\infty\}} \int_{[0,p]\times[0,\hat{\tau}]} e^{-\beta \hat{T}(t)} f(\alpha) \hat{m}(d\alpha, dt)\right]. \end{aligned} \tag{58}$$

Combining (55), (56), and (58) we now get

$$\limsup_{h\to 0} J^h(z_h, U^h) \leq \mathbb{E} \int_{[0,\hat{\tau}]} e^{-\beta \hat{T}(t)} \left(\int_{[0,p]} f(\alpha) d\hat{m}_t(d\alpha) \right) d\hat{T}(t).$$

We next consider the time inversion. Recall that $\tau \doteq \inf\{t : Z(t) \notin \mathbb{S}^o\}$. Note that $\tau \geq \hat{T}(\hat{\tau})$. Using this inequality and Lemma 4.3 we have

$$\mathbb{E} \int_{[0,\hat{\tau}]} e^{-\beta \hat{T}(t)} \left(\int_{[0,p]} f(\alpha) d\hat{m}_t(d\alpha) \right) d\hat{T}(t) \leq \mathbb{E} \int_{[0,\tau]} e^{-\beta t} \left(\int_{[0,p]} f(\alpha) dm_t(d\alpha) \right) dt.$$

Inequality (53) now follows from the above inequality and Remark 5.7. □

We now proceed to the proof of the reverse inequality

$$\liminf_{h\to 0} V^h(z_h) \geq V_{\ell,p}(z). \tag{59}$$

We begin with the following lemma which allows us to restrict our attention to controls that have several simplifying features. The proof is contained in the appendix. Recall the definition of an admissible control above Proposition 2.2 and the corresponding cost defined above (9).

LEMMA 5.9. *Let Φ be a probability system and $U \in \mathcal{A}_p(\Phi, z)$ be a control with corresponding cost function $J_\ell(z, U)$. Then for any $\delta > 0$ there exists $U_\delta \in \mathcal{A}_p(\Phi, z)$ such that $|J_\ell(z, U) - J_\ell(z, U_\delta)| < \delta$ and U_δ satisfies the following:*

1. *There exists $T \in (0, \infty)$ such that $M_\delta(t) = M_\delta(T), N_\delta(t) = N_\delta(T)$, and $C_\delta(t) = 0$ for all $t \geq T$.*

2. *There exists* $L \in (0,\infty)$ *such that*

$$\sup_{0\le t<\infty}\sup_{\omega}(M_\delta(t\wedge\tau,\omega)+N_\delta(t\wedge\tau,\omega))\le L.$$

3. *There exist* $\eta,\theta\in(0,\infty)$ *and* $K\in\mathbb{N}$ *such that* $C(t),M(t),N(t)$ *take values in the finite set* $\{k\eta : k=0,1,2,\dots,K\}$. *Furthermore,* $C,M,$ *and* N *are piecewise constant with possible time points of change being* $\{0,\theta,2\theta,3\theta,\dots\}$.
4. *There exists a* $\gamma\in(0,\infty)$ *such that* θ *is an integer multiple of* γ *and the chosen control* $U=(C,M,N)$ *satisfies the following equality for* $m\ge 1$*:*

$$\begin{aligned}&\mathbb{P}[(C(m\theta),\delta M(m\theta),\delta N(m\theta))=k\eta|U(s),s<m\theta;W(s),s\le m\theta]\\&\quad=\mathbb{P}[(C(m\theta),\delta M(m\theta),\delta N(m\theta))=k\eta|U(n\theta),n<m;W(l\gamma),l\gamma\le m\theta],\end{aligned}\tag{60}$$

where $k=(k_1,k_2,k_3)$ *and* k_1,k_2,k_3 *are integers, at most one of which is nonzero.*
5. *Denoting for* $m\ge 1$, $\Psi(m)\doteq\{C(n\theta),\delta M(n\theta),\delta N(n\theta),n<m\}$, $\mathcal{W}(m)\doteq\{W(l\gamma),l\gamma\le m\theta\}$, *and* $k\doteq(k_1,k_2,k_3)$, *rewrite the above probability as*

$$\begin{aligned}&\mathbb{P}[C(m\theta)=k_1\eta,\delta M(m\theta)=k_2\eta,\delta N(m\theta)=k_3\eta|\Psi(m),\mathcal{W}(m)]\\&\quad\doteq q_{m,k}(\Psi(m),z,\mathcal{W}(m)).\end{aligned}\tag{61}$$

Denote $\mathbb{P}[U(0)=(k_1\eta,k_2\eta,k_3\eta)]$ *by* $q_{0,k}(z)$. *For each* $m\ge 0$, *the function* $q_{m,k}$ *can be chosen so that the function* $(z,w)\to q_{m,k}(\psi,z,w)$ *is continuous for every* ψ.

Construction of asymptotically near optimal admissible controls for the MDP. Fix a probability system Φ, $z\in\mathbb{S}_\ell$, and a sequence $\{z_h\}$ such that $z_h\in\mathbb{S}^h$ and $z_h\to z$ as $h\to 0$. Let $\epsilon>0$ be arbitrary. Let $U\in\mathcal{A}_p(\Phi,z)$ be such that U satisfies properties 1–5 of Lemma 5.9 and $V_{\ell,p}(z)-\epsilon\le J_\ell(z,U)$. For each $h>0$, we construct from U an admissible control sequence $\{U_n^h,n\ge 0\}$ for the MDP in Definition 3.1 with initial condition z_h such that the cost for U^h asymptotically agrees with the cost of U. We outline below the main steps in the construction of such a control sequence. Let $\mathcal{K}\doteq\{(k_1,k_2,k_3) : k_i=0,1,\dots,K; i=1,2,3$ such that at most one of k_1,k_2,k_3 is positive$\}$.

Step 1. We begin by taking a random draw, denoted by $\kappa=(\kappa_1,\kappa_2,\kappa_3)$, from the probability distribution $\{q_{0,k}(z_h),k\in\mathcal{K}\}$. Set $\tilde{U}_0^h=\kappa\eta$, $Z_0^h=z_h$, and $\Psi^h(1)\doteq\tilde{U}_0^h$. Also set $\tilde{n}_0=0$. Note that at most one of κ_2 and κ_3 will be nonzero. If both κ_2 and κ_3 are 0, set $n_1=0$, skip Step 2 below and go to Step 3. Otherwise proceed to Step 2.

Step 2(A). Recall the kernel p_h defined in (21). If $\kappa_2>0$, let $U_0^h=(1,0)$ and take a random draw, denoted by Z_1^h, from $p_h(Z_0^h,U_0^h,d\tilde{z})$. We express this as "the chain takes a purchase control step." If $Z_1^h\in\partial_{\mathrm{R}}^h$, we set $U_1^h=(3,0)$ and draw Z_2^h from $p_h(Z_1^h,U_1^h,d\tilde{z})$; i.e., the chain takes a reflection step. Otherwise set $U_1^h=(1,0)$ and draw Z_2^h from $p_h(Z_1^h,U_1^h,d\tilde{z})$. Define (U_n^h,Z_n^h), $n=1,2,\dots$, recursively by taking either a purchase control step or, if needed, a reflection step, until a total of $[\kappa_2\eta/h]$ purchase control steps have been taken. Denote by n_1 the index of the state after the last purchase control has been exercised.

Step 2(B). If $\kappa_3>0$, let $U_0^h=(2,0)$ (that is, the chain takes a sales control step) and proceed as in Step 2(A) above, alternating sales control steps and reflection

steps (when needed) until $[\kappa_3\eta/h]$ sales control steps have been taken. Again, let n_1 denote the index of the state after the last sales control has been exercised.

Step 3. If $Z^h_{n_1} \in \partial^h_{\mathrm{R}}$, set $U^h_{n_1} = (3,0)$, and the chain takes a reflection step. Otherwise set $U^h_{n_1} = (0, \kappa_1\eta)$ and draw $Z^h_{n_1+1}$ from $p_h(Z^h_{n_1}, U^h_{n_1}, d\tilde{z})$; i.e., the chain takes a diffusion step with $c = \kappa_1\eta$. Let t^h_n be as defined below (23). Define $(U^h_n, Z^h_{n+1}, n = n_1, n_1+1, n_1+2, \ldots, \tilde{n}_1 - 1$ recursively, where $\tilde{n}_1 \doteq \inf\{n : t^h_n \geq \theta\}$, as follows. If $Z^h_n \in \partial^h_{\mathrm{R}}$, set $U^h_n = (3,0)$; otherwise set it to be $(0, \kappa_1\eta)$. Draw Z^h_{n+1} from $p_h(Z^h_n, U^h_n, d\tilde{z})$.

Step 4. Next we define the "pre-Wiener process" that is needed to obtain the control at the next step. Let $\{\nu_n, n \geq 1\}$ be an independent, identically distributed (i.i.d.) sequence of $N(0,1)$ random variables, independent of $(U^h_n, Z^h_{n+1})_{n=0}^{\tilde{n}_1-1}$. Define S^h_n for $n \leq \tilde{n}_1 - 1$ as in (26); here we consider only the second component $S^h_{n,2}$. Set $\mathcal{S}^h_0 \equiv 0$, and for $\tilde{n}_0 < n \leq \tilde{n}_1 - 1$,

$$\mathcal{S}^h_n \doteq \frac{S^h_{n+1,2} - S^h_{n,2}}{\sigma Y^h_n} 1_{\{|Y^h_n| \neq 0\}} + \nu_n \sqrt{\Delta^h_n} 1_{\{|Y^h_n| = 0\}}.$$

Next define $W^h_{\tilde{n}_0} = 0$ and $W^h_n \doteq W^h_{\tilde{n}_0} + \sum_{i=0}^{n-1} \mathcal{S}^h_i$. Now define for $0 \leq t \leq \theta$,

$$W^h(t) \doteq W^h_{n^h(t)} - W^h_{\tilde{n}_0}. \tag{62}$$

Finally, define $\mathcal{W}^h(1) \doteq \{W^h(l\gamma), l \in \mathbb{N}_0, l\gamma \leq \theta\}$.

Step 5. Suppose we have, for $j = 1, \ldots, m$, defined $\tilde{n}_j \doteq \inf\{n : t^h_n \geq j\theta\}$; (Z^h_{n+1}, U^h_n), $n = 0, 1, \ldots, \tilde{n}_j - 1$; $\Psi^h(j)$; and $\mathcal{W}^h(j)$. Consider now the case $j = m+1$. Take a random draw, denoted once more by $\kappa = (\kappa_1, \kappa_2, \kappa_3)$, from $\{q_{m,k}(\Psi^h(m), z_h, \mathcal{W}^h(m)),\ k \in \mathcal{K}\}$. Set $\tilde{U}^h_m = \kappa\eta$ and $\Psi^h(m+1) = (\tilde{U}^h_0, \ldots, \tilde{U}^h_m)$. Follow Step 2 with $\tilde{U}^h_0$ replaced by $\tilde{U}^h_m$ and the starting index of U^h replaced with $\tilde{n}_m$. Denote by n_{m+1} the index of the state obtained after the last singular control step in Step 2. Follow Step 3 with n_1 replaced by n_{m+1}. Let $\tilde{n}_{m+1} = \inf\{n : t^h_n \geq (m+1)\theta\}$. This defines $(Z^h_{n+1}, U^h_n), i = 0, 1, \ldots, \tilde{n}_{m+1} - 1$. Define $W^h(t) - W^h(m\theta)$, for $t \in [m\theta, (m+1)\theta)$, by the right side of (62) as in Step 4 with $\tilde{n}_0, \tilde{n}_1$ replaced by $\tilde{n}_m, \tilde{n}_{m+1}$, respectively. Now set $\mathcal{W}^h(m+1) \doteq \{W^h(l\gamma), l \in \mathbb{N}, l\gamma \leq (m+1)\theta\}$.

Noting that $\tilde{n}_m$ is strictly increasing in m, we obtain the controlled chain $\{(Z^h_n, U^h_n)$, $n = 0, 1, 2, \ldots\}$ via the recursion:

$$(\{(Z^h_{n+1}, U^h_n)\}_{n=0}^{\tilde{n}_m-1}, \Psi^h(m), \mathcal{W}^h(m)) \to (\{(Z^h_{n+1}, U^h_n)\}_{n=0}^{\tilde{n}_{m+1}-1}, \Psi^h(m+1), \mathcal{W}^h(m+1)).$$

The main step in the proof of (59) is showing that if interpolated processes (Z^h, U^h) using the above control sequence are defined as below (25) and W^h is defined by (62), then as $h \to 0$,

$$(Z^h, U^h, W^h) \text{ converges weakly in } D([0,\infty) : \mathbb{R}^5) \text{ to } (Z, U, W), \tag{63}$$

where W is a standard Brownian motion and Z is defined by (7) with the initial condition $Z(0-) = z$. This convergence is established by proving convergence over the period $[j\theta, (j+1)\theta)$ for each j in a recursive manner. Note that given the initial condition $Z(j\theta-) = z$ and the control value $U(j\theta) = k\eta$, the dynamics of Z for $t \in [j\theta, (j+1)\theta)$ are particularly simple and are given as

$$\begin{aligned} X(t) &= x + (1-\mu)\kappa_3\eta - (1+\lambda)\kappa_2\eta + \int_{j\theta}^t (rX(s) - \kappa_1\eta)ds, \\ Y(t) &= y + \kappa_2\eta - \kappa_3\eta + \int_{j\theta}^t bY(s)ds + \int_{j\theta}^t \sigma Y(s)dW(s). \end{aligned} \tag{64}$$

The following lemma provides the convergence of (Z^h, U^h, W^h) over one fixed period $[j\theta, (j+1)\theta)$ given the initial data at $j\theta$. The proof follows via straightforward weak convergence arguments and thus is omitted.

LEMMA 5.10. *Fix $z \in \mathbb{S}_\ell$ and let $k = (k_1, k_2, k_3) \in \mathcal{K}$. Let $(Z(t), W(t))$ given on some probability system Φ be defined for $t \in [0, \theta]$ by (7) with $Z(0-) = z$ and $(C(t), M(t), N(t)) = k\eta$ for $t \in [0, \theta]$. Consider a sequence $\{z_h\}$ such that $z_h \in \mathbb{S}^h$ and $z_h \to z$ as $h \to 0$. Define $\tilde{n}_1$ and the sequence $\{U_n^h, Z_n^h\}_{n=0}^{\tilde{n}_1}$ via Steps 2 and 3 and $\{S_n^h, W_n^h\}_{n=0}^{\tilde{n}_1}$ by Step 4. Let $\{\delta_h\}$ be a sequence of nonnegative reals such that $\delta_h \to 0$ as $h \to 0$. Define, for $t \in [0, \theta]$, the interpolated process (Z^h, W^h, U^h, E^h) as before (see (62) and below (26)) with the change that $\Delta^h(Z_0^h, U_0^h)$ is replaced with $\delta_h + \Delta^h(Z_0^h, U_0^h)$. Denote the laws of (Z^h, W^h, U^h) and (Z, W, U) on $D([0, \theta] : \mathbb{R}^5)$ by Π_h^{k,δ_h} and Π^k, respectively. Then $\Pi_h^{k,\delta_h} \to \Pi^k$ as $h \to 0$.*

In the following proposition we show that the cost, $J^h(z_h, U^h)$, corresponding to the above constructed control sequence, converges to $J_\ell(z, U)$ as $h \to 0$. The desired inequality in (59) then follows since $V^h(z_h) \geq J^h(z_h, U^h)$, $J_\ell(z, U) \geq V_{\ell,p}(z) - \epsilon$,and $\epsilon > 0$ is arbitrary.

PROPOSITION 5.11. *Let $\epsilon > 0$ be arbitrary and fix $z \in \mathbb{S}_\ell$. Let Φ be a probability system and $U \equiv U(\epsilon) \in \mathcal{A}_p(\Phi, z)$ be such that U satisfies properties 1–5 of Lemma 5.9 and $V_{\ell,p}(z) - \epsilon \leq J_\ell(z, U)$. Let, for each $h > 0$, $\{U_n^h\}$ be an admissible control sequence as constructed via Steps 1–5 above. Then $J^h(z_h, U^h) \to J_\ell(z, U)$ as $h \to 0$, and consequently (59) holds.*

Proof. For $t \geq 0$, let $n^h(t), Z^h(t), M^h(t), N^h(t), C^h(t)$ be as defined below (25). Define $U^h \equiv (M^h, N^h, C^h)$ and let W^h be as in (62). We begin by establishing (63). Define for $j \in \mathbb{N}_0$ and $t \in [0, \theta)$,

$$U_j^h(t) \doteq U^h(t + j\theta),\ Z_j^h(t) \doteq Z^h(t + j\theta),\ W_j^h(t) \doteq W^h(t + j\theta) - W^h(j\theta),$$

and set $(U_j^h(\theta), Z_j^h(\theta), W_j^h(\theta)) = (U_j^h(\theta-), Z_j^h(\theta-), W_j^h(\theta-))$. Define processes U_j, Z_j, W_j, $j \in \mathbb{N}_0$, in a similar manner. Recall the sequence $\{U_j^h\}$ constructed above Lemma 5.10 and let $\zeta_j^h \doteq (\tilde{U}_j^h, U_j^h, W_j^h, Z_j^h)$, $\zeta_j \doteq (\tilde{U}_j, U_j, W_j, Z_j)$, where $\tilde{U}_j \doteq (C(j\theta), \delta M(j\theta), \delta N(j\theta))$, $j \in \mathbb{N}_0$. Due to the piecewise constant feature of the control U, in order to prove (63), it suffices to show that

$$\text{for all } n \in \mathbb{N}_0,\ \Upsilon_h^n \doteq \{\zeta_j^h\}_{j=0}^n \text{ converges weakly to } \Upsilon \doteq \{\zeta_j\}_{j=0}^n \text{ as } h \to 0. \tag{65}$$

We will prove (65) via induction (on n). The case $n = 0$ is immediate from Lemma 5.10 and continuity of the kernel $q_{0,k}$ on noting that for $k \in \mathcal{K}$ and $E \in \mathcal{B}(D([0, \theta], \mathbb{R}^5)$,

$$\mathbb{P}(\tilde{U}_0^h = k\eta, (U_0^h, W_0^h, Z_0^h) \in E) = q_{0,k}(z_h)\Pi_h^{k,0}(E),$$
$$\mathbb{P}(\tilde{U}_0 = k\eta, (U_0, W_0, Z_0) \in E) = q_{0,k}(z)\Pi^k(E).$$

Suppose now that (65) holds for $n = 0, \ldots, m$ and consider the case $n = m + 1$. Denote the law of Υ_n^h, Υ_n by ϖ_n^h and ϖ, respectively. By the induction hypothesis, $\varpi_m^h \to \varpi_m$ as $h \to 0$. Furthermore, ϖ_{m+1}^h can be expanded in terms of ϖ_m^h as follows:

$$d\varpi_{m+1}^h(\upsilon_{m+1}) = \sum_{k \in \mathcal{K}} q_{m+1,k}(z_h, \tilde{u}^{*,m}, w^{*,m}) \Pi_h^{k,\delta_h(\upsilon_m)}(\varsigma_{m+1}) d\varpi_m^h(\upsilon_m), \tag{66}$$

where $\upsilon_m = \{\varsigma_j\}_{j=0}^m$; $\varsigma_j = (\tilde{u}_j, u_j, w_j, z_j)$; $\tilde{u}_j = k\eta, k \in \mathcal{K}$; $(u_j, w_j, z_j) \in D([0, \theta] : \mathbb{R}^5)$; $\tilde{u}^{*,m} = \{\tilde{u}_j\}_{j=0}^m$; $w^{*,m} = \{w_j(l\gamma), l \in \mathbb{N}, l\gamma \leq \theta\}_{j=0}^m$; and δ_h is a measurable map from the state space of Υ_m to $[0, 1]$ satisfying $0 \leq \delta_h(\upsilon_m) \leq \tilde{\Delta}_*^h$.

From the continuity properties of the kernel $\{q_{m+1,k}\}$ and the weak convergence of ϖ_m to ϖ, we have for all continuous and bounded functions F_1, F_2 defined on suitable spaces, as $h \to 0$,

$$\int F_1(\upsilon_m) \sum_{k\in\mathcal{K}} q_{m+1,k}(z, \tilde{u}^{*,m}, w^{*,m}) \left(\int F_2(\zeta_{m+1}) d\Pi^k(\zeta_{m+1}) \right) d\varpi_m^h(\upsilon_m)$$

$$\to \int F_1(\upsilon_m) \sum_{k\in\mathcal{K}} q_{m+1,k}(z, \tilde{u}^{*,m}, w^{*,m}) \left(\int F_2(\zeta_{m+1}) d\Pi^k(\zeta_{m+1}) \right) d\varpi_m(\upsilon_m). \tag{67}$$

Next, from Lemma 5.10, for all sequences $\{\delta_h\}$ converging to 0 and compact sets E (of Euclidean space of appropriate dimension), as $h \to 0$,

$$\sup_{k\in\mathcal{K}, \tilde{u}^{*,m}\in\eta\mathcal{K}^{m+1}, w^{*,m}\in E} \left| \int F_2(\zeta_{m+1}) q_{m+1,k}(z_h, \tilde{u}^{*,m}, w^{*,m}) d\Pi_h^{k,\delta_h}(\zeta_{m+1}) \right.$$

$$\left. - \int F_2(\zeta_{m+1}) q_{m+1,k}(z_h, \tilde{u}^{*,m}, w^{*,m}) d\Pi^k(\zeta_{m+1}) \right| \to 0. \tag{68}$$

The weak convergence of ϖ_{m+1}^h to ϖ_{m+1} now follows on combining (66), (67), and (68). This proves (63).

We now address convergence of the cost functions. First, let T be as in Lemma 5.9. Recall $\eta^h = \inf\{n \geq 0 : Z_n^h \in \partial^h\}$ and $\tau^h = t_{\eta^h}^h$. Note that $\tau^h = \inf\{t \geq 0 : Z^h(t) \in \partial^h\}$ due to the piecewise constant nature of $Z^h(t)$.

Let $\tau = \inf\{t \geq 0 : Z(t) \notin \mathbb{S}^o\}$. It can be shown in a manner similar to that used in the proof of Lemma 5.1 that $\tau^h \to \tau$ as $h \to 0$ on the set $\{\tau = \infty\}$. Also, on the set $\{\tau < \infty\}$ for every $\delta > 0$ there exists $t \in [\tau, \tau + \delta)$ and $\epsilon > 0$ such that $\mathrm{dist}(Z(t), \mathbb{S}) > \epsilon$. Furthermore, $|Z^h(t) - Z(t)|$ uniformly on $[0, T]$ and $\partial^h \to \partial\mathbb{S}$ as $h \to 0$. Together these three facts imply $\tau^h \wedge T \to \tau \wedge T$ as $h \to 0$. Therefore, since $(z_h, Z^h, U^h, W^h, \tau^h) \to (z, Z, U, W, \tau)$, by the dominated convergence theorem,

$$J^h(z_h, U^h) = \mathbb{E} \int_{[0,\tau^h\wedge T)} e^{-\beta t} f(C^h(t)) dt \to \mathbb{E} \int_{[0,\tau\wedge T)} e^{-\beta t} f(C(t)) dt = J_\ell(z, U). \qquad \square$$

Combining Corollary 2.5 and Propositions 5.8 and 5.11 we have the following theorem.

THEOREM 5.12. *Let $z \in \mathbb{S}$ and let $\{z_h, h > 0\}$ be a sequence with $z_h \in \mathbb{S}_\ell^h$ such that $z_h \to z$ as $h \to 0$. Then $\lim_{p\to\infty} \lim_{\ell\to\infty} \lim_{h\to 0} V^h(z^h) = \lim_{p\to\infty} \lim_{\ell\to\infty} V_{\ell,p}(z) = V(z)$.*

6. Computational methods for the MDP. The convergence results in the previous section ensure that for small values of h, the MDP defined in section 3 provides a good approximation to the diffusion control problem defined in section 2. In this section we outline the numerical methods for solving the MDP. Specifically, we provide the algorithm through which we compute the value function (25) and the associated optimal control for each initial state $z_h \in \mathbb{S}_\ell^h$. In practice, we fix a value of h and use the associated MDP to provide approximations to the diffusion control problem. Thus, for the remainder of the section, we will take h as a fixed value and suppress it in the notation.

Specifying the controlled Markov chain. In section 3, we specified a choice of transition probabilities and interpolation intervals which satisfy the local consistency criteria; see (15)–(17) and (20). Many variations of this choice are possible; when

specifying the particular controlled Markov chain, consideration must be given to the numerical implementation. For example, note that neither the denominators of the probabilities nor the interpolation intervals in (15) depend on the value of c. This was accomplished by allowing the self-transition (x, y) to (x, y). Also, we have separated the pure diffusion effects from the effects of the consumption control. That is, as consumption always decreases wealth, we associate it with only the transition from (x, y) to $(x - h, y)$. Recall that $\Delta(z, u) = 0$ for all z if $u = (3, c)$; that is, the interpolation interval is 0 if reflection occurs. Hence, using (24), a reflection step incurs no cost and thus $V(\ell + h, y) = V(\ell, y), V(x, \ell + h) = V(x, \ell)$. It is a consequence of Definition 3.1 that $\mathbb{P}[I_n = 3 | Z_n \in \partial_{\mathrm{R}} \setminus \partial^h] = 1$ for all n; that is, reflection is the only admissible action for states in the reflecting boundary. Therefore, by adjusting the transition probabilities associated with the diffusion and singular controls, it is possible to eliminate states in the reflecting boundary without affecting the cost function. This modification helps in speeding up the convergence of the numerical scheme.

In what follows, we will assume that the reflecting boundary states have been eliminated and the appropriate adjustments to the transition probabilities made. Thus, the state space of the controlled Markov chain used in the numerical schemes is given by $\mathbb{S}_{\mathrm{MDP}} \doteq \mathbb{S}_\ell^{h+} \setminus \partial_{\mathrm{R}}$, and the control space is $\mathcal{U}_{\mathrm{MDP}} \doteq \{0, 1, 2\} \times [0, p]$.

Dynamic programming equation. Let $U \equiv \{U_n, n = 0, 1, 2, \dots\}$ be an admissible control sequence (see Definition 3.1) for the MDP with state space $\mathbb{S}_{\mathrm{MDP}}$, control space $\mathcal{U}_{\mathrm{MDP}}$, and initial state z. For the numerical methods it is convenient to work with the cost function

$$J(z, U) = \mathbb{E} \sum_{n=0}^{\eta-1} e^{-\beta t_n} f(C_n) \tilde{\Delta}(Z_n, C_n) 1_{\{I_n = 0\}}, \tag{69}$$

which is asymptotically equivalent to (24). Recall that the value function is given as $V(z) = \sup_{U \in \mathcal{A}(z)} J(z, U)$.

We now present the dynamic programming equation that characterizes the value function. We begin by introducing the class of feedback controls. A *feedback control* is a measurable function $\mathbf{u} : \mathbb{S}_{\mathrm{MDP}} \to \mathcal{U}_{\mathrm{MDP}}$. We write $\mathbf{u} = (\mathbf{i}, \mathbf{c})$, where $\mathbf{i}$ and $\mathbf{c}$ are the two coordinates of the function $\mathbf{u}$. Using such a function one can construct an admissible control pair (Z_n, U_n) recursively by setting $Z_0 = z_0$, $U_n = \mathbf{u}(Z_n), n \geq 0$, and

$$\mathbb{P}[Z_{n+1} \in \cdot | Z_0, \dots, Z_n, U_0, \dots, U_n] = p(Z_n, U_n, \cdot).$$

With an abuse of terminology we will refer to this sequence $\{U_n\}$ as a feedback control as well. Note that $U_n \equiv (I_n, C_n) = (\mathbf{i}(Z_n), \mathbf{c}(Z_n))$.

If $U = \{U_n, n = 0, 1, 2, \dots\}$ is a feedback control, then one can easily check that the pair (Z_n, U_n) is a Markov chain, from which it follows that for all $z \in \mathbb{S}_{\mathrm{MDP}}$,

$$J(z, U) = \sum_{\tilde{z} \in \mathbb{S}_{\mathrm{MDP}}} r(z, \mathbf{u}(z), \tilde{z}) J(\tilde{z}, U) + f(\mathbf{c}(z)) \Delta^h(z, \mathbf{u}(z)), \tag{70}$$

where $r(z, \mathbf{u}(z), \tilde{z}) = e^{-\beta \Delta(z, \mathbf{u}(z))} p(z, \mathbf{u}(z), \tilde{z})$. Observing that $J(z, U) = 0$ for all $z \in \partial^h$, the summation above can be taken over $\tilde{z} \in \mathbb{S}^* \doteq \mathbb{S}_{\mathrm{MDP}} \setminus \partial^h$.

We can write the above equality in matrix form as follows. Let $|\mathbb{S}^*| = s$ and fix an ordering of all the states in $\mathbb{S}^*$, i.e., $\mathbb{S}^* = \{z_1, \dots, z_s\}$. Let $F(\mathbf{u})$ be an $s \times 1$ vector

whose ith entry is $f(\mathbf{c}(z_i))\Delta(z_i,\mathbf{u}(z_i)), i=1,\dots,s$. Let $R(\mathbf{u})$ be the $s\times s$ matrix with the (i,j)th entry as $r(z_i,\mathbf{u}(z_i),z_j)$. Finally, let $J(\mathbf{u})$ be the $s\times 1$ vector with the ith entry being $J(z_i,U)$. Then using these matrices, (70) can be written as

$$J(\mathbf{u}) = R(\mathbf{u})J(\mathbf{u}) + F(\mathbf{u}). \tag{71}$$

Next, for $u\in\mathcal{U}_{\mathrm{MDP}}$ let $R(u)$ be the $s\times s$ matrix with (i,j)th entry $r(z_i,u,z_j)$. From standard arguments (cf. section 5.8 in [19]) it follows that the value function V satisfies the following dynamic programming equation:

$$V = \sup_{u\in\mathcal{U}_{\mathrm{MDP}}} R(u)V + F(u), \tag{72}$$

where in the above equation V is interpreted as an $s\times 1$ vector whose ith entry is $V(z_i)$, and the supremum on the right-hand side above is taken row by row.

The following contraction property is central in the characterization of the value function via the dynamic programming equation in (72). The proof of the following lemma relies on the fact that the cost is of the discounted form with a strictly positive discount factor at all diffusion control steps, and although the discount is zero for singular control steps, such steps tend to push the process towards the boundary of the domain and thus cannot occur "too often."

LEMMA 6.1. *For all feedback controls* $\mathbf{u}$, $R^n(\mathbf{u})\to 0$ *as* $n\to\infty$.

Proof. Let $\mathbf{u}$ be a feedback control and denote $R^n(\mathbf{u})$ by simply R^n with entries r^n_{ij}, $i,j=1,\dots,s$. It suffices to show $\sum_{j=1}^s r^n_{ij}\to 0$ as $n\to\infty$, for each $i=1,\dots,s$.

Let (Z_n,U_n) be the controlled Markov chain associated with feedback control $\mathbf{u}$ and a transition kernel as defined in section 3 with the modifications discussed in this section, and let $\Delta_k\equiv\Delta(Z_k,U_k)$ be the associated interpolation intervals. Let $\eta\doteq\inf\{n: Z_n\in\partial^h\}$. A simple calculation yields for all $i=1,\dots,s$:

$$\sum_{j=1}^s r^n_{ij} = \mathbb{E}_i\left[e^{-\beta\sum_{k=0}^{n-1}\Delta_k}\sum_{j=1}^s 1_{\{Z_n=z_j\}}\right],$$

where $\mathbb{E}_i$ denotes the expectation given that $Z_0=z_i$.

Since the states in ∂^h are not included in $\mathbb{S}^*$ and $p(z,\mathbf{u}(z),z)=1$ for $z\in\partial^h$, we have $1_{\{Z_n=z_j\}}=0$ for $j=1,\dots,s$ when $n\ge\eta$. Thus we have

$$\sum_{j=1}^s r^n_{ij} = \mathbb{E}_i\left[1_{\{n<\eta\}}\sum_{j=1}^s e^{-\beta\sum_{k=0}^{n-1}\Delta_k}1_{\{Z_n=z_j\}}\right]. \tag{73}$$

Fix $a\in\mathbb{Z}_+$; conditions on a will be specified later. Define

$$\tilde{d}\doteq\#\{\theta\in\{1,2,\dots,[n/a]\}: U_m=(0,\cdot)\text{ for some } m\in[(\theta-1)a,\theta a)\}.$$

Set $\tilde{d}'=[n/a]-\tilde{d}$. The integer a is used to group the steps of the chain from 1 to n together into intervals. The quantity $\tilde{d}$, ($\tilde{d}'$) counts the number of such intervals with at least one diffusion step (respectively, no diffusion steps). By (14) there is a $\delta>0$ such that $\Delta_n\ge\delta$ for all diffusion steps (i.e., all n such that $U_n=(0,\cdot)$). Also, recall that $\Delta_n=0$ if step n is not a diffusion step. Combining these observations, we have

$$\mathbb{E}_i\left[1_{\{n<\eta\}}e^{-\beta\sum_{k=0}^{n-1}\Delta_k}\sum_{j=1}^s 1_{\{Z_n=z_j\}}\right] \le e^{-\beta\frac{\delta}{4}[\frac{n}{a}]} + \mathbb{E}_i\left[1_{\{n<\eta\}}1_{\{\tilde{d}'>\frac{3}{4}[\frac{n}{a}]\}}\right]. \tag{74}$$

We will utilize the behavior of the singular controls to bound the second term on the right-hand side of the line above. Let $E_1, E_2, \ldots, E_{\tilde{d}'}$ denote the intervals containing no diffusion steps, each of size a. Let K_d denote the number of purchase control steps in E_d; then $a - K_d$ is the number of sales control steps in E_d. Due to the finiteness of the state space, the maximum number of successive transitions to the left is bounded; in particular, it is bounded by $B \doteq 2(\ell/h+1)$. Similarly, B is a bound on the maximum number of downward transitions in a row. From (17) we see that each sales control always pushes the chain downward. Thus the application of too many sales controls in a row will cause the chain to hit the boundary. However, by (16) a purchase control potentially pushes the chain upward. Similarly, a purchase control always pushes the chain to the left, while a sales control has the potential to push the chain to the right. Thus in order to avoid hitting the boundary, the number of sales controls must be properly balanced by the number of purchase controls. More precisely, if $n < \eta$, we must have $|K_d - (a - K_d)| < B$; that is, $(a-B)/2 < K_d < (a+b)/2,\ d = 1, \ldots, \tilde{d}'$. For $m \in E_d$ define $\tilde{L}_m \doteq 1_{\{Z_{m+1} - Z_m = (-h,0)', U_m = (1,0)\}}$. The random variable $\tilde{L}_m$ indicates if the chain moves strictly to the left at step m given that a purchase control is applied. Since on E_d there are no diffusion steps and movement to the left is possible only at purchase control steps, the number of increments δZ_k equal to $(-h, 0)'$ on E_d is given by $\sum_{m \in E_d} \tilde{L}_m \doteq L_d$.

Let ϵ be chosen to satisfy $0 < \epsilon < p/2$. Recall that at a purchase control step the chain moves to the left with probability $q \doteq \lambda/(1+\lambda)$. Thus by Cramer's theorem (see Theorem 2.1.24 in [8]) there exists a $\kappa \equiv \kappa(\epsilon)$ such that

$$\mathbb{P}[L_d < K_d(q-\epsilon)] \le \mathbb{P}\left[\left|\frac{L_d}{K_d} - q\right| > \epsilon\right] \le e^{-K_d\kappa} \le e^{-\kappa(a-B)/2}, \tag{75}$$

where the last inequality follows from the bound on K_d. We claim that for each $d = 1, \ldots, \tilde{d}'$, $\{\eta < n\} \cap \{L_d > qK_d/2\} = \emptyset$. To see the claim suppose that $L_d > qK_d/2$. Then the number of upward steps ($\delta Z_k = (\cdot, h)'$) in E_d, given by $K_d - L_d$, is at most $(1 - q/2)K_d$. The number of downward steps ($\delta Z_k = (\cdot, -h)'$) in E_d equals the number of sales control steps, $a - K_d$. Thus using the bounds on K_d, we have

$$\#\{\text{down steps in } E_d\} - \#\{\text{up steps in } E_d\} \ge a - K_d - (1-q/2)K_d \ \ge\ aq/4 - (1+q/4)B,$$

which is greater than B for $a > B(4+q)/q$. Henceforth fix such an a. On the other hand, on the set $\{\eta < n\}$ we must have $|\#\{\text{down steps in } E_d\} - \#\{\text{up steps in } E_d\}| < B$; otherwise, the chain would hit the boundary. This leads to a contradiction and thus the claim holds. Combining this with (75) we have that

$$\begin{aligned}
\mathbb{E}_i\left[1_{\{n<\eta\}} 1_{\{\tilde{d}' > \frac{3}{4}[\frac{n}{a}]\}}\right] &= \mathbb{E}_i\left[1_{\{n<\eta, \tilde{d}' > \frac{3}{4}[\frac{n}{a}]\}} \prod_{d=1}^{\tilde{d}'} 1_{\{L_d \le qK_d/2\}}\right] \\
&\le \mathbb{E}_i\left[\prod_{d=1}^{\frac{3}{4}[\frac{n}{a}]} 1_{\{L_d \le qK_d/2\}}\right] \\
&\le e^{-\kappa \frac{a-B}{2} \frac{3}{4}[\frac{n}{a}]}.
\end{aligned}$$

Finally, by (73), (74), and the above, we have $\sum_{j=1}^{s} r_{ij}^n \le e^{-\beta \frac{\delta}{4}[\frac{n}{a}]} + e^{-\kappa \frac{a-B}{2} \frac{3}{4}[\frac{n}{a}]}$. The result now follows on noting that the term on the right approaches 0 as $n \to \infty$. □

An immediate consequence of the lemma (cf. section 2.3 of [19]) is the following.

THEOREM 6.2. *For any feedback control* $\mathbf{u}$, $J(\mathbf{u})$ *is the unique solution to the equation* $v = R(\mathbf{u})v + F(\mathbf{u})$. *Furthermore, the value function* $\{V(z), z \in \mathbb{S}^*\}$ *defined below* (69) *is the unique solution of* (72). *Denoting the* arg max *for the ith row maximization on the right side of* (72) *by* $\mathbf{u}(z_i)$ *and the control sequence corresponding to the feedback control* $\mathbf{u}$ *by* $U = \{U_n, n = 0, 1, 2, \dots\}$, *we have that* U *is an optimal control, i.e.,* $J(z, U) = V(z)$ *for all* $z \in \mathbb{S}^*$.

From the above theorem it follows that in order to compute the value function and the optimal control it suffices to solve (72).

Numerical methods. We will use classical iterative methods to find the optimal control by solving the dynamic programming equation (72). A sketch of the algorithm is provided here. Details can be found in Chapter 6 of [19].

The following theorem provides the basis for the numerical approximation of the optimal control. We refer the reader to Theorem 6.2.1 in [19] for a proof.

THEOREM 6.3. *Let* $\mathbf{u}_0$ *be a feedback control. Define a sequence of feedback controls* $\{\mathbf{u}_n, n \geq 1\}$ *and costs* $\{J(\mathbf{u}_n), n \geq 1\}$ *recursively as follows. Given* $\mathbf{u}_n$, *define*

$$J(\mathbf{u}_n) = R(\mathbf{u}_n)J(\mathbf{u}_n) + F(\mathbf{u}_n), \tag{76}$$

$$\mathbf{u}_{n+1} \doteq \arg\max_{u \in \mathcal{U}_{\mathrm{MDP}}} R(u)J(\mathbf{u}_n) + F(u), \tag{77}$$

where the arg max *on the right-hand side is computed row by row. Then* $J(\mathbf{u}_n) \to V$ *as* $n \to \infty$.

Given some control, (77) provides a way of "updating" the control in the search for the optimal control. However, this requires solving (76) to obtain the cost associated with the given control. Finding an exact solution to this equation can be numerically intensive since it involves the inversion of an $s \times s$ matrix. Thus we use instead an approximation to the cost function $J(\mathbf{u}_n)$ in (77). The following theorem provides a method for obtaining such an approximation. We refer the reader to Theorem 6.2.2 in [19] for the proof.

THEOREM 6.4. *Let* $\mathbf{u}$ *be an admissible feedback control. Then for any initial* $s \times 1$ *vector* $\tilde{J}_0$, *the sequence defined recursively by*

$$\tilde{J}_{n+1} = R(\mathbf{u})\tilde{J}_n + F(\mathbf{u}) \tag{78}$$

converges to $J(\mathbf{u})$.

The numerical method for finding the optimal control is obtained by combining Theorems 6.3 and 6.4 as follows.

Policy iteration: Having determined an approximation to $J(\mathbf{u}_n)$, denoted as $\tilde{J}(\mathbf{u}_n)$, one obtains $\mathbf{u}_{n+1}$ by solving the minimization problem in (77) by replacing $J(\mathbf{u}_n)$ there with $\tilde{J}(\mathbf{u}_n)$.

Value iteration: Given $\mathbf{u}_n$, iterate (78) a large number of times (say, m) with $R(\mathbf{u})$ there replaced by $R(\mathbf{u}_n)$ and initial value $\tilde{J}_0$ replaced by $\tilde{J}(\mathbf{u}_{n-1})$. Set $\tilde{J}(\mathbf{u}_n) \doteq \tilde{J}_m$.

The numerical algorithm alternates between policy iterations and value iterations until some suitable stopping criterion is met. Several modifications of (78) are often used to improve numerical efficiency; see section 6.2.4 of [19] for details.

7. Numerical study. We now present the results of a small pilot study using the method described in section 6. We consider one of the examples in [31]. As in that reference, we set $r = 0.07$, $b = 0.12$, $\sigma = 0.40$, and $\beta = 0.10$. We consider the case $\lambda = \mu = 0.01$ and the utility function $f(c) = 2\sqrt{c}$. We take $\ell = 10$ as in [31] and $p = 10$. The discretization parameter is taken to be $h = 0.25$. (Note that [31] uses $h = 0.025$.)

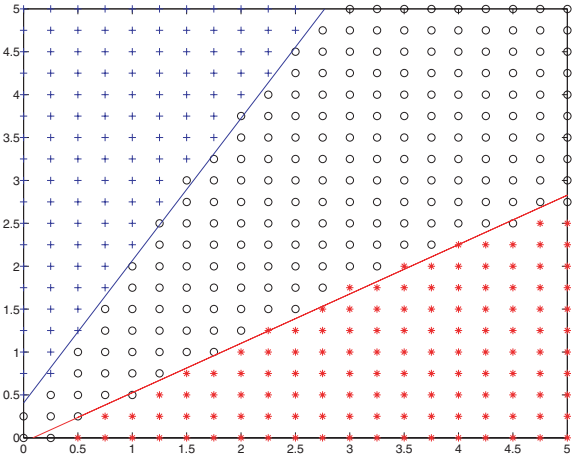


FIG. 2. *Numerically computed optimal control.*

To implement the numerical algorithm, we choose an initial feedback control matrix $\mathbf{u}_0$ given by, for $z \in \mathbb{S}^*$,

$$\mathbf{u}_0(z) = \begin{cases} (0,p), & x \geq 0, y \geq 0, \\ (1,0), & x \geq 0, y < 0, \\ (2,0), & \text{otherwise.} \end{cases}$$

Based on this control, the no-transaction region is the first quadrant of $\mathbb{R}^2$, and we always exercise the maximum amount of possible consumption. For $z \in \mathbb{S}^*$ we take $\tilde{J}_0(z)$ to be 75% of the value function computed in the absence of transaction costs; see equation (2.5) in [6].

We ran the algorithm described in the previous section. Figure 2 displays the first quadrant of the state space and illustrates the optimal control for this region. We see that the no-transaction region looks roughly like a cone. Consumption states are represented by the circles, purchase states by the plus signs, and sales states by the asterisks. The estimated boundaries of the no-transaction region (the solid lines in the figure) are given by the lines $y = 0.575x - 0.050$ (the boundary of the "buy" region) and $y = 1.659x + 0.405$ (the boundary of the "sell" region). The estimated sell boundary of the no-transaction region is similar to that obtained by Tourin and Zariphopoulou (see Figure 1 in [31]). However, the slope of our buy boundary appears to be lower than the slope illustrated in Figure 1 of [31]. A possible reason for this could be the difference in the discretization parameter. We used $h = 0.25$ to produce the test results provided here. Within the no-transaction region, consumption remains at a fairly constant percentage of wealth, 11.5%, which is very close to the constant percentage of consumption in the case of no-transaction costs (see Theorem 2.1 in [6]). We also compare the value function computed by the algorithm versus the value function in the case of no-transaction costs (again, see Theorem 2.1 in [6]). In general, the optimal value for an initial state computed in the presence of transaction costs is roughly 97% of the optimal value for the same state in the absence of transaction costs. However, when the initial wealth is small, this percentage tends to be lower (roughly 80%–90%).

8. Appendix. This section contains the proofs of Lemmas 4.4, 5.1, and 5.9. We refer the reader to the corresponding sections for the relevant notation and definitions.

Proof of Lemma 4.4. Without loss of generality, assume $h \in (0,1)$. Define

$$Y_i^h(t) \doteq \sum_{k=0}^{n^h(t)\wedge\eta_h-1} \delta Z_k^h 1_{\{I_k^h=i\}}, \quad n_i^h(t) \doteq \sum_{k=0}^{n^h(t)\wedge\eta_h-1} 1_{\{I_k^h=i\}}, \ i=1,2.$$

Writing $Y_i^h \equiv (Y_{i,1}^h, Y_{i,2}^h)'$, it follows from (16) and (17) that

$$\mathbb{E}Y_{1,2}^h(t) = h\frac{1}{1+\lambda}\mathbb{E}[n_1^h(t)], \ \ \mathbb{E}Y_{2,1}^h(t) = h(1-\mu)\mathbb{E}[n_2^h(t)].$$

A straightforward calculation shows $|B^h(t)| \le c_1(1+t)$ and $\mathbb{E}|S^h(t)| \le c_2(1+t)$, where the constants c_1, c_2 are independent of h and t. From (16), (17) we see that $hn_1^h(t) = M^h(t) = Y_{1,1}^h(t)$ and $hn_2^h(t) = N^h(t) = Y_{2,2}^h(t)$. Thus from (28) there is $\tilde{c}_1 \in (0,\infty)$ such that

$$hn_1^h(t) \le \tilde{c}_1(1+t) + |S_1^h(t)| + Y_{2,1}^h(t), \ hn_2^h(t) \le \tilde{c}_1(1+t) + |S_2^h(t)| + Y_{1,2}^h(t). \tag{79}$$

Combining the above inequalities we have, for some $c_3 \in (0,\infty)$, $h\mathbb{E}[n_1^h(t)] \le c_3(1+t) + h(1-\mu)\mathbb{E}[n_2^h(t)]$ and $h\mathbb{E}[n_2^h(t)] \le c_3(1+t) + h\mathbb{E}[n_1^h(t)]/(1+\lambda)$. It follows that $h\mathbb{E}[n_1^h(t)]$ and $h\mathbb{E}[n_2^h(t)]$ are "close" to each other. More precisely, there exist constants $\alpha \ge 1$, $c_4 > 0$, $L_0 > 0$ such that for $L \ge L_0$,

$$h(\mathbb{E}[n_1^h(t)] \vee \mathbb{E}[n_2^h(t)]) > L \Rightarrow h(\mathbb{E}[n_1^h(t)] \wedge \mathbb{E}[n_2^h(t)]) > \alpha L - c_4.$$

In particular, we have $\sup_h h\mathbb{E}[n_1^h(t)] = \infty$ if and only if $\sup_h h\mathbb{E}[n_2^h(t)] = \infty$. Now suppose $\sup_h h\mathbb{E}[n_1^h(t)] = \infty$ and $\sup_h h\mathbb{E}[n_2^h(t)] = \infty$. By Cramer's theorem (see Theorem 2.1.24 in [8]), for all $\delta > 0$ there exists a constant $c(\delta) \in (0,\infty)$ such that for all $k_0 \in \mathbb{N}_0$ and $h > 0$,

$$\max\Big\{\mathbb{P}[|Y_{2,1}^h - h(1-\mu)n_2^h(t)| > \delta h n_2^h(t), n_2^h(t) = k_0], \\ \mathbb{P}[|Y_{1,2}^h - h(1/(1+\lambda))n_1^h(t)| > \delta h n_1^h(t), n_1^h(t) = k_0]\Big\} \le c(\delta)e^{-k_0 c(\delta)}.$$

Choose δ such that $\mu + \delta < 1$ and $1/(1+\lambda) - \delta > 0$ (which is possible since $\mu \in (0,1)$ and $\lambda \in (0,\infty)$). Define $\alpha_1 = 1 - (1-\mu-\delta)(1/(1+\lambda)-\delta) < 1$ and $\theta = \alpha_1/4$. Fix $\epsilon \in (0,1)$ and choose K large enough so that

$$\frac{c(\delta)}{1-e^{-c(\delta)}}e^{-c(\delta)(K+1)} < \frac{\epsilon}{8} \ \text{ and } \ \frac{c_2(1+t)}{\theta K - \tilde{c}_1(1+t)} < \frac{\epsilon}{8}. \tag{80}$$

Since by assumption, $\sup_h h\mathbb{E}[n_1^h(t)] = \sup_h h\mathbb{E}[n_2^h(t)] = \infty$, there exists $h' \le 1$ such that

$$\mathbb{P}\left[n_1^{h'}(t) > \frac{K}{h'}, n_2^{h'}(t) > \frac{K}{h'}\right] > \epsilon. \tag{81}$$

Then for all $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\left[|Y_{2,1}^{h'}(t) - h'(1-\mu)n_2^{h'}(t)| > \delta h' n_2^{h'}(t), n_2^{h'}(t) > \frac{K}{h'}\right] \\
&= \sum_{j=[K/h']+1}^{\infty} \mathbb{P}[|Y_{2,1}^{h'}(t) - h'(1-\mu)n_2^{h'}(t)| > \delta h' n_2^{h'}(t), n_2^{h'}(t) = j] \\
&\leq \sum_{j=[K/h']+1}^{\infty} c(\delta)e^{-c(\delta)j} \quad = \frac{c(\delta)}{1-e^{-c(\delta)}} e^{-c(\delta)([\frac{K}{h'}]+1)} < \frac{\epsilon}{8},
\end{aligned}$$

where the last inequality follows from the choice of K in (80). Similarly, $\mathbb{P}[|Y_{1,2}^{h'}(t) - \frac{h'}{1+\lambda} n_1^{h'}(t)| > \delta h' n_1^{h'}(t), n_1^{h'}(t) > \frac{K}{h'}] < \frac{\epsilon}{8}$. Hence, in view of (81) we have

$$\begin{aligned}
\mathbb{P}\Bigg[&\min\{n_1^{h'}(t), n_2^{h'}(t)\} > \frac{K}{h'}\ ,\ |Y_{2,1}^{h'}(t) - h'(1-\mu)n_2^{h'}(t)| \leq \delta h' n_2^{h'}(t), \\
&\left|Y_{1,2}^{h'}(t) - \frac{h'}{1+\lambda} n_1^{h'}(t)\right| \leq \delta h' n_1^{h'}(t)\Bigg] > \frac{\epsilon}{2}.
\end{aligned}$$

Let E denote the event in the equation above. From (79) and (80),

$$\begin{aligned}
\mathbb{P}[h' n_1^{h'}(t) - Y_{2,1}^{h'}(t) \geq \theta K] &\leq \mathbb{P}[|S_1^{h'}(t)| \geq \theta K - \tilde{c}_1(1+t)] \\
&\leq \frac{\mathbb{E}|S_1^{h'}(t)|}{\theta K - \tilde{c}_1(1+t)} \\
&\leq \frac{c_2(1+t)}{\theta K - \tilde{c}_1(1+t)} < \frac{\epsilon}{8}.
\end{aligned}$$

Similarly, $\mathbb{P}[h' n_2^{h'}(t) - Y_{1,2}^{h'}(t) \geq \theta K] < \frac{\epsilon}{8}$. Thus

$$\begin{aligned}
\frac{\epsilon}{2} < \mathbb{P}[E] &\leq \mathbb{P}[E, h' n_1^{h'}(t) - Y_{2,1}^{h'}(t) < \theta K, h' n_2^{h'}(t) - Y_{1,2}^{h'}(t) < \theta K] \\
&\quad + \mathbb{P}[h' n_1^{h'}(t) - Y_{2,1}^{h'}(t) \geq \theta K] + \mathbb{P}[h' n_2^{h'}(t) - Y_{1,2}^{h'}(t) \geq \theta K] \\
&\leq \mathbb{P}[\tilde{E}] + \frac{\epsilon}{8} + \frac{\epsilon}{8},
\end{aligned}$$

where $\tilde{E}$ is the event in the first term on the right side above. It follows that $\mathbb{P}[\tilde{E}] > \epsilon/4$ and thus $\tilde{E}$ is nonempty. Now for any $\omega \in \tilde{E}$, we have from the definition of $\tilde{E}$ that

$$h' n_1^{h'}(t) - h'(1-\mu-\delta)n_2^{h'}(t) < \theta K, \ \ h' n_2^{h'}(t) - h'(1/(1+\lambda) - \delta) n_1^{h'}(t) < \theta K.$$

A straightforward calculation using these inequalities shows that for such ω,

$$h' n_1^{h'}(t) \leq \frac{2\theta}{1 - (1-\mu-\delta)(1/(1+\lambda) - \delta)} K = \frac{K}{2}.$$

However, this contradicts the fact that $h' n_1^{h'}(t) > K$ on $\tilde{E}$. Thus we must have $\sup_h h\mathbb{E}[n_1^h(t)] < \infty$, $\sup_h h\mathbb{E}[n_2^h(t)] < \infty$. The result now follows on recalling that $M^h(t) = h n_1^h(t)$ and $N^h(t) = h n_2^h(t)$. □

Proof of Lemma 5.1. The inequality $V(z) \leq \bar{V}(z)$ is immediate since every exact control can be expressed as a relaxed control. Consider now the reverse inequality.

Let Φ be a probability system and $\bar{U} = (\bar{m}, M, N) \in \bar{\mathcal{A}}(\Phi, z)$ be such that $\bar{V}(z) \leq \bar{J}(z, \bar{U}) + \epsilon$. From the boundedness of the cost function it follows that, without loss of generality, we can assume that there is a $T \in (0, \infty)$ such that $M(t) = M(t \wedge T)$ and $N(t) = N(t \wedge T)$ for all $t \in (0, \infty)$, and $\bar{m}_t(d\alpha) = \delta_p$ for all $t \geq T$. Also, T can be chosen large enough so that $f_*(p)e^{-\beta T}/\beta < \epsilon$.

Let Z be defined via (7) with $C(t) \doteq \int_{[0,p]} \alpha \bar{m}_t(d\alpha)$, and τ as before. Then

$$\bar{J}(z, \bar{U}) \leq \mathbb{E} \int_{[0,p]\times[0,T\wedge\tau)} e^{-\beta t} f(\alpha) m(d\alpha, dt) + \epsilon. \tag{82}$$

Also, by modifying m, M, and N if needed, we can assume that

$$M(t) = M(T \wedge \tau), \;\; N(t) = N(T \wedge \tau), \text{ and } \bar{m}_t(d\alpha) = \delta_p \text{ for all } t \geq T \wedge \tau. \tag{83}$$

Following the proof of Theorem 1.2.1 in [2] one can show that there exists a sequence of exact controls $C_n \in \mathcal{A}_p(\Phi, z)$ which satisfy

$$\sup_{0\leq t\leq T_1} \left| \int_{[0,p]\times[0,t]} e^{-\beta t} f(\alpha) m(d\alpha, dt) - \int_{[0,t]} e^{-\beta t} f(C_n(t)) dt \right| \to 0 \text{ a.s., and} \tag{84}$$

$$\sup_{0\leq t\leq T_1} \left| \int_{[0,t]} (C_n(s) - C(s)) ds \right| \to 0 \text{ a.s.} \tag{85}$$

as $n \to \infty$ for all $T_1 \in (0, \infty)$. In fact the cited theorem shows that, for each n, C_n can be chosen such that it takes values in a finite set and there is a sequence $0 < t_1^n < t_2^n \cdots$ such that C_n is constant over $[t_k^n, t_{k+1}^n)$ for all $k \in \mathbb{N}_0$.

Let Z_n be defined via (7) with C replaced by C_n and with M and N as introduced above. A straightforward application of Gronwall's inequality and (85) shows that for each $T_1 \in (0, \infty)$, there is a $c \equiv c(T_1) \in (0, \infty)$ such that

$$\sup_{0\leq t\leq T_1} \mathbb{E}|Z_n(t) - Z(t)| \leq c \sup_{0\leq t\leq T_1} \mathbb{E} \left| \int_{[0,t]} (C_n(s) - C(s)) ds \right|. \tag{86}$$

Hence $Z_n \to Z$, in probability, uniformly on $[0, T]$.

If $\tau = \infty$, $Z(t) \in \mathbb{S}^o$ for all $t \geq 0$. Thus, (86) implies that there exists N_0 such that if $n > N_0$, then $Z_n(t) \in \mathbb{S}^o$ for all $t \geq 0$, and therefore $\tau_n = \infty$ for all $n > N_0$. Then clearly $\tau_n \to \tau$ a.s. as $n \to \infty$ on the set $\{\tau = \infty\}$. Next note that, a.s. on the set $\{\tau < \infty\}$ and for every $\delta > 0$, there exist $t \in [\tau, \tau + \delta)$ and $\epsilon > 0$ such that $\text{dist}(Z(t), \mathbb{S}) > \epsilon$. This is because, in view of (83), $Z(\tau + t)$, $t \geq 0$, is described via (7) with initial condition $Z(\tau)$ and $M \equiv N \equiv 0$. If $Z(\tau) = 0$, the property is satisfied trivially since $C(t) = p$ for all $t \geq \tau$. Otherwise, the property follows from a standard argument based on the law of the iterated logarithm for Brownian motion (cf. pages 260–261 in [19]). This, along with the convergence of Z_n to Z, shows that $\tau_n \doteq \inf\{t : Z_n(t) \notin \mathbb{S}^o\}$ converges to τ a.s. as $n \to \infty$. In proving this statement we also use the observation that if $Z(t-) \notin \mathbb{S}$, then $Z(t) \notin \mathbb{S}$. Thus $\tau_n \wedge T \to \tau \wedge T$. Combining these observations with (84), we obtain

$$\mathbb{E} \int_{[0,\tau_n\wedge T]} e^{-\beta t} f(C_n(t)) dt \to \mathbb{E} \int_{[0,p]\times[0,\tau\wedge T]} e^{-\beta t} f(\alpha) m(d\alpha, dt). \tag{87}$$

The result now follows on using this observation in (82). □

Proof of Lemma 5.9. Let $\epsilon > 0$ be arbitrary and let $T \in (0,\infty)$ be such that $f_*(p)e^{-\beta T}/\beta < \epsilon$. Consider $\tilde{U} = (\tilde{C}, \tilde{M}, \tilde{N})$ given by $\tilde{C}(t) = C(t)1_{t<T}$, $\tilde{M}(t) = M(t\wedge T)$, and $\tilde{N}(t) = N(t\wedge T)$, $t \geq 0$. Clearly $\tilde{U} \in \mathcal{A}_p(z)$, and it is easy to check that $|J_\ell(z,U) - J_\ell(z,\tilde{U})| < \epsilon$. This proves 1. Henceforth we will assume, without loss of generality, that 1 holds for U in the statement of the lemma. Using (7) and the bounds on the state and control space, it is easy to show $\sup_{0\leq t\leq T}[M(t\wedge\tau,\omega)+N(t\wedge\tau,\omega)] \leq c_1 + c_2 \sup_{0\leq t\leq T}|W(t)|$, where c_1 and c_2 are nonnegative constants that may depend on T. Let $L \in (0,\infty)$ be large enough so that $c_2\mathbb{E}\sup_{0\leq t\leq T}|W(t)|/(L-c_1) < \epsilon$. Define $\tilde{U}$ by $\tilde{C} \equiv C$, $\tilde{M}(t) \doteq M(t)\wedge L$, $\tilde{N}(t) \doteq N(t)\wedge L$. Let $\tilde{Z}$ be the corresponding controlled process and $\tilde{\tau}$ the corresponding hitting time. Let T be as in part 1, and define $A \doteq \{\sup_{0\leq t\leq T}[M(t\wedge\tau)+N(t\wedge\tau)] < L\}$. Then

$$J_\ell(z,\tilde{U}) = \mathbb{E}\left[1_A \int_{[0,\tilde{\tau}]} e^{-\beta t} f(\tilde{C}(t))dt\right] + \mathbb{E}\left[1_{A^c}\int_{[0,\tilde{\tau}]} e^{-\beta t} f(\tilde{C}(t))dt\right]. \tag{88}$$

Using the bound on f, the choice of L, and Markov's inequality, the second term on the right side of the above inequality is bounded by ϵ. Also, since on the set A, $M(t) < L$ and $N(t) < L$ for all $t \leq \tau\wedge T$, we have that the evolution of $\tilde{Z}$ is the same as that of Z. Therefore $\tilde{\tau}$ in the first expression on the right side of (88) can be replaced with τ. This shows that $|J_\ell(z,\tilde{U}) - J_\ell(z,U)| \leq 2\epsilon$ and hence 2 follows.

We now consider 3. Let $U \equiv (C,M,N)$ be an admissible control satisfying properties 1 and 2 and let Z be the solution to (7) under (C,M,N) defined on some probability system. Following Theorem 1.2.1 of [2] (see comments below (85)) we can assume without loss of generality that C takes values in a finite set, is RCLL, and piecewise constant with finitely many points of change over $[0,T]$. We also assume without loss of generality (by modifying controls if needed) that $M(t) = M(t\wedge\tau)$, $N(t) = N(t\wedge\tau)$, and $C(t) = C(t)1_{t<\tau} + p1_{t\geq\tau}$. Fix $\eta,\theta \in (0,\infty)$ and define the piecewise constant processes $C_{\eta,\theta}$, $M_{\eta,\theta}$, and $N_{\eta,\theta}$ as follows. For $m = 0$, define $M_{\eta,\theta}(m\theta) = M_{\eta,\theta}(0) = k\eta$ if $M(0) \in [k\eta, k\eta+\eta)$. For $m \geq 1$, set $\delta M_{\eta,\theta}(m\theta) = k\eta$ if $M(m\theta) - M(m\theta-\theta) \in [k\eta, k\eta+\eta)$. By property 2, we need only consider the finite set $\{k\eta : k = 0,1,2,\ldots,K\}$, where K is some positive integer. Then let $M_{\eta,\theta}(t) = M_{\eta,\theta}(m\theta)$ for $t \in [m\theta, m\theta+\theta)$. Define $N_{\eta,\theta}$ analogously based on N. Define $C_{\eta,\theta}(m\theta) = k\eta$ if $C(m\theta) \in [k\eta, k\eta+\eta)$ and $C_{\eta,\theta}(t) = C_{\eta,\theta}(m\theta)$ for $t \in [m\theta, m\theta+\theta)$. The constructed process $U_{\eta,\theta} \equiv (C_{\eta,\theta}, M_{\eta,\theta}, N_{\eta,\theta})$ is an admissible control. Let $Z_{\eta,\theta}$ denote the solution to (7) under this control, defined on some probability system, and let $\tau_{\eta,\theta}$ denote the first time this process exits $\mathbb{S}^o_\ell$. Choose a sequence (η_k,θ_k) such that $\eta_k,\theta_k \to 0$ as $k\to\infty$. Denote Z_{η_k,θ_k} by Z_k. Similar abbreviations are used for $U_{\eta_k,\theta_k}, \tau_{\eta_k,\theta_k}$. One can easily check that $(Z_k,U_k) \to (Z,U)$ in $D([0,T],\mathbb{R}^2\times[0,p]\times[0,L]\times[0,L])$ (in probability) as $k\to\infty$. If $\tau = \infty$, the uniform convergence $Z_k \to Z$ implies that there exists K_0 such that for all $k > K_0$ we have $Z_k(t) \in \mathbb{S}^o$ for all $t \geq 0$, and thus $\tau_k = \infty$ for all $k > K_0$. Therefore $\tau_k \to \tau$ a.s. as $k\to\infty$ on the set $\{\tau = \infty\}$. Next note that, a.s. on the set $\{\tau < \infty\}$, for every $\delta > 0$ there exists $t \in [\tau, \tau+\delta)$ and $\epsilon > 0$ such that $\mathrm{dist}(Z(t),\mathbb{S}) > \epsilon$. This is because, on this set, by our choice of U, $Z(\tau+t)$, $t \geq 0$, is described via (7) with $M \equiv N \equiv 0$ and initial condition $Z(\tau)$. In the case $Z(\tau) = 0$ the property is satisfied trivially since $C(t) = p$ for all $t \geq \tau$. Otherwise, the property follows from an argument analogous to the proof of Theorem 9.4.3 of [19] (see pp. 260–261). Next, recalling that $Z_k \to Z$, $Z(t)$ is continuous for all $t \geq \tau$ and the observation that $Z(\tau-) \in (\mathbb{S}^o_\ell)^c \Rightarrow Z(\tau) \in (\mathbb{S}^o_\ell)^c$,

we conclude $\tau_k \wedge T \to \tau \wedge T$ in probability. The convergence of $J(z, U_k)$ to $J(z, U)$ now follows. This proves 3.

The proofs of 4 and 5 are quite standard and we provide only a sketch; the reader is referred to the proof of Theorem 10.3.1 of [19] (pages 285–287) for details. Assume that U satisfies properties 1–3, and let $\gamma > 0$. Part 4 is essentially a consequence of the martingale convergence theorem on noting that the σ-fields $\mathcal{G}_\gamma \doteq \sigma\{U(n\theta), n < m; W(l\gamma), l\gamma \leq, \theta\}$ increase to the σ-field $\mathcal{G} \doteq \sigma\{U(n\theta), n < m; W(s), s \leq, \theta\}$ as $\gamma \downarrow 0$. The main idea is to define controls U_γ and controlled processes Z_γ recursively over intervals $[m\theta, (m+1)\theta)$ by using the right side of (60) in defining the law of U_γ over $[m\theta, (m+1)\theta)$. Proving the weak convergence of (Z_γ, C_γ) to (Z, C) is straightforward. The convergence of hitting times is argued as in the proof of part 3. Finally, part 5 is proved by convolving $q_{m,k}$, defined in (61), in the (z, w) variables by a parametrized family of mollifiers and arguing weak convergence of the resulting controlled pair to (Z, C) as the mollifying parameter approaches a suitable limit. Convergence of stopping times is argued, once more, as in 3. □

Acknowledgments. The authors would like to thank the anonymous referees for their suggestions, which have greatly improved the presentation of the paper.

REFERENCES

[1] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., John Wiley & Sons, Inc., New York, 1999.
[2] V. S. Borkar, *Control of Diffusion Processes*, John Wiley & Sons, Inc., New York, 1989.
[3] A. Budhiraja and A. Ghosh, *Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function*, Ann. Appl. Probab., to appear.
[4] P.-L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.
[5] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N. S.), 27 (1992), pp. 1–67.
[6] M. H. A. Davis and A. R. Norman, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676–713.
[7] P. Dupuis and H. Ishii, *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics/Stochastics Rep., 35 (1991), pp. 31–62.
[8] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston, 1993.
[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley & Sons, Inc., New York, 1986.
[10] B. G. Fitzpatrick and W. H. Fleming, *Numerical methods for an optimal investment-consumption model*, Math. Oper. Res., 16 (1991), pp. 823–841.
[11] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.
[12] J. M. Harrison, *Brownian models of queueing networks with heterogeneous customer population*, in Stochastic Differential Systems, Stochastic Control Theory, and Their Applications, Springer, New York, 1988, pp. 147–186.
[13] J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302–308.
[14] A. Hindy, C. Huang, and S. Zhu, *Numerical analysis of a free-boundary singular control problem in financial economics*, J. Econom. Dynam. Control, 21 (1997), pp. 297–327.
[15] H. Ishii and S. Koike, *Boundary regularity and uniqueness for an elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 8 (1983), pp. 317–346.
[16] I. Karatzas, J. Lehoczky, S. Sethi, and S. Shreve, *Explicit solution of a general investment-consumption problem*, Math. Oper. Res., 11 (1986), pp. 261–294.
[17] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[18] T. G. Kurtz, *Random time changes and convergence in distribution under the Meyer-Zheng conditions*, Ann. Probab., 19 (1991), pp. 1010–1034.

[19] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer-Verlag, New York, 2001.
[20] H. J. Kushner and L. F. Martins, *Numerical methods for stochastic singular control problems*, SIAM J. Control Optim., 29 (1991), pp. 1443–1475.
[21] L. F. Martins and H. J. Kushner, *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209–1233.
[22] L. F. Martins, S. E. Shreve, and H. M. Soner, *Heavy traffic convergence of a controlled, multi-class queuing system*, SIAM J. Control Optim., 34 (1996), pp. 2133–2171.
[23] H. Liu, *Optimal consumption and investment with transaction costs and multiple risky assets*, J. Finance, 59 (2004), pp. 289–338.
[24] M. J. P. Magill and G. M. Constantinides, *Portfolio selection with transaction costs*, J. Econom. Theory, 13 (1976), pp. 245–263.
[25] R. C. Merton, *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3 (1971), pp. 373–413.
[26] J. L. Menaldi and M. Robin, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.
[27] P. A. Meyer and W. A. Zheng, *Tightness criteria for laws of semimartingales*, Ann. Inst. H. Poincaré Probab. Statist., 20 (1984), pp. 353–372.
[28] P. Protter, *Stochastic Integration and Differential Equations*, 2nd ed., Springer-Verlag, New York, 2004.
[29] S. E. Shreve and H. M. Soner, *Optimal investment and consumption with transaction costs*, Ann. Appl. Probab., 4 (1991), pp. 609–692.
[30] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton–Jacobi equations*, J. Differential Equations, 56 (1985), pp. 345–390.
[31] A. Tourin and T. Zariphopoulou, *Viscosity solutions and numerical schemes for investment/consumption models with transaction costs*, in Numerical Methods in Finance, Cambridge University Press, Cambridge, 1997, pp. 245–269.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2207–2223

PERFORMANCE RECOVERY IN DIGITAL IMPLEMENTATION OF ANALOGUE SYSTEMS*

GUOFENG ZHANG†, TONGWEN CHEN‡, AND XIANG CHEN§

Abstract. In this paper, the *generalized* bilinear transformation (GBT) is proposed. Compared with the traditional bilinear, zero-order hold (ZOH) and first-order hold transformations, one advantage of GBT is that it may convert unstable poles (zeros) to stable poles (zeros). It is proved that controllability and observability are invariant under GBT. After that, it is shown that the performance of a sampled-data system obtained via GBT approaches that of the analogue system as the underlying sampling period goes to zero. Performance studied here is characterized in terms of internal stability and ℓ_p induced norms for all $1 \leq p \leq \infty$. This results extends the main results in [G. Zhang and T. Chen, *Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, Suppl.*, 2003, pp. 28–33] and [G. Zhang and T. Chen, *Automatica J. IFAC*, 40 (2004), pp. 327–330] from SISO to MIMO and also removes the limitation on the "A" matrix of the system. Finally, an example is employed to compare digital implementations via GBT and the ZOH transformation.

Key words. fractional-order hold transformation, generalized bilinear transformation, graph metric, Hankel singular values, performance recovery, unstable zeros

AMS subject classifications. 34K20, 34K25, 46E15, 47A07, 93B03, 93B07

DOI. 10.1137/050643416

1. Introduction.

Sampling a continuous-time system is a fundamental problem in a variety of scientific areas, such as computer control, system identification, and signal processing. It is becoming even more conspicuous in light of the huge success of computer-aided processing and networking [1], [2], [6], [13], [22], [23], [18], [24], [35], [17]. There are many intriguing problems related to sampling. This paper focuses on the problem of performance recovery in digital implementation of analogue systems. Consider the continuous-time closed-loop system Σ_1 shown in Figure 1.1, where W is a finite-dimensional linear time-invariant (FDLTI) stable block, and both G and K are FDLTI. Suppose that Σ_1 is internally stable [6, p. 241] and also satisfies some input-output performance specifications. Now conduct a digital implementation as in Figure 1.2 by using K_h obtained via some discretization method, with h being the underlying sampling period. For convenience, we denote the system in Figure 1.2 by Σ_2. Note that Σ_2 is a sampled-data control system. Within this set-up, the following question could be raised: Under what condition will Σ_2 also be internally stable and have similar input-output properties as those of Σ_1?

This problem has been studied extensively in recent years. In Chen and Francis [5], [6], an upper bound for the sampling period h is derived, for which the sampled-data control system Σ_2 is ℓ_p input-output stable; furthermore, it is shown that performance of Σ_2 converges to that of Σ_1 as h approaches zero. The digital implementation

*Received by the editors October 24, 2005; accepted for publication (in revised form) August 4, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/sicon/45-6/64341.html

†School of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang, China 310038 (gfzhang@hdu.edu.cn).

‡Department of Electrical and Computer Engineering, University of Alberta Edmonton, Alberta T6G 2V4, Canada (tchen@ece.ualberta.ca). The research of this author was partially supported by an NSERC Discovery grant.

§Corresponding author. Department of Electrical and Computer Engineering, University of Windsor, Windsor, Ontario N9B 3P4, Canada (xchen@uwindsor.ca). The research of this author was partially supported by an NSERC Discovery grant.

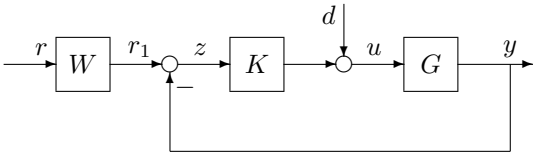


FIG. 1.1. Σ_1: *analogue system.*

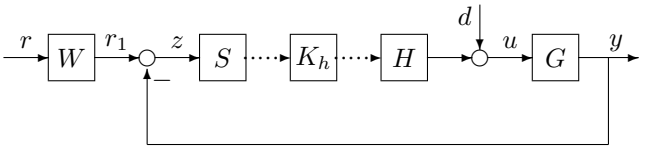


FIG. 1.2. Σ_2: *digital implementation of* Σ_1.

of an analogue controller will generally degrade the performance of the closed-loop system. Two examples are given in [21] to illustrate how this degradation is quantified as a function of the underlying sampling period. Some sufficient conditions that guarantee uniform-in-time input-output performance convergence of sampled-data control systems are proposed in [28]. A new discretization method is proposed in [16] based on the principle of controller approximation, where an upper bound is derived to ensure closed-loop stability. Quite contrary to the above, Oishi [25] illustrates via concrete examples that, in some circumstances, even as the sampling period tends to zero, the best sampled-data closed-loop performance may not necessarily converge to the best analogue closed-loop performance.

All the above results are derived for digital implementation of analogue systems via the step-invariant transformation (the zero-order hold (ZOH) equivalent). Let K_d and K_{bt} denote two different versions of K_h, obtained via the step-invariant and bilinear transformations respectively. By showing that the ℓ_p induced norm of $K_d - K_{bt}$ approaches zero for all $1 \le p \le \infty$ as h goes to zero, Zhang and Chen [38] and [39] proved that performance recovery of digital implementation of Σ_1 via the bilinear transformation still holds. However, it is assumed in [38] and [39] that G in Figure 1.1 is a single-input–single-output (SISO) stable block, and furthermore, the "A" matrix is diagonalizable with all real eigenvalues. Actually, the result holds for $p = \infty$ no matter whether G is multi-input–multi-output (MIMO) or G is diagonalizable with all real eigenvalues. One of the main purposes of this paper is to generalize the results in [38] and [39] to the cases of $1 \le p < \infty$. The remainder of this paper is outlined as follows.

In section 2, we propose the generalized bilinear transformation (GBT) that contains the (traditional) bilinear transformation as a special case. Denoting the resulting digital controller by K_{gbt}, we explore the relation between the poles and zeros of K and K_{gbt}. One advantage of the GBT is that it may be able to convert unstable zeros or poles to stable zeros or poles. This is quite desirable since unstable zeros and poles always impose various limitations on system performance [7], [15], [27], [44], process identification algorithms [19], [30], network communication bandwidth [33], [11], and stability of networked control systems [40]. Then we show that controllability, as well as observability, is preserved under the GBT. More important, the GBT provides a class of discretizations. When control engineers have some practical digital implementation problem at hand, instead of only the ZOH, first-order hold,

or traditional bilinear transformations, now they have infinite many choices at their disposal. Therefore, for a specified control performance, he may select a particular discretization method by tuning the parameter in the GBT. Clearly, this new degree of freedom will enhance the capability of a control engineer.

In section 3, we investigate the limiting behavior of K_{gbt}. More specifically, we prove the following result: Given K stable, the ℓ_p induced norm of $K_{gbt} - K_d$ approaches zero for all $1 \leq p \leq \infty$ as the underlying sampling period h tends to zero. This result holds no matter whether K is SISO or MIMO, thus generalizing Theorem 1 in [39]. This result plays a key role in the development in section 4.

Based on the results in section 3, we establish in section 4 that all existing stability and performance results for the step-invariant transformation (the ZOH equivalent) can be translated into those for the GBT. Using the methods proposed in this paper, limiting properties and performance recovery of other types of digital implementations such as the Al-Alaoui transformation can be dealt with in a similar way.

Section 5 is a case study. An example is used to compare digital implementations of analogue systems via the ZOH transformation and the GBT. It demonstrates that GBT converts some unstable system zeros (poles) to stable ones, therefore bringing in a closed-loop system which has much less performance limitations than that using ZOH.

Our main purpose in this paper is to show performance recovery in digital implementations of analogue systems via the GBT in fast sampling. However, it is worthwhile to point out that in some circumstance fast sampling is dangerous. For example, if a quantizer of fixed sensitivity is inserted into a control loop consisting of a system and an unstable controller, the simulation in [3] demonstrates that system performance will grow *unbounded* as the sampling period goes to zero. Hence, in the framework of network-based control, where signal quantization is prevalent, very fast sampling will in general degrade control performance [20]. This understanding warn us that we should be careful in choosing a suitable sampling period when dealing with practical problems.

Finally, some words about notation. The norm symbol $\|\cdot\|$ represents the Euclidean norm if $\cdot$ is a vector, or its largest singular value if $\cdot$ is a matrix; $\|\cdot\|_{\ell_p}$ is the ℓ_p norm if applied to a vector and is the ℓ_p induced norm if applied to a system. Following the convention, for a discrete-time transfer function $H(z)$ with a state-space realization (A, B, C, D), define

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] := D + C(zI - A)^{-1}B.$$

2. GBT. The bilinear transformation is motivated by considering the trapezoidal approximation of an integrator. Given an integrator $1/s$ with input u and output y, the trapezoidal approximation of

$$y(kh + h) = y(kh) + \int_{kh}^{kh+h} u(\tau)d\tau \tag{2.1}$$

is

$$y\left(kh + h\right) = y\left(kh\right) + h\frac{u\left(kh + h\right) + u\left(kh\right)}{2};$$

i.e., the integral is approximated by the average value of $u(kh + h)$ and $u(kh)$. Now we approximate the integral in (2.1) using some other combination of $u(kh + h)$ and

$u(kh)$. More specifically, consider

$$y(kh+h) = y(kh) + h(\alpha u(kh+h) + (1-\alpha)u(kh)), \tag{2.2}$$

where $\alpha \in (0,1)$. The transfer function of (2.2) is (in z transform)

$$\frac{y(z)}{u(z)} = h\frac{\alpha z + (1-\alpha)}{z-1}.$$

This motivates us to introduce the GBT

$$\frac{1}{s} = h\frac{\alpha z + (1-\alpha)}{z-1},$$

that is,

$$s = \frac{1}{h}\frac{z-1}{\alpha z + (1-\alpha)}. \tag{2.3}$$

Therefore, under the generalized bilinear transformation, the continuous-time transfer function $K(s)$ in Figure 1.1 is mapped to $K_{gbt}(z)$, where

$$K_{gbt}(z) = K\left(\frac{1}{h}\frac{z-1}{\alpha z + (1-\alpha)}\right). \tag{2.4}$$

In terms of state-space data, take a minimal realization for $K(s)$, namely, (A,B,C,D), and assume $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$; it is straightforward to derive that $K_{gbt}(z)$ has a state-space model $(A_{gbt}, B_{gbt}, C_{gbt}, D_{gbt})$ with

$$A_{gbt} = (I-\alpha hA)^{-1}(I+(1-\alpha)hA), \quad B_{gbt} = (1-\alpha)(I-\alpha hA)^{-1}hB,$$
$$C_{gbt} = C\left(I + \frac{\alpha}{1-\alpha}A_{gbt}\right), \quad D_{gbt} = D + \frac{\alpha}{1-\alpha}CB_{gbt}.$$

When $\alpha = 1/2$, we recover the traditional bilinear transformation. In what follows, without loss of generality, it is assumed that the matrix B is of full column rank, i.e., $\operatorname{rank}(B) = m$.

(Conversely, given a discrete-time system $K_{gbt}(z)$ with state-space data $(A_{gbt}, B_{gbt}, C_{gbt}, D_{gbt})$, one can also get a continuous-time system K, one of whose state-space realizations is

$$A = -\Omega(I - A_{gbt}),\ B = \Omega B_{gbt},\ C = C_{gbt}(I + \alpha h\Omega(I - A_{gbt})),$$
$$D = D_{gbt} - C_{gbt}\alpha h\Omega B_{gbt},$$

where $\Omega = ((1-\alpha)hI + \alpha hA_{gbt})^{-1}$.)

In what follows we study properties of the generalized bilinear transformation. Let us first look at poles and invariant zeros. Equation (2.3) yields

$$z = \frac{1+(1-\alpha)hs}{1-\alpha hs}. \tag{2.5}$$

Hence, if s is an eigenvalue of the matrix A, then z in (2.5) is that of A_{gbt}. The same is true for invariant zeros [43]. Actually we have a stronger conclusion.

THEOREM 2.1. *If s is an invariant zero of K, then z in* (2.5) *is an invariant zero of K_{gbt}. Furthermore, if $D = 0$, then $z = -\frac{1-\alpha}{\alpha}$ is also a zero of K_{gbt}.*

Proof. We begin with the first part. Let s be an invariant zero of K. Then there exist $0 \neq x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ such that

$$\begin{aligned} (A - sI)\, x + Bu &= 0, \\ Cx + Du &= 0. \end{aligned} \tag{2.6}$$

It suffices to show that

$$\begin{aligned} (A_{gbt} - zI)\, \tilde{x} + B_{gbt}\tilde{u} &= 0, \\ C_{gbt}\tilde{x} + D_{gbt}\tilde{u} &= 0, \end{aligned} \tag{2.7}$$

where

$$\tilde{u} = u, \quad \tilde{x} = \frac{1}{1 + \frac{\alpha}{1-\alpha} z} x. \tag{2.8}$$

In terms of (2.8), the left-hand side of (2.7) is equivalent to

$$\begin{aligned} (A_{gbt} - zI)\, \tilde{x} + B_{gbt}\tilde{u} &= \left((I - \alpha hA)^{-1} (I + (1-\alpha)hA) - zI \right) \tilde{x} \\ &\quad + (1-\alpha) (I - \alpha hA)^{-1} hB\tilde{u} \\ &= (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA - (I - \alpha hA)\, z)\, \tilde{x} + (1-\alpha)hBu \right\}, \\ C_{gbt}\tilde{x} + D_{gbt}\tilde{u} &= C \left(I + \frac{\alpha}{1-\alpha} A_{gbt} \right) \tilde{x} + \left(D + \frac{\alpha}{1-\alpha} C B_{gbt} \right) \tilde{u} \\ &= C\tilde{x} + Du \\ &\quad + \frac{\alpha}{1-\alpha} C (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA)\, \tilde{x} + (1-\alpha)hBu \right\}. \end{aligned} \tag{2.9}$$

Rewriting (2.6) as $Bu = (sI - A)\, x$, $-Cx = Du$, and substituting it into the first equation of (2.9), we have

$$\begin{aligned} &(A_{gbt} - zI)\, \tilde{x} + B_{gbt}\tilde{u} \\ &\quad = (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA - (I - \alpha hA)\, z)\, \tilde{x} + (1-\alpha)h\, (sI - A)\, x \right\} \\ &\quad = (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA)\, \tilde{x} + (1-\alpha)h\, (sI - A)\, x - (I - \alpha hA)\, z\tilde{x} \right\}. \end{aligned}$$

Therefore, we need only to show that

$$(I + (1-\alpha)hA)\, \tilde{x} + (1-\alpha)h\, (sI - A)\, x = (I - \alpha hA)\, z\tilde{x} \tag{2.10}$$

and

$$\begin{aligned} &C\tilde{x} + Du + \frac{\alpha}{1-\alpha} C (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA)\, \tilde{x} + (1-\alpha)hBu \right\} \\ &= C (\tilde{x} - x) + \frac{\alpha}{1-\alpha} C (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA)\, \tilde{x} + (1-\alpha)h\, (sI - A)\, x \right\} = 0. \end{aligned} \tag{2.11}$$

Clearly, $\tilde{x}$ in (2.8) satisfies (2.10). In light of this, the left-hand side of (2.11) is converted to

$$\begin{aligned} &C\tilde{x} + Du + \frac{\alpha}{1-\alpha} C (I - \alpha hA)^{-1} \left\{ (I + (1-\alpha)hA)\, \tilde{x} + (1-\alpha)hBu \right\} \\ &= C (\tilde{x} - x) + \frac{\alpha}{1-\alpha} C (I - \alpha hA)^{-1} (I - \alpha hA)\, z\tilde{x} = C (\tilde{x} - x) + \frac{\alpha}{1-\alpha} Cz\tilde{x} \\ &= C \left(\left(1 + \frac{\alpha}{1-\alpha} z \right) \tilde{x} - x \right) = 0. \end{aligned}$$

In summary, (2.7) holds.

Next, assuming $D = 0$, we prove the second part of the theorem. Since $D = 0$, (2.7) becomes

$$\begin{aligned} (A_{gbt} - zI)\,\tilde{x} + B_{gbt}\tilde{u} &= 0, \\ C\left(I + \frac{\alpha}{1-\alpha}A_{gbt}\right)\tilde{x} + \frac{\alpha}{1-\alpha}CB_{gbt}\tilde{u} &= 0, \end{aligned} \tag{2.12}$$

which yields

$$C\left[1 + \frac{\alpha}{1-\alpha}z\right]\tilde{x} = 0. \tag{2.13}$$

Clearly, $z = -\frac{1-\alpha}{\alpha}$ satisfies (2.13). Substitution of it into (2.12) gives rise to

$$\begin{aligned} (A_{gbt} - zI)\tilde{x} + B_{gbt}\tilde{u} &= 0, \\ \tilde{x} &= zhB\tilde{u}. \end{aligned} \tag{2.14}$$

Because B is of full column rank, the dimension of $span(B)$, namely, the space spanned by the column vectors of B, is m. Hence, there exist vectors $\tilde{u}_1, \ldots, \tilde{u}_m$ in $\mathbb{R}^m$ such that $B\tilde{u}_1, \ldots, B\tilde{u}_m$ constitute a basis of $span(B)$. Define $\tilde{x}_i$ via the second equation of (2.14), $i = 1, \ldots, n-m$. Then $(\tilde{x}_i, \tilde{u}_i)$, $(i = 1, \ldots, n-m)$ are solutions to (2.14) for $z = -\frac{1-\alpha}{\alpha}$. The theorem is proved. □

Remark 1. When $D = 0$, $s = \infty$ is actually a zero of K. By (2.5), $z = -(1-\alpha)/\alpha$ is indeed the corresponding zero of K_{gbt}. In this regard, Theorem 2.1 tells us that the generalized bilinear transformation may generate new zeros which have no *finite* counterparts in the continuous-time domain. Following conventional notation, we call these zeros *discretization zeros*. Similarly, given a finite zero in the continuous-time domain, a finite zero in the discrete-time domain can be obtained via (2.5). We call such zeros *intrinsic zeros*. Suppose that K is a SISO system whose transfer function has relative degree $n - m > 0$ (note that in this case, $D = 0$). If K is discretized via the step-invariant transformation, it is proved in [1] that if $n - m > 2$, some discretization zero(s) will become unstable at a sufficiently small sampling period h. Even when $n - m = 2$, as $h \to 0$, the unique discretization zero approaches the point $(-1, 0)$ on the plane from outside the unit disk under a certain condition (Theorem 5 in [14]). As to the first-order hold transformation, for $n - m \geq 2$, there are unstable discretization limiting zeros for sufficiently small h (Theorem 9 in [14]). As for the traditional bilinear transformation, namely, $\alpha = 1/2$ in (2.4), Theorem 2.1 indicates that there are $n-m$ discretization zeros at $(-1, 0)$ on the plane, so the resulting discrete-time system is at most marginally stable. However, for $\alpha \in (1/2, 1)$, $z = -\frac{1-\alpha}{\alpha}$ is within the unit circle, implying stable discretization zeros. It is worth noting that discretization zeros are independent of the sampling period for the generalized bilinear transformation. Therefore, as far as system zeros are concerned, the generalized bilinear transformation with $\alpha \in (1/2, 1)$ is the best.

In the following, we investigate the stability of poles (resp., zeros) of the discrete-time system K_{gbt} obtained from the continuous-time system K. Suppose that $s = a + jb$ is a pole (resp., zero) of K, where $j = \sqrt{-1}$. Then by (2.5), we have

$$\left|\frac{1 + (1-\alpha)hs}{1 - \alpha hs}\right| < 1 \Leftrightarrow (1 + ah(1-\alpha))^2 + h^2b^2(1-\alpha)^2 < (1 - a\alpha h)^2 + b^2h^2\alpha^2,$$

which is equivalent to

$$\frac{2a}{a^2+b^2} < (2\alpha - 1)\, h. \tag{2.15}$$

We have the following observations.

1. When $\alpha = 1/2$, the GBT reduces to the traditional bilinear transformation. In this case, inequality (2.15) always holds, provided that $a < 0$; i.e., the traditional bilinear transformation maps stable poles (resp., zeros) of K to stable poles (resp., zeros) of K_{gbt}. If $D = 0$, Remark 1 shows that discretization zeros are at $(-1, 0)$ on the plane.
2. When $\alpha \in (0, 1/2)$, $(2\alpha - 1) < 0$. Hence, for such a value of α, the resulting GBT may map a stable pole (resp., zero) of K to an unstable pole (resp., zero) of K_{gbt}. If $D = 0$, discretization zeros lie outside the unit circle, thus leading to a nonminimum phase discrete-time system.
3. When $\alpha \in (1/2, 1)$, $(2\alpha - 1) > 0$. Hence, for such a value of α, the resulting generalized bilinear transformation may map an unstable pole (resp., zero) of K to a stable pole (resp., zero) of K_{gbt}. If $D = 0$, discretization zeros are all stable.

Now we see that the traditional bilinear transformation stands in the middle of the GBT; on the one side lie transformations that may convert a stable zero (resp., pole) to an unstable pole (resp., zero) (for $0 < \alpha < 1/2$); on the other side stand transformations that may transform an unstable pole (resp., zero) to a stable zero (resp., pole) (for $1/2 < \alpha < 1$). As is well known, unstable zeros or poles are generally undesirable in a control system because they always impose various kinds of limitations on system performance, such as frequency-dependent constraints on the complementary sensitivity function of discrete-time systems [7], the achievable transient performance of tracking and rejection to disturbances applied to the plant output in some servomechanism problem [27], a discrete-time loop transfer recovery procedure [44], and optimal prediction and estimation algorithms for stochastic models [19], [30]. Therefore, it is always desirable to avoid unstable poles and zeros. In this regard, the generalized bilinear transformation with $1/2 < \alpha < 1$ has advantages over the traditional bilinear transformation ($\alpha = 1/2$) because it may be able to eliminate unstable zeros or poles of an analogue systems provided that the sampling period is not too small, which is always the case in practice, as is commented on in section 1. Moreover, in light of Theorem 2.1, discretization zeros generated via the GBT with $1/2 < \alpha < 1$ are all stable, which is also an advantage over the traditional bilinear transformation. Actually, some other types of bilinear transformations have been used in the literature. For instance, in [4], a general bilinear relationship is applied to the problem of deriving discrete analogues of continuous singular perturbation and direct truncation model reduction. We remark that a similar application can be done for the GBT presented in this paper.

Because controllability and observability are two fundamental properties of any control system, in the following we establish that the generalized bilinear transformation preserve these two features.

THEOREM 2.2. *Assume that $\alpha \in (0, 1)$. K is controllable (resp., observable) if and only if K_{gbt} is controllable (resp., observable).*

Proof. We first prove preservation of observability by contradiction. Suppose that K_{gbt} is not observable; then there exist a nonzero vector $w \in \mathbb{R}^n$ and a scalar $z \in \mathbb{R}$

such that

$$A_{gbt}w = zw \tag{2.16}$$

and

$$C_{gbt}w = 0. \tag{2.17}$$

There are two possible cases as follows.

Case 1($z = 0$). In this case, (2.16) is equivalent to

$$A_{gbt}w = 0. \tag{2.18}$$

Substituting it into (2.17), we have $Cw = 0$. According to (2.18), $(I + (1-\alpha)hA)\, w = 0$, that is, $(A - \frac{1}{(\alpha-1)h}I)w = 0$. Clearly, $x = w \neq 0$ and $s = \frac{1}{(\alpha-1)h}$ satisfy $Ax = sx$ and $Cx = 0$.

Case 2($z \neq 0$). In this case, define $x = (I + \frac{\alpha}{1-\alpha}A_{gbt})w$. Then $Cx = 0$. However, we need to verify that $x \neq 0$. If $x = 0$, then $w = -\frac{\alpha}{1-\alpha}A_{gbt}w$. Combining it with (2.16), we have $(1 + \frac{\alpha z}{1-\alpha})A_{gbt}w = 0$. Since $A_{gbt}w \neq 0$ (otherwise $z = 0$), $1 + \frac{\alpha z}{1-\alpha} = 0$, i.e., $z = \frac{\alpha-1}{\alpha}$. According to (2.16), $(I + (1-\alpha)hA)\, w = z(I - \alpha hA)w = \frac{\alpha-1}{\alpha}(I - \alpha hA)w$, which implies $w = 0$ and contradicts our hypothesis. As a consequence, $x \neq 0$. Define $s = \frac{1}{h}\frac{z-1}{1-\alpha+\alpha h}$. Then it is straightforward to verify that $Ax = sx$ and $Cx = 0$.

The "sufficient" part can be proved by reverting the preceding procedure. Consequently the case of observability is proved. Preservation of controllability can be readily proved in a similar way. □

Remark 2. We have studied such properties of the GBT as the mapping of zeros and poles and preservation of controllability and observability. Certainly, there are more to be investigated. For example, Chen and Weller [9] studied the problem of how the finite and infinite zero structures, as well as invertibility structures, of a general continuous-time LTI multivariable system are mapped to those of its discrete-time counterpart under the bilinear transformation. Similar results can be drawn for the GBT.

Remark 3. From our point of view, the real significance of GBT is that it induces a class of controller discretizations. The parameter α provides an extra degree of freedom in the course of digital implementation of analogue controllers. Therefore, it is possible for a designer to utilize α to achieve better control performance. For example, it is now possible to adjust α to assign the poles of the digital controller to some prespecified place. Take step tracking as another example. For a sampled-data system obtained by approximating the analogue controller via either the step-invariant transformation or the bilinear transformation, it is well known that it is step-tracking if and only if the analogue system is step-tracking. However, due to digital approximation, the transient response in the sampled-data system is worse than that of the analogue system. Transient response is primarily determined by closed-loop poles. Hence, with the aid of GBT, we may expect better transient response by choosing α carefully. In summary, there are a lot of issues to be investigated regarding GBT. Due to space limitation, in what follows we discuss only one type of performance, namely, performance recovery characterized by ℓ_p induced norms of systems. In section 3, we show that the ℓ_p induced norms of $K_d - K_{gbt}$ approach zero as the sampling period h goes to zero. Then based on this result, in section 4 we prove performance recovery in digital implementation of analogue systems via the GBT.

3. Convergence in ℓ_p induced norms. Given an analogue system K with a state-space realization (A, B, C, D), its step-invariant transformation K_d has a state-space realization $K_d(z) = (A_d, B_d, C, D)$, where

$$A_d = e^{Ah}, \quad B_d = \int_0^h e^{A\tau} d\tau B.$$

In this section, assuming that K is stable, we investigate the convergence properties between K_d and K_{gbt} in terms of ℓ_p induced norms when the sampling period h goes to zero. Some preliminaries are necessary for further development.

LEMMA 3.1. *Assume that the pair (A, B) is stabilizable. Then there exists a constant matrix F such that $A+BF$ and $\frac{1}{1-\alpha}A+BF$ ($\alpha \in (0,1)$) are all asymptotically stable.*

Proof. Since (A, B) is stabilizable, it is well known that there exists a positive definite matrix P such that

$$A'P + PA - PBB'P < 0.$$

Define $F = -\frac{1}{2}\frac{1}{1-\alpha}B'P$. It can be verified that

$$(A + BF)' P + P (A + BF) \le A'P + PA - PBB'P < 0$$

and

$$\left(\frac{1}{1-\alpha}A + BF\right)' P + P\left(\frac{1}{1-\alpha}A + BF\right) = \frac{1}{1-\alpha}(A'P + PA - PBB'P) < 0.$$

By the Lyapunov theorem, $A + BF$ and $\frac{1}{1-\alpha}A + BF$ are both asymptotically stable. □

LEMMA 3.2. *If $A+BF$ is asymptotically stable (in continuous time), $A_d + B_dF$, $A_{gbt} + B_{gbt}F$, and $A_{gbt} + \frac{1}{1-\alpha}B_{gbt}F$ are all asymptotically stable (in discrete time) for sufficiently small h.*

Proof. The stability of $A_d + B_dF$ can be proved using the power series. Observe that

$$\begin{aligned} A_d + B_dF &= e^{Ah} + \int_0^h e^{A\tau} d\tau BF \\ &= e^{Ah} + \left(hI + \frac{1}{2}Ah^2 + o(h^2)\right) BF \\ &= I + h(A + BF) + o(h), \end{aligned}$$

where $o(h)$ satisfies $\lim_{h\to 0} \frac{o(h)}{h} = 0$. Because $A + BF$ is asymptotically stable, so is $A_d + B_dF$ for sufficiently small h. To prove the stability of $A_{gbt} + B_{gbt}F$, in addition to the power series, Lemma 3.1 is also required. Observe that

$$\begin{aligned} A_{gbt} + B_{gbt}F &= (I - \alpha hA)^{-1} (I + (1-\alpha)hA) + (1-\alpha)h (I - \alpha hA)^{-1} BF \\ &= (I - \alpha hA)^{-1} (I + (1-\alpha)h(A + BF)) \\ &= (I + \alpha hA + o(h)) (I + (1-\alpha)h(A + BF)) \\ &= I + h(A + (1-\alpha)BF) + o(h) \\ &= I + (1-\alpha)h\left(\frac{1}{1-\alpha}A + BF\right) + o(h). \end{aligned}$$

Coupling this with Lemma 3.1 yields the asymptotic stability of $(\frac{1}{1-\alpha}A + BF)$ for sufficiently small h. Hence, $A_{gbt} + B_{gbt}F$ is asymptotically stable for sufficiently small h. Finally, we show the asymptotic stability of $A_{gbt} + \frac{1}{1-\alpha}B_{gbt}F$. Observe that

$$\begin{aligned} A_{gbt} + \frac{1}{1-\alpha}B_{gbt}F &= (I - \alpha hA)^{-1}(I + (1-\alpha)hA) + (I - \alpha hA)^{-1} hBF \\ &= (I - \alpha hA)^{-1}(I - \alpha hA + h(A + BF)) \\ &= I + (I - \alpha hA)^{-1} h(A + BF) \\ &= I + (I + \alpha hA + o(h))\, h(A + BF) \\ &= I + h(A + BF) + o(h). \end{aligned}$$

As a consequence, $A_{gbt} + \frac{1}{1-\alpha}B_{gbt}F$ is asymptotically stable. □

Now we have a set-up for the following lemma that is crucial in further development.

LEMMA 3.3. *In the graph metric* [34], [26], $K_d - K_{gbt}$ *converges to zero as the sampling period* h *goes to zero.*

This result can be easily derived based on the above two lemmas as well as the proof of Theorem 2 in [39] by replacing

$$\begin{bmatrix} M_{bt}(z) \\ N_{bt}(z) \end{bmatrix} = \left[\begin{array}{c|c} A_{bt} + 2B_{bt}F & B_{bt} \\ \hline 2F & I \\ \hline C_{bt} + 2D_{bt}F & D_{bt} \end{array}\right]$$

with

$$\begin{bmatrix} M_{gbt}(z) \\ N_{gbt}(z) \end{bmatrix} = \left[\begin{array}{c|c} A_{bt} + \frac{1}{1-\alpha}B_{gbt}F & B_{gbt} \\ \hline \frac{1}{1-\alpha}F & I \\ \hline C_{gbt} + \frac{1}{1-\alpha}D_{gbt}F & D_{gbt} \end{array}\right]$$

for a right coprime factorization of $K_{gbt} = N_{gbt}(M_{gbt})^{-1}$.

If K is asymptotically stable and $\alpha \in (1/2, 1)$, then it follows from Lemma 3.2 that both K_d and K_{gbt} are asymptotically stable. In this case, the graph metric induces the same topology as that induced by the H_∞ norm. We have the following.

COROLLARY 3.4. *If* K *is stable and* $\alpha \in (1/2, 1)$, $\lim_{h\to 0^+} \|K_d(z) - K_{gbt}(z)\|_{\ell_2} = 0$.

With the aid of Corollary 3.4, we are now in a position to prove the main result of this section.

THEOREM 3.5. *Assume that* K *is stable and* $\alpha \in (1/2, 1)$. *Then*

$$\lim_{h\to 0^+} \|K_d - K_{gbt}\|_{\ell_p} = 0$$

for all $1 \le p \le \infty$.

Proof. Observe that

$$K_d(z) - K_{gbt}(z) = \left[\begin{array}{cc|c} A_d & 0 & B_d \\ 0 & A_{gbt} & B_{gbt} \\ \hline C & -C_{gbt} & -\alpha hC(I - \alpha hA)^{-1}B \end{array}\right].$$

Let

$$K_{cl}(z) := \left[\begin{array}{cc|c} A_d & 0 & B_d \\ 0 & A_{gbt} & B_{gbt} \\ \hline C & -C_{gbt} & 0 \end{array}\right], \quad D_{cl} := -\alpha hC(I - \alpha hA)^{-1}B.$$

Then

$$K_d(z) - K_{gbt}(z) = K_{cl}(z) + D_{cl}.$$

By Corollary 3.4,

$$\lim_{h\to 0^+} \|K_{cl}(z)\|_{\ell_2} = 0.$$

Let the Hankel singular values of $K_{cl}(z)$ be $\sigma_H^h = \{\sigma_1^h, \cdots, \sigma_N^h\}$, where $\sigma_1^h \geq \cdots \geq \sigma_N^h \geq 0$. According to the discrete-time counterpart of Theorem 7.8 in [43], we have

$$\|K_{cl}(z)\|_{\ell_1} \leq 2\sum_{k=1}^{N} \sigma_i^h \leq 2N\sigma_1^h.$$

As a result,

$$\lim_{h\to 0^+} \|K_{cl}(z)\|_{\ell_1} = 0.$$

Clearly, $D_{cl} \to 0$ as $h \to 0$. Therefore

$$\lim_{h\to 0^+} \|K_d(z) - K_{gbt}(z)\|_{\ell_1} = 0.$$

Furthermore, according to the discrete-time version of Theorem 9.1.2 in [6],

$$\|K_d - K_{gbt}\|_{\ell_p} \leq \|K_d - K_{gbt}\|_{\ell_1} \quad \text{for all } 1 \leq p \leq \infty.$$

The result is proved. □

Remark 4. The above theorem is applicable to MIMO systems. The restriction of the "A" matrix being diagonalizable with all real eigenvalues in [39] is also removed. Thus it generalizes Theorem 1 of [39]. In this way, along with Theorem 9.4.1 in [6], if the step-invariant or GBT is applied to an analogue system, internal stability as well as other performance specifications of the analogue system can be recovered as the sampling period tends to zero. This is the topic of the next section.

Remark 5. As commented before, when $\alpha = 1/2$, the GBT in (2.4) reduces to the traditional one. For convenience, we denote it by K_{bt}. It can be shown that $K(s)$ and $K_{bt}(z)$ have the same Hankel singular values. Unfortunately, the GBT does not enjoy this nice property.

4. Performance recovery. In this section, we discuss an application of Theorem 3.5. It is well known that internal stability of an analogue control system can be recovered if the controller is implemented via the step-invariant transformation. Similar results can be proved for other types of performance specifications. More concretely, consider the feedback system Σ_1 in Figure 1.1. Now we perform digital implementation using the step-invariant transformation; then K_h in Figure 1.2 becomes K_d. Assume that both G and K_d are FDLTI and strictly causal. Furthermore, suppose that W is FDLTI, strictly causal, and stable, introduced as a prefilter before the sampler for later digital implementation. This closed-loop system is said to be *internally stable* [6] if the mapping

$$\begin{bmatrix} I & G \\ -K & I \end{bmatrix}^{-1} : \begin{bmatrix} r \\ d \end{bmatrix} \longmapsto \begin{bmatrix} z \\ u \end{bmatrix}$$

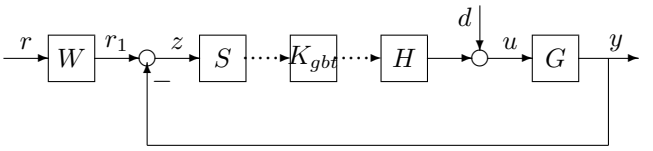


FIG. 4.1. *Digital system* Σ_3 *via the GBT.*

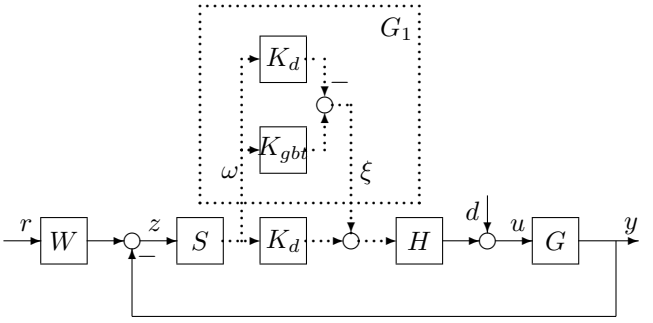


FIG. 4.2. *The equivalent system of* Σ_3.

is bounded $\mathcal{L}_2(\mathbb{R}_+) \to \mathcal{L}_2(\mathbb{R}_+)$.

The following proposition will be used later.

PROPOSITION 4.1 (see [6]). *Suppose that* Σ_1 *is internally stable and satisfies* $\sup_{\|r\|_{\ell_p}\le 1} \|z\|_{\ell_p} < \epsilon$, *for some* $\epsilon > 0$. *Then*

(a) Σ_2 *is internally stable as* $h \to 0$;

(b) $\sup_{\|r\|_{\ell_p}\le 1} \|z\|_{\ell_p} < \epsilon$ *as* $h \to 0$.

If we replace K_d in Figure 1.2 with K_{gbt}, then we get another sampled-data system, Σ_3, as shown in Figure 4.1. We have the following result that is the counterpart of Proposition 4.1.

THEOREM 4.2. *This pertains to the configuration of Figure* 4.1. *Suppose that* Σ_1 *is internally stable and satisfies* $\sup_{\|r\|_{\ell_p}\le 1} \|z\|_p < \epsilon$ *for some* $\epsilon > 0$. *Furthermore, assume that* K *is strictly causal and stable. Then*

(a) Σ_3 *is internally stable as* $h \to 0$;

(b) $\sup_{\|r\|_{\ell_p}\le 1} \|z\|_{\ell_p} < \epsilon$ *as* $h \to 0$.

In order to prove this theorem, we observe the following fact: Let G be a stable, FDLTI, discrete-time system. Then

$$\|G\|_{\ell_p} \le \|G\|_{\ell_1} \quad \text{for all } p \in \mathbb{N} \cup \infty. \tag{4.1}$$

We can now proceed with the proof of Theorem 4.2.

Proof. First we reconfigure system Σ_3 in the way shown in Figure 4.2. The system G_1 can be viewed as a perturbation to the system under it, which is exactly system Σ_2 in Figure 1.2. Therefore, if the norm from ω to ξ tends to zero as $h \to 0$, then according to the small gain theorem, we can ascertain that system Σ_3 is internally stable. Hence, the problem is reduced to showing property (b) for system Σ_3. For that, and in view of (4.1), it is sufficient to show $\lim_{h\to 0} \|G_1\|_{\ell_1} = 0$, which is given by Theorem 3.5. ∎

5. Discussion on performance limitations of digital implementations of analogue systems. In this section, we discuss digital implementation of analogue systems with examples. If an analogue plant $G(s)$ is preceded by a zero-order holder and followed by an ideal sampler, then one gets its discrete-time counterpart, $G_d(z)$. Now suppose there is a certain holder as well as some sampler such that the discrete-time version of $G(s)$ is $G_{gbt}(z)$ obtained via (2.4), i.e., the GBT is physically implementable. Next, we will employ an example to compare these two implementations in terms of various performance limitations of closed-loop systems, such as limitations of feedback control given by Bode or Poisson integrals as well as bandwidth limitations and quantization in network-based control or communication.

Suppose that we have a continuous-time system $G(s)$ whose transfer function is

$$G(s) = \frac{s^3 - 3s + 12s - 10}{s^4 - 10s^3 - 7s^2 + 64s + 60}.$$

$G(s)$ has three unstable zeros: $z_1 = 1 + 3i$, $z_2 = 1 + 3i$, and $z_3 = 1$, where $i = \sqrt{-1}$. The poles of $G(s)$ are $p_1 = 10$, $p_2 = 3$, $p_3 = -2$, and $p_4 = -1$. Note that the first two are unstable. Let the sampling period be $h = 0.5$. Assume that $G_d(z)$ is the discretization of $G(z)$ via the ZOH transformation, and $G_{gbt}(z)$ is that of $G(z)$ via a GBT at $\alpha = 7/8$. Then

$$G_d(z) = \frac{12.76z^3 - 46.34z^2 + 75.95z - 63.65}{z^3 - 153.9z^3 + 814.3z^2 - 682.2z + 148.4}$$

and

$$G_{gbt}(z) = \frac{0.1765z^4 - 0.3691z^3 + 0.1594z^2 - 0.1595z - 0.02719}{z^4 + 3.163z^3 - 2.656z^2 - 0.744z + 0.5568}.$$

It is easy to see that the zeros of $G_d(z)$ are

$$z_1^d = 0.8779 + 1.3742i,\ z_2^d = 0.8779 - 1.3742i,\ z_3^d = 1.8761,$$

and its poles are

$$p_1^d = 148.4132,\ p_2^d = 4.4817,\ p_3^d = 0.6065,\ p_4^d = 0.3679.$$

Hence, $G_d(z)$ has three unstable zeros. Note that the relative degree of $G(s)$ is 1; thus there are no discretization zeros [1]. $G_d(z)$ has two unstable poles, too. The zeros of $G_{gbt}(z)$ are

$$z_1^{gbt} = 1.8889,\ z_2^{gbt} = -0.1429,\ z_3^{gbt} = 0.1724 + 0.7356i,\ z_4^{gbt} = 0.1724 - 0.7356i,$$

and its poles are

$$p_1^{gbt} = -3.8000,\ p_2^{gbt} = -0.4815,\ p_3^{gbt} = 0.4667,\ p_4^{gbt} = 0.6522.$$

It is worthwhile to observe that $z_2^{gbt} = -0.1429$ is a discretization zero, i.e., $-\frac{1-\alpha}{\alpha} = -\frac{1-7/8}{7/8} = -1/7 \approx -0.1429$ corresponding to the zero at ∞ as indicated by Theorem 2.1. $G_{gbt}(z)$ has one unstable zero as well as one unstable pole. In the following, we will compare $G_d(z)$ and $G_{gbt}(z)$ in terms of performance limitations, such as those characterized by sensitivity and complementary sensitivity functions as well as bandwidth requirement in network-based control.

The benefit of feedback is that it compensates for the influence of various sources of uncertainties and disturbances. For example, a sensitivity function quantifies a system's ability to attenuate output disturbance, while a complementary sensitivity function describes a system's ability to attenuate noise response as well as maintain stability robustness. Such performance limitations can be characterized by Bode and Poisson integrals of sensitivity and complementary sensitivity functions. Due to their fundamental importance, in what follows we compare G_d and G_{gbt} in this respect.

First we discuss G_{gbt}. It is easy to show that G_{gbt} is stabilizable by a strictly proper stable controller, F, so to speak (see p. 115 of [6] for one choice of such a controller). Hence, the resulting open-loop transfer function $G_{gbt}F$ is strictly proper too. In light of this, according to Theorem 2 of [36] (resp., Theorem 3.3 of [8]), the Bode integral of the sensitivity function from $-\pi$ to π is $2\pi \cdot (log(p_1^{gbt})) \approx 2\pi \cdot 1.3350 \approx 8.3881$. As to G_d, it is not easy to find directly a strictly proper stable controller M that ensures stability of the closed-loop system. However, by mapping G_d into the continuous-time domain via the traditional bilinear transformation, one can use Theorem 1 in [37] to design a strictly proper stable controller, say M. Indeed, the procedure described by Corollary 3 in [37] can be implemented to find such a controller. Therefore, the open-loop transfer function G_dM is also strictly proper, and its unstable poles are given uniquely by G_d. According to Theorem 2 of [36] (resp., Theorem 3.3 of [8]), the Bode integral of the sensitivity function from $-\pi$ to π is around 40.8407, which is much bigger than that of the sensitivity function $G_{gbt}F$. Due to space limitation, we omit the calculation of Bode integrals for complementary functions of G_dM and $G_{gbt}F$. We also omit the Poisson integrals for the sensitivity function and the complementary function of G_dM as well as those of $G_{gbt}F$ (interested readers may follow the discussions on pages 74–83 of [29]). However, our calculation shows that those performance limitations for G_{gbt} are much less severe than those of G_d because of the unstable poles and zeros structures.

Signal quantization is widely adopted in network-based control. Given a plant G, quantization of state space of G will surely affect the control law to be designed. Insofar as state feedback is concerned, unstable poles turn out to determine primarily the quantization of the system state space. For example, it is shown in [10] that the presence of unstable poles causes nonexistence of a stabilizing state feedback law under any fixed state quantization. A new network data transmission strategy has recently been proposed, and its dynamics have been analyzed in depth in [40], [41], and [42], where it is proved that the presence of unstable poles leads to no stabilizing feedback laws under this new networked control scheme. In [12], a parameter ρ is employed to describe the coarseness of quantization. In general, the smaller ρ is, the coarser the quantization is. Coarser quantization indicates that less control effort is required. Now we compare the optimal ρ values for both cases of the ZOH and GBT transformations. According to Theorem 2.2 of [12], the coarsest quantizer ρ associated with $G_d(z)$ is 0.6266, while that of $G_{gbt}(z)$ is 0.5833. Therefore quantization for $G_{gbt}(z)$ can be coarser, signifying less control effort.

Furthermore, a Bode integral formula has been generalized to the case of networked control systems. For example, in a Gaussian network, it is shown in [11] that the Bode integral is equal to some average directed information [32] that describes the required transmission rate of a stable communication scheme (Theorem 4.6 of [11]). It turns out that the Bode integral for the sensitivity function is given by unstable poles of the plant to be controlled via the network. More concretely, for $G_d(z)$, this

quantity is

$$\ln\left(p_1^d\right) + \ln\left(p_2^d\right) = 6.5000,$$

while that of $G_{gbt}(z)$ is

$$\ln\left(p_1^{gbt}\right) = 1.3350.$$

Clearly, a lower transmission rate is required if the original continuous-time plant G is implemented via the generalized bilinear transformation with $\alpha = 7/8$. Interestingly, it is shown [33], [31] that the above value is also the minimal bandwidth required to guarantee the asymptotic observability and asymptotic stabilizability of a discrete-time linear system in network-based control, which are argued to be two fundamentally important concepts in network-based control.

6. Conclusion. In this paper, we have proposed the generalized bilinear transformation (GBT). We have proved that GBT preserves both controllability and observability. We have shown that, compared with the traditional bilinear transformation, as well as the zero-order hold (ZOH) transformation, one advantage of GBT is its ability to convert unstable poles (resp., zeros) to stable poles (resp., zeros). We have also proved that GBT and ZOH converge to one another as the sampling period goes to zero in the sense of ℓ_p induced norms for all $1 \leq p \leq \infty$. Finally, we have established performance recovery of digital implementations of continuous-time systems via GBT. We hope that the brief discussion of the GBT in this paper can lay the groundwork for a subsequent study of its possible applications in the fields of digital and network-based control, and signal processing.

Acknowledgments. The authors are grateful to the associate editor and the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

[1] K. J. Åström, P. Hagander, and J. Sternby, *Zeros of sampled systems*, Automatica J. IFAC, 20 (1984), pp. 21–38.
[2] K. J. Åström and B. Wittenmark, *Computer-Controlled Systems*, 3rd ed., Prentice–Hall, Upper Saddle River, NJ, 1997.
[3] B. Bamieh, *Intersample and finite wordlength effects in sampled-data problems*, IEEE Trans. Automat. Control, 48 (2003), pp. 639–643.
[4] B. Clapperton, F. Crusca, and M. Aldeen, *Bilinear transformation and generalized singular perturbation model reduction*, IEEE Trans. Automat. Control, 41 (1996), pp. 589–593.
[5] T. Chen and B. Francis, *Input-output stability of sampled-data control systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 50–58.
[6] T. Chen and B. Francis, *Optimal Sampled-Data Control Systems*, Springer, London, 1995.
[7] J. Chen, *Integral constraints and performance limits on complementary sensitivity: Discrete-time systems*, Systems Control Lett., 39 (2000), pp. 45–53.
[8] J. Chen, *Sensitivity integrals and transformation techniques: A new perspective*, IEEE Trans. Automat. Control, 42 (1997), pp. 1037–1042.
[9] B. M. Chen and S. R. Weller, *Mappings of the finite and infinite zero structures and invertibility structures of general linear multivariable systems under the bilinear transformation*, Automatica J. IFAC, 34 (1998), pp. 111–124.
[10] D. Delchamps, *Stabilizing a linear system with quantized state feedback*, IEEE Trans. Automat. Control, 35 (1990), pp. 916–924.
[11] N. Elia, *When Bode meets Shannon: Control-oriented feedback communication schemes*, IEEE Trans. Automat. Control, 49 (2004), pp. 1477–1488.

[12] N. Elia and S. K. Mitter, *Stabilization of linear systems with limited information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1384–1400.
[13] A. Feuer and G. C. Goodwin, *Sampling in Digital Signal Processing and Control*, Birkhäuser Boston, Boston, MA, 1996.
[14] T. Hagiwara, T. Yuasa, and M. Araki, *Stability of the limiting zeros of sampled-data systems with zero- and first-order holds*, Internat. J. Control, 58 (1993), pp. 1325–1346.
[15] L. B. Jemaa and E. J. Davison, *Limiting performance of optimal linear discrete filters*, Automatica J. IFAC, 39 (2003), pp. 1221–1226.
[16] J. P. Keller and B. D. O. Anderson, *A new approach to the discretization of continuous-time controllers*, IEEE Trans. Automat. Control, 37 (1992), pp. 214–223.
[17] Z. Kowalczuk, *Discrete approximation of continuous-time systems: A survey*, Proc. IEE-G, 144 (1993), pp. 264–278.
[18] B. Kuo, *Digital Control Systems*, 2nd ed., Saunders College Publishing, Philadelphia, 1992.
[19] E. K. Larsson, *Limiting sampling results for continuous-time ARMA systems*, Internat. J. Control, 78 (2005), pp. 461–473.
[20] F.-L. Lian, J. Moyne, and D. Tilbury, *Network design consideration for distributed control systems*, IEEE Trans. Control Systems Technol., 10 (2002), pp. 297–307.
[21] G. M. H. Leung, T. P. Perry, and B. A. Francis, *Performance analysis of sampled-data control systems*, Automatica J. IFAC, 27 (1991), pp. 699–704.
[22] F. Marvasti, ed., *Nonuniform Sampling: Theory and Practice*, Kluwer Academic, New York, 2001.
[23] M. Morari and E. Zagiriou, *Robust Process Control*, Prentice–Hall, Englewood Cliffs, NJ, 1989.
[24] R. H. Middleton and G. C. Goodwin, *Digital Control and Estimation: A United Approach*, Prentice–Hall, Englewood Cliffs, NJ, 1990.
[25] Y. Oishi, *Asymptotic properties of the best sampled-data control performance*, IEEE Trans. Automat. Control, 46 (2001), pp. 1132–1137.
[26] L. Qiu and E. Davison, *Pointwise gap metrics on transfer matrices*, IEEE Trans. Automat. Control, 37 (1992), pp. 741–758.
[27] L. Qiu and E. J. Davison, *Performance limitations of non-minimum phase systems in the servomechanism problem*, Automatica J. IFAC, 29 (1993), pp. 337–349.
[28] C. A. Rabbath, N. Hori, and N. Lechevin, *Convergence of sampled-data models in digital redesign*, IEEE Trans. Automat. Control, 49 (2004), pp. 850–855.
[29] M. M. Seron, J. H. Braslavsky, and G. C. Goodwin, *Fundamental Limitations in Filtering and Control*, Springer, London, 1997.
[30] T. Söderström, *On zero locations for sampled stochastic systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 1249–1253.
[31] S. Tatikonda and S. K. Mitter, *Control over noisy channels*, IEEE Trans. Automat. Control, 49 (2004), pp. 1196–1201.
[32] S. Tatikonda, A. Sahai, and S. K. Mitter, *Stochastic linear control over a communication channel*, IEEE Trans. Automat. Control, 49 (2004), pp. 1549–1561.
[33] S. Tatikonda and S. Mitter, *Control under communication constraints*, IEEE Trans. Automat. Control, 49 (2004), pp. 1056–1068.
[34] M. Vidyasagar, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, 29 (1984), pp. 403–418.
[35] B. Wittenmark, K. J. Åström, and K. E. Årzén, *Computer Control: An Overview*, IFAC Professional Brief, 2002. Available online at http:www.ifac-control.org/publications/pbriefs/PB_wittenmark_etal_final.pdf
[36] B.-F. Wu and E. A. Jonckheere, *A simplified approach to Bode's theorem for continuous-time and discrete-time systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1797–1802.
[37] D. C. Youla, J. J. Bongiorno, and C. N. Lu, *Single-loop feedback-stabilization of linear multivariable dynamical plants*, Automatica J. IFAC, 10 (1974), pp. 159–173.
[38] G. Zhang and T. Chen, *A convergence result for digital implementation via the bilinear and step-invariant transformations*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, suppl., 2003, pp. 28–33.
[39] G. Zhang and T. Chen, *Comparing digital implementation via the bilinear and step-invariant transformations*, Automatica J. IFAC, 40 (2004), pp. 327–330.
[40] G. Zhang and T. Chen, *Networked control systems, a perspective from chaos*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 3075–3101.
[41] G. Zhang, G. Chen, T. Chen, and Y. Lin, *Analysis of a type of nonsmooth dynamical systems*, Chaos Solitons Fractals, 30 (2006), pp. 1153–1164.

[42] G. Zhang, G. Chen, T. Chen, and M. D'Amico, *Dynamical analysis of a networked control system*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., to appear.
[43] K. Zhou and J. Doyle, *Essentials of Robust Control*, Prentice–Hall, Upper Saddle River, NJ, 1998.
[44] Z. Zhang and J. Freudenberg, *Discrete-time loop transfer recovery for systems with non-minimum phase zeros and time delays*, Automatica J. IFAC, 29 (1993), pp. 351–363.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2224–2256

PATHWISE STOCHASTIC CONTROL PROBLEMS AND STOCHASTIC HJB EQUATIONS*

RAINER BUCKDAHN† AND JIN MA‡

Abstract. In this paper we study a class of *pathwise stochastic control problems* in which the optimality is allowed to depend on the paths of exogenous noise (or information). Such a phenomenon can be illustrated by considering a particular investor who wants to take advantage of certain extra information but in a completely legal manner. We show that such a control problem may not even have a "minimizing sequence," but nevertheless the (Bellman) dynamical programming principle still holds. We then show that the corresponding Hamilton–Jacobi–Bellman equation is a stochastic partial differential equation, as was predicted by Lion and Souganidis [*C. R. Acad. Sci. Paris Sér. I Math.*, 327 (1998), pp. 735–741]. Our main device is a Doss–Sussmann-type transformation introduced in our previous work [*Stochastic Process. Appl.*, 93 (2001), pp. 181–204] and [*Stochastic Process. Appl.*, 93 (2001), pp. 205–228]. With the help of such a transformation we reduce the pathwise control problem to a more standard relaxed control problem, from which we are able to verify that the value function of the pathwise stochastic control problem is the unique stochastic viscosity solution to this stochastic partial differential equation, in the sense of [*Stochastic Process. Appl.*, 93 (2001), pp. 181–204] and [*Stochastic Process. Appl.*, 93 (2001), pp. 205–228].

Key words. pathwise stochastic control, dynamical programming, Bellman principle, Doss–Sussmann transformation, stochastic viscosity solutions

AMS subject classifications. 60H10, 34F05, 90A12

DOI. 10.1137/S036301290444335X

1. Introduction.

In this paper we are interested in the so-called *pathwise stochastic control problem*, originally proposed by P.-L. Lions and Souganidis [13]. A version of such a problem can be described by the following example of optimazation problem in finance, in which the underlying risky asset follows a "hidden Markovian" stochastic volatility model:

$$\begin{cases} dS_t = S_t[\mu(t,S_t)dt + \widehat{\sigma}(Y_t)dW_t], \\ dY_t = h(t,Y_t)dt + \sigma_1(t,Y_t)dW_t + \sigma_2(t,Y_t)\circ dB_t, \\ S_0 = \xi, Y_0 = y. \end{cases} \tag{1.1}$$

where W and B are two independent Brownian motions, S_t is the asset value at t, and $\widehat{\sigma}(Y)$ is the volatility process. Here the Stratonovic differential $\circ dB$ in (1.1) is used to simplify our future discussion; it can be replaced by an Itô-type integral if needed.

We note that the extra noise B in (1.1) can be thought of as some extra information that cannot be detected in the market in general, but is available to the particular investor. The problem then is to show how this investor can take advantage of such extra information to optimize his/her utility, but by taking actions that are

*Received by the editors April 27, 2004; accepted for publication (in revised form) August 20, 2006; published electronically February 2, 2007.
http://www.siam.org/journals/sicon/45-6/44335.html

†Département de Mathématiques, Université de Bretagne Occidentale, F-29285 Brest Cedex, France (Rainer.Buckdahn@univ-brest.fr.)

‡Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395 (majin@math.purdue.edu). This author was supported in part by NSF grants 0204332 and 0505427. Part of this work was completed while this author was visiting the Department of Mathematics, University of Bretagne-Occidentale, France, whose hospitality is greatly appreciated.

completely "legal," in the sense that the investor has to choose the optimal strategy in the usual class of the admissible portfolios.

Mathematically, we can formulate such an optimization problem for this investor as follows. First recall that if the short rate of the riskless asset in this market is $\{r_t,\ t \geq 0\}$, and we assume that the (self-financing) portfolios $\{\pi_t,\ t \geq 0\}$ are all the $\{\mathcal{F}_t^W\}$-progressively measurable, square-integrable processes, then the dynamics of the wealth process, $\{\mathcal{V}_t, t \geq 0\}$, of the investor satisfies the SDE

$$\begin{cases} d\mathcal{V}_t = [r_t\mathcal{V}_t + \pi_t(\mu(t,S_t) - r_t)]dt + \pi_t\widehat{\sigma}(Y_t)dW_t, \\ \mathcal{V}_0 = v. \end{cases} \tag{1.2}$$

Next, denoting $\{\mathcal{F}_t^B\}_{t\geq 0}$ to be the filtration generated by B, we define the following "cost functional" (given the extra information):

$$J(\pi) = E\left\{H(\mathcal{V}_T) + \int_0^T \ell(t,\mathcal{V}_t,\pi_t)dt\Big|\mathcal{F}_T^B\right\}. \tag{1.3}$$

Clearly, the purpose of conditioning on $\mathcal{F}_T^B$ means that we are seeking optimization given all the possible extra information (some of them might be anticipating!), while the restriction that all strategies are $\{\mathcal{F}_t^W\}$-adapted indicates that they are completely "legal."

To make (1.2) and (1.3) fit more into a stochastic control framework, we combine (1.1) and (1.2) as follows. Define $X \triangleq (Y,S,\mathcal{V})^T$ and

$$\begin{aligned} b(t,X_t,\pi_t) &\triangleq [h(t,Y_t), \mu(t,S_t)S_t, r_t\mathcal{V}_t + \pi_t(\mu(t,S_t) - r_t)]^T, \\ \sigma(t,X_t,\pi_t) &\triangleq [\sigma_1(t,Y_t), \hat{\sigma}(Y_t)S_t, \hat{\sigma}(Y_t)\pi_t]^T, \\ \theta(t,X_t) &\triangleq [\sigma_2(t,Y_t), 0, 0]^T. \end{aligned}$$

Then the new "state" process X satisfies the following SDE:

$$\begin{cases} dX_t = b(t,X_t,\pi_t)dt + \sigma(t,X_t,\pi_t)dW_t + \theta(t,X_t)\circ dB_t, \\ X_s = (y,\xi,v)^T, \end{cases} \qquad 0 \leq s \leq t \leq T, \tag{1.4}$$

and cost functional (1.3) can be rewritten as

$$J(s,(y,\xi,v);\pi) = E_{s,(y,\xi,v)}\left\{H(X_T) + \int_s^T \ell(t,X_t,\pi_t)dt\Big|\mathcal{F}_T^B\right\}. \tag{1.5}$$

Our *pathwise stochastic control problem* is then to minimize the cost functional (1.5) over the set of "admissible controls" $\mathcal{A}$, which is by definition all the $\{\mathcal{F}_t^W\}$-adapted strategies. The value function of this stochastic control problem is defined by

$$V(s,(y,\xi,v)) = \operatorname*{essinf}_{\pi\in\mathcal{A}} J(s,(y,\xi,v)). \tag{1.6}$$

Here the essential infimum should be understood as one with respect to the indexed family of random variables (see, e.g., [5, 7] or [10, Appendix A]; detailed definition will be given in section 2).

At this point we would like to point out that the pathwise stochastic control problem of this kind was one of the motivations for the study of the "stochastic

viscosity solution" for fully nonlinear stochastic PDEs (see Lions and Souganidis [13]). However, while it has long been predicted that the Hamilton–Jacobi–Bellman (HJB) equation for such stochastic control problem is a fully nonlinear stochastic PDE, to date the mathematical content of this problem has not been fully explored. One of the main purposes of this paper is to try to establish a rigorous framework for the pathwise stochastic control problem and provide some necessary machinery for future study. It turns out that many of them are interesting in their own right. Among other things, we shall prove the Bellman dynamic programming principle in this particular situation, from which we then prove that the value function is a *stochastic viscosity* solution of a stochastic HJB equation, in the sense of our previous works (see Buckdahn and Ma [2, 3]). It should be noted that the special measurability issue involved in the "legality" of our admissible controls has not been studied before.

We should note that the pathwise control problem defined above is quite different from a standard stochastic control problem, or even those with partial observations (see, e.g., Bensoussan [1]). The most essential difference is that the cost functional is now a random field instead of a deterministic function, and therefore so is the value function. Consequently, the usual infimum (or supremum) involved in the optimization problem should naturally be replaced by the "essential infimum" (or "essential supremum"). Such a seemingly "routine" change, together with the "legality" requirements for the admissible controls, turns out to be the source of many substantial difficulties, both from a mathematical point of view and from the control theoretical point of view. In fact, in the appendix we shall provide an example which shows that in general there does not exist a minimizing sequence for our pathwise stochastic control problem. The lack of such a sequence seems to be fatal for the dynamic programming method. To overcome this difficulty we introduce an intermediate stochastic control problem (called the "wider-sense" control problem in what follows), in which the stochastic integral against the Brownian motion B is eliminated. We show that this wider-sense control problem is in some sense equivalent to a traditional stochastic control problem, and we can use it as a bridge to reach our goal.

Another immediate problem is the stochastic HJB equation itself, mainly in various forms of measurability issues including the definition of the stochastic integration (one should appreciate the fact that an HJB equation is always "backward"!). One of the main reasons that we insist on using the Stratonovic stochastic integral with regard to the Brownian motion B is that it is "insensitive" to the direction of integration, as we can show. Finally, we would like to remark that the notion of stochastic viscosity solutions has been studied by Lions and Souganidis [12, 13, 14, 15] and Buckdahn and Ma [2, 3, 4]. The definition of the stochastic viscosity solution in this paper is consistent with our previous works, with slight modifications to suit the present situation. We note that such a modification will not alter the uniqueness result from [3].

The rest of the paper is organized as follows. In section 2 we formulate the problem more formally and provide some necessary preliminaries. In section 3 we introduce some wider-sense control problems and establish some properties and relationship among them. Section 4 is devoted to a proof of the Bellman principle. Finally, in section 5 we prove that the value function of the corresponding wider-sense control problem is the viscosity solution to a randomized HJB equations, and in section 6 we extend the result to the original problem.

2. Problem formulation and preliminaries. In this section we give a detailed formulation of our pathwise control problem. Let $W = (W^1, \dots, W^d)$ and $B = (B^1, \dots, B^m)$ be two independent, standard Brownian motions defined on some

complete probability space $(\Omega, \mathcal{F}, P)$, and let $T > 0$ be a given finite time horizon. We denote $\mathbf{F}^W = \{\mathcal{F}_t^W\}_{t\in[0,T]}$ and $\mathbf{F}^B = \{\mathcal{F}_t^B\}_{t\in[0,T]}$ to be the two filtrations generated by W and B, respectively, and augmented by the P-null sets in $\mathcal{F}$ so that they satisfy the *usual hypotheses* (see, e.g., [16]). The following two filtrations will be frequently used in the future:

$$\begin{cases} \mathbf{F} = \{\mathcal{F}_t\}_{t\in[0,T]} \triangleq \{\mathcal{F}_t^W \vee \mathcal{F}_t^B\}_{t\in[0,T]} = \mathbf{F}^W \vee \mathbf{F}^B, \\ \mathbf{G} = \{\mathcal{G}_t\}_{t\in[0,T]} \triangleq \{\mathcal{F}_t^W \vee \mathcal{F}_T^B\}_{t\in[0,T]} = \mathbf{F}^W \vee \mathcal{F}_T^B. \end{cases} \tag{2.1}$$

Here, for two σ-fields $\mathcal{F}$ and $\mathcal{G}$, $\mathcal{F} \vee \mathcal{G}$ denotes $\sigma(\mathcal{F} \cup \mathcal{G})$ as usual. The meaning for those involving filtrations is obvious.

Throughout this paper we let $\mathbb{E}$ be a generic Euclidean space, with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$. We denote $C^k(\mathbb{E}; \mathbb{E}_1)$ to be the usual space of $\mathbb{E}_1$-valued, k-times continuously differentiable functions defined on $\mathbb{E}$. Furthermore, we denote $C_b^k(\mathbb{E}; \mathbb{E}_1) \subset C^k(\mathbb{E}; \mathbb{E}_1)$ to be all functions that have uniformly bounded partial derivatives and $C_p^k(\mathbb{E}; \mathbb{E}_1) \subset C^k(\mathbb{E}; \mathbb{E}_1)$ to be all functions whose partial derivatives are of at most polynomial growth. The spaces $C^{k,\ell}([0,T] \times \mathbb{E}; \mathbb{E}_1)$, $C_b^{k,\ell}([0,T] \times \mathbb{E}; \mathbb{E}_1)$, $C_p^{k,\ell}([0,T] \times \mathbb{E}; \mathbb{E}_1)$ are defined similarly. Finally, if $\mathbb{E}_1 = \mathbb{R}$, we shall omit $\mathbb{E}_1$ in the notation above.

Now let $\mathbb{B}$ be a generic Banach space and $\mathbf{H} = \{\mathcal{H}_t\}_{t\in[0,T]}$ a generic Filtration on $(\Omega, \mathcal{F}, P)$. We shall denote the following:

- For any $1 \le p < \infty$, $L^p(\mathbf{H}, [0,T]; \mathbb{B})$ denotes all $\mathbb{B}$-valued, $\mathbf{H}$-progressively measurable processes ψ, such that $E\int_0^T \|\psi_t\|_{\mathbb{B}}^p dt < \infty$. In particular, we denote $L^0(\mathbf{H}, [0,T]; \mathbb{B})$ to be all $\mathbb{B}$-valued, $\mathbf{H}$-progressively measurable processes and $L^\infty(\mathbf{H}, [0,T]; \mathbb{B})$ to be a subset of $L^0(\mathbf{H}, [0,T]; \mathbb{B})$ in which all processes are uniformly bounded.
- For any $1 \le p < \infty$, $L_{loc}^p(\mathbf{H}, [0,T] \times \mathbb{E}; \mathbb{E}_1)$ denotes the space of all $\mathbb{E}_1$-valued random fields η defined on $[0,T] \times \Omega \times \mathbb{E}$, such that for fixed $x \in \mathbb{E}$, the mapping $(t, \omega) \mapsto \eta(t, \omega, x)$ is $\mathbf{H}$-progressively measurable, and that $E\int_0^T |\eta(t, \cdot, x)|^p dt < \infty$. In particular, we denote $L_{loc}^p(\mathbf{H}, [0,T]; C^k(\mathbb{E}, \mathbb{E}_1))$ to be all $C^k(\mathbb{E}, \mathbb{E}_1)$-valued, $\mathbf{H}$-progressively measurable random fields η such that η and all the partial derivatives $D_{x_i}\eta$ are elements of $L_{loc}^p(\mathbf{H}, [0,T] \times \mathbb{E}; \mathbb{E}_1)$.
- $C^{k,\ell}(\mathbf{H}, [0,T] \times \mathbb{E}; \mathbb{E}_1)$ (resp., $C_b^{k,\ell}(\mathbf{H}, [0,T] \times \mathbb{E}; \mathbb{E}_1)$, $C_p^{k,\ell}(\mathbf{H}, [0,T] \times \mathbb{E}; \mathbb{E}_1)$) denotes the space of all $\mathbb{E}_1$-valued random fields φ, defined on $[0,T] \times \Omega \times \mathbb{E}$ such that for P-a.e. $\omega \in \Omega$, $\varphi(\cdot, \omega, \cdot) \in C^{k,\ell}([0,T] \times \mathbb{E}; \mathbb{E}_1)$ (resp., $C_b^{k,\ell}([0,T] \times \mathbb{E}; \mathbb{E}_1)$, $C_p^{k,\ell}([0,T] \times \mathbb{E}; \mathbb{E}_1)$), and that for fixed $x \in \mathbb{E}$, the process $\varphi(\cdot, \cdot, x)$ is $\mathbf{H}$-progressively measurable.
- $\mathcal{M}_{0,T}(\mathbf{H})$ denotes all the $\mathbf{H}$-stopping times τ such that $0 \le \tau \le T$, P-a.s., and $\mathcal{M}_{0,\infty}(\mathbf{H})$ denotes all $\mathbf{H}$-stopping times that are almost surely finite.

To formulate our control problem let us first specify the *admissible control* sets. Let U be a compact metric space. We denote $\mathcal{A}$ to be the set of all $\mathbf{F}^W$-progressively measurable processes $\alpha : [0,T] \times \Omega \to U$, and denote $\widetilde{\mathcal{A}}$ to be the set of all $\mathbf{G}$-progressively measurable processes $\beta : [0,T] \times \Omega \to U$. We shall refer to the elements in $\mathcal{A}$ as the admissible controls and those in $\widetilde{\mathcal{A}}$ as the *admissible controls in a wider sense.*

For $(t,x) \in [0,T] \times \mathbb{R}^n$ and $\alpha \in \mathcal{A}$, let us consider the following controlled stochastic system: for $0 \le s \le t \le T$,

$$X_t = x + \int_s^t b(r, X_r, \alpha_r)\, dr + \int_s^t \sigma(r, X_r, \alpha_r)\, dW_r + \int_s^t \theta(r, X_r) \circ dB_r. \tag{2.2}$$

The solution of SDE (2.2) will be denoted by $X^{\alpha,s,x}$. The superscripts $^{(\alpha,s,x)}$ will often be dropped for notational simplicity if the context is clear.

The *cost functional* of the pathwise control problems is defined by

$$J(\alpha;s,x)\stackrel{\Delta}{=}E\left\{H(X_T^{\alpha,s,x})+\int_s^T \ell(t,X_t^{\alpha,s,x},\alpha_t)dt\Big|\mathcal{F}_T^B\right\}, \tag{2.3}$$

where $(s,x)\in[0,T]\times\mathbb{R}^n$, and $\alpha\in\mathcal{A}$, and the *value function* is defined by

$$V(s,x)\stackrel{\Delta}{=}\operatorname*{essinf}_{\alpha\in\mathcal{A}} J(\alpha,s,x). \tag{2.4}$$

We note that the definition of essential infimum for a family of nonnegative random variables can be found in [7] and [10, Appendix A]; we recast it here for ready reference.

DEFINITION 2.1. *Let $\mathcal{X}$ be a nonempty family of nonnegative random variables defined on a probability space $(\Omega,\mathcal{F},P)$. The essential infimum of $\mathcal{X}$, denoted by* essinf $\mathcal{X}$*, is a random variable X^* satisfying the folowing:*

(i) *for all $X\in\mathcal{X}$, $X^*\le X$, P-a.s.; and*

(ii) *if Y is a random variable such that $Y\le X$ for all $X\in\mathcal{X}$, then $Y\le X^*$, P-a.s.*

Throughout this paper we shall make use of the following standing assumptions:

(H1) The functions $b:\mathbb{R}^n\times U\mapsto\mathbb{R}^n$, $\sigma:\mathbb{R}^n\times U\mapsto\mathbb{R}^{n\times d}$ are bounded, uniformly continuous, and uniformly Lipschitz with respect to $x\in\mathbb{R}^n$, uniformly in $u\in U$.

(H2) The function θ belongs to $C^4_{\ell,b}(\mathbb{R}^n;\mathbb{R}^{n\times m})$.

(H3) The function $H:\mathbb{R}^n\mapsto\mathbb{R}$ is uniformly bounded and continuous.

The Lipschitz constants in (H1)–(H3) will be denoted by a generic one $K>0$.

We would like to remark here that in (2.4) the value function $V(\cdot,\cdot)$ is obtained by taking an "essinf" instead of an "inf" as in the usual stochastic control problem. Such a change on the one hand is necessary due to the randomness of the cost functionals, as it is often seen in optimization problems involving random objectives (see, e.g., [10]); it does, on the other hand, generate a great deal of subtleties to the otherwise standard control problem. In fact, in the Appendix we shall provide a counterexample which shows that with such an "essinf" one does not even have a "minimizing sequence" to the control problem. This gives rise to some substantial difficulties in proving the dynamic programming principle, as well as in deriving the stochastic HJB equations. We will show how to get around of this difficulty by studying the related "wider-sense control problems" in section 4.

We should also note that since we do not require any "nondegeneracy" condition on the coefficient σ (or $\sigma\sigma^T$) in this framework, we can apply a standard treatment to reduce the system to a time-homogeneous one and to eliminate the running cost ℓ in (2.3) by adding the extra states

$$X_t^0=t\qquad\text{and}\qquad X_t^{n+1}=\int_s^t \ell(r,X_r,\alpha_r)dr.$$

For example, in this case the cost functional can be written as

$$J(\alpha;s,(s,x,0))=E\{\widetilde{H}(X_T^{\alpha,s,(s,x,0)})\big|\mathcal{F}_T^B\},$$

where $\widetilde{H}(x^0,x,x^{n+1})=H(x)+x^{n+1}$. Therefore, in the rest of the paper we shall

consider the following simplified version of (2.2), (2.3), and (2.4):

$$X_t = x + \int_s^t b\,(X_r, \alpha_r)\,dr + \int_s^t \sigma\,(X_r, \alpha_r)\,dW_r + \int_s^t \theta\,(X_r) \circ dB_r, \tag{2.5}$$

$$J(\alpha; s, x) \triangleq E\left\{H(X_T^{\alpha,s,x})\middle|\mathcal{F}_T^B\right\}, \tag{2.6}$$

and

$$V(s, x) \triangleq \operatorname*{essinf}_{\alpha\in\mathcal{A}} J(\alpha, s, x). \tag{2.7}$$

To conclude this section we remark that under (H1) and (H2), for any admissible control $\alpha \in \mathcal{A}$ the SDE (2.2) has a unique ($\mathbf{F}$-adapted) solution. But if $\alpha \in \widetilde{\mathcal{A}}$, the situation would be much more complicated, although it could be made sensible if we allow the integral $\int \theta(X_r) \circ dB_r$ to be the anticipating Stratonovic integral. But we shall avoid such complexity by introducing the wider-sense problems. Finally, we note that the value function defined by (2.7) is an $\mathcal{F}_T^B$-measurable random field. We shall prove that it is indeed $\mathcal{F}_{s,T}^B$-measurable for any $s \in [0, T]$, and hence the stochastic HJB equation, which is "backwardly" defined (given the terminal condition at time T), would simply be a time-reversed stochastic PDE in the usual sense.

3. A Doss–Sussmann-type transformation. In this section we introduce the first step towards our wider-sense control problem. In light of the idea of "stochastic characteristics" (cf. Lions and Souganidis [12, 13, 14, 15]) and/or "Doss–Sussmann" transformation (cf. Buckdahn and Ma [2, 3, 4]), we would like to remove the stochastic integral "$\int \theta(X) \circ dB$" so it becomes less "problematic." We note that such a step is essential in the study of stochastic viscosity solutions as well.

We proceed as follows. First consider the following SDE with parameters: for any $0 \le s \le T$, and $z \in \mathbb{R}^n$,

$$\eta_t^s(z) = z + \int_s^t \theta(\eta_r^s(z)) \circ dB_r \qquad s \le t \le T. \tag{3.1}$$

We note that this SDE can be converted easily into the following "Itô form":

$$\eta_t^s(z) = z + \int_s^t \theta(\eta_r^s(z))dB_r + \frac{1}{2}\int_s^t [D_x\theta \otimes \theta](\eta_r^s(z))dr, \qquad s \le t \le T. \tag{3.2}$$

Here $[D_x\theta \otimes \theta]$ denotes the product of the tensor $D_x\theta$ and the matrix θ, defined by

$$[D_x\theta \otimes \theta]_i \triangleq \sum_{j=1}^m \sum_{k=1}^n \frac{\partial\theta^{ij}}{\partial x^k}\theta^{kj} = \operatorname{tr}\{[D_x\theta^i]\theta^T\}, \qquad i = 1, \dots, n, \tag{3.3}$$

where θ^i is the ith row vector of the matrix θ. We note that in this paper the dimensions of the Brownian motion are not essential. Thus, to simplify notation, in what follows we shall assume $m = 1$. Thus the tensor product is simplified to the usual matrix product: $D\theta \otimes \theta = [D\theta]\theta$.

Next, we note that the stochastic flow $z \mapsto \eta_t^s(z)$ is a diffeomorphism for all $s \le t \le T$, P-a.s., and

$$D_z\eta_t^s(z) = I + \int_s^t D_z\eta_r^s(z) D\theta(\eta_r^s(z)) \circ dB_r, \quad s \le t \le T. \tag{3.4}$$

Consequently, $[D_z\eta_t^s]^{-1}(\cdot)$ exists; and it can be shown (see, e.g., [2]) that the inverse flow of $\eta_t^s(\cdot)$, denoted by $\zeta_t^s(\cdot)$, exists and satisfies the first order stochastic PDE:

$$\zeta_t^s(z) = z - \int_s^t D_z\zeta_r^s(z)\theta(z)\circ dB_r, \qquad 0\le s\le t\le T,\ z\in\mathbb{R}^n, \tag{3.5}$$

and it holds that $\eta_t^s(\zeta_t^s(z)) = \zeta_t^s(\eta_t^s(z)) = z$ for $0\le s\le t\le T$, $z\in\mathbb{R}^n$, P-a.s. Let us now define the following random fields: for $(t,z,u)\in[0,T]\times\mathbb{R}^n\times U$,

$$\widetilde{\sigma}(t,z,u) \triangleq [D_z\eta_t^0(z)]^{-1}\sigma(\eta_t^0(z),u); \tag{3.6}$$

$$\begin{aligned}\widetilde{b}(t,z,u) \triangleq [D_z\eta_t^0(z)]^{-1}\Big\{ & b(\eta_t^0(z),u)\\ & -\frac{1}{2}\mathrm{tr}\left\{[D_z\eta_t^0(z)]^{-1}\sigma\sigma^T(\eta_t^0(z),u)[(D_z\eta_t^0(z))^T]^{-1}[D_{zz}^2\eta_t^0(z)]\right\}\Big\}.\end{aligned} \tag{3.7}$$

Clearly, $\widetilde{b}$ and $\widetilde{\sigma}$ are $\mathbf{F}$-progressively measurable. The following lemma is essential.

LEMMA 3.1. *Assume* (H1) *and* (H2). *Then for any* $0<\gamma<\frac{1}{12}$, *there exist a sequence of* $\mathbf{F}^B$*-stopping times* $0\le\tau_1\le\tau_2\le\cdots\le\tau_\ell\le\cdots$ *satisfying* $P\{\tau_\ell = T\}\uparrow 1$, *as* $\ell\to\infty$, *and some constant* $C>0$, *depending only on the coefficients* b, σ, *and* θ, *such that for all* $\ell\ge 1$ *and* $t\in[0,T]$, $z,z'\in\mathbb{R}^n$, *and* $u\in U$, *it holds* P*-a.s. on* $\{\tau_\ell = T\}$ *that*

$$\begin{aligned}|\widetilde{\sigma}(t,z,u)| &\le C\ell(1+|z|^2)^\gamma \le C\ell(1+|z|);\\ |\widetilde{b}(t,z,u)| &\le C\ell^4(1+|z|^2)^{4\gamma} \le C\ell^4(1+|z|);\\ |\widetilde{\sigma}(t,z,u)-\widetilde{\sigma}(t,z',u)| &\le C\ell^3(1+|z|^2+|z'|^2)^{3\gamma}|z-z'|\\ &\le C\ell^3(1+|z|+|z'|)|z-z'|;\\ |\widetilde{b}(t,z,u)-\widetilde{b}(t,z',u)| &\le C\ell^6(1+|z|^2+|z'|^2)^{6\gamma}|z-z'|\\ &\le C\ell^6(1+|z|+|z'|)|z-z'|.\end{aligned}$$

Proof. First, let $\mathbb{E}$ be any Euclidean space, and let $f\in L^2_{loc}(\mathbf{G};[0,T],C^1(\mathbb{R}^n;\mathbb{E}))$ (that is, $f(t,\omega,\cdot)\in C^1(\mathbb{R}^n;\mathbb{E})$, and all components of f and D_zf belong to the space $L^2(\mathbf{G};[0,T]\times\mathbb{R}^n;\mathbb{E})$). For any constant $\gamma>0$ and $p\ge 2\vee n$, we can apply the Sobolev imbedding theorem (cf. section 7.10 of [9]), the Burkholder–Gundy–Davis inequality, and the Hölder inequality to conclude that there exists a constant $C_{p,\gamma}>0$ (which we allow to vary from line to line) such that, for all $s\in[0,T]$,

$$\begin{aligned}&E\left\{\sup_{\substack{t\in[s,T]\\ z\in\mathbb{R}^n}}\left[(1+|z|^2)^{-\gamma}\Big|\int_s^t f(r,z)dB_r\Big|\right]^2\right\}\\ &\le C_{p,\gamma}E\left\{\sup_{t\in[s,T]}\int_{\mathbb{R}^n}(1+|z|^2)^{-p\gamma}\left\{\Big|\int_s^t f(r,z)dB_r\Big|^p+\Big|\int_s^t D_zf(r,z)dB_r\Big|^p\right\}dz\right\}^{\frac{2}{p}}\\ &\le C_{p,\gamma}\left\{\int_{\mathbb{R}^n}(1+|z|^2)^{-p\gamma}\left\{E\left[\sup_{t\in[s,T]}\Big|\int_s^t f(r,z)dB_r\Big|^p\right]\right.\right.\\ &\qquad\left.\left.+E\left[\sup_{t\in[s,T]}\Big|\int_s^t D_zf(r,z)dB_r\Big|^p\right]\right\}dz\right\}^{\frac{2}{p}}\\ &\le C_{p,\gamma}\left\{\int_{\mathbb{R}^n}(1+|z|^2)^{-p\gamma}\left\{E\left[\int_s^T|f(r,z)|^2\,dr\right]^{\frac{p}{2}}+E\left[\int_s^T|D_zf(r,z)|^2\,dr\right]^{\frac{p}{2}}\right\}dz\right\}^{\frac{2}{p}}.\end{aligned} \tag{3.8}$$

Here we note that, as we pointed out before, if $\mathbb{E} = \mathbb{R}^{n\times m}$, then $D_z f$ should be understood as a "tensor." But such a notational complexity does not cause any substantial difficulty; therefore to simply presentation in the rest of the proof we consider only the case $n = m = 1$.

Now, differentiating (3.1) twice we have, for $s \le t \le T$,

$$D_z\eta_t^s(z) = 1 + \int_s^t \theta'(\eta_r^s(z))D_z\eta_r^s(z) \circ dB_r,$$
$$D_z^2\eta_t^s(z) = \int_s^t \{D_z^2\eta_r^s(z)\theta'(\eta_r^s(z)) + \theta''(\eta_r^s(z))[D_z\eta_t^s(z)]^2\} \circ dB_r.$$

Here and in what follows D_z^k denotes the kth derivative of the flow $z \mapsto \eta_t^s(z)$. Thus, noting assumption (H2) it is readily seen that for any $q \ge 2$,

$$E\left\{\sup_{t\in[s,T]} \left(|D_z\eta_t^s(z)|^q + |D_z^2\eta_t^s(z)|^q\right)\right\} \le \widetilde{C}_q \qquad \forall z \in \mathbb{R}^n. \tag{3.9}$$

Setting $f(s,\cdot,z) = \theta'(\eta_t^s(z))D_z\eta_t^s(z)$ in (3.8) and using (3.9) we obtain that

$$\begin{aligned} &E\left\{\sup_{\substack{t\in[s,T]\\ z\in\mathbb{R}^n}} \left[\left(1+|z|^2\right)^{-\gamma} \left|\int_s^t \theta'(\eta_t^s(z))D_z\eta_t^s(z)dB_r\right|\right]^2\right\} \\ &\le C_{p,\gamma}\left\{\int_{\mathbb{R}^n}(1+|z|^2)^{-p\gamma}\left[E\left(\int_s^T \left|\theta'(\eta_r^s(z))\right|^2 dr\right)^p\right]^{1/2} dz\right\}^{2/p} \\ &\quad + C_{p,\gamma}\left\{\int_{\mathbb{R}^n}(1+|z|^2)^{-p\gamma}\left[E\left(\int_s^T \left|\theta''(\eta_r^s(z))\right|^2 dr\right)^p\right]^{1/2} dz\right\}^{2/p} \\ &\le C_{p,\gamma}\left\{\int_{\mathbb{R}^n}(1+|z|^2)^{-p\gamma}dz\right\}^{2/p} \le C_{p,\gamma}, \end{aligned} \tag{3.10}$$

provided $2p\gamma > 1$. Moreover, from (H2) we also have

$$E\left\{\sup_{\substack{t\in[s,T]\\ z\in\mathbb{R}^n}} \left[\left(1+|z|^2\right)^{-\gamma} \left|\int_s^t [\theta'\theta]'(\eta_r^s(z))D_z\eta_r^s(z)dr\right|\right]^2\right\} \le C_{p,\gamma}. \tag{3.11}$$

This, together with (3.10), gives that

$$E\left\{\sup_{\substack{t\in[s,T]\\ z\in\mathbb{R}^n}} \left[(1+|z|^2)^{-\gamma}\left|\int_s^t \theta'(\eta_r^s(z))D_z\eta_r^s(z) \circ dB_r\right|\right]^2\right\} \le C_{p,\gamma}, \tag{3.12}$$

and consequently

$$E\left\{\sup_{\substack{t\in[s,T]\\ z\in\mathbb{R}^n}} \left[\left(1+|z|^2\right)^{-\gamma} |D_z\eta_t^s(z)|\right]^2\right\} \le C_{p,\gamma}. \tag{3.13}$$

Repeating the similar argument one also shows that, for all $\gamma > 0$ and $2p\gamma > 1$, there is some constant $C_{p,\gamma} > 0$ such that

$$E\left\{\sup_{\substack{s\in[t,T]\\ z\in R^n}}\left[(1+|z|^2)^{-\gamma}\left\{|[D_z\eta_t^s]^{-1}(z)|+|D_z^2\eta_t^s(z)|+|D_z^3\eta_t^s(z)|\right\}\right]^2\right\}\le C_{p,\gamma}. \tag{3.14}$$

We now fix $0 < \gamma < 1/12$, and define a sequence of $\mathbf{F}$-stopping times: for $\ell \ge 1$,

$$\tau_\ell \stackrel{\triangle}{=} \inf\left\{s\ge 0:\sup_{z\in\mathbb{R}^n}(1+|z|^2)^{-\gamma}\left[|[D_z\eta_s^0]^{-1}(z)|+\sum_{i=1}^3|D_z^i\eta_s^0(z)|\right]>\ell\right\}\wedge T. \tag{3.15}$$

Clearly, by virtue of (3.13) and (3.14), this sequence of stopping times $\{\tau_\ell\}$ satisfies the following properties:

(i) $0 \le \tau_1 \le \tau_2 \le \cdots$;
(ii) $\{\tau_\ell = T\} \uparrow \Omega$, P-a.s., as $\ell \to +\infty$;
(iii) for each $s \in [t, \tau_\ell)$, $z \in \mathbb{R}^n$, and $\ell \ge 1$, it holds P-a.s. that

$$|[D_z\eta_s^0]^{-1}(z)|+|D_z\eta_s^0(z)|+|D_z^2\eta_s^0(z)|+|D_z^3\eta_s^0(z)|\le \ell(1+|z|^2)^\gamma. \tag{3.16}$$

We can now easily derive the desired properties of $\widetilde{\sigma}$ and $\widetilde{b}$ by combining the definitions (3.6), (3.7) and the estimate (3.16), proving the lemma. □

A direct consequence of Lemma 3.1 is that we can now consider the following "transformed" control system of (2.2). Note that W and B are independent, and W is a $\mathbf{G}$-Brownian motion as well. Also, since both $\widetilde{b}$ and $\widetilde{\sigma}$ are $\mathbf{G}$-progressively measurable random fields, for any β in the wider-sense admissible control set $\widetilde{\mathcal{A}}$ the following SDE is well-defined:

$$F_t^{s,x,\beta} = x + \int_s^t \widetilde{b}(r, F_r^{s,x,\beta}, \beta_r)dr + \int_s^t \widetilde{\sigma}(r, F_r^{s,x,\beta}, \beta_r)dW_r, \qquad t\in[s,T]. \tag{3.17}$$

The following result validates the name "Doss–Sussmann transformation."

LEMMA 3.2. *Assume* (H1) *and* (H2). *Let* $\{\eta_t^s(z) : (t,z) \in [s,T]\times\mathbb{R}^n\}$ *be the stochastic flow given by* (3.1) *or* (3.2), $\{X_t^{s,x,\alpha} : t \in [s,T]\}$ *the solution of the SDE* (2.5), *and* $\{F_t^{s,x,\beta} : t \in [s,T]\}$ *the solution of* (3.17). *Then the following hold:*

(i) *For each* $\alpha \in \mathcal{A}$, *the solution* $F^{s,x,\alpha}$ *is* $\mathbf{F}$*-progressively measurable. Moreover, it holds that*

$$\eta_t^s(F_t^{s,x,\alpha}) = X_t^{s,\eta_s^0(x),\alpha}, \qquad t\in[s,T],\ P\text{-a.s.},$$

or, equivalently, $\eta_t^s(F_t^{s,\zeta_s^0(x),\alpha}) = X_t^{s,x,\alpha}$, $t\in[s,T]$, P*-a.s.*

(ii) *For each* $\beta \in \widetilde{\mathcal{A}}$, *we define* $X_t^{s,x,\beta} \stackrel{\triangle}{=} \eta_t^s(F_t^{s,\zeta_s^0(x),\beta})$, $t\in[s,T]$. *Then there exists a sequence of* $\mathbf{F}$*-stopping times* $\{\tau_\ell\}_{\ell\ge1}$ *with* $P\{\tau_\ell = T\}\uparrow 1$, *such that*

$$\sup_{\beta\in\widetilde{A}} E\left\{\sup_{t\in[s,\tau_\ell]}\left(\left|X_t^{s,x,\beta}\right|^p+\left|X_t^{s,\eta_s^0(x),\beta}\right|^p\right)\right\}<+\infty \quad \forall\ p\ge1.$$

Proof. Again, we shall assume $n = 1$ to simplify the presentation. The higher-dimensional case can be treated in exactly the same way without substantial difficulties.

(i) Let $\alpha \in \mathcal{A}$, and denote $F_t^\alpha = F_t^{s,x,\alpha}$ and $\eta_t = \eta_t^0$, $t \geq s$, for simplicity. Define $X_t^\alpha = \eta_t(F_t^\alpha)$, $t \geq s$. Applying the Itô–Ventzell formula (cf., e.g., [11]) in a differential form we get, for $t \in [s,T]$,

$$\begin{aligned}
dX_t^\alpha &= d\eta_t(F_t^\alpha) + D_z\eta_t(F_t^\alpha)dF_t^\alpha + \frac{1}{2}D_z^2\eta_t(F_t^\alpha)d[F^\alpha]_t \\
&= \theta(\eta_t(F_t^\alpha)) \circ dB_t + D_z\eta_t(F_t^\alpha)\widetilde{\sigma}(t,F_t^\alpha,\alpha_t)dW_t + D_z\eta_t(F_t^\alpha)\widetilde{b}(t,F_t^\alpha,\alpha_t)dt \\
&\quad + \frac{1}{2}D_z^2\eta_t(F_t^\alpha)|\widetilde{\sigma}(t,F^\alpha,\alpha_t)|^2dt \\
&= \theta(\eta_t(F_t^\alpha)) \circ dB_t + \sigma(\eta_t(F_t^\alpha),\alpha_t)dW_t \\
&\quad + \left\{ b(\eta_t(F_t^\alpha),\alpha_t) - \frac{1}{2}\left[\left| [D_z\eta_t(F_t^\alpha)]^{-1} \right|^2 |\sigma(\eta_t(F_t^\alpha),\alpha_t)|^2 D_z^2\eta_t(F_t^\alpha) \right] \right\} dt \\
&\quad + \frac{1}{2}\left| [D_z\eta_t(F_t^\alpha)]^{-1} \right|^2 |\sigma(\eta_t(F_t^\alpha),\alpha_t)|^2 D_z^2\eta_t(F_s^\alpha)dt \\
&= \theta(X_t^\alpha) \circ dB_t + \sigma(X_t^\alpha,\alpha_t)dW_t + b(X_t^\alpha,\alpha_t)dt.
\end{aligned}$$

Furthermore, at $t = s$, one has $X_s^\alpha = \eta_s^0(F_s^{s,x,\alpha}) = \eta_s^0(x)$. Thus by uniqueness of the SDE one must have $X^\alpha \equiv X^{s,\eta_s^0(z),\alpha}$.

(ii) We now let $\beta \in \widetilde{\mathcal{A}}$, and define $X_s^\beta = \eta_t(F_t^\beta)$, $t \in [s,T]$. Note that for all $(t,z) \in [s,T] \times \mathbb{R}^n$ we can write

$$\eta_t(z) = \eta_t(0) + \int_0^1 D_z\eta_t(\theta z)d\theta.$$

One can easily show, thanks to Lemma 3.1, that for all $\gamma > 0$, there exists an increasing sequence of $\mathbb{F}$-stopping times $(\tau_\ell)_{\ell \geq 1}$ with $P\{\tau_\ell = T\} \uparrow 1$, as $\ell \to \infty$, such that all the estimates of Lemma 3.1 are satisfied, and it holds furthermore that

$$|\eta_s(z)|^2 \leq \ell\left(1 + |z|^{2+\gamma}\right), \qquad s \in [t,\tau_\ell],\ z \in \mathbb{R}^n,\ P\text{-a.s.}$$

Consequently, we see that the unique $\mathbb{G}$-adapted solution F^β of (3.17) must satisfy

$$E\left\{ \sup_{s\in[t,\tau_\ell]} |F_s^\beta|^p \right\} < +\infty \qquad \forall\ p > 1.$$

Finally, it is readily seen that $X_t^\beta = \eta_t(F_t^\beta)$, $t \in [s,T]$, is a $\mathbb{G}$-adapted continuous process and it satisfies that $E\{\sup_{s\in[t,\tau_\ell]} |X_s^\beta|^p\} < +\infty$ for all $p > 1$. The other estimate is similar. The proof is complete. □

4. Wider-sense control problems. In this section we introduce two types of wider-sense stochastic control problems which will help us attack the original pathwise control problem. We begin by considering the state equation after the Doss–Sussmann transformation:

$$F_t^\beta = x + \int_s^t \widetilde{\sigma}(r,F_r^\beta,\beta_r)dW_r + \int_s^t \widetilde{b}(r,F_r^\beta,\beta_r)dr, \quad s \in [t,T], \tag{4.1}$$

where $\beta \in \widetilde{\mathcal{A}}$. Lemma 3.1 guarantees the well-posedness of this SDE, and Lemma 3.2 enables us to rewrite the cost functional (2.6) as

$$J(s,x;\alpha) = E\left\{H\left(X_T^{s,x,\alpha}\right)\middle|\, \mathcal{F}_T^B\right\} = E\left\{H\left(\eta_T^0(F_T^{s,\zeta_0^s(x),\alpha})\right)\middle|\, \mathcal{F}_T^B\right\}, \qquad P\text{-a.s.},$$

for all $\alpha \in \mathcal{A}$. Let us now define the following two new "cost functionals":

$$\widetilde{J}(s,x,\beta) = E\{H(\eta_T^0(F_T^{s,x,\beta}))|\mathcal{F}_T^B\}, \qquad \beta \in \widetilde{\mathcal{A}}, \tag{4.2}$$

and

$$\widehat{J}(s,x,\beta) = E\{\widetilde{J}(s,x,\beta)\} = E\{H(\eta_T^0(F_T^{s,x,\beta}))\}, \quad \beta \in \widetilde{\mathcal{A}}. \tag{4.3}$$

Then Lemma 3.2(i) tells us that

$$\widetilde{J}(s,x,\alpha) = J(s,\eta_s^0(x),\alpha) \qquad \forall \alpha \in \mathcal{A}. \tag{4.4}$$

In other words $\widetilde{J}$ is in some sense an "extension" of J to the wider-sense admissible control set $\widetilde{\mathcal{A}}$. The following two "wider-sense" stochastic control problems are the building blocks of our method.

Wider-sense control problem I (WSCP-I).

- State: $F^{s,x,\beta}$, $\beta \in \widetilde{\mathcal{A}}$.
- Cost functional: $\widetilde{J}(s,x;\beta)$, $\beta \in \widetilde{\mathcal{A}}$.
- Value function:

$$\widetilde{V}(t,x) = \operatorname*{essinf}_{\beta \in \widetilde{\mathcal{A}}} \widetilde{J}(t,x,\beta). \tag{4.5}$$

Remark 4.1. (i) The main purpose of introducing WSCP-I is to remove the "problematic" term involving the Brownian motion B from the state equation. However, a closer look at the cost functional and the value function should lead to the understanding that it is still a far cry from a standard stochastic control problem. For example, the cost functional not only still contains a conditional expectation, the terminal cost function is actually random (via the flow $\eta_\cdot^0$). As a consequence the value function $\widetilde{V}(s,x)$ is still a random field(!), and thus the "pathwise" nature of the problem remains.

(ii) Although the value function in WSCP-I still involves an "essinf," this time it is much more benign than the original one. The main difference is that in this case there do exist minimizing sequences to this problem.

We make a further modification to completely eliminate the "pathwise" nature of the control problem.

Wider-sense control problem II (WSCP-II).

- State: $F^{s,x,\beta}$, $\beta \in \widetilde{\mathcal{A}}$.
- Cost functional: $\widehat{J}(s,x;\beta)$, $\beta \in \widetilde{\mathcal{A}}$.
- Value function:

$$\widehat{V}(s,x) = \inf_{\beta \in \widetilde{\mathcal{A}}} \widehat{J}(s,x,\beta). \tag{4.6}$$

It is readily seen that WSCP-II looks almost like a standard stochastic control problem, except for the form of the terminal cost function (it is still random via the flow η^0). But it is much easier to handle than the previous two control problems. In the rest of this section we analyze the relationship among the two wider-sense stochastic control problems and the original pathwise control problem.

We begin by observing some more or less obvious facts. First, it is clear that for any $\beta \in \widetilde{\mathcal{A}}$, it holds that

$$E\{\widetilde{V}(s,x)\} = E\left\{\operatorname*{essinf}_{\beta \in \widetilde{\mathcal{A}}} \widetilde{J}(s,x;\beta)\right\} \le E\{\widetilde{J}(s,x;\beta)\}.$$

Thus we must have

$$E\{\widetilde{V}(s,x)\} \le \hat{V}(s,x), \qquad (s,x) \in [0,T] \times \mathbb{R}^n. \tag{4.7}$$

Next, from (4.4) we see that for all $\alpha \in \mathcal{A}$ and $(s,x) \in [0,T] \times \mathbb{R}^n$, it holds that

$$V(s,\eta_s^0(x)) = \operatorname*{essinf}_{\alpha\in\mathcal{A}} J(s,\eta_s^0(x);\alpha) = \operatorname*{essinf}_{\alpha\in\mathcal{A}} \widetilde{J}(s,x;\alpha) \ge \widetilde{V}(s,x). \tag{4.8}$$

We now give the main result of this section. Among other things, we show that the equalities in (4.7) and (4.8) both hold, and we construct a minimizing sequence for WSCP-I, which is essential for our future discussion.

THEOREM 4.2. *Assume* (H1)–(H3). *Then the following statements hold:*

(i) $\widetilde{V}(s,x) = V(s,\eta_s^0(x))$, *for all* $(s,x) \in [0,T] \times \mathbb{R}^n$, P-*a.s.*

(ii) *There exists some sequence* $\{\beta^k\}_{k\ge1} \subset \widetilde{\mathcal{A}}$ *such that*

$$\widetilde{V}(s,x) = \lim_{k\to\infty} \downarrow \widetilde{J}(s,x;\beta^k), \qquad P\text{-a.s.}$$

Here and in what follows " $\lim\downarrow$*" stands for the monotone decreasing limit.*

(iii) $E\{\widetilde{V}(s,x)\} = \widehat{V}(s,x)$ *for all* $(s,x) \in [0,T] \times \mathbb{R}^n$.

Proof. (i) From (4.8) we know that $\widetilde{V}(s,x) \le V(s,\eta_s^0(x))$. We need only show the reverse inequality. To this end, consider a subset $\widetilde{\mathcal{A}}_0 \subset \widetilde{\mathcal{A}}$ that consists of all elements of the form

$$\beta_t(\omega) = \sum_{1\le i,j\le N} u_{i,j} \mathbf{1}_{[t_{i-1},t_i)\times(A_{i,j}\cap B_{i,j})}(t,\omega), \qquad 0 \le s \le t,$$

where $N \ge 1$, $s = t_0 < t_1 < \cdots < t_N = T$, $u_{i,j} \in U$, $A_{i,j} \in \mathcal{F}_{t_{i-1}}^W$, $B_{i,j} \in \mathcal{F}_T^B$ with $B_{i,j} \cap B_{i,k} = \emptyset$ for $j \ne k$, and $\cup_{j=1}^N B_{i,j} = \Omega$.

It is not hard to check that $\widetilde{\mathcal{A}}_0$ is a dense subset of $\widetilde{\mathcal{A}}$ in the space $L^2([s,T]\times\Omega;U)$. That is, for any $\beta \in \widetilde{\mathcal{A}}$ one can find a sequence $(\beta^\ell)_{\ell\ge1} \subset \widetilde{\mathcal{A}}_0$ such that

$$E\left\{\int_s^T |\beta_r - \beta_r^\ell|^2 dr\right\} \to 0 \qquad \text{as } \ell \to \infty.$$

It follows that $\{F^{s,x,\beta^\ell}\}_{\ell\ge1}$ converges to $F^{s,x,\beta}$ in $L^0(\Omega;C([0,T];\mathbb{R}^n))$ as $\ell \to \infty$, thanks to Lemma 3.1, and that $\widetilde{J}(s,x;\beta^\ell) \to \widetilde{J}(s,x,\beta)$, in probability, as $\ell \to \infty$. Consequently, one has $\operatorname{essinf}_\ell \widetilde{J}(s,x;\beta^\ell) \le \widetilde{J}(s,x;\beta)$, P-a.s., and thus it suffices to show that

$$V(s,\eta_s^0(x)) \le \widetilde{J}(s,x;\beta) \qquad \forall\ \beta \in \widetilde{\mathcal{A}}_0. \tag{4.9}$$

To this end, we fix arbitrarily a $\beta \in \widetilde{\mathcal{A}}_0$. Denote by Λ the set of all finite sequences $\lambda = \{\lambda_i\}_{1\le i\le N}$, where λ_i's take values in the finite set $\{1,2,\dots,N\}$. For each $\lambda \in \Lambda$, we denote $B_\lambda = \cap_{i=1}^N B_{i,\lambda_i}$, and

$$\alpha_t^\lambda(\omega) = \sum_{1\le i\le N} \mathbf{1}_{[t_{i-1},t_i)}(t) u_{i,\lambda_i} \mathbf{1}_{A_{i,\lambda_i}}(\omega), \qquad (t,\omega) \in [s,T]\times\Omega.$$

Then it is readily seen that all B_λ's are $\mathcal{F}_T^B$-measurable sets, and all α^λ's are elements of the original admissible control set $\mathcal{A}$(!).

Now let us rewrite β as follows:

$$\beta_t(\omega) = \sum_{\lambda\in\Lambda} \alpha_t^\lambda(\omega)\mathbf{1}_{B_\lambda}(\omega), \qquad (t,\omega)\in[s,T]\times\Omega.$$

Observe that $\sum_{\lambda\in\Lambda} B_\lambda = \Omega$, and the SDE (4.1) does not contain a stochastic integral with respect to the Brownian motion B. Using the total probability formula and the uniqueness of solution to the SDE, one can show that the following decomposition holds:

$$F_t^{s,x,\beta} = \sum_{\lambda\in\Lambda} F_t^{s,x,\alpha^\lambda}\mathbf{1}_{B_\lambda}, \qquad t\in[s,T],\ P\text{-a.s.}$$

Similarly, applying the same arguments and noting Lemma 3.2(i) we also have

$$\eta_t^0(F_t^{s,x,\beta}) = \eta_t^0\left(\sum_{\lambda\in\Lambda} F_t^{s,x,\alpha^\lambda}\mathbf{1}_{B_\lambda}\right) = \sum_{\lambda\in\Lambda}\eta_t^0(F_t^{s,x,\alpha^\lambda})\mathbf{1}_{B_\lambda} = \sum_{\lambda\in\Lambda} X_t^{s,\eta^0(x),\alpha^\lambda}\mathbf{1}_{B_\lambda}$$

for all $t\in[s,T]$. Therefore, we have

$$\begin{aligned}\widetilde{J}(s,x,\beta) &= E\left\{H\Big(\eta_T^0(F_T^{s,x,\beta})\Big)\Big|\mathcal{F}_T^B\right\} = \sum_{\lambda\in\Lambda} E\left\{H(X_T^{s,\eta_s^0(x),\alpha^\lambda})\Big|\mathcal{F}_T^B\right\}\mathbf{1}_{B_\lambda}\\ &= \sum_{\lambda\in\Lambda} J(s,\eta^0(x),\alpha^\lambda)\mathbf{1}_{B_\lambda} \geq V(s,\eta^0(x)),\end{aligned}$$

proving (4.9), whence (i).

(ii) Let $(s,x)\in[0,T]\times\mathbb{R}^n$ be fixed. We first choose a sequence $\{\widetilde{\beta}^k\}_{k\geq1}\subset\widetilde{\mathcal{A}}$ such that

$$\widetilde{V}(s,x) = \operatorname*{essinf}_{k\geq1}\widetilde{J}(s,x;\widetilde{\beta}^k).$$

To do this, we borrow the idea of [8]. For any $\varepsilon>0$, define

$$\delta = \inf_{\{\beta^k\}\subset\widetilde{\mathcal{A}}} P\left\{\widetilde{V}(s,x)\leq\operatorname*{essinf}_{k\geq1}\widetilde{J}(s,x;\beta^k)-\varepsilon\right\}. \tag{4.10}$$

(Note that the infimum above is taken over all the sequences $\{\beta^k\}$ in $\widetilde{\mathcal{A}}$!) We claim the following two facts:

(a) the infimum in (4.10) is always attained, and

(b) $\delta=0$.

To prove (a), we first use the definition of "inf" to find, for $n=1,2,\ldots$, sequences $\{\beta^{n,k}\}_{k\geq1}$, $n=1,2,\ldots$, such that

$$\delta\leq P\left\{\widetilde{V}(s,x)\leq\operatorname*{essinf}_{k\geq1}\widetilde{J}(s,x;\beta^{n,k})-\varepsilon\right\}<\delta+\frac{1}{n}.$$

Let us consider the sequence $\{\beta^{n,k}\}_{n\geq1,k\geq1}$, and denote

$$\begin{aligned}A_\varepsilon^n &\triangleq \left\{\omega:\widetilde{V}(s,x)\leq\operatorname*{essinf}_{k\geq1}\widetilde{J}(s,x;\beta^{n,k})-\varepsilon\right\};\\ \bar{A}_\varepsilon &\triangleq \left\{\omega:\widetilde{V}(s,x)\leq\operatorname*{essinf}_{n\geq1,k\geq1}\widetilde{J}(s,x;\beta^{n,k})-\varepsilon\right\}.\end{aligned}$$

Then it is easily seen that $\bar{A}_\varepsilon \subseteq A^n_\varepsilon$ for all n. This, together with the definition of δ, leads to

$$\delta \le P\{\bar{A}_\varepsilon\} \le P\{A^n_\varepsilon\} = P\left\{\widetilde{V}(s,x) \le \operatorname*{essinf}_{k\ge 1} \widetilde{J}(s,x;\beta^{n,k}) - \varepsilon\right\} < \delta + \frac{1}{n}$$

for all $n \ge 1$. Letting $n \to \infty$ we obtain that

$$\delta = P\left\{\widetilde{V}(s,x) \le \operatorname*{essinf}_{n\ge 1,k\ge 1} \widetilde{J}(s,x;\beta^{n,k}) - \varepsilon\right\}.$$

This proves (a).

(b) By a rearrangement of indices let us denote the minimizer in part (a) by $\{\widetilde{\beta}^k\}_{k\ge 1}$. For any $\varepsilon > 0$ and $\beta \in \widetilde{\mathcal{A}}$ we denote

$$\bar{A}_\varepsilon \triangleq \left\{\widetilde{V}(s,x) \le \operatorname*{essinf}_{k\ge 1} \widetilde{J}(s,x;\widetilde{\beta}^k) - \varepsilon\right\}, \qquad A_\varepsilon(\beta) \triangleq \left\{\widetilde{V}(s,x) \le \widetilde{J}(s,x;\beta) - \varepsilon\right\}.$$

Suppose that $P(\bar{A}_\varepsilon) = \delta > 0$. Then we claim that for each $\varepsilon > 0$ there exists a $\bar{\beta} \in \widetilde{\mathcal{A}}$ such that

$$P\{\bar{A}_\varepsilon \setminus A_{\frac{\varepsilon}{2}}(\bar{\beta})\} > 0. \tag{4.11}$$

Indeed, if for all $\beta \in \widetilde{\mathcal{A}}$ one has $P\{\bar{A}_\varepsilon \setminus A_{\frac{\varepsilon}{2}}(\beta)\} = 0$, then for all $\beta \in \widetilde{\mathcal{A}}$ one must have

$$\widetilde{V}(s,x) + \frac{\varepsilon}{2}\mathbf{1}_{\bar{A}_\varepsilon} \le \widetilde{J}(s,x;\beta), \qquad P\text{-a.s.}$$

But then it follows that

$$\widetilde{V}(s,x) + \frac{\varepsilon}{2}\mathbf{1}_{\bar{A}_\varepsilon} \le \operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} \widetilde{J}(s,x;\beta) = \widetilde{V}(s,x), \qquad P\text{-a.s.},$$

contradicting $P(\bar{A}_\varepsilon) = \delta > 0$. Hence (4.11) must hold, and consequently

$$\begin{aligned} P\left\{\widetilde{V}(s,x) \le \widetilde{J}(s,x;\bar{\beta}) \wedge \operatorname*{essinf}_{k\ge 1} \widetilde{J}(s,x;\widetilde{\beta}^k) - \varepsilon\right\} &\le P\{\bar{A}_e \cap A_\varepsilon(\bar{\beta})\} \\ &\le P\{\bar{A}_e \cap A_{\frac{\varepsilon}{2}}(\bar{\beta})\} = P\{\bar{A}_\varepsilon\} - P\{\bar{A}_\varepsilon \setminus A_{\frac{\varepsilon}{2}}(\bar{\beta})\} < \delta. \end{aligned} \tag{4.12}$$

Let us now modify the sequence $\{\widetilde{\beta}^k\}$ slightly: define

$$\bar{\beta}^\kappa \triangleq \begin{cases} \widetilde{\beta}^k & \text{on } \{\widetilde{J}(s,x;\widetilde{\beta}^k) \le \widetilde{J}(s,x;\bar{\beta})\}, \\ \bar{\beta} & \text{on } \{\widetilde{J}(s,x;\widetilde{\beta}^k) > \widetilde{J}(s,x;\bar{\beta})\}, \end{cases} \qquad k \ge 1. \tag{4.13}$$

Since $J(s,x,\bar{\beta})$ and $J(s,x,\widetilde{\beta}^k)$'s are all $\mathcal{F}^B_T$-measurable, $\{\bar{\beta}^k\}_{k\ge 1} \subset \widetilde{\mathcal{A}}$. Furthermore, by definition it is readily seen that

$$\widetilde{J}(s,x;\bar{\beta}^k) \le \widetilde{J}(s,x;\bar{\beta}) \wedge \widetilde{J}(s,x;\widetilde{\beta}^k) \qquad \forall k \ge 1.$$

Thus, using (4.12) we obtain that

$$P\left\{\widetilde{V}(s,x) \le \operatorname*{essinf}_{k\ge 1} \widetilde{J}(s,x;\bar{\beta}^k) - \varepsilon\right\} < \delta.$$

This contradicts the definition of δ, proving (b).

To conclude the proof we use a similar technique to construct the desired minimizing sequence inductively as follows. Let $\beta^1 = \widetilde{\beta}^1$, and for $k \geq 2$, we define

$$\beta^\kappa \triangleq \begin{cases} \widetilde{\beta}^k & \text{on } \{\widetilde{J}(s,x;\widetilde{\beta}^k) \leq \widetilde{J}(s,x;\beta^{k-1})\}, \\ \beta^{k-1} & \text{on } \{\widetilde{J}(s,x;\widetilde{\beta}^k) > \widetilde{J}(s,x;\beta^{k-1})\}, \end{cases} \qquad k \geq 1. \tag{4.14}$$

Again, we have $\{\beta^k\}_{k\geq 1} \subset \widetilde{\mathcal{A}}$ and $\widetilde{J}(s,x;\beta^k) \leq \widetilde{J}(s,x;\beta^{k-1}) \wedge \widetilde{J}(s,x;\widetilde{\beta}^k)$ for all $k \geq 1$. Consequently, we have

$$\lim_{k\to\infty} \downarrow \widetilde{J}(s,x;\beta^k) \leq \operatorname*{essinf}_{k\geq 1} \widetilde{J}(s,x;\widetilde{\beta}^k) = \widetilde{V}(s,x).$$

Finally, that $\widetilde{V}(s,x) \leq \lim_{k\to\infty} \downarrow \widetilde{J}(s,x;\beta^k)$ is obvious. We proved (ii).

(iii) is a direct consequence of (ii). Thus proof is complete. □

We remark that part (ii) in Theorem 4.2 provides us the first version of a "minimizing sequence"(!). As we can see, the construction of such a sequence depends heavily on $\widetilde{J}$, hence $\mathcal{F}_T^B$-measurable. Thus the wider-sense admissible class $\widetilde{\mathcal{A}}$ is essential. The counterexample in the appendix shows that this cannot be relaxed.

To end this section let us take a brief look at the existence of optimal control. Note that WSCP-II is now a rather standard optimal control problem; therefore with a possible change of probability space, one should always be able to find an optimal control, at least in a "relaxed" form (cf., e.g., El Karoui, Nguyen, and Jeanblanc-Picqué [8]). We do not pursue this issue here due to the length of the paper. However, we give the following corollary of Theorem 4.2 that more or less explains the benefit of introducing the wider-sense controls.

COROLLARY 4.3. *Assume* (H1)–(H3). *Then any optimal control β^* for the WSCP-*II *is also an optimal control for WSCP-*I. *That is, if for $(s,x) \in [0,T] \times \mathbb{R}^n$, $\beta^* \in \widetilde{\mathcal{A}}$ is such that $\widehat{V}(s,x) = \widehat{J}(s,x;\beta^*)$, then it must hold that*

$$\widetilde{V}(s,x) = \widetilde{J}(s,x;\beta^*), \qquad P\text{-a.s.}$$

Proof. Let $\beta^* \in \widetilde{\mathcal{A}}$ be an optimal control for WSCP-II, that is, $\widehat{V}(s,x) = \widehat{J}(s,x,\beta^*)$. We show that it actually holds that

$$\widetilde{J}(s,x,\beta^*) \leq \widetilde{J}(s,x,\beta) \qquad \forall\ \beta \in \widetilde{\mathcal{A}}. \tag{4.15}$$

To see this let $\beta \in \widetilde{\mathcal{A}}$ be arbitrary. We define, as before, a new control $\widetilde{\beta}$ be such that

$$\widetilde{\beta} = \beta \mathbf{1}_{\{\widetilde{J}(s,x,\beta) < \widetilde{J}(s,x,\beta*)\}} + \beta^* \mathbf{1}_{\{\widetilde{J}(s,x,\beta) \geq \widetilde{J}(s,x,\beta^*)\}}.$$

Again, we have $\widetilde{\beta} \in \widetilde{\mathcal{A}}$, and the optimality of β^* leads to

$$\begin{aligned} E[\widetilde{J}(s,x,\beta^*)] &\leq E[\widetilde{J}(s,x,\widetilde{\beta})] \\ &= E\left[\widetilde{J}(s,x,\beta)\mathbf{1}_{\{\widetilde{J}(s,x,\beta) < \widetilde{J}(s,x,\beta^*)\}} + \widetilde{J}(s,x,\beta^*)\mathbf{1}_{\{\widetilde{J}(s,x,\beta) \geq \widetilde{J}(s,x,\beta^*)\}}\right], \end{aligned}$$

or, equivalently, $E[(\widetilde{J}(s,x,\beta) - \widetilde{J}(s,x,\beta^*))\mathbf{1}_{\{\widetilde{J}(s,x,\beta) < \widetilde{J}(s,x,\beta^*)\}}] \geq 0$. Thus we obtain that $P\{\widetilde{J}(s,x,\beta) < \widetilde{J}(s,x,\beta^*)\} = 0$, proving (4.15), whence the corollary. □

5. Properties of the value function $\widetilde{V}$. In this section we take a closer look at the value function of WSCP-I, $\widetilde{V}(s,x)$. To be more precise we would like to derive some finer results on its measurability as well as regularity, as a random field. These properties will be important for us to derive the Bellman principle, and ultimately the stochastic HJB equation.

First note that by definition we know immediately that $\widetilde{V}(s,x)$ is $\mathcal{F}_T^B$-measurable for all $(s,x)\in[0,T]\times\mathbb{R}^n$, and therefore it is $\mathbf{G}$-progressively measurable. On the other hand, note that

$$\begin{aligned}\widetilde{J}(s,x,\beta) &= E\left\{H(\eta_T^0(F_T^{s,x,\beta})\Big|\mathcal{F}_T^B\right\} = E\left\{E\left\{H\left(\eta_T^0(F_T^{s,x,\beta})\right)\Big|\mathcal{G}_s\right\}\Big|\mathcal{F}_T^B\right\}\\ &\geq E\left\{\operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\left\{H\Big(\eta_T^0(F_T^{s,x,\beta})\Big)\Big|\mathcal{G}_s\right\}\Big|\mathcal{F}_T^B\right\}.\end{aligned}$$

Therefore we can only have

$$\widetilde{V}(s,x)\geq E\left\{\operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\left\{H\Big(\eta_T^0(F_T^{s,x,\beta})\Big)\Big|\mathcal{G}_s\right\}\Big|\mathcal{F}_T^B\right\}.$$

Let us first establish a stronger result than the relation above and construct another version of minimizing sequence, which is essential for us to derive the Bellman principle.

THEOREM 5.1. *Assume* (H1)–(H3). *For all* $(s,x)\in[0,T]\times\mathbb{R}^n$, *there exists a sequence* $\{\beta^k\}\subset\widetilde{\mathcal{A}}$, *such that*

$$\widetilde{V}(s,x)=\operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\left\{H\left(\eta_T^0(F_T^{s,x,\beta})\right)\Big|\mathcal{G}_s\right\}=\lim_{k\to\infty}\downarrow E\left\{H\left(\eta_T^0(F_T^{s,x,\beta^k})\right)\Big|\mathcal{G}_s\right\},\ P\text{-a.s.}$$

Proof. We first show that, for any $(s,x)\in[0,T]\times\mathbb{R}^n$,

$$\widetilde{V}(s,x)\leq\operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\left\{H(\eta_T^0(F_T^{s,x,\beta}))\Big|\mathcal{G}_s\right\},\qquad P\text{-a.s.}\tag{5.1}$$

To this end, we fix $(s,x)\in[0,T]\times\mathbb{R}^n$ and $\beta\in\widetilde{\mathcal{A}}$, and construct a special sequence $\{\beta^k\}_{k\geq1}$ that approximates β in $L^2([0,T]\times\Omega)$. We proceed as follows. First, let us denote for each $k\geq1$ and $0\leq a<b\leq T$

$$D_k^{a,b}\triangleq\left\{a+\frac{i}{2^k}(b-a):i=0,1,2,\dots,2^k\right\},$$

and denote the generic elements of $D_k^{a,b}$ to be $t_i^{k,a,b}$ $(=a+\frac{i}{2^k}(b-a))$, $i=0,1,2,\dots,2^k$. To simplify notation we shall denote $D_k^a=D_k^{0,a}$, and $t_i^{k,a}=t_i^{k,0,a}$ for $a\in[0,T]$. Finally we define $D_k=D_k^s\cup D_k^{s,T}$.

Now consider the probability space $([0,T]\times\Omega,\mathcal{P},\mu)$, where $\mathcal{P}$ is the $\mathbf{G}$-predictable σ-field on $[0,T]\times\Omega$ and $\mu(dtd\omega)=\frac{1}{T}dtP(d\omega)$. Let $\mathcal{G}_t^s\triangleq\sigma\{W_r-W_s,\ r\in[s,t]\}\vee\mathcal{F}_T^B$, $t\geq s$, and introduce the following σ-fields:

$$\begin{aligned}\mathcal{G}_\ell^{k,s}&\triangleq\sigma\{\Delta W_{t_i^{k,s}},\ \Delta B_{t_j^{k,T}}:\ 0\leq i\leq\ell-1;\ 0\leq j\leq2^k-1\},\\ \widetilde{\mathcal{G}}_\ell^{k,s,T}&\triangleq\sigma\{\Delta W_{t_i^{k,s}},\ \Delta W_{t_{i'}^{k,s,T}},\ \Delta B_{t_j^{k,T}}:\ 0\leq i'\leq\ell-1;\ 0\leq i,j\leq2^k-1\}.\end{aligned}$$

where $\Delta\xi_{t_i^{k,a,b}} \triangleq \xi_{t_{i+1}^{k,a,b}} - \xi_{t_i^{k,a,b}}$, $\xi = W, B$. Now let

$$\mathcal{P}_k \triangleq \sigma\left\{(t_\ell^{k,s}, t_{\ell+1}^{k,s}] \times A_\ell;\ (t_\ell^{k,s,T}, t_{\ell+1}^{k,s,T}] \times \widetilde{A}_\ell :\ A_\ell \in \mathcal{G}_\ell^{k,s},\ \widetilde{A}_\ell \in \widetilde{\mathcal{G}}_\ell^{k,s,T}\right\}.$$

It is then clear that $\{\mathcal{P}_k\}_{k\geq 1}$ is an increasing family of σ-fields, and $\mathcal{P}_k \uparrow \mathcal{P}$ as $k \to \infty$ (cf., e.g., Dellacherie and Meyer [6]). Furthermore, if we define $\beta^k \triangleq E^\mu\{\beta|\mathcal{P}_k\}$, then we must have $\beta^k \in \widetilde{\mathcal{A}}$, and $\beta^k \to \beta$ in $L^2([0,T]\times\Omega)$ as $k \to \infty$. Consequently, possibly along a subsequence (we may again denote it by $\{\beta^k\}$), one has

$$E\left\{H(\eta_T^0(F_T^{s,x,\beta^k}))\Big|\mathcal{G}_s\right\} \to E\left\{H(\eta_T^0(F_T^{s,x,\beta}))\Big|\mathcal{G}_s\right\}, \qquad P\text{-a.s.} \quad \text{as } k \to \infty.$$

Now let us use the sequence $\{\beta^k\}$ to prove (5.1). First note that by a monotone-class argument and the structure of the σ-fields $\mathcal{P}_k$'s, it can be shown that the processes β^k's have the following representations:

$$\beta_t^k(\omega) = \gamma^k(t, W_{t_1^{k,s}}(\omega), W_{t_2^{k,s}}(\omega), \dots, W_{t_{2^k-1}^{k,s}}(\omega), \omega), \qquad (t,\omega) \in [0,T]\times\Omega,$$

where, for $y = (y_1, \dots, y_{2^k}) \in \mathbb{R}^{2^k d}$,

$$\begin{aligned}\gamma^k(t,y,\omega) = &\sum_{\ell=0}^{2^k-1} f_{k,\ell}(y1, \dots, y_\ell, B_{t_1^{k,T}}(\omega), \dots, B_{t_{2^k-1}^{k,T}}(\omega))\mathbf{1}_{(t_\ell^{k,s}, t_{\ell+1}^{k,s}]}(t)\\ &+ \sum_{\ell=0}^{2^k-1} g_{k,\ell}(y, W_{t_1^{t,s,T}}, \dots, W_{t_{2^k-1}^{k,s,T}}(\omega), B_{t_1^{k,T}}(\omega), B_{t_{2^k-1}^{k,T}}(\omega))\mathbf{1}_{(t_\ell^{k,s,T}, t_{\ell+1}^{k,s,T}]}(t),\end{aligned}$$

where $f_{k,\ell} : \mathbb{R}^{\ell d} \times \mathbb{R}^{2^k m} \mapsto U$ and $g_{k,\ell} : \mathbb{R}^{2^k d} \times \mathbb{R}^{\ell d} \times \mathbb{R}^{2^k m} \mapsto U$ are Borel measurable functions. Note that $\gamma^k : [0,T]\times\mathbb{R}^{2^k d}\times\Omega \mapsto \mathbb{R}^n$ is a $\mathbf{G}^s$-progressive random field (that is, for each $y \in \mathbb{R}^{2^k n}$, the mapping $(t,\omega) \mapsto \gamma^k(t,y,\omega)$ is $\mathcal{G}_t^s$-progressively measurable). We see that for each $y \in \mathbb{R}^{2^k d}$, the random variable $H(\eta_T^0(F_T^{s,x,\gamma^k(y,\cdot)}))$ is independent of $\mathcal{F}_s^W$. Therefore,

$$\begin{aligned}E\left\{H\left(\eta_T^0(F_T^{s,x,\beta})\right)\Big|\mathcal{G}_s\right\} &\geq \operatorname*{essinf}_{k\geq 1} E\left\{H\left(\eta_T^0(F_T^{s,x,\beta^k})\right)\Big|\mathcal{G}_s\right\}\\ &\geq \operatorname*{essinf}_{k\geq 1}\inf_{y\in\mathbb{R}^{2^k d}} E\left\{H\left(\eta_T^0(F_T^{s,x,\gamma^k(y,\cdot)})\right)\Big|\mathcal{G}_s\right\}\\ &= \operatorname*{essinf}_{k\geq 1}\inf_{y\in\mathbb{R}^{2^k d}} E\left\{H\left(\eta_T^0(F_T^{s,x,\gamma^k(y,\cdot)})\right)\Big|\mathcal{F}_T^B\right\}\\ &\geq \operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\left\{H\left(\eta_T^0(F_T^{s,x,\beta})\right)\Big|\mathcal{F}_T^B\right\} = \widetilde{V}(s,x), \qquad P\text{-a.s.}\end{aligned}$$

We now prove the reversed inequality of (5.1) and construct another "minimizing sequence." In fact we will show that there exists a sequence $\{\beta^k\} \subset \widetilde{\mathcal{A}}$ such that

$$\widetilde{V}(s,x) = \lim_{k\to\infty} \downarrow E\left\{H\left(\eta_T^0(F_T^{s,x,\beta^k})\right)\Big|\mathcal{G}_s\right\}, \qquad P\text{-a.s.} \tag{5.2}$$

Since the right side above is obviously no less than $\operatorname{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\{H(\eta_T^0(F_T^{s,x,\beta}))|\mathcal{G}_s\}$, the reversed inequality of (5.1), whence the theorem, will follow.

To construct the desired minimizing sequence, let us first choose a sequence $\{\hat\beta^k\}_{k\ge1}\subset\widetilde{\mathcal{A}}$ such that $\widetilde V(s,x)=\lim_{k\to\infty}\downarrow\widetilde J(s,x,\hat\beta^k)$, P-a.s., thanks to Theorem 4.2(ii), and then modify it as follows. Let $\beta^1\triangleq\hat\beta^1$, and for $k\ge2$, we denote recursively that

$$A^k\triangleq\Big\{E\Big\{H\Big(\eta_T^0(F_T^{s,x,\hat\beta^k})\Big)\Big|\mathcal{G}_s\Big\}\le E\Big\{H\Big(\eta_T^0(F_T^{s,x,\beta^{k-1}})\Big)\Big|\mathcal{G}_s\Big\}\Big\},$$

and define, for $(t,\omega)\in[s,T]\times\Omega$,

$$\beta_t^k(\omega)=\hat\beta_t^k(\omega)\mathbf{1}_{([s,T]\times A^k)\cup([0,s]\times\Omega)}(t,\omega)+\beta_t^{k-1}(\omega)\mathbf{1}_{[s,T]\times[A^k]^c}(t,\omega).$$

Then obviously $\{\beta^k\}_{k\ge1}\subset\widetilde A$, since A^k's are all $\mathbf{G}$-progressively measurable. Further, it holds that

$$H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)=H\Big(\eta_T^0(F_T^{s,x,\hat\beta^k})\Big)\mathbf{1}_{A^k}+H\Big(\eta_T^0(F_T^{s,x,\beta^{k-1}})\Big)\mathbf{1}_{(A^k)^c}.$$

Therefore, we have

$$E\Big\{H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)\Big|\mathcal{G}_s\Big\}\le E\Big\{H\Big(\eta_T^0(F_T^{s,x,\hat\beta^k})\Big)\Big|\mathcal{G}_s\Big\}\wedge E\Big\{H\Big(\eta_T^0(F_T^{s,x,\beta^{k-1}})\Big)\Big|\mathcal{G}_s\Big\}. \tag{5.3}$$

Consequently, the sequence $\{E\{H(\eta_T^0(F^{s,x,\beta^k}))|\mathcal{G}_s\}:k\ge1\}$ is monotone decreasing, and by (5.1),

$$\widetilde V(s,x)\le\lim_{k\to\infty}\downarrow E\Big\{H\Big(\eta_T^0(F_T^{t,x,\beta^k})\Big)\Big|\mathcal{G}_s\Big\},\qquad P\text{-a.s.} \tag{5.4}$$

But on the other hand, noting the definition of $\{\hat\beta^k\}_{k\ge1}$, the fact (5.3), and applying the monotone convergence theorem, we also have

$$\begin{aligned}
&E[\widetilde V(s,x)]=E\Big\{\lim_{k\to\infty}\downarrow\widetilde J(s,x,\hat\beta^k)\Big\}=\lim_{k\to\infty}\downarrow E\{\widetilde J(s,x,\hat\beta^k)\}\\
&=\lim_{k\to\infty}\downarrow E\Big\{E\Big\{H\Big(\eta_T^0(F_T^{s,x,\hat\beta^k})\Big)\Big|\mathcal{G}_s\Big\}\Big\}\ge\lim_{k\to\infty}\downarrow E\Big\{E\Big\{H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)\Big|\mathcal{G}_s\Big\}\Big\}\\
&=E\Big\{\lim_{k\to\infty}\downarrow E\Big\{H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)\Big|\mathcal{G}_s\Big\}\Big\}.
\end{aligned}$$

This, together with (5.4), leads to (5.2), completing the proof. □

We now turn to the regularity of the value function $\widetilde V(s,x)$. It is well understood that in a standard stochastic control problem the value function is usually (locally) Lipschitz in the spatial variable, and Hölder-1/2 in the temporal variable. We will show in the next theorem that this is in principle still true, but with a slight modification.

We first give a lemma that concerns the solution $F^{s,x,\beta}$ of the SDE (3.17). Let us begin by recalling Lemma 3.1 and the sequence of $\mathcal{F}_T^B$-measurable random times (whence $\mathbf{G}$-stopping times!) $\{\tau_\ell\}_{\ell\ge1}$ there (see (3.15)):

$$\tau_\ell=\inf\left\{s\ge0:\sup_{z\in\mathbb{R}^n}(1+|z|^2)^{-\gamma}\left(\Big|[D_z\eta_s^0(z)]^{-1}\Big|+\sum_{i=1}^3\Big|D_z^i\eta_s^0(z)\Big|\right)>\ell\right\}\wedge T, \tag{5.5}$$

for $\ell=1,2,\dots$. We can localize $F^{s,x,\beta}$ even further: for any $M>0$, and fixed $(s,x)\in[0,T]\times\mathbb{R}^n$ and $\beta\in\widetilde{\mathcal{A}}$, we define

$$\tau_{\ell,M}^{s,x,\beta}=\inf\Big\{t\ge s:|F_t^{s,x,\beta}|>M,\ \tau_\ell=T\Big\}\wedge T. \tag{5.6}$$

Then it is clear that $\tau_{\ell,M}^{s,x,\beta}$ is again a $\mathbf{G}$-stopping time. We have the following lemma.

LEMMA 5.2. *Assume* (H1)—H(3). *Then, for any $p \geq 1$ and $\ell \geq 1$, there exists a constant $C_{\ell,p} > 0$ such that for any $(s,x),(s',x') \in [0,T] \times \mathbb{R}^n$, $\beta \in \widetilde{\mathcal{A}}$, and $M > 0$, it holds P-a.s. on the set $\{\tau_\ell = T\}$ that*

$$E\Big\{ \sup_{t\in[s,T]} |F_t^{s,x,\beta}|^p \Big| \mathcal{G}_s \Big\} \leq C_{\ell,p}(1+|x|^p); \tag{5.7}$$

$$P\Big\{ \tau_{\ell,M}^{s,x,\beta} < M \Big| \mathcal{G}_s \Big\} \leq \frac{1}{M^p} E\Big\{ \sup_{t\in[s,T]} |F_t^{s,x,\beta}|^p \Big| \mathcal{G}_s \Big\} \leq C_{\ell,p} \frac{(1+|x|^p)}{M^p}; \tag{5.8}$$

$$E\Big\{ \sup_{s\leq t\leq \tau_{\ell,M}^{s,x,\beta}\wedge\tau_{\ell,M}^{s,x',\beta}} |F_t^{s,x,\beta} - F_t^{s,x',\beta}|^p \Big| \mathcal{G}_s \Big\} \leq C_{\ell,p} e^{M^p} |x-x'|^p; \tag{5.9}$$

$$E\Big\{ \sup_{s\leq t\leq \tau_{\ell,M}^{s,x,\beta}\wedge\tau_{\ell,M}^{s',x,\beta}} |F_t^{s,x,\beta} - F_t^{s',x,\beta}|^p \Big| \mathcal{G}_s \Big\} \leq C_{\ell,p} e^{M^p} (1+|x|^p) |s-s'|^{\frac{p}{2}}. \tag{5.10}$$

Proof. Note that all the following discussions are restricted to the set $\{\tau_\ell = T\}$. By virtue of the (localized) Lipschitz conditions and the linear growth properties of the coefficients $\widetilde{b}$ and $\widetilde{\sigma}$, thanks to Lemma 3.1, the estimates (5.7), (5.9), and (5.10) follow easily from some standard arguments for SDEs. Further, the estimate (5.8) is a direct consequence of definition (5.6), Chebyshev's inequality, and (5.7). We leave the details to the reader. □

The second main result of this section is the following. One should note that the parameter $\varepsilon > 0$ makes our result different from similar ones in standard stochastic control theory.

THEOREM 5.3. *Assume* (H1)–(H3). *Then, for any $\ell \geq 1$, $R > 0$, and $\varepsilon > 0$, there exists a $C_{\ell,R,\varepsilon} > 0$, such that for all $(s,x),(s',x') \in [0,T] \times \mathbb{B}_R^n$, it holds P-a.s. on $\{\tau_\ell = T\}$ that*

$$|\widetilde{V}(s,x) - \widetilde{V}(s',x')| \leq \varepsilon + C_{\ell,R,\varepsilon} \|H\|_\infty (|s-s'|^{\frac{1}{2}} + |x-x'|). \tag{5.11}$$

Here $\mathbb{B}_R^n \triangleq \{z \in \mathbb{R}^n : |z| \leq R\}$, and $\|\cdot\|_\infty$ is the sup-norm of $L^\infty(\mathbb{R}^n)$.

Proof. We first fix $R > 0$ and let $(s,x),(s',x') \in [0,T] \times \mathbb{B}_R^n$ be arbitrarily given. Then by Theorem 5.1 we can choose a sequence $\{\beta^k\}_{k\geq 1}$ such that

$$\widetilde{V}(s,x) = \lim_{k\to\infty} \downarrow E\Big\{ H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big) \Big| \mathcal{G}_s \Big\}, \qquad P\text{-a.s.} \tag{5.12}$$

Since H is uniformly continuous by (H3), for any $\varepsilon > 0$ we can find $\delta > 0$ such that

$$|H(z) - H(z')| < \frac{\varepsilon}{4}, \qquad \text{whenever } |z-z'| < \delta. \tag{5.13}$$

Now fix ℓ such that $P\{\tau_\ell = T\} > 0$. For each $M > 0$ and $k > 0$, we recall the stopping times defined by (5.6) and denote, for simplicity, that

$$\tau_{k,M} \triangleq \tau_{\ell,M}^{s,x,\beta^k}, \qquad \tau'_{k,M} \triangleq \tau_{\ell,M}^{s,x',\beta^k}, \qquad \tau''_{k,M} \triangleq \tau_{\ell,M}^{s',x',\beta^k},$$

and define $\bar{\tau}_{k,M} \triangleq \tau_{k,M} \wedge \tau'_{k,M} \wedge \tau''_{k,M}$. Then using (5.8) (with $p=2$) we have

$$P\{\bar{\tau}_{k,M} < T | \mathcal{G}_{s\wedge s'}\} \leq \frac{1}{M^2} C_{\ell,2}(1+|x|^2+|x'|^2) \leq C_{\ell,2} \frac{(1+2R^2)}{M^2}. \tag{5.14}$$

Bearing in mind that $\widetilde{V}(s,x)$ is $\mathcal{F}_T^B$-measurable, and $\mathcal{F}_T^B \subset \mathcal{G}_t$ for all $t \geq 0$, we can easily check that the minimizing sequence in Theorem 5.1 also satisfies

$$\widetilde{V}(s,x) = \lim_{k\to\infty} \downarrow E\left\{H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)\Big|\mathcal{G}_{s\wedge s'}\right\}, \qquad P\text{-a.s.}$$

On the other hand, we observe that

$$\widetilde{V}(s',x') = \operatorname*{essinf}_{\beta\in\widetilde{\mathcal{A}}} E\left\{H\Big(\eta_T^0(F_T^{s',x',\beta})\Big)\Big|\mathcal{G}_{s\wedge s'}\right\}, \qquad P\text{-a.s.}$$

Keeping these in mind we now define

$$A_{\varepsilon,\ell}^k = \left\{\tau_\ell = T; \widetilde{V}(s,x) \geq E\left\{H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)\Big|\mathcal{G}_{s\wedge s'}\right\} - \frac{\varepsilon}{4}\right\}.$$

Then $\lim_{k\to\infty} P\{A_{\varepsilon,\ell}^k\} = P\{\tau_\ell = T\} > 0$, and hence $P\{A_{\varepsilon,\ell}^k\} > 0$ for k large enough. Also, on the set $A_{\varepsilon,\ell}^k$ we have

$$\begin{aligned}(5.15)\qquad &\widetilde{V}(s',x') - \widetilde{V}(s,x)\\ &\leq \frac{\varepsilon}{4} + \left|E\left\{H\Big(\eta_T^0(F_T^{s',x',\beta^k})\Big)\Big|\mathcal{G}_{s\wedge s'}\right\} - E\left\{H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big)\Big|\mathcal{G}_{s\wedge s'}\right\}\right|\end{aligned}$$

We now analyze the right-hand side above. Again let us simplify the notation a little bit. Denote

$$\begin{aligned}&\Delta F_T^k \triangleq F_T^{s',x',\beta^k} - F_T^{s,x,\beta^k}, \quad \Delta\eta(F_T^k) \triangleq \eta_T^0(F_T^{s',x',\beta^k}) - \eta_T^0(F_T^{s,x,\beta^k}),\\ &\Delta H(\eta(F_T^k)) \triangleq H\left(\eta_T^0(F_T^{s',x',\beta^k})\right) - H\left(\eta_T^0(F_T^{s,x,\beta^k})\right).\end{aligned}$$

Recalling the definition of $\bar{\tau}_{k,M}$ and δ we see that the second term on the right-hand side of (5.15) becomes

$$\begin{aligned}&\left|E\left\{\Delta H(\eta(F_T^k))\Big|\mathcal{G}_{s\wedge s'}\right\}\right|\\ &= \left|E\left\{\Delta H(\eta(F_T^k))\mathbf{1}_{\{\bar{\tau}_{k,M}<T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} + E\left\{\Delta H(\eta(F_T^k))\mathbf{1}_{\{\bar{\tau}_{k,M}=T\}}\Big|\mathcal{G}_{s\wedge s'}\right\}\right|\\ &\leq E\left\{\left|\Delta H(\eta(F_T^k))\right|\mathbf{1}_{\{\bar{\tau}_{k,M}<T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} + E\left\{\left|\Delta H(\eta(F_T^k))\right|\mathbf{1}_{\{|\Delta\eta(F_T^k)|<\delta,\bar{\tau}_{k,M}=T\}}\Big|\mathcal{G}_{s\wedge s'}\right\}\\ &\quad + E\left\{\left|\Delta H(\eta(F_T^k))\right|\mathbf{1}_{\{|\Delta\eta(F_T^k)|\geq\delta,\bar{\tau}_{k,M}=T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} = I_1 + I_2 + I_3,\end{aligned}$$

(5.16)

where I_1, I_2, and I_3 are defined in an obvious way. Clearly, by (5.14) we see that

$$\begin{aligned}I_1 &= E\left\{\left|\Delta H(\eta(F_T^k))\right|\mathbf{1}_{\{\bar{\tau}_{k,M}<T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} \leq 2\|H\|_\infty P\left\{\bar{\tau}_{k,M} < T\Big|\mathcal{G}_{s\wedge s'}\right\}\\ &\leq C_{\ell,2}\frac{1+2R^2}{M^2}2\|H\|_\infty. \qquad (5.17)\end{aligned}$$

Also, in light of (5.13) we see that $|\Delta H(\eta(F_T^k))| < \frac{\varepsilon}{4}$ on the set $\{|\Delta\eta(F_T^k)| < \delta\}$, which implies that

$$I_2 = E\left\{\left|\Delta H(\eta(F_T^k))\right|\mathbf{1}_{\{|\Delta\eta(F_T^k)|<\delta,\bar{\tau}_{k,M}=T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} < \frac{\varepsilon}{4}. \qquad (5.18)$$

Now, first applying the Chebyshev inequality and then applying Lemma 5.2 with $p=1$, we have, P-a.s. on $\{\tau_\ell = T\}$,

$$\begin{aligned} I_3 &= E\left\{\left|\Delta H(\eta(F_T^k))\right| \mathbf{1}_{\{|\Delta\eta(F_T^k)|\geq\delta,\bar\tau_{k,M}=T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} \\ &\leq 2\|H\|_\infty P\left\{|\Delta\eta(F_T^k)|\geq\delta,\bar\tau_{k,M}=T\Big|\mathcal{G}_{s\wedge s'}\right\} \qquad (5.19)\\ &\leq \frac{2}{\delta}\|H\|_\infty E\left\{\ell\left(1+|F_T^{s',x',\beta^k}|+|F_T^{s,x,\beta^k}|\right)|\Delta F_T^k|\mathbf{1}_{\{\bar\tau_{k,M}=T\}}\Big|\mathcal{G}_{s\wedge s'}\right\} \\ &\leq \frac{2(1+2M)\ell}{\delta}\|H\|_\infty E\left\{|F_{\bar\tau^k}^{s,x,\beta^k}-F_{\bar\tau^k}^{s,x',\beta^k}|+|F_{\bar\tau^k}^{s',x',\beta^k}-F_{\bar\tau^k}^{s,x',\beta^k}|\Big|\mathcal{G}_{s\wedge s'}\right\} \\ &\leq \frac{2(1+2M)C_{\ell,1}\ell e^M(1+R)}{\delta}\|H\|_\infty(|s-s'|^{\frac12}+|x-x'|). \end{aligned}$$

Plugging (5.16)–(5.19) into (5.15) we see that on the set $A_{\varepsilon,\ell}^k$ one has

$$\begin{aligned} \widetilde V(s',x')-\widetilde V(s,x) &< \frac{\varepsilon}{2}+C_{\ell,2}\frac{1+2R^2}{M^2}2\|H\|_\infty \\ &\quad+\frac{2(1+2M)C_{\ell,1}\ell e^M(1+R)}{\delta}\|H\|_\infty(|s-s'|^{\frac12}+|x-x'|). \end{aligned} \qquad (5.20)$$

We note that since $\{\tau_\ell = T\} = \cup_{k\geq1}A_{\varepsilon,\ell}^k$, (5.20) actually holds on the set $\{\tau_\ell = T\}$.

To conclude, for fixed $\varepsilon>0$, $\ell>0$, and $R>0$, we first choose $\delta=\delta(\varepsilon)>0$ such that (5.13) holds, and then choose $M = M_{\ell,R,\varepsilon} > 2\sqrt{C_{\ell,2}(1+2R^2)\|H\|_\infty/\varepsilon}$. Denoting

$$C_{\ell,R,\varepsilon} \triangleq \frac{2(1+2M)C_{\ell,1}\ell e^M(1+R)}{\delta(\varepsilon)},$$

it is then easily seen that (5.20) becomes

$$\widetilde V(s',x')-\widetilde V(s,x) < \varepsilon + C_{\ell,R,\varepsilon}\|H\|_\infty(|s-s'|^{\frac12}+|x-x'|), \quad P\text{-a.s. on } \{\tau_\ell=T\}.$$

Reversing the role of (s,x) and (s',x'), we have proved the theorem. □

A closer look at the proof of Theorem 5.3 would lead to the following dependence result for the wider-sense state process $X_t^{s,x,\beta}=\eta_t^0(F_t^{s,x,\beta})$, $t\in[s,T]$, $\beta\in\widetilde{\mathcal{A}}$.

COROLLARY 5.4. *Assume* (H1)–(H3). *Then for any $\ell\geq1$, $R>0$, and $\varepsilon>0$ there exists some constant $C_{\ell,R,\varepsilon}>0$, depending only on ℓ, R, and ε, such that for all $\beta\in\widetilde{\mathcal{A}}$ and all $(s,x),(s',x')\in[0,T]\times\mathbb{R}^n$ with $x,x'\in\mathbb{B}_R$, it holds P-a.s. on $\{\tau_\ell=T\}$ that*

$$\left|E\left\{X_T^{s',x',\beta}-X_T^{s,x,\beta}\Big|\mathcal{G}_{s\wedge s'}\right\}\right| \leq \varepsilon + C_{\ell,R,\varepsilon}((|s-s'|^{\frac12}+|x-x'|). \qquad (5.21)$$

Proof. Noting that $X_T^{s,x,\beta}=\eta_T^0(F_T^{s,x,\beta})$, the proof follows from similar arguments of Theorem 5.3. We leave it to the interested reader. □

To conclude this section we note that Theorem 5.3 does not provide us the usual regularity of the value function (that is, Lipschitz in x, and Hölder-1/2 in s). In fact, as a random field, even the following property of $\widetilde V(\cdot,\cdot)$ is not completely trivial.

THEOREM 5.5. *Assume* (H1)–(H3). *The random field $\widetilde V(s,x)$ possesses a continuous version on $[0,T]\times\mathbb{R}^n$.*

Proof. We would like to construct a P-null exceptional set beyond which the function $(s,x)\mapsto\widetilde V(s,x,\omega)$ is continuous for all fixed ω.

To this end, let us denote $\mathbf{Q}$ to be all the rationals in $\mathbb{R}$. Denote also $\mathbf{Q}_+^* = \mathbf{Q}\cap(0,\infty)$ and $\mathbf{Q}_T = \mathbf{Q}\cap[0,T]$. The notation for the n-dimensional version $\mathbf{Q}^n \subset \mathbb{R}^n$ is defined in an obvious way.

For $\ell \geq 1$ we define the following subset of Ω:

$$\Omega_\ell \stackrel{\triangle}{=} \bigcap_{\substack{R\in\mathbf{Q}_+^*\\ \varepsilon\in\mathbf{Q}_+^*}} \{\tau_\ell = T \text{ and (5.11) holds for all } (s,x),(s',x') \in \mathbf{Q}_T \times \mathbb{B}_R^n \cap \mathbf{Q}^n\},$$

$$\widetilde{\Omega} = \cup_{\ell=1}^\infty \Omega_\ell.$$

Applying Theorem 5.3 we see that for fixed $\ell > 0$, $R > 0$, and $\varepsilon > 0$, (5.11) holds P-a.s. on the set $\{\tau_\ell = T\}$. In other words, one must have $\Omega_\ell \subseteq \{\tau_\ell = T\}$ and $P\{\{\tau_\ell = T\} \setminus \Omega_\ell\} = 0$ for all ℓ. Since $\cup_{\ell=1}^\infty \{\tau_\ell = T\} = \Omega$, modulo a P-null set, one has

$$\Omega \setminus \widetilde{\Omega} \subseteq \bigcup_{\ell=1}^\infty \{\{\tau_\ell = T\} \setminus \Omega_\ell\}, \qquad \text{almost surely.}$$

It then follows that $P\{\widetilde{\Omega}\} = 1$.

Now let us fix $\omega \in \widetilde{\Omega}$. By (5.11) we see that the mapping $(s,x) \mapsto \widetilde{V}(s,x,\omega)$ is continuous on $\mathbf{Q}_T \times \mathbf{Q}^n$. For general $(s,x) \in [0,T]\times\mathbb{R}^n$, we choose a sequence $\{(s_k,x_k)\}_{k\geq 1} \subset \mathbf{Q}_T \times \mathbf{Q}^n$, such that $(s_k,x_k) \to (s,x)$, as $k \to \infty$. Applying (5.11) again we see that $\{\widetilde{V}(s_k,x_k)\}_{k\geq 1}$ is Cauchy, and we can define the limit by $\bar{V}(s,x,\omega)$ and show that it is independent of the choice of the sequence $\{(s_k,x_k)\}$. Now define a random field

$$\bar{V}(s,x,\omega) = \begin{cases} \lim\limits_{k\to\infty} \widetilde{V}(s_k,x_k,\omega), & \mathbf{Q}_T \times \mathbf{Q}^n \ni (s_k,x_k) \to (s,x),\ \omega \in \widetilde{\Omega}, \\ 0, & \omega \in \Omega \setminus \widetilde{\Omega}. \end{cases}$$

Then, using a standard "3ε-argument," one shows that $\bar{V}$ is continuous on $[0,T]\times\mathbb{R}^n$ for all $\omega \in \widetilde{\Omega}$. It remains to verify that $\bar{V}$ is a version of $\widetilde{V}$. But by continuity of $\bar{V}$ we need only check that $\bar{V}$ is a modification of $\widetilde{V}$. To wit, for any $(s,x) \in [0,T]\times\mathbb{R}^n$, it holds that $\bar{V}(s,x) = \widetilde{V}(s,x)$, P-a.s. But by virtue of Theorem 5.3 and the definition of $\bar{V}$ one can check that for fixed (s,x) and any sequence $\{(s_k,x_k)\}_{k\geq 1} \subset \mathbf{Q}_T \times \mathbf{Q}^n$ such that $(s_k,x_k) \to (s,x)$, it must hold that

$$\widetilde{V}(s,x) = \lim_{k\to\infty} \widetilde{V}(s_k,x_k) = \bar{V}(s,x), \qquad P\text{-a.s.}$$

We leave the details to the readers. The proof is now complete. □

6. The Bellman principle. We are now ready to establish the first fundamental result of this paper— the "Bellman principle" of dynamic programming for the wider-sense control problem. We should note that our reduction of the original problem to problems WSCP-I and WSCP-II enables us to find different versions of minimizing sequences, which will be essential in our discussions in this section.

Our main result of this section is the following.

THEOREM 6.1 (Bellman principle). *Assume* (H1)–(H3). *For any* $(s,x) \in [0,T)\times\mathbb{R}^n$ *and* $h > 0$ *such that* $0 \leq s \leq s+h \leq T$, *it holds that*

$$\widetilde{V}(s,x) = \operatorname*{essinf}_{\beta\in\widetilde{A}} E\left\{\widetilde{V}(s+h, F_{s+h}^{s,x,\beta})\Big|\mathcal{G}_s\right\}, \qquad P\text{-a.s.}$$

The proof of Theorem 6.1 is rather lengthy, so we will split it into two lemmas, each taking care of one direction of the inequality, and each using a special technique.

LEMMA 6.2. *Assume* (H1)–(H3). *For all* $(s,x) \in [0,T) \times \mathbb{R}^n$, $h > 0$ *such that* $s+h \le T$, *it holds that*

$$\widetilde{V}(s,x) \le \operatorname*{essinf}_{\beta \in \widetilde{\mathcal{A}}} E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta}) \Big| \mathcal{G}_s \right\}. \tag{6.1}$$

Proof. Clearly, we need only show that for all $(s,x) \in [0,T) \times \mathbb{R}^n$, $h > 0$ such that $s+h \le T$, and $\beta \in \widetilde{A}$, it holds that

$$\widetilde{V}(s,x) \le E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta}) \Big| \mathcal{G}_s \right\}, \qquad P\text{-a.s.} \tag{6.2}$$

The proof depends heavily on the estimates established in Theorem 5.3. First let us fix $\ell > 0$ and $\varepsilon > 0$, and let $0 \le s < s+h \le T$ and $x \in \mathbb{R}^n$ be given and fixed as well. Applying Lemma 5.2 and the Chebyshev inequality, we have, for any $R > 0$, P-a.s. on $\{\tau_\ell = T\}$,

$$P\left\{ |F_{s+h}^{s,x,\beta}| > R \Big| \mathcal{G}_t \right\} \le \frac{1}{R^2} E\left\{ \sup_{t \in [s,T]} |F_t^{s,x,\beta}|^2 \Big| \mathcal{G}_t \right\} \le \frac{C_{\ell,2}}{R^2}(1+|x|^2).$$

Let us now fix $R = R_{\ell,x,\varepsilon} > \sqrt{C_{\ell,2}(1+|x|^2)\|H\|_\infty/\varepsilon}$, so that

$$P\left\{ |F_{s+h}^{s,x,\beta}| > R \Big| \mathcal{G}_s \right\} < \left(\frac{\varepsilon}{\|H\|_\infty} \right) \wedge \varepsilon, \quad P\text{-a.s. on } \{\tau_\ell = T\}. \tag{6.3}$$

Next, following the main ideas of the proof of Theorem 5.3, we can find a constant $C_{\ell,R,\varepsilon}$ such that for all $y, y' \in \mathbb{B}_R^n$, it holds almost surely on $\{\tau_\ell = T\}$ that

$$\begin{aligned} &\left| E\left\{ H\Big(\eta_T^0(F_T^{s+h,y,\beta})\Big) \Big| \mathcal{G}_{s+h} \right\} - E\left\{ H\Big(\eta_T^0(F_T^{s+h,y',\beta})\Big) \Big| \mathcal{G}_{s+h} \right\} \right| \\ &\qquad \le \frac{\varepsilon}{2} + C_{\ell,R,\varepsilon}\|H\|_\infty |y-y'| \end{aligned} \tag{6.4}$$

and

$$\left| \widetilde{V}(s+h,y) - \widetilde{V}(s+h,y') \right| \le \frac{\varepsilon}{2} + C_{\ell,R,\varepsilon}\|H\|_\infty |y-y'|. \tag{6.5}$$

Now let us fix $\delta = \delta_{\ell,x,\varepsilon} < \frac{\varepsilon}{2C_{\ell,R,\varepsilon}\|H\|_\infty}$, and choose a finite set of open balls $\{\mathcal{O}_\delta(y^k)\}_{k=1}^N$, centered at y^k's and with radius δ such that it covers the (compact) ball $\mathbb{B}_R^n$. That is, $\mathbb{B}_R^n \subset \bigcup_{k=1}^N \mathcal{O}_\delta(y^k)$. Clearly, on the set $\{\tau_\ell = T\}$, whenever $y \in \mathcal{O}_\delta(y^k)$ we have

$$\left| E\left\{ H\Big(\eta_T^0(F_T^{s+h,y,\beta})\Big) \Big| \mathcal{G}_{s+h} \right\} - E\left\{ H\Big(\eta_T^0(F_T^{s+h,y^k,\beta})\Big) \Big| \mathcal{G}_{s+h} \right\} \right| < \varepsilon \tag{6.6}$$

and

$$\left| \widetilde{V}(s+h,y) - \widetilde{V}((s+h,y^k)) \right| < \varepsilon. \tag{6.7}$$

Furthermore, we define a partition of $\mathbb{B}_R^n$ as follows:

$$\Gamma_{R,1} \stackrel{\triangle}{=} \mathcal{O}_\delta(y^1) \cap \mathbb{B}_R^n; \qquad \Gamma_{R,k} \stackrel{\triangle}{=} \left(\mathcal{O}_\delta(y^k) \setminus \bigcup_{l=1}^{k-1} \mathcal{O}_\delta(y^l) \right) \bigcap \mathbb{B}_R^n, \qquad k > 1.$$

Thus it follows from (6.7) that, for any $\beta \in \widetilde{\mathcal{A}}$, P-a.s. on $\{|F^{s,x,\beta}_{s+h}| \le R, \tau_\ell = T\}$,

$$\begin{aligned}\widetilde{V}(s+h, F^{s,x,\beta}_{s+h}) &= \sum_{k=1}^{N} \widetilde{V}(s+h, F^{s,x,\beta}_{s+h}) \mathbf{1}_{\Gamma_{R,k}}(F^{s,x,\beta}_{s+h}) \\ &\ge \sum_{k=1}^{N} \widetilde{V}(s+h, y^k) \mathbf{1}_{\Gamma_{R,k}}(F^{s,x,\beta}_{s+h}) - \varepsilon.\end{aligned} \tag{6.8}$$

Now for each k we apply Theorem 5.1 (or, in particular, (5.2)) to find a $\beta^k \in \widetilde{\mathcal{A}}$ such that the set

$$A_{\varepsilon,k} \stackrel{\triangle}{=} \left\{\omega : \widetilde{V}(s+h, y^k, \omega) < E\left\{H\Big(\eta^0_T(F^{s+h,y^k,\beta^k}_T)\Big)\Big|\mathcal{G}_{s+h}\right\}(\omega) - \varepsilon\right\}$$

satisfies $P(A_{\varepsilon,k}) \le 2^{-k}\varepsilon$. Next, define

$$\widetilde{A}_{\ell,\varepsilon,k} \stackrel{\triangle}{=} \{|F^{s,x,\beta}_{s+h}| \le R, \tau_\ell = T\} \cap A^c_{\varepsilon,k} \quad \text{and} \quad \widetilde{\Omega}_{\ell,\varepsilon} \stackrel{\triangle}{=} \bigcap_{k=1}^{N} \widetilde{A}_{\ell,\varepsilon,k}. \tag{6.9}$$

Then

$$\begin{aligned}P\{\widetilde{\Omega}^c_{\ell,\varepsilon}\} &\le P\left\{\{\tau_\ell < T\} \cup \{|F^{s,x,\beta}_{s+h}| > R, \tau_\ell = T\} \cup \left(\bigcup_{k=1}^{N} A_{\varepsilon,k}\right)\right\} \\ &\le P\{\tau_\ell < T\} + P\{|F^{s,x,\beta}_{s+h}| > R, \tau_\ell = T\} + \sum_{k=1}^{N} P(A_{\varepsilon,k}) \\ &\le P\{\tau_\ell < T\} + 2\varepsilon \to 0 \qquad \text{as } \ell \uparrow \infty \text{ and } \varepsilon \downarrow 0.\end{aligned} \tag{6.10}$$

Let us now restrict ourselves to the set $\widetilde{\Omega}_{\ell,\varepsilon}$. Clearly, almost surely on $\widetilde{\Omega}_{\ell,\varepsilon}$ we have

$$\widetilde{V}(s+h, y^k) \ge E\left\{H\Big(\eta^0_T(F^{s+h,y^k,\beta^k}_T)\Big)\Big|\mathcal{G}_{s+h}\right\} - \varepsilon.$$

Thus on the set $\widetilde{\Omega}_{\ell,\varepsilon} \in \mathcal{G}_{s+h}$ the estimate (6.8) can further be written as

$$\widetilde{V}(s+h, F^{s,x,\beta}_{s+h}) \ge \sum_{k=1}^{N} E\left\{H\Big(\eta^0_T(F^{s+h,y^k,\beta^k}_T)\Big)\Big|\mathcal{G}_{s+h}\right\} \mathbf{1}_{\{F^{s,x,\beta}_{s+h} \in \Gamma_{R,k}\}} - 2\varepsilon. \tag{6.11}$$

Now let $\beta \in \widetilde{\mathcal{A}}$ be any control. We modify β as follows: for $(t,\omega) \in [s,T] \times \Omega$,

$$\widetilde{\beta}_t(\omega) = \begin{cases} \beta^k_t(\omega) & \text{if } t \ge s+h, \omega \in \{\tau_\ell = T\} \cap \{F^{s,x,\beta}_{s+h} \in \Gamma_{R,k}\},\ 1 \le k \le N, \\ \beta_t(\omega) & \text{otherwise.} \end{cases}$$

Then clearly $\widetilde{\beta} \in \widetilde{\mathcal{A}}$ as well, and, moreover, the pathwise uniqueness of SDE (3.17), together with the estimate (6.6), shows that for each k, P-a.s. on $\{\tau_\ell = T\}$,

$$\begin{aligned}&E\left\{H\Big(\eta^0_T(F^{s,x,\widetilde{\beta}}_T)\Big)\Big|\mathcal{G}_{s+h}\right\} \mathbf{1}_{\{F^{s,x,\beta}_{s+h} \in \Gamma_{R,k}\}} \\ &= E\left\{H\Big(\eta^0_T(F^{s+h,F^{s,x,\beta}_{s+h},\beta^k}_T)\Big)\mathbf{1}_{\{F^{s,x,\beta}_{s+h} \in \Gamma_{R,k}\}}\Big|\mathcal{G}_{s+h}\right\} \\ &\le E\left\{H\Big(\eta^0_T(F^{s+h,y^k,\beta^k}_T)\Big)\mathbf{1}_{\{F^{s,x,\beta}_{s+h} \in \Gamma_{R,k}\}}\Big|\mathcal{G}_{s+h}\right\} + \varepsilon \mathbf{1}_{\{F^{s,x,\beta}_{s+h} \in \Gamma_{R,k}\}}.\end{aligned} \tag{6.12}$$

Thus, noting that $\cup_{k=1}^N \Gamma_{R,k} = \mathbb{B}_R^n$, we can continue from (6.11) to get

$$\begin{aligned}
&\widetilde{V}(s+h, F_{s+h}^{s,x,\beta}) \\
&\geq \sum_{k=1}^N E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) \mathbf{1}_{\{F_{s+h}^{s,x,\beta} \in \Gamma_{R,k}\}} \Big| \mathcal{G}_{s+h} \right\} - 2\varepsilon - \varepsilon \sum_{k=1}^N \mathbf{1}_{\{F_{s+h}^{s,x,\beta} \in \Gamma_{R,k}\}} \\
&= E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) \Big| \mathcal{G}_{s+h} \right\} \sum_{k=1}^N \mathbf{1}_{\{F_{s+h}^{s,x,\beta} \in \Gamma_{R,k}\}} - 2\varepsilon - \varepsilon \sum_{k=1}^N \mathbf{1}_{\{F_{s+h}^{s,x,\beta} \in \Gamma_{R,k}\}} \\
&= \left[E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) \Big| \mathcal{G}_{s+h} \right\} - \varepsilon \right] - 2\varepsilon = E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) \Big| \mathcal{G}_{s+h} \right\} - 3\varepsilon.
\end{aligned}$$

Note that the above inequality holds only on the $\widetilde{\Omega}_{\ell,\varepsilon}$. Taking conditional expectation $E\{\cdot|\mathcal{G}_s\}$ over $\widetilde{\Omega}_{\ell,\varepsilon} \in \mathcal{G}_{s+h}$ on both sides, we have

$$\begin{aligned}
&E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta}) \Big| \mathcal{G}_s \right\} \geq E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta}) \mathbf{1}_{\widetilde{\Omega}_{\ell,\varepsilon}} \Big| \mathcal{G}_s \right\} \\
&\geq E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) \mathbf{1}_{\widetilde{\Omega}_{\ell,\varepsilon}} \Big| \mathcal{G}_s \right\} - 3\varepsilon = E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) [1 - \mathbf{1}_{\widetilde{\Omega}_{\ell,\varepsilon}^c}] \Big| \mathcal{G}_s \right\} - 3\varepsilon \\
&\geq E\left\{ H\Big(\eta_T^0(F_T^{s,x,\widetilde{\beta}})\Big) \Big| \mathcal{G}_s \right\} - \|H\|_\infty P\{\widetilde{\Omega}_{\ell,\varepsilon}^c | \mathcal{G}_s\} - 3\varepsilon \\
&\geq \widetilde{V}(s,x) - \|H\|_\infty P\{\widetilde{\Omega}_{\ell,\varepsilon}^c | \mathcal{G}_s\} - 3\varepsilon.
\end{aligned}$$

Letting $\ell \to \infty$ and then $\varepsilon \downarrow 0$, we obtain (6.2), whence (6.1). □

The next lemma will show the reversed inequality.

LEMMA 6.3. *Assume* (H1)–(H3). *Then, for all* $(s,x) \in [0,T) \times \mathbb{R}^n$, *and* $s+h \leq T$,

$$\widetilde{V}(s,x) \geq \operatorname*{essinf}_{\beta \in \widetilde{\mathcal{A}}} E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta}) \Big| \mathcal{G}_s \right\}. \tag{6.13}$$

Proof. This time we shall prove that for all $(s,x) \in [0,T) \times \mathbb{R}^n$, and $s+h \leq T$, there exists a sequence $\{\beta^k\} \subset \widetilde{A}$, such that

$$\widetilde{V}(s,x) = \lim_{k \to \infty} \downarrow E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta^k}) \Big| \mathcal{G}_s \right\}, \qquad P\text{-a.s.} \tag{6.14}$$

We note that this "minimizing sequence" is different from all the previous ones, and that (6.13) follows from (6.14) trivially.

To prove (6.14), we fix $(s,x) \in [0,T) \times \mathbb{R}^n$ and $h > 0$ such that $s+h \leq T$. First applying Theorem 5.1 we know that there exists a sequence $\{\beta^k\} \subset \widetilde{\mathcal{A}}$ such that

$$\widetilde{V}(s,x) = \lim_{k \to \infty} \downarrow E\left\{ H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big) \Big| \mathcal{G}_s \right\}.$$

Next, applying Lemma 6.2 we have

$$\begin{aligned}
\widetilde{V}(s,x) &\leq E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta^k}) \Big| \mathcal{G}_s \right\} \\
&\leq E\left\{ E\Big\{ H\Big(\eta_T^0\Big(F_T^{s+h, F_{s+h}^{s,x,\beta^k}, \beta^k}\Big)\Big) \Big| \mathcal{G}_{s+h} \Big\} \Big| \mathcal{G}_s \right\} \\
&= E\left\{ H\Big(\eta_T^0(F_T^{s,x,\beta^k})\Big) \Big| \mathcal{G}_s \right\}, \qquad P - \text{a.s., } \forall k \geq 1.
\end{aligned} \tag{6.15}$$

We will use the by now standard technique to modify the sequence $\{\beta^k\}$ to derive the desired minimizing sequence. Let $\widetilde{\beta}^1 = \beta^1$. For $k > 1$, we define, for $t \in [s, T]$,

$$\widetilde{\beta}_t^k = \begin{cases} \beta_t^k & \text{on } \left\{ E\Big\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta^k}) \Big| \mathcal{G}_s \Big\} \le E\Big\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\widetilde{\beta}^{k-1}}) \Big| \mathcal{G}_s \Big\} \right\}, \\ \widetilde{\beta}_t^{k-1} & \text{otherwise,} \end{cases}$$

and set $\widetilde{\beta}_t^k \equiv \widetilde{\beta}_t^1$ for all k for $t \in [0, s)$. Then, clearly $\{\widetilde{\beta}^k\} \subset \widetilde{\mathcal{A}}$, and for all $k \ge 1$, one has

$$\begin{aligned} & E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\widetilde{\beta}^k}) \Big| \mathcal{G}_s \right\} \\ & \le E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\widetilde{\beta}^{k-1}}) \Big| \mathcal{G}_s \right\} \wedge E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\beta^k}) \Big| \mathcal{G}_s \right\} \\ & \le E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\widetilde{\beta}^{k-1}}) \Big| \mathcal{G}_s \right\} \wedge E\left\{ H\Big(\eta_T^0(F_T^{s,x,\beta^k}) \Big) \Big| \mathcal{G}_s \right\}, \qquad P\text{-a.s.} \end{aligned}$$

Consequently we have

$$\begin{aligned} \widetilde{V}(s,x) & \le \lim_{k\to\infty} \downarrow E\left\{ \widetilde{V}(s+h, F_{s+h}^{s,x,\widetilde{\beta}^k}) \Big| \mathcal{G}_s \right\} \\ & \le \lim_{k\to\infty} \downarrow E\left\{ H\Big(\eta_T^0(F_T^{s,x,\beta^k}) \Big) \Big| \mathcal{G}_s \right\} = \widetilde{V}(s,x), \end{aligned}$$

thanks to Theorem 5.1 again. This proves the lemma. □

Remark 6.4. By modifying the proof slightly one can easily show that the Bellman principle (Theorem 6.1) also holds if the initial states (s, x) and the increment h are replaced by $\mathcal{F}_T^B$-measurable random variables, since $\mathcal{F}_T^B \subset \mathcal{G}_t$ for all $t \ge 0$. In particular, if (s, x) is replaced by $\mathcal{F}_T^B$-measurable random variables (τ, ξ) and $h > 0$ remains a deterministic constant, then we make the convention that $\tau + h = (\tau + h) \wedge T$.

7. Stochastic HJB equation. In this section we shall derive two versions of the stochastic HJB equations: one for the value function of WSCP-I and the other for the original pathwise control problem.

7.1. HJB equation for WSCP-I. We first consider the HJB equation for the value function $\widetilde{V}$. Let us denote

$$\widetilde{\mathcal{L}}_{s,x,u} \triangleq \frac{1}{2} \mathrm{tr}\left\{ \widetilde{\sigma}\widetilde{\sigma}^T(s,x,u) D_{xx}^2 \right\} + \widetilde{b}(t,x,u) D_x.$$

Note that in the above the coefficients $\widetilde{b}$ and $\widetilde{\sigma}$ are $\mathcal{F}_T^B$-measurable random fields, and thus $\widetilde{\mathcal{L}}$ is a "random" differential operator. We then consider the following (random) PDE:

$$\begin{cases} -\dfrac{\partial v}{\partial s}(s,x) - \operatorname*{essinf}_{u\in U} [\widetilde{\mathcal{L}}_{s,x,u} v](s,x) = 0, & (s,x) \in [0,T] \times \mathbb{R}^n; \\ \\ v(T,x) = H\left(\eta_T^0(x) \right), & x \in \mathbb{R}^n. \end{cases} \tag{7.1}$$

We should note that since this randomized PDE does not involve any stochastic integrals, it could be studied in an ω-wise manner. However, the "adaptedness" nature

of the random field is by no means obvious from such a treatment. As an alternative, in what follows we shall consider the random PDE as a special (degenerate) stochastic PDE with a time reversal, and introduce a definition of "stochastic viscosity solution" of this equation that is in the spirit of Buckdahn and Ma [2, 3, 4].

DEFINITION 7.1. *A continuous, $\mathcal{B}([0,T]\times\mathbb{R}^n)\otimes\mathcal{F}_T^B$-measurable random field v is a stochastic viscosity subsolution (resp., supersolution) if*

(i) $v(T,x)\le$ *(resp.,* $\ge$*)* $H\left(\eta_T^0(x)\right)$, $x\in\mathbb{R}^n$;

(ii) *for all* $\tau\in L^0_{\mathcal{F}_T^B}(\Omega;[0,T])$, $\xi\in L^2_{\mathcal{F}_T^B}(\Omega;\mathbb{R}^n)$, *and* $\varphi\in C^{1,2}_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n)$, *it holds that*

$$-\frac{\partial\varphi}{\partial s}(\tau,\xi)-\operatorname*{essinf}_{u\in U}[\widetilde{\mathcal{L}}_{\tau,\xi,u}\varphi](\tau,\xi)\le 0\quad\textit{(resp., }\ge 0\textit{)}, \tag{7.2}$$

on the (ω-)set $\{v-\varphi$ *achieves a local maximum (resp., minimum) at* $(\tau,\xi)\}\in\mathcal{F}_T^B$.

A random field v is called a stochastic viscosity solution if it is both a stochastic subsolution and a stochastic supersolution.

We remark here that the main difference between Definition 7.1 and those of [2, 3, 4], besides the time reversal, is that in [2, 3, 4] we require τ to be a stopping time and ξ to be an $\mathcal{F}_\tau^B$-measurable random variable. Due to the special structure here we assume only that τ is an $\mathcal{F}_T^B$-measurable random time. But one should appreciate again that any $\mathcal{F}_T^B$-measurable random time τ is a $\mathbf{G}$-stopping time!

Our first main result of this section is the following theorem.

THEOREM 7.2. *Assume* (H1)—(H3). *The value function $\widetilde{V}$ is a stochastic viscosity solution of* (7.1) *in the sense of Definition* 7.1. *Furthermore, the solution is unique in the class $C_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n)$.*

Proof. We first show that $\widetilde{V}$ is a subsolution. Let $\tau\in L^0_{\mathcal{F}_T^B}(\Omega;[0,T])$, $\xi\in L^2_{\mathcal{F}_T^B}(\Omega;\mathbb{R}^n)$, and $\varphi\in C^{1,2}_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n)$ be given. We assume that the set

$$\Gamma=\Gamma^{\tau,\xi}\triangleq\{\widetilde{V}-\varphi\text{ achieves a local maximum at }(\tau,\xi)\}$$

satisfies $P(\Gamma)>0$ (for otherwise there is nothing to prove).

We begin by the following "localization" procedure. First, let $\{\tau_\ell\}$ be the sequence of $\mathbf{G}$-stopping times defined by (3.15). We can then choose $\ell>0$ and $\varepsilon>0$ such that the set

$$\Gamma_{\ell,\varepsilon}\triangleq\{\widetilde{V}-\varphi\text{ achieves in }\mathbb{B}_\varepsilon^{n+1}(\tau,\xi)\text{ its maximum at }(\tau,\xi),\quad\tau_\ell=T\} \tag{7.3}$$

also satisfies $P(\Gamma_{\ell,\varepsilon})>0$, where $\mathbb{B}_\varepsilon^{n+1}(s,x)$ denotes the (closed) ball in $\mathbb{R}^{n+1}$ centered at (s,x) with radius ε.

Next, noting that the random fields $\widetilde{V}$ and φ are both $\mathcal{B}([0,T]\times\mathbb{R}^n)\otimes\mathcal{F}_T^B$-measurable, and $\widetilde{V}$ is uniformly bounded (by $\|H\|_\infty$!), we can modify the value of φ outside of $\mathbb{B}_{\varepsilon/2}^{n+1}(\tau,\xi)$ so that φ is uniformly bounded,

$$P\Big\{\sup_{(s,x)\in[0,T]\times\mathbb{R}^n}\Big[|D_t\varphi(s,x)|+|D_x\varphi(s,x)|+|D^2_{xx}\varphi(s,x)|\Big]<\infty\Big\}=1,$$

and $\Gamma_{\ell,\varepsilon}\subset\{(\tau,\xi)\in\operatorname{argmax}\{\widetilde{V}-\varphi\}\}\cap\{\tau_\ell=T\}$. Consequently, we can find $\rho>0$ small enough, and a set $\Gamma^\rho_{\ell,\varepsilon}\in\mathcal{F}_T^B$, such that

(i) $P(\Gamma^\rho_{\ell,\varepsilon})>P(\Gamma_{\ell,\varepsilon})-\rho>0$;

(ii) $\Gamma^{\rho}_{\ell,\varepsilon} \subseteq \{(\tau,\xi) \in \operatorname{argmax}\{\widetilde{V} - \varphi\}\} \cap \{\tau_\ell = T\}$;

(iii) $D_t\varphi(\tau,\xi)$, $D_x\varphi(\tau,\xi)$, $D^2_{xx}\varphi(\tau,\xi)$ are all bounded on $\Gamma^{\rho}_{\ell,\varepsilon}$.

Now let us recall from Remark 6.4 that the Bellman principle (Theorem 6.1) can be extended to the case where the initial state (τ,ξ) is a pair of $\mathcal{F}^B_T$-measurable random variables (with the convention that $\tau+h = (\tau+h)\wedge T$). Thus, for any $h>0$ and $\beta_s \equiv u \in U$ we can apply such a version of Theorem 6.1 and Itô's formula to get the following: on the set $\Gamma^{\rho}_{\ell,\varepsilon}$,

$$
\begin{aligned}
0 &\le \frac{1}{h} E\left\{ \widetilde{V}(\tau+h, F^{\tau,\xi,\beta}_{\tau+h}) - \widetilde{V}(\tau,\xi) \Big| \mathcal{G}_\tau \right\} \le \frac{1}{h} E\left\{ \varphi(\tau+h, F^{\tau,\xi,\beta}_{\tau+h}) - \varphi(\tau,\xi) \Big| \mathcal{G}_\tau \right\} \\
&= \frac{1}{h} E\left\{ \int_\tau^{\tau+h} \left[\partial_t + \widetilde{\mathcal{L}}_{s, F^{\tau,\xi,\beta}_s, u} \right] \varphi(s, F^{\tau,\xi,\beta}_s) ds \Big| \mathcal{G}_\tau \right\} \qquad (7.4) \\
&\le \left[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u} \right] \varphi(\tau,\xi) + r_{\tau,\xi}(h),
\end{aligned}
$$

where

$$
r_{\tau,\xi}(h) \stackrel{\triangle}{=} E\left\{ \sup_{s\in[\tau,\tau+h]} \left| \left[\partial_t + \widetilde{\mathcal{L}}_{s, F^{\tau,\xi,\beta}_s, u} \right] \varphi(s, F^{\tau,\xi,\beta}_s) - \left[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u} \right] \varphi(\tau,\xi) \right| \Big| \mathcal{G}_\tau \right\}.
$$

Note that $\Gamma^{\rho}_{\ell,\varepsilon} \subset \{\tau_\ell = T\}$, and applying Lemma 3.1 and 3.2(ii) we have, on the set $\Gamma^{\rho}_{\ell,\varepsilon}$,

$$
\begin{aligned}
&E\left\{ \sup_{s\in[\tau,T]} |(\partial_t + \widetilde{\mathcal{L}}_{s, F^{\tau,\xi,\beta}_s, u})\varphi(s, F^{\tau,\xi,\beta}_s) \right\} \\
&\le C\Big(1 + E\Big\{ \sup_{s\in[\tau,T]} (|\widetilde{\sigma}(s, F^{\tau,\xi,\beta}_s, u)|^2 + |\widetilde{b}(s, F^{\tau,\xi,\beta}_s, u)|) \Big\} \Big) \\
&\le C\ell^2 \Big(1 + E\Big\{ \sup_{s\in[\tau,T]} |F^{\tau,\xi,\beta}_s|^2 \Big\} \Big) < \infty.
\end{aligned}
$$

Thus, applying the dominated convergence theorem we see that $\lim_{h\downarrow 0} r_{\tau,\xi}(h) = 0$, in probability, on the set $\Gamma^{\rho}_{\ell,\varepsilon}$, and (7.4) leads to

$$
\operatorname*{essinf}_{u\in U} \left[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u} \right] \varphi(\tau,\xi) \ge 0, \qquad P\text{-a.s. on the set } \Gamma^{\rho}_{\ell,\varepsilon}.
$$

Therefore, first letting $\rho \to 0$, then $\varepsilon \to 0$, and then $\ell \to \infty$, we obtain that

$$
-\operatorname*{essinf}_{u\in U} \left[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u} \right] \varphi(\tau,\xi) \le 0, \quad P\text{-a.s. on } \Gamma.
$$

In other words, $\widetilde{V}$ is a stochastic subsolution.

The proof that $\widetilde{V}$ is also a supersolution is a little more involved. We again fix $\tau \in L^0_{\mathcal{F}^B_T}(\Omega;[0,T])$, $\xi \in L^2_{\mathcal{F}^B_T}(\Omega;\mathbb{R}^n)$ and $\varphi \in C^{1,2}_{\mathcal{F}^B_T}([0,T]\times\mathbb{R}^n)$. Also, for any $\varepsilon>0$, $\ell>0$, $\rho>0$, we again find $\Gamma^{\rho}_{\ell,\varepsilon}$ that satisfies (i)–(iii) before, with "max" being replaced by "min."

To derive the desired inequality, we argue slightly differently. First we modify Lemma 6.3 to obtain a sequence $\{\beta^k\} \subset \widetilde{\mathcal{A}}$ such that (6.14) holds for the $\mathcal{F}^B_T$-measurable pair (τ,ξ) and $0<h\ll 1$, that is,

$$
\widetilde{V}(\tau,\xi) = \lim_{k\to\infty} \downarrow E\left\{ \widetilde{V}(\tau+h, F^{\tau,\xi,\beta^k}_{\tau+h}) \Big| \mathcal{G}_\tau \right\}, \quad P\text{-a.s.}
$$

Here we note that the convention $\tau + h = (\tau + h) \wedge T$ is used again and that the sequence $\{\beta^k\}$ may depend on h. Let us denote, for fixed $h > 0$ and each $k \geq 1$,

$$A_h^k \triangleq \left\{\omega : E\left\{\widetilde{V}(\tau + h, F_{\tau+h}^{\tau,\xi,\beta^k})\Big|\, \mathcal{G}_\tau\right\}(\omega) - \widetilde{V}(\tau, \xi, \omega) \leq h^2\right\}.$$

Then by definition of $\{\beta^k\}$ we have $\lim_{k\to\infty} P\{\Omega \setminus A_h^k\} = 0$. Since $P(\Gamma_{\ell,e}^\rho) > 0$, we must also have $\lim_{k\to\infty} P\{A_h^k \cap \Gamma_{\ell,\varepsilon}^\rho\} = P(\Gamma_{\ell,e}^\rho) > 0$.

Now, using the definition of φ and applying Itô's formula, one has, P-a.s. on the set $A_h^k \cap \Gamma_{\ell,\varepsilon}^\rho$,

$$\begin{aligned} h &\geq \frac{1}{h} E\left\{\widetilde{V}(\tau + h, F_{\tau+h}^{\tau,\xi,\beta^k}) - \widetilde{V}(\tau, \xi)\Big|\, \mathcal{G}_\tau\right\} \\ &\geq \frac{1}{h} E\left\{\varphi(\tau + h, F_{\tau+h}^{\tau,\xi,\beta^k}) - \varphi(\tau, \xi)\Big|\, \mathcal{G}_\tau\right\} \\ &= \frac{1}{h} E\left\{\int_\tau^{\tau+h} \Big(\partial_t + \widetilde{\mathcal{L}}_{s,F_s^{\tau,\xi,\beta^k},\beta^k}\Big)\varphi(s, F_s^{\tau,\xi,\beta^k})ds\,\Big|\, \mathcal{G}_\tau\right\} \qquad (7.5) \\ &\geq \operatorname*{essinf}_{u\in U}\Big[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u}\Big]\varphi(\tau, \xi) - \widetilde{r}_{\tau,\xi}(h), \end{aligned}$$

where

$$\widetilde{r}_{\tau,\xi}(h) \triangleq \sup_{k\geq 1} E\left\{\sup_{s\in[\tau,\tau+h]}\left|\Big[\partial_t + \widetilde{\mathcal{L}}_{s,F_s^{\tau,\xi,\beta^k},\beta_s^k}\Big]\varphi(s, F_s^{\tau,\xi,\beta^k}) - \Big[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,\beta_s^k}\Big]\varphi(\tau, \xi)\right|\Big|\, \mathcal{G}_\tau\right\}.$$

Let us now denote $\Psi = (\partial_t\varphi, D_x\varphi, D_{xx}^2\varphi)$ and for fixed (τ, ξ),

$$G_{\tau,\xi}^k(s) \triangleq |\widetilde{\sigma}\widetilde{\sigma}^T(s, \xi, \beta_s^k) - \widetilde{\sigma}\widetilde{\sigma}^T(\tau, \xi, \beta_s^k)| + |\widetilde{b}(s, \xi, \beta_s^k) - \widetilde{b}(\tau, \xi, \beta_s^k)|, \qquad s \in [0, T].$$

Clearly, Ψ is bounded and continuous on $\Gamma_{\ell,\varepsilon}^\rho$. Furthermore, a tedious but straightforward argument using Lemmas 3.1 and 5.2 ((5.7) in particular) and the Chebyshev inequality, one can show that for any $h > 0, k \geq 1$, and $\delta > 0$, there exists a constant $C_{\ell,\varepsilon,\rho} > 0$ such that, P-a.s., on $\Gamma_{\ell,\varepsilon}^\rho$, it holds that

$$\begin{aligned} &E\left\{\sup_{s\in[\tau,\tau+h]}\Big[\partial_t + \widetilde{\mathcal{L}}_{s,F_s^{\tau,\xi,\beta^k},\beta_s^k}\Big]\varphi(s, F_s^{\tau,\xi,\beta^k}) - \Big[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,\beta_s^k}\Big]\varphi(\tau, \xi)\Big|\mathcal{G}_\tau\right\} \\ &\leq C_{\ell,\varepsilon,\rho}\left(1 + \frac{1}{\delta^2}\right)\big(1 + |\xi|^4\big)\left\{\sup_{\substack{s\in[\tau,\tau+h],\\ |x|\leq\delta}} |\Psi(s, \xi + x) - \Psi(\tau, \xi)| + \sup_{\tau\leq s\leq\tau+h} E\{G_{\tau,\xi}^k(s)|\mathcal{G}_\tau\}\right\}. \end{aligned}$$

Consequently, for all $h > 0, \delta > 0$, it holds, P-a.s. on $\Gamma_{\ell,\varepsilon}^\rho$ that

$$\begin{aligned} \widetilde{r}_{\tau,\xi}(h) &\leq C_{\ell,\varepsilon\rho}\left(1 + \frac{1}{\delta^2}\right)(1 + |\xi|^4) \\ &\quad \times \left\{\sup_{\substack{s\in[\tau,\tau+h],\\ |x|\leq\delta}} |\Psi(s, \xi + x) - \Psi(\tau, \xi)|\right\} + \sup_{\substack{s\in[\tau,\tau+h]\\ k\geq 1}} E\{G_{\tau,\xi}^k(s)|\mathcal{G}_\tau\}\right\}. \end{aligned}$$

Now from the definition of $\widetilde{\sigma}$ and $\widetilde{b}$ ((3.6) and (3.7)) and the assumption (H1) we see that $G_{\tau,\xi}^k(s)$ is continuous in s, uniformly in k. Thus applying the dominated convergence theorem we have

$$\lim_{h\downarrow 0} \sup_{s\in[\tau,\tau+h],k\geq 1} E\{G_{\tau,\xi}^k(s)|\mathcal{G}_\tau\} = 0, \qquad P\text{-a.s.}$$

It then follows that $\widetilde{r}_{\tau,\xi}(h) \to 0$, P-a.s. on $\Gamma^{\rho}_{\ell,\varepsilon}$ as $h \downarrow 0$. In other words, we obtain

$$\operatorname*{essinf}_{u\in U}\left[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u}\right]\varphi(\tau,\xi) \le 0, \qquad P\text{-a.s. on } \Gamma^{\rho}_{\ell,\varepsilon}.$$

Again, first sending $\rho \to 0$, then $\varepsilon \to 0$, and then $\ell \to \infty$, we obtain that

$$-\operatorname*{essinf}_{u\in U}\left[\partial_t + \widetilde{\mathcal{L}}_{\tau,\xi,u}\right]\varphi(\tau,\xi) \ge 0, \quad P\text{-a.s.}$$

That is, $\widetilde{V}$ is a stochastic viscosity supersolution of (7.1).

To conclude, note that the uniqueness of the viscosity solution can be proved by following the almost identical idea of Buckdahn and Ma [3], with even easier arguments since in the present case there is no "martingale" term in the degenerated stochastic PDF (7.1). The proof is complete. □

7.2. HJB equation for the original control problem. We now turn to our last objective of this paper: to derive the stochastic HJB equation for the original value function V. The idea is straightforward, that is, we shall apply the "inverse" Doss–Sussmann transformation and see what will happen to the stochastic HJB equation (7.1). We first observe the following simple fact: if $\varphi \in C^{1,2}_{\mathcal{F}^B_T}([0,T]\times R^n)$, then by the Itô–Ventzell formula (cf., e.g., Kunita [11]), one has (recall the process $\{\zeta^s_t\}$ from (3.5))

$$\begin{aligned} d[\varphi(t,\zeta^0_t(x))] &= \frac{\partial\varphi}{\partial t}(t,\zeta^0_t(x))dt + (D_x\varphi)(t,\zeta^0_t(x))\circ d\zeta^0_t(x) \\ &= \frac{\partial\varphi}{\partial t}(t,\zeta^0_t(x))dt - (D_x\varphi)\,(t,\zeta^0_t(x))D_z\zeta^0_t(x)\theta(x)\circ dB_t \\ &= \frac{\partial\varphi}{\partial t}(t,\zeta^0_t(x))dt - D_x[\varphi(t,\zeta^0_t(x))]\theta(x)\circ dB_t. \end{aligned} \tag{7.6}$$

Now recall from Theorem 4.2(i) that $\widetilde{V}(s,x) = V(s,\eta^0_s(x))$, and hence $V(s,x) = \widetilde{V}(s,\zeta^0_s(x))$. Thus if $\widetilde{V} \in C^{1,2}_{\mathcal{F}^B_T}$, then in light of (7.6) we should have

$$dV(s,x) = d[\widetilde{V}(s,\zeta^0_s(x))] = \frac{\partial\widetilde{V}}{\partial t}(t,\zeta^0_s(x))ds - D_x[\widetilde{V}(t,\zeta^0_s(x))]\theta(x)\circ dB_s. \tag{7.7}$$

Combining (7.1) for $\widetilde{V}$ and relation (7.7) we can then formally write down the stochastic HJB equation for the original value function V:

$$\begin{cases} dV(s,x) + \operatorname*{essinf}_{u\in U}[\mathcal{L}_{x,u}V](s,x)ds + D_xV(s,x)\theta(x)\circ dB_s = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (s,x)\in[0,T)\times\mathbb{R}^n, \\ V(T,x) = H(x), \qquad x\in\mathbb{R}^n. \end{cases} \tag{7.8}$$

Again, we remark that this stochastic PDE is a "terminal value" problem; therefore it would become a complicated issue if we were to seek an $\{\mathcal{F}^B_t\}$-adapted solution. However, the following modification of the definition of stochastic viscosity solution for the stochastic PDE (7.8) given in [2, 3, 4] proves sufficient for our purpose.

DEFINITION 7.3. *A random field $v \in C_{\mathcal{F}^B_T}([0,T]\times\mathbb{R}^n)$ is said to be a (stochastic) viscosity subsolution (resp., supersolution) of equation* (7.8) *if*

(i) *$v(T,x) \le$ (resp., $\ge$) $H(x)$ for all $x\in\mathbb{R}^n$;*

(ii) *for any* $(\tau,\xi)\in L^2(\mathcal{F}_T^B;[0,T]\times\mathbb{R}^n)$, *and* $\varphi\in C^{1,2}_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n)$, *it holds that*

$$-\frac{\partial\varphi}{\partial t}\left(\tau,\zeta_s^0(\xi)\right)\leq \text{(resp., }\geq)\operatorname*{essinf}_{u\in U}\mathcal{L}_{\tau,\xi,u}\left[\varphi(\tau,\zeta_\tau^0(\cdot))\right](\xi),$$

P-a.s. on the set

$$\Gamma^\varphi_{\tau,\xi}\stackrel{\triangle}{=}\{\omega:(\tau(\omega),\xi(\omega))\in \operatorname{argmax}_{loc}[v(s,x,\omega)-\varphi(s,\zeta_0^s(x),\omega)]\ \text{(resp., } \operatorname{argmin}_{loc})\}$$

We have the following theorem.

THEOREM 7.4. *Assume* (H1)–(H3). *Then the value function* V *is the unique stochastic viscosity solution of* (7.8).

Proof. Having proved Theorem 7.2, we need only prove the following equivalence relation: a random field $v\in C_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n)$ is a stochastic viscosity subsolution (resp., supersolution) of the stochastic HJB equation (7.8) if and only if

$$\widetilde{v}(t.x)=v(t,\eta_t^0(x)),\qquad (t,x)\in[0,T]\times\mathbb{R}^n,$$

is a stochastic viscosity subsolution (supersolution) of the (random) HJB equation (7.1).

To this end, we observe that for any given $(\tau,\xi)\in L^2(\mathcal{F}_T^B;[0,T]\times\mathbb{R}^n)$ and $\varphi\in C^{1,2}_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n\}$,

$$\begin{aligned}\Gamma^\varphi_{\tau,\xi}&=\{\omega:(\tau(\omega),\xi(\omega))\in\operatorname{argmax}_{loc}[v(s,x,\omega)-\varphi(s,\zeta_s^0(x),\omega)]\}\\&=\{\omega:(\tau(\omega),\zeta_\tau^0(\xi)(\omega))\in\operatorname{argmax}_{loc}[v(s,\eta_s^0(x),\omega)-\varphi(s,x,\omega)]\}\\&=\{\omega:(\tau(\omega),\zeta_\tau^0(\xi)(\omega))\in\operatorname{argmax}_{loc}[\widetilde{v}(s,x,\omega)-\varphi(s,x,\omega)]\}.\end{aligned}$$

On the other hand, it is easy to check that

$$\mathcal{L}_{x,u}[\varphi(t,\zeta_t^0(\cdot))](x)=(\widetilde{\mathcal{L}}_{t,\zeta_t^0(x),u}\varphi)(t,\zeta_t^0(x))\qquad\forall(t,x,u)\in[0,T]\times\mathbb{R}^n\times U;$$

we see that $\widetilde{v}$ is a viscosity subsolution of equation (7.1) if and only if for all $(\tau,\xi)\in L^2(\mathcal{F}_T^B;[0,T]\times\mathbb{R}^n)$ and $\varphi\in C^{1,2}_{\mathcal{F}_T^B}([0,T]\times\mathbb{R}^n\}$, it holds, P-a.s. on the set $\Gamma^\varphi_{\tau,\xi}$, that

$$-\left[\left(\frac{\partial}{\partial t}+\operatorname*{essinf}_{u\in U}\widetilde{\mathcal{L}}_{\tau,\zeta_\tau^0(\xi),u}\right)\varphi\right](\tau,\zeta_\tau^0(\xi))\leq 0.$$

But this amounts to saying that for any such given (τ,ξ) and φ, it holds, P-a.s. on $\Gamma^\varphi_{\tau,\xi}$, that

$$-\frac{\partial\varphi}{\partial t}\left(\tau,\zeta_0^s\left(\xi\right)\right)\leq\operatorname*{essinf}_{u\in U}\mathcal{L}_{\tau,\xi,u}\left[\varphi\left(\tau,\zeta_0^\tau(\cdot)\right)\right](\xi).$$

But by Definition 7.3 it shows exactly that v is a stochastic viscosity subsolution of (7.8). The proof for the supersolutions is analogous. Thus the theorem follows from Theorem 7.2. □

8. Appendix (a counterexample). We give an example showing that the original pathwise stochastic control problem does not possess any minimizing sequence in $\mathcal{A}$. Let us assume that $n=1$, $U=[0,1]$, and $(s,x)=(0,0)$. Assume also that

$b(t,x,u) \equiv u$, $\sigma \equiv 0$, and $\theta(x) \equiv x$. Finally, let $H(x) = |x-1|$. That is, the system dynamics is

$$\begin{cases} dX_t^\alpha = \alpha_t dt + X_t^\alpha \circ dB_t, & t \in [0,T], \\ X_0^\alpha = 0. \end{cases}$$

Furthermore, the Doss–Sussmann transformation is $\eta_t^0(z) = z + \int_0^t \eta_r^0(z) \circ dB_r$, which can be solved explicitly as $\eta_t^0(z) = ze^{B_t}$, $t \in [0,T]$. Therefore, the "transformed" coefficients are $\widetilde{\sigma} \equiv 0$ and $\widetilde{b}(t,z,u) = \left[D_z \eta_t^0(z)\right]^{-1} b(\eta_t^0(z),u) = e^{-B_t}u$; and the "transformed" system equation becomes

$$dF_t^\beta = \widetilde{b}(t, F_t^\beta, \beta_t)dt = e^{-B_t}\beta_t dt, \qquad \beta \in \widetilde{A},$$

which can also be solved explicitly as $F_t^\beta = \int_0^t e^{-B_r}\beta_r dr$. Consequently we have $\eta_T^0(F_T^\beta) = \int_0^T e^{B_T - B_r}\beta_r dr$.

We can now easily check that, at $(s,x) = (0,0)$,

$$V(0,0) = \widetilde{V}(0,0) = \operatorname*{essinf}_{\beta \in \widetilde{A}} E\left\{ |\eta_T^0(F_T^\beta) - 1| \,\middle|\, \mathcal{F}_T^B \right\} = \left[1 - \int_0^T e^{B_T - B_r} dr\right]^+,$$

with a wider sense optimal control being

$$\beta_t^* = \begin{cases} \left(\int_0^T e^{B_T - B_r} dr\right)^{-1} & \text{on } \left\{\int_0^T e^{B_T - B_r} dr > 1\right\}, \\ \\ 1 & \text{on } \left\{\int_0^T e^{B_T - B_r} dr \le 1\right\}. \end{cases}$$

We should note that such an optimal control is unique on the set $\{\int_0^T e^{B_T - B_r} dr \le 1\}$.

On the other hand, we observe that if $\alpha \in \mathcal{A}$, then so is $E[\alpha]$, and by Jensen's inequality we have

$$\begin{aligned} J(0,0;\alpha) &= E\left\{ \left| \int_0^T e^{B_T - B_r}\alpha_r dr - 1 \right| \middle| \mathcal{F}_T^B \right\} \ge \left| \int_0^T e^{B_T - B_r} E[\alpha_r] dr - 1 \right| \\ &= J(0,0,E[\alpha]) \ge V(0,0). \end{aligned}$$

Now let us assume that there exists a minimizing sequence $\{\alpha^k\}_{k \ge 1} \subset \mathcal{A}$. Then it holds, P-a.s., that $\lim_{k\to\infty} J(0,0;\alpha^k) = V(0,0)$. But the above argument shows that

$$J(0,0,E[\alpha^k]) = \left| \int_0^T e^{B_T - B_r} E[\alpha_r^k] dr - 1 \right| \longrightarrow V(0,0) = \left[1 - \int_0^T e^{B_T - B_r} dr\right]^+ \tag{8.1}$$

as well. Since $0 \le E[\alpha_t^k] \le 1$ for all t, we see that on the set $\{\int_0^t e^{B_T - B_t} dr \le 1\}$ it must hold that

$$\int_0^T e^{B_T - B_r} E[\alpha_r^k] dr \longrightarrow \int_0^T e^{B_T - B_r} dr \qquad \text{as } k \to \infty,$$

or, equivalently, $\int_0^T e^{B_T - B_r}[1 - E[\alpha_r^k]] dr \to 0$, on $\{\int_0^t e^{B_T - B_t} dr \le 1\}$. But again notice that $0 \le E[\alpha_t^k] \le 1$, which implies that $E[\alpha_\cdot^k] \longrightarrow 1$, as $k \to \infty$ in (Lebesgue)

measure, and consequently one must have

$$\left|\int_0^T e^{B_T-B_r}E[\alpha_r^k]dr-1\right| \longrightarrow \left|\int_0^T e^{B_T-B_r}dr-1\right|, \quad P\text{-a.s.}$$

Since

$$\left|\int_0^T e^{B_T-B_r}dr-1\right| > V(0,0) \qquad \text{on} \quad \left\{\int_0^T e^{B_T-B_r}dr>1\right\},$$

and $P\{\int_0^T e^{B_T-B_r}dr>1\}>0$, we obtain that

$$\lim_{k\to\infty} J(0,0,E[\alpha^k]) = \left|\int_0^T e^{B_T-B_r}dr-1\right| > V(0,0) \qquad \text{on } \left\{\int_0^T e^{B_T-B_r}dr>1\right\}.$$

This contradicts (8.1). Therefore such a minimizing sequence cannot exist.

REFERENCES

[1] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.

[2] R. Buckdahn and J. Ma, *Stochastic viscosity solutions for nonlinear stochastic partial differential equations. Part* I, Stochastic Process. Appl., 93 (2001), pp. 181–204.

[3] R. Buckdahn and J. Ma, *Stochastic viscosity solutions for nonlinear stochastic partial differential equations. Part* II, Stochastic Process. Appl., 93 (2001), pp. 205–228.

[4] R. Buckdahn and J. Ma, *Pathwise stochastic Taylor expansion for solutions of stochastic differential equations*, Ann. Probab., 30 (2002), pp. 1131–1171.

[5] C. Dellacherie, *Sur l'existence de certains essinf et esssup de familles de processus mesurables; Séminaire Probab.* XII*, Univ. Strasbourg* 1976/77, Lecture Notes in Math. 649, Springer-Verlag, Berlin, 1978, pp. 512–514.

[6] C. Dellacherie and P. Meyer, *Probabilities and Potential*, North-Holland, Amsterdam, New York, 1978.

[7] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I: *General Theory*, Wiley-Interscience, New York, 1957.

[8] N. El Karoui, D. H. Nguyen, and M. Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: Existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 1998.

[10] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Appl. Math. (N.Y.) 39, Springer-Verlag, New York, 1998.

[11] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1990.

[12] P.-L. Lions and P. E., Souganidis, *Fully nonlinear stochastic partial differential equations*, C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1085–1092.

[13] P.-L. Lions and P. E. Souganidis, *Fully nonlinear stochastic partial differential equations: Non-smooth equation and applications*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 735–741.

[14] P.-L. Lions and P. E. Souganidis, *Fully nonlinear stochastic pde with semilinear stochastic dependence*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 617–624.

[15] P.-L. Lions and P. E. Souganidis, *Uniqueness of weak solutions of fully nonlinear stochastic partial differential equations*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 783–790.

[16] P. Protter, *Stochastic Integration and Stochastic Differential Equations—A New Approach*, Springer-Verlag, New York, Berlin, Heidelberg, 1990.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2257–2276

ADAPTIVE DIGITAL CONTROL OF HAMMERSTEIN NONLINEAR SYSTEMS WITH LIMITED OUTPUT SAMPLING*

FENG DING†, TONGWEN CHEN‡, AND ZENTA IWAI§

Abstract. This paper is motivated by the practical control considerations that nonlinearity is abundant in industrial processes and output sampling rates are often limited due to hardware constraints. In particular, for a Hammerstein nonlinear sampled-data system in which the output sampling period is an integer multiple of the input updating period, we derive, by using a polynomial transformation technique, a mathematical model which is suitable for parameter estimation with dual-rate measurement data. Further, we present an adaptive control scheme for such a dual-rate nonlinear system; the parameter estimation–based adaptive algorithm can achieve virtually asymptotically optimal control and ensure that the closed-loop system is stable and globally convergent. The simulation results are included.

Key words. Hammerstein systems, sampled-data systems, multirate modeling, self-tuning regulator, adaptive control, dual-rate systems, convergence properties, least squares methods

AMS subject classifications. 93C40, 93E10, 93C10

DOI. 10.1137/05062620X

1. Introduction.

This paper focuses on adaptive control problems of a class of nonlinear systems, namely, Hammerstein nonlinear systems, with limited output sampling frequencies; in particular, the Hammerstein systems of interest have output sampling periods which are integer multiples of input updating periods. These systems are sometimes referred to as *dual-rate sampled-data systems*. Such slow output sampled systems often arise in industry due to hard limits on sensoring devices; see the industrial processes described in, e.g., [26, 31, 34]. Without assuming knowledge of the nonlinear model involved, our main idea in this work is to perform closed-loop adaptive control through an online identification scheme, which estimates the parameters of the nonlinear model by making use of only the available dual-rate data.

In the identification area of nonlinear systems, there exists a large amount of research exploring different approaches; see, e.g., [3, 4, 5, 6, 35, 36, 37]. Vörös presented a recursive half-substitution algorithm for discontinuous nonlinear systems [36]. For Hammerstein–Wiener nonlinear models, Bai reported a two-stage identification algorithm based on the recursive least squares and the singular value decomposition [3] and a blind identification approach [5]. Recently, Cerone and Regruto derived the parameter error bounds in the Hammerstein models [8] by assuming that the output measurement error was bounded; Ding and Chen developed an iterative method

*Received by the editors March 7, 2005; accepted for publication (in revised form) August 15, 2006; published electronically February 13, 2007. This research was completed while the second author visited Kumamoto University as a Guest Professor holding a JSPS Research Fellowship. Financial support from the Japan Society for the Promotion of Science (JSPS), the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the National Natural Science Foundation of China (grants 60574051 and 60528007) is gratefully acknowledged.

http://www.siam.org/journals/sicon/45-6/62620.html

†Corresponding author. Control Science and Engineering Research Center, Southern Yangtze University, Wuxi, Jiansu 214122, People's Republic of China (fding@sytu.edu.cn, fding@ece.ualberta.ca).

‡Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada (tchen@ece.ualberta.ca).

§Department of Mechanical Engineering and Materials Science, Faculty of Engineering, Kumamoto University, 2-39-1 Kurokami, Kumamoto 860-8555, Japan (iwai@gpo.kumamoto-u.ac.jp).

and a recursive least squares identification method for Hammerstein nonlinear autoregression and moving average with exogenous input (ARMAX) systems [12]. In this paper, we will extend the noniterative method of Chang and Luus [7] for Hammerstein models to present a parameter estimation–based adaptive algorithm for Hammerstein nonlinear *dual-rate* systems.

In the field of adaptive control of linear dual-rate systems, Albertos, Salt, and Tormero studied adaptive control schemes for dual-rate systems [1]; Ishitobi, Kawanaka, and Nishi presented a least squares–based self-tuning control algorithm [28]; and Kanniah, Malik, and Hope proposed a control algorithm based on a parameterized model with its autoregression (AR) coefficients corresponding to the fast sampling rate and the moving average (MA) coefficients to the slow sampling rate [29]. Though these are earlier contributions in this area, the reported adaptive control schemes have two main limitations as follows (in view of the goal in our work):

- First, the prediction and control were both based on the slow sampling rate; the desired fast-rate system performance could not be achieved in these schemes in design, and there might be poor intersample behavior even if the behavior at the slow sampling instants is acceptable.
- Second, these schemes used *linear* models as the primary assumption; for nonlinear Hammerstein models, which capture a wide class of industrial systems and processes, new research is required.

In order to overcome the two limitations for dual-rate Hammerstein models, we develop an estimation scheme to obtain *fast-rate* models which can be used to implement an adaptive control law updated at the *fast rate*, even if the output is sampled at a relatively slow rate.

To the best of our knowledge, few contributions have addressed such adaptive control problems involving Hammerstein nonlinear dual-rate systems, especially the convergence problem of the algorithms involved, which are the focus of this work. Therefore, the objective of this paper is to extend the control scheme of dual-rate systems in [10] to the nonlinear case, in particular, in the following ways:

- By a polynomial transformation technique, derive models suitable for dual-rate identification of Hammerstein nonlinear sampled-data systems.
- Based on the models derived, propose a parameter estimation–based adaptive control scheme.
- Study the convergence properties of the proposed adaptive control algorithms.

Briefly, the paper is organized as follows. Section 2 introduces the problem formulation related to Hammerstein nonlinear dual-rate systems and an adaptive control scheme. Section 3 uses a polynomial transformation technique to derive a suitable mathematical model and present an adaptive control algorithm based on parameter estimation. Sections 4 and 5 analyze the output tracking performance and global stability of the closed-loop systems under the proposed adaptive control. Section 6 presents an illustrative example demonstrating the effectiveness of the proposed algorithm. Finally, section 7 offers some concluding remarks.

2. Problem formulation. In many chemical processes [26, 31, 34], due to hardware sensoring constraints, only infrequent and scarce output sampling is available; but the input updating can be achieved by actuators and digital computers at relatively fast speeds [10, 11, 30]. This gives rise to a class of *dual-rate* systems in which the input updating rate is an integer multiple of the output sampling rate. This paper focuses on adaptive control of a class of Hammerstein nonlinear dual-rate sampled-data systems—a static nonlinearity $f(\cdot)$ followed by a linear dynamic subsystem $G(z)$,

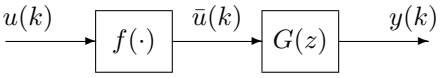


FIG. 1. *A Hammerstein nonlinear system.*

as depicted in Figure 1, where $u(k)$ is the system input, the inner variable $\bar{u}(k)$ is the output of the nonlinear block (unmeasurable), $y(k)$ is the system output but is available only every qth sampling instant, q being a positive integer ($q > 1$). Thus, the available input-output data set consists of

- $\{u(k) : \ k = 0, 1, 2, \dots\}$ at the fast rate, and
- $\{y(kq) : \ k = 0, 1, 2, \dots\}$ at the slow rate.

Here, we refer to the unavailable intersample outputs, $y(kq+j)$, $j = 1, 2, \dots, q-1$, as the *missing* output samples, and to $\{u(k), y(kq)\}$ as the *dual-rate* measurement data.

Often the static nonlinear part in the Hammerstein model is assumed to be a polynomial of a known order m in the input as follows (see, e.g., [7, 22, 33]):

$$\bar{u}(k) = c_1 u(k) + c_2 u^2(k) + \cdots + c_m u^m(k),$$

or, more generally, to be a nonlinear function of a known basis $(\gamma_1, \gamma_2, \dots, \gamma_m)$ as follows [8]:

$$(1) \quad \bar{u}(k) = f(u(k)) = c_1\gamma_1(u(k)) + c_2\gamma_2(u(k)) + \cdots + c_m\gamma_m(u(k)) = \sum_{i=1}^{m} c_i\gamma_i(u(k)),$$

where the c_i's are unknown parameters. Of course, the nonlinearity considered here also covers cubic spline nonlinearity [9], support vector machines [23, 24], and piecewise linear functions with discontinuities [35, 36], in addition to the single-parameter nonlinearities [4]. The linear block is a time-invariant system (at the fast rate) with a known order n and takes the following real-rational form:

$$(2) \qquad y(k) = \frac{B(z)}{A(z)}\,\bar{u}(k), \ G(z) := \frac{B(z)}{A(z)}$$

with

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n},$$
$$B(z) = b_1 z^{-1} + b_2 z^{-2} + \cdots + b_n z^{-n}.$$

Here, z^{-1} represents a unit backward shift operator at the fast rate, $z^{-1}u(k) = u(k-1)$, and $G(z)$ is the transfer function of the linear dynamical part with one unit delay.

Notice that in the characterization of the Hammerstein model shown in Figure 1, $f(u)$ and $G(z)$ are actually not unique from the identification point of view. Any pair $(cf(u), G(z)/c)$ for some nonzero and finite constant c would produce identical input and output measurements. In other words, none of the identification schemes can distinguish between $(f(u), G(z))$ and $(cf(u), G(z)/c)$. Therefore, to get a unique parameterization, without loss of generality, one of the gains of $f(u)$ and $G(z)$ has to be fixed. There are several ways to normalize the gains [3, 8, 22]. Here, we adopt the assumption used in [5, 22]: The first coefficient of the function $f(\cdot)$ equals 1; i.e., $c_1 = 1$. (In case $c_1 = 0$, one can use any nonzero c_i and normalize it to 1.)

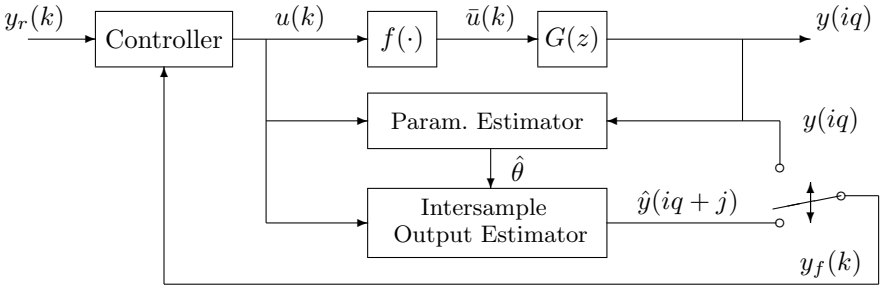


FIG. 2. *The adaptive control scheme* ($j = 1, 2, \ldots, q-1$).

For dual-rate sampled-data control systems, we expect that the control law be updated at the fast rate even if the output is sampled at the slow rate. To implement this, the adaptive control scheme we propose is shown in Figure 2, where $y_r(k)$ denotes a deterministic reference input or desired output signal. The control scheme makes use of a parameter estimator, or an identification algorithm, which generates online the estimates $\hat{\theta}$ of the unknown system parameters based on the dual-rate data $\{u(k), y(kq)\}$; based on $\hat{\theta}$, and thus the fast-rate model, and the input $u(k)$, the intersample (missing) outputs $\hat{y}$ are estimated—see the intersample output estimator in Figure 2. In order to feed back to the controller a fast rate signal $y_f(k)$, representing the output $y(k)$, we use the slow sampled output $y(iq)$ at every q period, giving $y(0)$, $y(q)$, and $y(2q)$, etc., and use the estimated output $\hat{y}(iq+j)$ to fill in the missing samples in $y(k)$. Hence in Figure 2, $y_f(k)$ connects to $y(iq)$ at times $k = iq$, and connects to $\hat{y}(iq+j)$ at $k = iq+j$, $j = 1, 2, \ldots, (q-1)$; thus the output of the switch, $y_f(k)$, is a fast rate signal. The operation of the periodic switch can be expressed in the following equation:

$$y_f(k) = \begin{cases} y(iq), & k = iq, \\ \hat{y}(iq+j), & k = iq+j,\ j = 1, 2, \ldots, (q-1). \end{cases} \tag{3}$$

To summarize, the dual-rate adaptive control scheme uses a fast single-rate controller and a periodic switch. It is conceptually simple, easy to implement in digital computers, and practical for industry.

In connection with Figure 2, the objectives of this paper can be restated more clearly as follows:

- Establish a dual-rate model which is suitable for identification using the given dual-rate data.
- Propose an algorithm to estimate the intersample outputs $\{y(kq+j) :\ j = 1, 2, \ldots, (q-1)\}$.
- Design an adaptive controller so that the output $y(k)$ tracks the given desired output $y_r(k)$ by minimizing the tracking error criterion given by

$$J[u(k)] = [y_f(k+1) - y_r(k+1)]^2 \tag{4}$$

for deterministic systems, or

$$J[u(k)] = \mathrm{E}\{[y_f(k+1) - y_r(k+1)]^2 | \mathcal{F}_{k-1}\} \tag{5}$$

for stochastic systems, and study the properties of the closed-loop system.

Here, $\{\mathcal{F}_k\}$ is the σ algebra sequence generated by the observations up to and including time k.

3. Control algorithm description. Let us introduce some notation first. The symbol I stands for an identity matrix of appropriate dimensions. The superscript T denotes the matrix transpose. For a square matrix X, $|X| = \det[X]$ represents the determinant; $\lambda_i[X]$ denotes the ith eigenvalue of X; and $\lambda_{\max}[X]$ and $\lambda_{\min}[X]$ represent the maximum and minimum eigenvalues of X, respectively. The norm of a matrix M is defined by $\|M\|^2 = \mathrm{tr}[MM^{\mathrm{T}}]$. The notation $\mathbf{1}_n$ denotes an n-dimensional column vector whose elements are all 1; $f(k) = o(g(k))$ represents $f(k)/g(k) \to 0$ as $k \to \infty$; for $g(k) \geq 0$, we write $f(k) = O(g(k))$ or $f(k) \sim g(k)$ if there exists a positive constant δ_1 such that $|f(k)| \leq \delta_1 g(k)$.

The model in (1)–(2) is not suitable for dual-rate adaptive control because it would involve the unavailable outputs $\{y(kq+j) : \ j = 1, 2, \ldots, (q-1)\}$. To obtain a model that we can use directly on the dual-rate data, by a polynomial transformation technique, $G(z)$ can be converted into a form so that the denominator is a polynomial in z^{-q} instead of z^{-1}.

For a general discussion, let the roots of $A(z)$ be z_i to get

$$A(z) = \prod_{i=1}^{n}(1 - z_i z^{-1}).$$

Define

$$\phi_q(z) := \prod_{i=1}^{n}(1 + z_i z^{-1} + z_i^2 z^{-2} + \cdots + z_i^{q-1} z^{-q+1}) = \prod_{i=1}^{n} \frac{1 - z_i^q z^{-q}}{1 - z_i z^{-1}}.$$

Multiplying the numerator and denominator of $G(z)$ by $\phi_q(z)$ and using the polynomial transformation formula [27, 32],

$$1 - x^q = (1 - x)(1 + x + x^2 + \cdots + x^{q-1}),$$

we get a new model:

$$P(z) = \frac{B(z)\phi_q(z)}{A(z)\phi_q(z)} =: \frac{\beta(z)}{\alpha(z)},$$

or

$$\alpha(z)y(k) = \beta(z)\bar{u}(k) \tag{6}$$

with

$$\alpha(z) = 1 + \alpha_1 z^{-q} + \alpha_2 z^{-2q} + \cdots + \alpha_n z^{-qn},$$
$$\beta(z) = \beta_1 z^{-1} + \beta_2 z^{-2} + \cdots + \beta_{qn} z^{-qn}.$$

Equation (6) has the advantage that the denominator is a polynomial of z^{-q}; arising from here is a recursive equation using only slowly sampled outputs. The control algorithm we propose later for dual-rate systems will be based on this model, which does not involve the unavailable intersample outputs.

From the model in (6), we easily get the recursive equation:

$$\begin{aligned} y(k) &= -\sum_{i=1}^{n} \alpha_i y(k - iq) + \sum_{i=1}^{qn} \beta_i \bar{u}(k - i) \\ &= -\sum_{i=1}^{n} \alpha_i y(k - iq) + \sum_{i=1}^{qn} \beta_i \sum_{j=1}^{m} c_j \gamma_j(u(k - i)). \end{aligned}$$

Define the parameter vector θ and information vector $\varphi(k)$ as

$$\theta = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ c_2\boldsymbol{\beta} \\ \vdots \\ c_m\boldsymbol{\beta} \end{bmatrix} \in \mathbb{R}^{n_0}, \quad \varphi(k-1) = \begin{bmatrix} -y(k-q) \\ -y(k-2q) \\ \vdots \\ -y(k-nq) \\ \psi(k-1) \end{bmatrix} \in \mathbb{R}^{n_0}, \ n_0 := (mq+1)n,$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{qn} \end{bmatrix} \in \mathbb{R}^{qn}, \quad \boldsymbol{c} = \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_m \end{bmatrix} \in \mathbb{R}^{m-1},$$

$$\psi(k-1) = \begin{bmatrix} \psi_1(k-1) \\ \psi_2(k-1) \\ \vdots \\ \psi_m(k-1) \end{bmatrix} \in \mathbb{R}^{mqn},$$

$$\psi_j(k-1) = \begin{bmatrix} \gamma_j(u(k-1)) \\ \gamma_j(u(k-2)) \\ \vdots \\ \gamma_j(u(k-nq)) \end{bmatrix} \in \mathbb{R}^{qn}, \ j = 1, 2, \ldots, m.$$

Then we have

$$y(k) = \varphi^{\mathrm{T}}(k-1)\theta. \tag{7}$$

Notice that if k is an integer multiple of q, then $\varphi(k-1)$ uses only available dual-rate data—only the past measurement outputs (slow rate) and inputs (fast rate).

From here, we can see that the number of parameters will be greatly increased after the parameterization of nonlinear systems with dual-rate sampling; this will lead to increased computational complexity, which may be reduced by using the hierarchical identification scheme [13, 14, 15, 16, 17, 18].

Let $y_r(k)$ be a desired output signal; define the output tracking error

$$\xi(k+1) = y(k+1) - y_r(k+1).$$

In the deterministic case, if the control signal $u(k)$ is chosen according to the equation $y_r(k+1) = \varphi^{\mathrm{T}}(k)\theta$ obtained by minimizing the criterion function in (4), then the tracking error $\xi(k+1)$ approaches zero asymptotically.

For stochastic systems, based on the model in (7), introducing a zero-mean white noise disturbance term $v(k)$, we have

$$y(k) = \varphi^{\mathrm{T}}(k-1)\theta + v(k). \tag{8}$$

Let $\hat{\theta}$ be the estimate of unknown parameter vector θ; then $\hat{y}(k+1) = \varphi^{\mathrm{T}}(k)\hat{\theta}$ is the output prediction, which is computed by the intersample output estimator in Figure 2. According to the certainty equivalence principle [25], or minimizing the criterion function in (5), the control law takes the following form:

$$y_r(k+1) = \varphi^{\mathrm{T}}(k)\hat{\theta}. \tag{9}$$

Replacing k in (8) with kq gives

$$y(kq) = \varphi^{\mathrm{T}}(kq-1)\theta + v(kq). \tag{10}$$

Then the recursive least squares algorithm may be used to produce the estimate $\hat{\theta}(kq)$ of θ at current time kq, and the algorithm is as follows:

$$\hat{\theta}(kq) = \hat{\theta}(kq-q) + P(kq)\varphi(kq-1)e(kq), \tag{11}$$

$$e(kq) = [y(kq) - \varphi^{\mathrm{T}}(kq-1)\hat{\theta}(kq-q)], \tag{12}$$

$$\hat{\theta}(i) = \hat{\theta}(kq),\ i = kq, kq+1, \ldots, kq+q-1, \tag{13}$$

$$P^{-1}(kq) = P^{-1}(kq-q) + \varphi(kq-1)\varphi^{\mathrm{T}}(kq-1), \tag{14}$$

$$\hat{\theta}(kq) = [\hat{\theta}_1(kq),\ \hat{\theta}_2(kq), \ldots,\ \hat{\theta}_{n_0}(kq)]^{\mathrm{T}}. \tag{15}$$

Based on (9), the control law is given by

$$\varphi^{\mathrm{T}}(kq+j)\hat{\theta}(kq) = y_r(kq+1+j),\quad j = 0, 1, \ldots, q-1. \tag{16}$$

To initialize the control algorithm in (11)–(16), we take $P(0) = p_0 I$, with p_0 normally a large positive number, e.g., $p_0 = 10^6$, and $\hat{\theta}(0) = \hat{\theta}_0$ some small real vector, e.g., $\hat{\theta}(0) = \mathbf{1}_{n_0}/p_0$. Notice that the parameter estimate $\hat{\theta}$ is updated every q (fast) samples, namely, at the slow rate, as is the covariance matrix P; in between the slow samples, we simply hold $\hat{\theta}$ unchanged. Thus, every time $\hat{\theta}$ is updated, we have q new input samples and one new output sample.

Defining the gain vector

$$L(kq) := P(kq)\varphi(kq-1)$$

and applying the matrix inversion lemma

$$(A+BC)^{-1} = A^{-1} - A^{-1}B(I+CA^{-1}B)^{-1}CA^{-1}$$

to (14), it is easy to get that the covariance matrix P can be updated as follows:

$$P(kq) = [I - L(kq)\varphi^{\mathrm{T}}(kq-1)]P(kq-q),$$
$$L(kq) = \frac{P(kq-q)\varphi(kq-1)}{1+\varphi^{\mathrm{T}}(kq-1)P(kq-q)\varphi(kq-1)}.$$

Note $c_1 = 1$; the estimates

$$\hat{\boldsymbol{\alpha}}(kq) = [\hat{\alpha}_1(kq),\ \hat{\alpha}_2(kq), \ldots,\ \hat{\alpha}_n(kq)]^{\mathrm{T}},$$
$$\hat{\boldsymbol{\beta}}(kq) = [\hat{\beta}_1(kq),\ \hat{\beta}_2(kq), \ldots,\ \hat{\beta}_{qn}(kq)]^{\mathrm{T}}$$

of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be read from the first n and second qn entries of $\hat{\theta}(kq)$, respectively. Let

$$\hat{\boldsymbol{c}}(kq) = [\hat{c}_2(kq),\ \hat{c}_3(kq), \ldots,\ \hat{c}_m(kq)]^{\mathrm{T}}$$

be the estimate of $\boldsymbol{c}$. Referring to the definition of θ, we get that the estimates of c_j, $j = 2, 3, \ldots, m$, may be computed by

$$\hat{c}_j = \frac{\hat{\theta}_{n+(j-1)qn+i}(kq)}{\hat{\beta}_i(kq)},\ j = 2, 3, \ldots, m;\ i = 1, 2, \ldots, qn.$$

From here, we can see that there is considerable redundancy in determination of each coefficient $\hat{c}_j$ in the nonlinear function $f(\cdot)$, since for each c_j we have qn estimates $\hat{c}_j$ for $i = 1, 2, \ldots, n$. Because we do not need such n estimates for $\hat{c}_j$, one simple way is to take their average as the estimate of c_j [12], i.e.,

$$\hat{c}_j(kq) = \frac{1}{qn} \sum_{i=1}^{qn} \frac{\hat{\theta}_{n+(j-1)qn+i}(kq)}{\hat{\beta}_i(kq)}, \ j = 2, 3, \ldots, m.$$

Of course, the singular value decomposition technique or least squares optimization methods can also be used to find the estimate of $\boldsymbol{c}$ [3, 7].

The control signal $u(kq + j)$ in (16) may be obtained by solving the following nonlinear equation:

$$\begin{aligned} (17) \qquad & \sum_{i=1}^{n} \hat{\theta}_i(kq) y(kq + j + 1 - iq) + \sum_{l=1}^{m} \sum_{i=1}^{nq} \hat{\theta}_{n+(l-1)qn+i}(kq) \gamma_l(u(kq + j + 1 - i)) \\ & = y_r(kq + j + 1), \qquad j = 0, 1, \ldots, q - 1. \end{aligned}$$

Here, a difficulty arises because over the interval $[kq, kq + q)$, except for $j = q - 1$, the expression on the right-hand side of the above equation contains the future and past missing outputs $y(kq + j + 1 - iq)$. So it looks impossible to compute the control law by (17) and to realize the algorithm in (11)–(16). Our solution is based on the adaptive control scheme stated in section 2—these unknown outputs $y(kq + j)$ in (17) are replaced by their estimates $\hat{y}(kq + j)$. Hence we have

$$\begin{aligned} (18) \qquad & \sum_{i=1}^{n} \hat{\theta}_i(kq) \hat{y}(kq + j + 1 - iq) + \sum_{l=1}^{m} \sum_{i=1}^{nq} \hat{\theta}_{n+(l-1)qn+i}(kq) \gamma_l(u(kq + j + 1 - i)) \\ & = y_r(kq + j + 1) \end{aligned}$$

or

$$\begin{aligned} & \sum_{l=1}^{m} \hat{\theta}_{n+(l-1)qn+1}(kq) \gamma_l(u(kq + j)) = y_r(kq + j + 1) - \sum_{i=1}^{n} \hat{\theta}_i(kq) \hat{y}(kq + j + 1 - iq) \\ (19) \qquad & - \sum_{l=1}^{m} \sum_{i=2}^{nq} \hat{\theta}_{n+(l-1)qn+i}(kq) \gamma_l(u(kq + j + 1 - i)), \qquad j = 0, 1, \ldots, q - 2. \end{aligned}$$

In fact, it is only when $j = q - 1$ that the control term $u(kq + j)$ does not involve the missing outputs and can be generated by

$$\begin{aligned} & \sum_{l=1}^{m} \hat{\theta}_{n+(l-1)qn+1}(kq) \gamma_l(u(kq + q - 1)) = y_r(kq + q) - \sum_{i=1}^{n} \hat{\theta}_i(kq) \hat{y}(kq + q - iq) \\ (20) \qquad & - \sum_{l=1}^{m} \sum_{i=2}^{nq} \hat{\theta}_{n+(l-1)qn+i}(kq) \gamma_l(u(kq + q - i)). \end{aligned}$$

In our adaptive control algorithm in (11)–(16) (or (11)–(15), (19), and (20)), based on the parameter estimation, the control signal $u(k)$ is computed by the model inverse of the Hammerstein model defined in (1) and (2) using past inputs $u(k - j)$ for $j = 1, 2, \ldots$, the current output $y(k)$ ($\hat{y}(k)$), past outputs $y(k - j)$ ($\hat{y}(k - j)$) for

$j = 1, 2, \ldots$, and desired output $y_r(k+1)$ (for one-step-ahead self-tuning control) as k increases, and the input $u(k)$ is made to drive the system output at time $k+1$ to the target value $y_r(k+1)$. In order to ensure that the input signal is bounded, this model inverse must be stable; thus the inverse of the nonlinearity $f(\cdot)$ must exist (i.e., the nonlinearity $f(\cdot)$ must be invertible) and $G(z)$ is minimum phase, exactly as in the case of linear models. The control signal $u(k)$ is also determined by using the following approach. Assuming $b_1 \neq 0$ or $\beta \neq 0$, we compute the control signal $u(k)$ by

$$u(kq+j) = \hat{f}^{-1}(\bar{u}(kq+j)),\ i = 0, 1, 2, \ldots, q-1, \tag{21}$$

where $\hat{f}^{-1}(\cdot)$ is the inverse of the estimated nonlinear function $\hat{f}(\cdot)$ defined by

$$\begin{aligned}\bar{u}(kq+j) = \hat{f}(u(kq+j)) = \gamma_1(u(kq+j)) + \hat{c}_2(kq)\gamma_2(u(kq+j)) \\ + \cdots + \hat{c}_m(kq)\gamma_m(u(kq+j)),\end{aligned} \tag{22}$$

and the intermediate signal $\bar{u}(k)$ is obtained from the inverse of the linear model, i.e.,

$$\begin{aligned}\bar{u}(kq+j) = \frac{1}{\hat{\beta}_1(kq)}\Bigg[y_r(kq+j+1) + \sum_{i=1}^{n}\hat{\alpha}_i(kq)\hat{y}(kq+j+1-iq) \\ - \sum_{i=2}^{nq}\hat{\beta}_i(kq)u(kq+j-i+1)\Bigg],\quad j = 0, 1, \ldots, q-2,\end{aligned}$$

$$\begin{aligned}\bar{u}(kq+q-1) = \frac{1}{\hat{\beta}_1(kq)}\Bigg[y_r(kq+q) + \sum_{i=1}^{n}\hat{\alpha}_i(kq)y(kq+q-iq) \\ - \sum_{i=2}^{nq}\hat{\beta}_i(kq)u(kq+j-i+1)\Bigg].\end{aligned}$$

The above two control law equations contain only $nq+n$ parameters (less than $mqn+n$ parameters); however, this increased computation is still tolerable and affordable.

Due to the nonlinear function $f(\cdot)$, the control signal $u(k)$ in (22) is usually solved by numerical methods.

The following assumption is required and is also reasonable: Given $\bar{u}(k)$, there exists a $u(k)$ satisfying (1). Note that we do not require that $u(k)$ be *uniquely* determined by given $\bar{u}(k)$, because all such $u(k)$'s produce the same $\bar{u}(k)$, which does not affect our output tracking performance.

4. Output tracking performance. Let us first introduce some definitions and assumptions. The sequence $\{v(k), \mathcal{F}_k\}$ is assumed to be a martingale difference sequence defined on a probability space $\{\Omega, \mathcal{F}, P\}$, where $\{\mathcal{F}_k\}$ is the σ algebra sequence generated by the observations up to and including time k [25]. The noise sequence $\{v(k)\}$ satisfies the following conditions:

(A1) $\mathrm{E}[v(k)|\mathcal{F}_{k-1}] = 0$ a.s.;

(A2) $\mathrm{E}[v^2(k)|\mathcal{F}_{k-1}] = \sigma^2(k) \leq \bar{\sigma}^2 < \infty$ a.s.;

(A3) $\limsup_{k\to\infty}\frac{1}{k}\sum_{i=1}^{k}v^2(i) \leq \bar{\sigma}^2 < \infty$ a.s.

That is, $\{v(k)\}$ is an independent random noise sequence with zero mean and bounded time-varying variance.

Define

$$r(kq) := \mathrm{tr}[P^{-1}(kq)],\ r(0) := \frac{n_0}{p_0}.$$

It follows easily that

$$\begin{aligned} r(kq) &= \lambda_1[P^{-1}(kq)] + \lambda_2[P^{-1}(kq)] + \cdots + \lambda_{n_0}[P^{-1}(kq)] \\ &\le n_0\lambda_{\max}[P^{-1}(kq)], \\ |P^{-1}(kq)| &= \lambda_1[P^{-1}(kq)]\lambda_2[P^{-1}(kq)]\cdots\lambda_{n_0}[P^{-1}(kq)] \\ &\le \lambda_{\max}^{n_0}[P^{-1}(kq)] \le r^{n_0}(kq), \\ \ln|P^{-1}(kq)| &= O(\ln r(kq)). \end{aligned} \tag{23}$$

In order to study the output tracking performance of the adaptive control algorithm proposed earlier, the following lemma is required.

LEMMA 1. *For the algorithm in* (11)–(16)*, the following inequality holds:*

$$\sum_{i=1}^{\infty} \frac{\varphi^{\mathrm{T}}(iq-1)P(iq)\varphi(iq-1)}{\{\ln r(iq)\}^c} < \infty \ \textit{a.s. for any}\ \ c > 1.$$

Proof. From the definition of $P(kq)$, we have

$$\begin{aligned} P^{-1}(iq-q) &= P^{-1}(iq) - \varphi(iq-1)\varphi^{\mathrm{T}}(iq-1) \\ &= P^{-1}(iq)[I - P(iq)\varphi(iq-1)\varphi^{\mathrm{T}}(iq-1)]. \end{aligned}$$

Taking determinants on both sides gives

$$\begin{aligned} |P^{-1}(iq-q)| &= |P^{-1}(kq)||I - P(kq)\varphi(kq-1)\varphi^{\mathrm{T}}(kq-1)| \\ &= |P^{-1}(iq)|[1 - \varphi^{\mathrm{T}}(iq-1)P(iq)\varphi(iq-1)]. \end{aligned}$$

Thus,

$$\varphi^{\mathrm{T}}(iq-1)P(iq)\varphi(iq-1) = \frac{|P^{-1}(iq)| - |P^{-1}(iq-q)|}{|P^{-1}(iq)|}. \tag{24}$$

Noting that $|P^{-1}(iq)|$ is a nondecreasing function of i, dividing both sides of (24) by $\{\ln r(kq)\}^c$ gives

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{\varphi^{\mathrm{T}}(iq-1)P(iq)\varphi(iq-1)}{\{\ln r(iq)\}^c} &= \sum_{i=1}^{\infty} \frac{|P^{-1}(iq)| - |P^{-1}(iq-q)|}{|P^{-1}(iq)|\{\ln r(iq)\}^c} \\ &\le n_0^c \sum_{i=1}^{\infty} \frac{|P^{-1}(iq)| - |P^{-1}(iq-q)|}{|P^{-1}(iq)|\{\ln|P^{-1}(iq)|\}^c} \\ &= n_0^c \sum_{i=1}^{\infty} \int_{|P^{-1}(iq-q)|}^{|P^{-1}(iq)|} \frac{\mathrm{d}x}{|P^{-1}(iq)|\{\ln|P^{-1}(iq)|\}^c} \\ &\le n_0^c \sum_{i=1}^{\infty} \int_{|P^{-1}(iq-q)|}^{|P^{-1}(iq)|} \frac{\mathrm{d}x}{x(\ln x)^c} \end{aligned}$$

$$
\begin{aligned}
&= n_0^c \int_{|P^{-1}(0)|}^{|P^{-1}(\infty)|} \frac{\mathrm{d}x}{x(\ln x)^c} \\
&= \frac{-n_0^c}{c-1} \cdot \frac{1}{\{\ln x\}^{c-1}} \Big|_{|P^{-1}(0)|}^{|P^{-1}(\infty)|} \\
&= \frac{n_0^c}{c-1} \left[\frac{1}{\{\ln |P^{-1}(0)|\}^{c-1}} - \frac{1}{\{\ln |P^{-1}(\infty)|\}^{c-1}} \right] \\
&< \infty \text{ a.s., } c > 1. \qquad \square
\end{aligned}
$$

We shall prove the main results of this paper by formulating a martingale process as in [13, 14, 19, 21] and by using stochastic process theory and the martingale convergence theorem (Lemma D.5.3 in [25]).

THEOREM 2. *For the system in* (10)*, assume that* (A1)–(A3) *hold,* $B(z)$ *is stable, and the reference input* $y_r(k)$ *is bounded in the sense of*

$$\text{(A4)} \quad |y_r(k)| < \infty.$$

Then the adaptive control algorithm in (11)–(17) *guarantees that the output tracking error at the output sampling instants has the property of minimum variance, i.e.,*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} [y_r(iq) - y(iq) + v(iq)]^2 = 0 \ \ a.s.;$$

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \mathrm{E}\{[y_f(iq) - y_r(iq)]^2 | \mathcal{F}_{iq-1}\} \le \bar{\sigma}^2 < \infty \ \ a.s.$$

Proof. Define the parameter estimation error vector as

$$\tilde{\theta}(kq) = \hat{\theta}(kq) - \theta.$$

Using (10) and (11), we have

$$
\begin{aligned}
\tilde{\theta}(kq) &= \tilde{\theta}(kq-q) + P(kq)\varphi(kq-1)[\varphi^{\mathrm{T}}(kq-1)\theta + v(kq) - \varphi^{\mathrm{T}}(kq-1)\hat{\theta}(kq-q)] \\
&:= \tilde{\theta}(kq-q) + P(kq)\varphi(kq-1)[-\tilde{y}(kq) + v(kq)], \qquad (25)
\end{aligned}
$$

where

$$\tilde{y}(kq) := \varphi^{\mathrm{T}}(kq-1)\tilde{\theta}(kq-q) = \varphi^{\mathrm{T}}(kq-1)\hat{\theta}(kq-q) - \varphi^{\mathrm{T}}(kq-1)\theta. \qquad (26)$$

By using (10) and (16), it follows that

$$\tilde{y}(kq) = y_r(kq) - y(kq) + v(kq).$$

Define a nonnegative definite function

$$V(kq) = \tilde{\theta}^{\mathrm{T}}(kq)P^{-1}(kq)\tilde{\theta}(kq).$$

Using (10), (25), and (26), we have

$$
\begin{aligned}
V(kq) &= \tilde{\theta}^{\mathrm{T}}(kq-q)P^{-1}(kq)\tilde{\theta}(kq-q) + 2\tilde{\theta}^{\mathrm{T}}(kq-q)\varphi(kq-1)[-\tilde{y}(kq) + v(kq)] \\
&\quad + \varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)[-\tilde{y}(kq) + v(kq)]^2 \\
&= \tilde{\theta}^{\mathrm{T}}(kq-q)[P^{-1}(kq-q) + \varphi^{\mathrm{T}}(kq-1)\varphi(kq-1)]\tilde{\theta}(kq-q) \\
&\quad + 2\tilde{y}(kq)[-\tilde{y}(kq) + v(kq)] + \varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)[-\tilde{y}(kq) + v(kq)]^2 \\
&= V(kq-q) - [1 - \varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)]\tilde{y}^2(kq) \\
&\quad + \varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)v^2(kq) \\
&\quad + 2[1 - \varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)]\tilde{y}(kq)v(kq). \qquad (27)
\end{aligned}
$$

Noting that $\tilde{y}(kq)$, $\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)$ are uncorrelated with $v(kq)$ and are $\mathcal{F}_{kq-1}$-measurable, taking the conditional expectation on both sides of (27) with respect to $\mathcal{F}_{kq-1}$ and using (A1)–(A2) give

$$\begin{aligned}\mathrm{E}[V(kq)|\mathcal{F}_{kq-1}] \le V(kq-q) &- [1-\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)]\tilde{y}^2(kq)\\ &+2\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)\bar{\sigma}^2.\end{aligned}$$

Let

$$W(kq) := \frac{V(kq)}{[\ln r(kq)||]^c},\ c>1.$$

Noting that $\ln r(kq)$ is nondecreasing, we have

$$\begin{aligned}\mathrm{E}[W(kq)|\mathcal{F}_{kq-1}] &\le \frac{V(kq-q)}{[\ln r(kq)]^c} - \frac{1-\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)}{[\ln r(kq)]^c}\ \tilde{y}^2(kq)\\ &\quad+\frac{2\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)}{[\ln r(kq)]^c}\ \bar{\sigma}^2\\ &\le W(kq-q) - \frac{1-\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)}{[\ln r(kq)]^c}\ \tilde{y}^2(kq)\\ &\quad+\frac{2\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)}{[\ln r(kq)]^c}\ \bar{\sigma}^2. \qquad (28)\end{aligned}$$

In terms of Lemma 1, we can see that the sum of the last right-hand term of (28) for k from $k=1$ to $k=\infty$ is finite. Since

$$1-\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1) = [1+\varphi^{\mathrm{T}}(kq-1)P(kq-q)\varphi(kq-1)]^{-1} \ge 0,$$

applying the martingale convergence theorem (Lemma D.5.3 in [25]) to (28), we conclude that $W(kq)$ converges a.s. to a finite random variable, say, W_0; i.e.,

$$W(kq) = \frac{V(kq)}{[\ln r(kq)]^c} \to W_0 < \infty \text{ a.s., or } V(kq) = O([\ln r(kq)]^c) \text{ a.s.},$$

and also

$$\sum_{k=1}^{\infty} \frac{1-\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1)}{[\ln r(kq)]^c}\ \tilde{y}^2(kq) < \infty \text{ a.s.}$$

Due to $\varphi^{\mathrm{T}}(kq-1)P(kq)\varphi(kq-1) = o(1)$, we have

$$\sum_{i=1}^{\infty} \frac{\tilde{y}^2(iq)}{[\ln r(iq)]^c} < \infty \text{ a.s.} \qquad (29)$$

As $r(kq)\to\infty$, using the Kronecker lemma (Lemma D.5.5 in [25]) yields

$$\lim_{k\to\infty} \frac{1}{[\ln r(kq)]^c} \sum_{i=1}^{k}\ \tilde{y}^2(iq) = 0 \text{ a.s.}$$

Since $[\ln r(kq)]^c = o(r(kq))$, we have

$$\lim_{k\to\infty} \frac{k}{r(kq)}\ \frac{1}{k}\sum_{i=1}^{k}\ \tilde{y}^2(iq) = 0 \text{ a.s.} \qquad (30)$$

Since $B(z)$ is stable, applying Lemma B.3.3 in [25] to (10) and using (A3) yields

$$\frac{1}{k}\sum_{i=1}^{k} u^2(iq) \le \frac{c_1}{k}\sum_{i=1}^{k} y^2(iq) + c_2 \text{ a.s.},$$

where c_i represent finite positive constants. According to the definitions of $r(kq)$ and $\varphi(kq)$, it is not difficult to get

$$\begin{aligned}\frac{r(kq)}{k} &\le \frac{c_3}{k}\sum_{i=1}^{k} y^2(iq) + c_4 \\ &= \frac{c_3}{k}\sum_{i=1}^{k} [y_r(iq) - \tilde{y}(iq) + v(iq)]^2 + c_4 \\ &\le \frac{c_5}{k}\sum_{i=1}^{k} \tilde{y}^2(iq) + c_6 \text{ a.s.}\end{aligned}$$

Thus, from (30)

$$0 = \lim_{k\to\infty} \frac{\frac{1}{k}\sum_{i=1}^{k} \tilde{y}^2(iq)}{\frac{r(kq)}{k}} \ge \lim_{k\to\infty} \frac{\frac{1}{k}\sum_{i=1}^{k} \tilde{y}^2(iq)}{\frac{c_5}{k}\sum_{i=1}^{k} \tilde{y}^2(iq) + c_6} \ge 0 \text{ a.s.},$$

and hence

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \tilde{y}^2(iq) = 0 \text{ a.s.},$$

or

$$\lim_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} [y_r(iq) - y(iq) + v(iq)]^2 = 0 \text{ a.s.} \tag{31}$$

Since

$$\begin{aligned}&\mathrm{E}\{[y_r(kq) - y(kq) + v(kq)]^2 | \mathcal{F}_{kq-1}\} \\ &\quad = \mathrm{E}[(y_r(kq) - y(kq))^2 + 2y_r(kq)v(kq) - 2y(kq)v(kq) + v^2(kq) | \mathcal{F}_{kq-1}] \\ &\quad = \mathrm{E}[(y_r(kq) - y(kq))^2 | \mathcal{F}_{kq-1}] + 0 - 2\sigma^2(kq) + \sigma^2(kq) \\ &\quad = \mathrm{E}[(y_r(kq) - y(kq))^2 | \mathcal{F}_{kq-1}] - \sigma^2(kq) \text{ a.s.},\end{aligned}$$

and $y_f(kq) = y(kq)$ at the output sampling instants, we have

$$\begin{aligned}\limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \mathrm{E}\{[y_f(iq) - y_r(iq)]^2 | \mathcal{F}_{iq-1}\} &= \limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \mathrm{E}\{[y(iq) - y_r(iq)]^2 | \mathcal{F}_{iq-1}\} \\ &= \limsup_{k\to\infty} \frac{1}{k}\sum_{i=1}^{k} \sigma^2(iq) \le \bar{\sigma}^2 \text{ a.s.}\end{aligned}$$

This proves Theorem 2. □

Since single-rate systems belong to a special class of dual-rate systems with $q = 1$, the results in Theorem 2 still hold for single-rate Hammerstein nonlinear systems, which, to the best of our knowledge, have not even been reported in the literature before.

5. Global convergence. In this section, we analyze global stability properties of the closed-loop system under the adaptive control proposed earlier.

From (3) and (10), we have

$$y_f(kq) = y(kq) = \varphi^{\mathrm{T}}(kq-1)\theta + v(kq), \tag{32}$$

$$y_f(kq+j) = \hat{y}(kq+j), \quad j = 1, 2, \ldots, q-1. \tag{33}$$

From Figure 2 and (10), since $v(kq)$ is a "white" noise, the best estimates of all missing output $y(kq+j)$ are given by

$$\hat{y}(kq+j+1) = \hat{\varphi}^{\mathrm{T}}(kq+j)\hat{\theta}(kq), \quad j = 0, 1, \ldots, q-2,$$

with

$$\hat{\varphi}(kq+j) = \begin{bmatrix} -\hat{y}(kq+j+1-q) \\ -\hat{y}(kq+j+1-2q) \\ \vdots \\ -\hat{y}(kq+j+1-nq) \\ \psi(kq+j) \end{bmatrix} \in \mathbb{R}^{n_0}, \ \psi(kq+j) = \begin{bmatrix} \psi_1(kq+j) \\ \psi_2(kq+j) \\ \vdots \\ \psi_m(kq+j) \end{bmatrix} \in \mathbb{R}^{mqn},$$

$$\psi_i(kq+j) = \begin{bmatrix} \gamma_i(u(kq+j)) \\ \gamma_i(u(kq+j-1)) \\ \vdots \\ \gamma_i(u(kq+j+1-nq)) \end{bmatrix} \in \mathbb{R}^{qn}, \ i = 1, 2, \ldots, m.$$

The missing output estimates $\hat{y}(kq+j)$ can also be computed from the recursive equation:

$$\begin{aligned} \hat{y}(kq+j+1) = &-\sum_{i=1}^{n} \hat{\theta}_i(kq)\hat{y}(kq+j+1-iq) \\ &+ \sum_{l=1}^{m}\sum_{i=1}^{nq} \hat{\theta}_{n+(l-1)qn+i}(kq)\gamma_l(u(kq+j+1-i)), \\ &j = 0, 1, \ldots, q-2. \end{aligned} \tag{34}$$

Comparing (18) with (34), we find that the missing intersample output estimates $\hat{y}(kq+j)$, $j = 1, 2, \ldots, q-1$, equal the desired outputs $y_r(kq+j)$; so we have

$$y_r(kq+j) = \hat{y}(kq+j) = \hat{\varphi}^{\mathrm{T}}(kq+j)\hat{\theta}(kq), \quad j = 1, 2, \ldots, q-1, \tag{35}$$

$$\hat{\varphi}(kq+j) = \begin{bmatrix} -y_r(kq+j+1-q) \\ -y_r(kq+j+1-2q) \\ \vdots \\ -y_r(kq+j+1-nq) \\ \psi(kq+j) \end{bmatrix}. \tag{36}$$

It is easy to understand that the unknown intersample outputs $y(kq+j)$ are replaced by the desired outputs $y_r(kq+j)$ because our goal is to make $y(k)$ track $y_r(k)$. Hence, combining (20) with (35)–(36) generates the control signal sequence $\{u(kq+j)$,

$j = 0, 1, \ldots, q-1\}$ based on the parameter estimates $\hat{\theta}(kq)$ obtained. Thus, the following theorem is easily established.

THEOREM 3. *Assume that the conditions of Theorem* 2 *hold, $A(z)$ and $B(z)$ both are stable, and $f(\cdot)$ is invertible. Then the adaptive control algorithm in* (11)–(15), (19), *and* (20) *ensures that the closed-loop system is stable and globally convergent with probability* 1*; moreover,*

- *the input and output variables are uniformly bounded, i.e.,*

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} [u^2(i) + y^2(i) + y_f^2(i)] < \infty \ a.s.;$$

- *the average output tracking error is less than or equal to $\bar{\sigma}^2/q$, i.e.,*

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \mathrm{E}\{[y_f(i) - y_r(i)]^2 | \mathcal{F}_{i-1}\} \le \frac{\bar{\sigma}^2}{q} \ a.s.$$

Proof. Since $y_r(k)$ is bounded, from Theorem 2 and condition (A3), it is easy to get that the outputs $y(kq)$ at the output sampling instants are uniformly bounded, i.e.,

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} y^2(iq) \le \delta_y < \infty \text{ a.s.}$$

Also, the intersample output estimates $\hat{y}(kq+j)$, $j = 1, 2, \ldots, (q-1)$, satisfy

$$\hat{y}(kq+j) = y_r(kq+j), \quad j = 1, 2, \ldots, q-1.$$

So $y_f(kq+j)$ is bounded. According to (32) and (33), $y_f(k)$ is bounded. Since $A(z)$ and $B(z)$ are stable, so are $\alpha(z)$ and $\beta(z)$, and $u(k)$ is bounded in terms of Lemma B.3.3 in [25]. Hence we have

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} u^2(i) < \infty \text{ a.s.}, \qquad \limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} y^2(i) < \infty \text{ a.s.},$$

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} y_f^2(i) < \infty \text{ a.s.},$$

which means that all the input and output variables are uniformly bounded. Also,

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \mathrm{E}\{[y_f(i) - y_r(i)]^2 | \mathcal{F}_{i-1}\} = \limsup_{k\to\infty} \frac{1}{kq} \sum_{i=1}^{k} \mathrm{E}\{[y(iq) - y_r(iq)]^2 | \mathcal{F}_{iq-1}\}$$

$$+ \limsup_{k\to\infty} \frac{1}{kq} \sum_{j=1}^{q-1} \sum_{i=1}^{k} \mathrm{E}\{[\hat{y}(iq+j) - y_r(iq+j)]^2 | \mathcal{F}_{iq+j-1}\}.$$

Since $y_f(i) = \hat{y}(i) = y_r(i)$ at the missing output sampling instants, the last term on the right-hand side is zero, and the first term is no more than $\bar{\sigma}^2/q$ from Theorem 2. This proves Theorem 3. □

Theorems 2 and 3 indicate that the proposed adaptive control scheme in the dual-rate setting can achieve the property of minimum variance at the output sampling instants. Between the output sampling instants, we have $\hat{y}_f(kq+j) = \hat{y}(kq+j) = y_r(kq+j)$, $j = 1, 2, \ldots, q-1$. The control scheme here is easily extended to non-

minimum phase cases by defining a new weighted criterion:

$$J[u(k)] = \mathrm{E}\{[P(z)y_f(k+d) - Q(z)y_r(k+d)]^2 + [R(z)u(k)]^2 | \mathcal{F}_{k-1}\}.$$

Here, $P(z)$, $Q(z)$, and $R(z)$ are polynomials in z^{-1}.

If the single-rate system considered is an AR model with exogenous input (ARX model), then by the model transform, the noise model should be $\phi_q(z)v(k)$ instead of $v(t)$. This leads to an ARMAX model. In this case, we may use the recursive extended least squares algorithm in [25] to estimate the parameters.

The persistent excitation condition is required for the convergence of the parameter estimation. Like in linear single-rate cases [25], adaptive control algorithms do not guarantee the convergence of the parameter estimation to their true values.

In order to avoid generating $u(k)$ with too large magnitudes, for a given small positive ε, if $|\hat{\beta}_1(kq)| < \varepsilon$, we take $\hat{\beta}_1(kq) = \mathrm{sgn}[\hat{\beta}_1(kq)]\varepsilon$, where the sign function is defined by

$$\mathrm{sgn}(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

6. Example. In this section, we illustrate the results reported with simulation examples, including an experimental water level system.

Example 1. For the Hammerstein nonlinear model shown in Figure 1, we select the dual-rate sampling ratio as $q = 2$, the fast-rate discrete-time (dynamic) model $G(z)$ as a second-order one with

$$G(z) = \frac{B(z)}{A(z)} = \frac{0.412z^{-1} + 0.309z^{-2}}{1 - 1.60z^{-1} + 0.80z^{-2}},$$

and the static nonlinear function $f(\cdot)$ as

$$\bar{u}(k) = f(u(k)) = c_1u(k) + c_2u^2(k) + c_3u^3(k) = u(k) + 0.5u^2(k) + 0.25u^3(k).$$

We choose the noise sequence $\{v(k)\}$ to be a white noise sequence with zero mean and variance $\bar{\sigma}^2 = 0.05^2$ and the desired output to be

$$y_r(400i + j) = (-1)^{i+1}, \quad i = 0, 1, 2, \ldots; \quad j = 1, 2, \ldots, 400.$$

The adaptive control algorithm in section 3 is applied to this system. The system output $y(k)$ and the desired output $y_r(k)$ are shown in Figure 3 with $q = 2$. Figure 4 with $q = 1$ is the simulated results of the Åström–Wittenmark self-tuning regulator (A–W STR) suitable for single-rate systems [2].

From Figures 3 and 4, we can see that the control algorithm proposed in this paper can achieve both less and more stationary average tracking errors than the A–W STR algorithm. This is due to $\hat{y}(kq + j) - y_r(kq + j) = 0$, $j = 1, 2, \ldots, q - 1$, between output sampling instants. Thus, the closed-loop tracking performance is satisfactory. This confirms the results reported earlier.

Example 2. This example is a computer-controlled experimental water tank system at the University of Alberta, and is shown in Figure 5. In this system, the manipulated variable is the position of the inlet water valve, denoted $u(t)$; the measured variable is the water level in the tank, denoted $y(t)$; the inner variable, namely, the water flow rate $\bar{u}(t)$ (the output of the water valve) is not measured. The map from $u(t)$ to $\bar{u}(t)$ is nonlinear and can be approximated by a third-order polynomial,

$$\bar{u}(t) = c_1u(t) + c_2u^2(t) + c_3u^3(t)$$

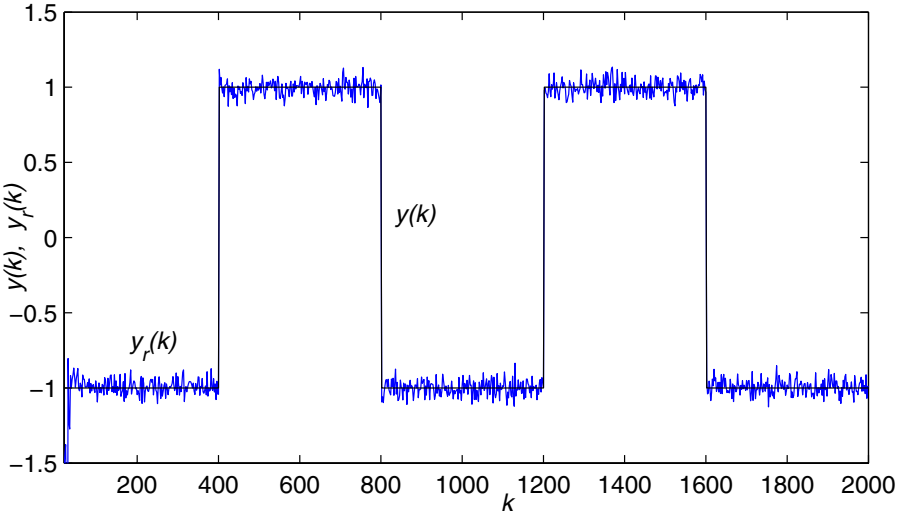


FIG. 3. $y(k)$ *and* $y_r(k)$ *versus* k $(q = 2)$.

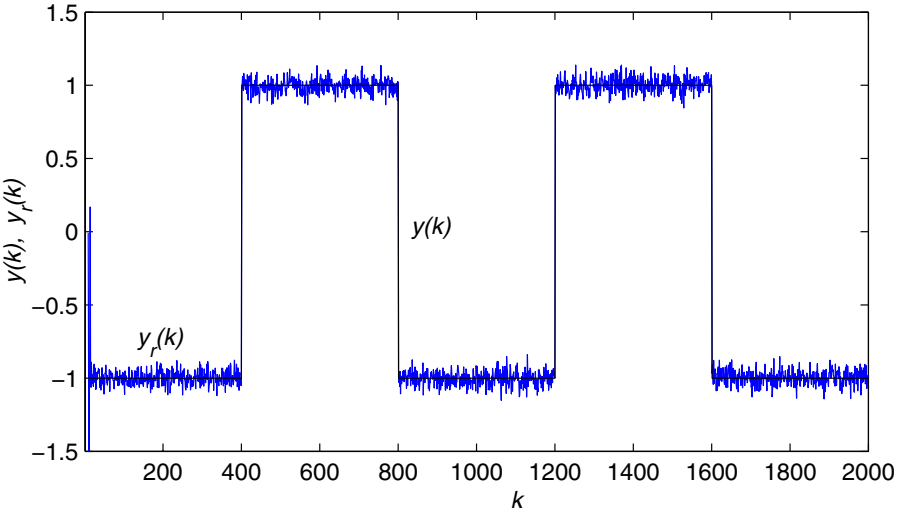


FIG. 4. $y(k)$ *and* $y_r(k)$ *versus* k $(q = 1)$.

(keeping the first few terms in the Taylor series expansion). For our study, the input $u(t)$ is updated every $h = 40$ s to get $u(k)$, and the output $y(t)$ is sampled every $qh = 80$ s (hence $q = 2$), yielding $y(kq)$. Around an operating point, we use a random binary sequence generated by MATLAB as the input signal. By data correlation analysis on the collected input-output data, we determined that the model order associated with $G(z)$ from $\bar{u}(k)$ to $y(k)$ is 2 with a delay of 2 samples, and hence $G(z)$ has the form of

$$G(z) = \frac{b_2 z^{-2} + b_3 z^{-3}}{1 + a_1 z^{-1} + a_2 z^{-2}}.$$

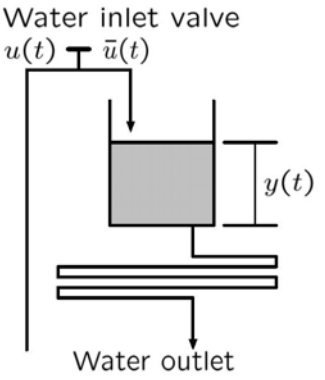


FIG. 5. *Experimental setup of a water tank.*

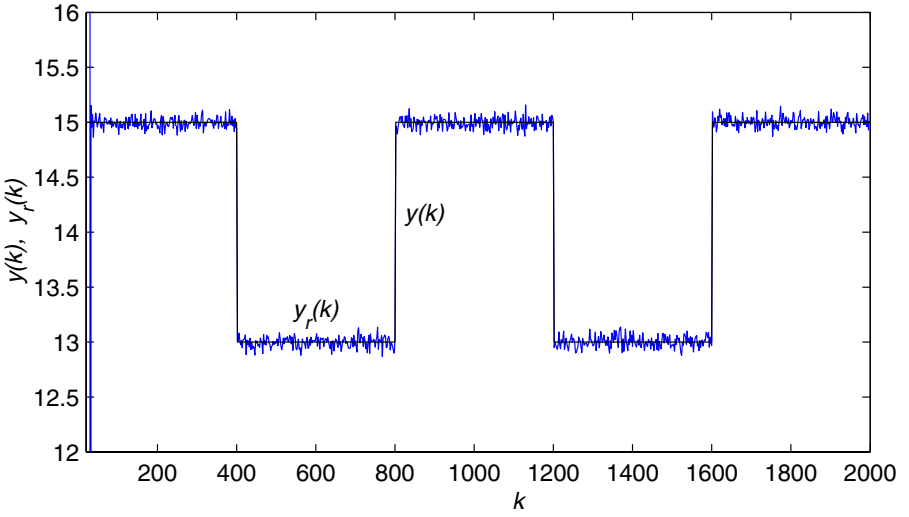


FIG. 6. *$y(k)$ and $y_r(k)$ versus k ($q = 2$).*

Using the polynomial transform, we estimated the new model from $\bar{u}(k)$ to $y(kq)$ as follows:

$$P(z) = \frac{\beta(z)}{\alpha(z)} = \frac{0.92691z^{-2} + 0.83503z^{-3} - 0.65555z^{-4} - 0.66384z^{-5}}{1 - 1.55795z^{-2} + 0.60728z^{-4}}.$$

The parameters of $P(z)$ were obtained by the dual-rate least squares algorithm in (11)–(15); see [20]. Similarly, we took $\{v(k)\}$ to be a white noise sequence with zero mean and variance $\bar{\sigma}^2 = 0.05^2$ and the desired output to be

$$y_r(400i + j) = 14 + (-1)^i, \quad i = 0, 1, 2, \ldots; \quad j = 1, 2, \ldots, 400.$$

Here, $y_r(k) = 14$ is the operating point. The adaptive control algorithm in section 3 was applied to this water tank system. The system output $y(k)$ and the desired output $y_r(k)$ are shown in Figure 6. From Figure 6, we conclude that satisfactory tracking performance has been achieved, and this shows that the algorithm reported is effective.

7. Conclusions. An adaptive control algorithm for Hammerstein nonlinear sampled-data systems with *limited* (slow) output sampling is presented based on only *slow-rate* output measurement data; the algorithm generates a relatively fast-rate control signal from an online parameter identification routine which estimates fast-rate models for Hammerstein systems involving dynamic linear blocks and static nonlinear blocks. Performance analysis in the stochastic framework indicates that the proposed control algorithm can achieve a desired tracking control objective under certain conditions. The simulation results also validate the theoretical findings.

REFERENCES

[1] P. Albertos, J. Salt, and J. Tormero, *Dual-rate adaptive control*, Automatica J. IFAC, 32 (1996), pp. 1027–1030.

[2] K. J. Åström and B. Wittenmark, *On self-tuning regulators*, Automatica J. IFAC, 9 (1973), pp. 185–199.

[3] E. W. Bai, *An optimal two-stage identification algorithm for Hammerstein-Wiener nonlinear systems*, Automatica J. IFAC, 34 (1998), pp. 333–338.

[4] E. W. Bai, *Identification of linear systems with hard input nonlinearities of known structure*, Automatica J. IFAC, 38 (2002), pp. 853–860.

[5] E. W. Bai, *A blind approach to the Hammerstein-Wiener model identification*, Automatica J. IFAC, 38 (2002), pp. 967–979.

[6] E. W. Bai, *Decoupling the linear and nonlinear parts in Hammerstein model identification*, Automatica J. IFAC, 40 (2004), pp. 671–676.

[7] F. Chang and R. Luus, *A noniterative method for identification using Hammerstein model*, IEEE Trans. Automat. Control, 16 (1971), pp. 464–468.

[8] V. Cerone and D. Regruto, *Parameter bounds for discrete-time Hammerstein models with bounded output errors*, IEEE Trans. Automat. Control, 48 (2003), pp. 1855–1860.

[9] E. J. Dempsey and D. T. Westwick, *Identification of Hammerstein models with cubic spline nonlinearities*, IEEE Trans. Biomedical Engineering, 51 (2004), pp. 237–245.

[10] F. Ding and T. Chen, *Least squares based self-tuning control of dual-rate systems*, Internat. J. Adapt. Control Signal Process., 18 (8) (2004), pp. 697–714.

[11] F. Ding and T. Chen, *Combined parameter and output estimation of dual-rate systems using an auxiliary model*, Automatica J. IFAC, 40 (2004), pp. 1739–1748.

[12] F. Ding and T. Chen, *Identification of Hammerstein nonlinear ARMAX systems*, Automatica J. IFAC, 41 (2005), pp. 1479–1489.

[13] F. Ding and T. Chen, *Hierarchical gradient-based identification of multivariable discrete-time systems*, Automatica J. IFAC, 41 (2005), pp. 315–325.

[14] F. Ding and T. Chen, *Hierarchical least squares identification methods for multivariable systems*, IEEE Trans. Automat. Control, 50 (2005), pp. 397–402.

[15] F. Ding and T. Chen, *Hierarchical identification of lifted state-space models for general dual-rate systems*, IEEE Trans. Circuits Syst. I Regul. Pap., 52 (2005), pp. 1179–1187.

[16] F. Ding and T. Chen, *Gradient based iterative algorithms for solving a class of matrix equations*, IEEE Trans. Automat. Control, 50 (2005), pp. 1216–1221.

[17] F. Ding and T. Chen, *On iterative solutions of general coupled matrix equations*, SIAM J. Control Optim., 44 (2006), pp. 2269–2284.

[18] F. Ding and T. Chen, *Iterative least squares solutions of coupled Sylvester matrix equations*, Systems Control Lett., 54 (2005), pp. 95–107.

[19] F. Ding and T. Chen, *Parameter estimation of dual-rate stochastic systems by using an output error method*, IEEE Trans. Automat. Control, 50 (2005), pp. 1436–1441.

[20] F. Ding and T. Chen, *Parameter estimation for dual-rate systems with finite measurement data*, Dyn. Contin. Discrete Impuls. Syst., Ser. B Appl. Algorithms, 11 (2004), pp. 101–121.

[21] F. Ding and T. Chen, *Performance analysis of estimation algorithms of non-stationary ARMA processes*, IEEE Trans. Signal Process., 54 (2006), pp. 1041–1053.

[22] P. G. Gallman, *A comparison of two Hammerstein model identification algorithms*, IEEE Trans. Automat. Control, 21 (1976), pp. 124–126.

[23] I. Goethals, K. Pelkmans, J. A. K. Suykens, and B. De Moor, *NARX identification of Hammerstein models using least squares support vector machines*, in Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS 2004), Elsevier, Amsterdam, 2005, pp. 507–512.

[24] I. Goethals, K. Pelkmans, J. A. K. Suykens, and B. De Moor, *Identification of MIMO Hammerstein models using least squares support vector machines*, Automatica J. IFAC, 41 (2005), pp. 1263–1272.

[25] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control*, Prentice–Hall, Englewood Cliffs, NJ, 1984.

[26] R. D. Gudi, S. L. Shah, and M. R. Gray, *Multirate state and parameter estimation in an antibiotic fermentation with delayed measurements*, Biotechnology and Bioengineering, 44 (1994), pp. 1271–1278.

[27] M. T. Guilandoust, A. J. Morris, and M. T. Tham, *Parameter adaptive inferential control using input-output approach*, Proc. IEE-D, 134 (1987), pp. 171–179.

[28] M. Ishitobi, M. Kawanaka, and H. Nishi, *Ripple-suppressed multirate adaptive control*, in Proceedings of the 15th World Congress of the International Federation of Automatic Control, (Barcelona, Spain), Elsevier, Amsterdam, 2002.

[29] J. Kanniah, O. P. Malik, and G. S. Hope, *Self-tuning regulator based on dual-rate sampling*, IEEE Trans. Automat. Control, 29 (1984), pp. 755–759.

[30] D. Li, S. L. Shah, and T. Chen, *Analysis of dual-rate inferential control systems*, Automatica, 38 (2002), pp. 1053–1059.

[31] D. Li, S. L. Shah, T. Chen, and K. Z. Qi, *Application of dual-rate modeling to CCR octane quality inferential control*, IEEE Trans. Control Systems Technology, 11 (2003), pp. 43–51.

[32] W. P. Lu and G. Fisher, *Least-squares output estimation with multirate sampling*, IEEE Trans. Automat. Control, 34 (1989), pp. 669–672.

[33] K. S. Narendra and P. G. Gallman, *An iterative method for the identification of nonlinear systems using a Hammerstein model*, IEEE Trans. Automat. Control, 11 (1966), pp. 546–550.

[34] M. Ohshima, I. Hashimoto, M. Takeda, T. Yoneyama, and F. Goto, *Multirate multivariable model predictive control and its application to a semi-commercial polymerization reactor*, in Proceedings of the 1992 American Control Conference, (Chicago), Vol. 2, IEEE, Piscataway, NJ, 1992, pp. 1576–1581.

[35] J. Vörös, *Parameter identification of discontinuous Hammerstein systems*, Automatica, 33 (1997), pp. 1141–1146.

[36] J. Vörös, *Recursive identification of Hammerstein systems with discontinuous nonlinearities containing dead-zones*, IEEE Trans. Automat. Control, 48 (2003), pp. 2203–2206.

[37] T. Wigren and A. E. Nordsjö, *Compensation of the RLS algorithm for output nonlinearities*, IEEE Trans. Automat. Control, 44 (1999), pp. 1913–1918.

[38] T. Wigren, *User choices and model validation in system identification using nonlinear Wiener models*, in Proceedings of the 13th IFAC Symposium on System Identification (SYSID), Elsevier, Amsterdam, 2003, pp. 863–868.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2277–2298

OUTPUT STABILIZATION VIA NONLINEAR LUENBERGER OBSERVERS*

L. MARCONI†, L. PRALY‡, AND A. ISIDORI§

Abstract. The present paper addresses the problem of the existence of an (output) feedback law that asymptotically steers to zero prescribed outputs, while keeping all state variables bounded, for any initial conditions in a given compact set. The problem can be viewed as an extension of the classical problem of semiglobally stabilizing the trajectories of a controlled system to a compact set. The problem also encompasses a version of the classical problem of output regulation. Under only a *weak minimum phase* assumption, it is shown that there exists a controller solving the problem at hand. The paper is deliberately focused on theoretical results regarding the existence of such a controller. Practical aspects involving the design and the implementation of the controller are left to a forthcoming work.

Key words. output stabilization, nonlinear output regulation, nonlinear observers, Lyapunov functions, nonminimum-phase systems, robust control

AMS subject classifications. 93D15, 93D21, 93C10, 93B52, 93C15, 93D05

DOI. 10.1137/050642344

1. Introduction.

The problem of controlling a system in such a way that some prescribed outputs converge to zero while all other state variables remain bounded is a relevant problem in control theory. It includes, as special cases, the problem of asymptotic stabilization of a fixed equilibrium point and the problem of asymptotic stabilization of a fixed invariant set. It also includes design problems in which some selected variables are required to asymptotically track (or to asymptotically reject) certain signals generated by an independent autonomous system. Problems of this kind, usually referred to as problems of "output regulation," have been extensively studied in the past for linear systems (see [10, 18, 17]) as well as, beginning with the seminal work [23], for nonlinear systems. As a matter of fact, these problems can be viewed as problems in which a "regulated" output of an "augmented system" (a system consisting of the controlled plant and the exogenous system generator) must be asymptotically steered to zero while all other state variables are kept bounded. As pointed out in [23], for instance, the basic challenges in a problem of this type are to create an invariant set on which the desired regulated output vanishes, and to render this set asymptotically attractive.

Even though paper [23] is limited in scope (the design method suggested therein being only meant to secure local, and nonrobust, regulation about an equilibrium point) it has the merit of highlighting a few basic concepts and ideas which shaped all subsequent developments in this area of research. These ideas include the fundamental link between the problem in question and the notion of "zero dynamics" (a concept introduced and studied earlier by the same authors), the necessity of the existence of

*Received by the editors October 10, 2005; accepted for publication (in revised form) September 3, 2006; published electronically February 13, 2007. This work was partially supported by MIUR and the Institute of Advanced Study (ISA) of the University of Bologna, by NSF grant ECS-0314004.
http://www.siam.org/journals/sicon/45-6/64234.html
†C.A.SY. – D.E.I.S., University of Bologna, Bologna, Italy (lmarconi@deis.unibo.it).
‡CAS, École des Mines de Paris, Fontainebleau, France (Laurent.Praly@ensmp.fr).
§D.I.S., Università di Roma "La Sapienza," Roma, Italy and E.S.E., Washington University, St. Louis, MO 63130 (isidori@ese.wustl.edu).

a (controlled) invariant set on which the desired regulated output vanishes, and an embryo of design philosophy based on the idea of making this invariant set locally (and exponentially) attractive.

In the past 15 years, the design philosophy of [23] was extended in several directions. One clear need was to move from "local" to "nonlocal" convergence, a goal which was pursued, for instance, in [25, 21, 30, 5], where different approaches (at increasing levels of generality) have been proposed. Another concern was to obtain design methods which are insensitive, or even robust, with respect to model uncertainties. This issue was originally addressed in [20], where it was shown how, under appropriate hypotheses, the property of (local) asymptotic regulation can be made robust with respect to plant parameter variations, extending in this way a celebrated property of linear regulators.

In the presence of plant parameter variations, the challenge is to design a (parameter-independent) controller in such a way that the closed-loop system possesses a (possibly parameter *dependent*) attractive invariant set on which the regulated output vanishes. The two issues of forcing the existence of such an invariant set and of making the invariant set (locally or nonlocally) attractive are of course interlaced, and this is precisely what, in the past, has determined the various scenarios under which different solutions to the problem have been proposed. In the paper [20], for instance, a solution was achieved by assuming that the set of all feed-forward controls which force the regulated output to be identically zero had to be generated by a single (parameter-independent) *linear* system. This assumption was weakened in [12], in [11], and subsequently in [6], where it was replaced with the assumption that the controls in question are generated by a single (parameter-independent) *nonlinear* system, uniformly observable in the sense of [19].

The crucial observation that made the advances in [6] and [11] possible was the realization that the two issues of forcing the existence of an invariant set (on which the regulated variable vanishes) and of making the invariant set attractive are intimately related to, and actually can be cast as, the problem of designing a (nonlinear) *observer*. As a matter of fact, the design method suggested in [6] was based almost entirely on the construction of a nonlinear "high-gain" observer following the methods of Gauthier and Kupka [19], while the design method suggested in [11] was based almost entirely on the construction of a nonlinear adaptive observer following the methods of Bastin and Gevers [3] and Marino and Tomei [24].

Having realized that the design of observers is instrumental in the design of controllers which solve the problem in question, researchers came to the idea of examining whether alternative options, in the design of observers, could be of some help in weakening the assumptions even further. This turns out to be true, as shown in the present paper, in the case when we adopt the approach to the design of nonlinear observers outlined by Kazantis and Kravaris [27] and further pursued by Kreisselmeier and Engel [28], Krener and Xiao [26], and Andrieu and Praly [1].

While in all earlier contributions it was assumed that the controls which force the regulated output to be identically zero could be interpreted as outputs of a (in general, nonlinear) system having special observability properties (which eventually became part of the controller), a crucial property highlighted in the proof of Theorem 3 of [1] shows that no assumption of this kind is actually needed. The controls in question can *always* be generated by means of a system of appropriate dimension whose dynamics are linear but whose output map is a nonlinear (and, in general, only continuous but not necessarily locally Lipschitzian) map. Once this system is embedded in the controller, boundedness of all closed-loop trajectories and convergence to the desired

invariant set can be guaranteed, as in the earlier contributions [6] and [11], by a somewhat standard paradigm which blends practical stabilization with a small-gain property for feedback interconnection of systems which are input-to-state stable (with restrictions).

The purpose of this paper is to provide a complete proof of how the results of [1] can be exploited for the design of a controller solving the problem, and also to show how some technical hypotheses used in the asymptotic analysis of [7] can be totally removed, yielding in this way a general theory cast only on a very simple and meaningful assumption. This paper is deliberately meant to present only those theoretical results needed to show the *existence* of the solution of the problem in question. Issues related to practical aspects involving constructive design and implementation will be dealt with in a forthcoming work.

The paper is organized as follows. In the next section the main framework under which the problem is solved is presented and discussed. Then section 3 presents an outline of the main results concerning the existence of the output feedback regulator. Section 4 concludes the paper with some with final remarks. Technical proofs of the results in section 3 are postponed to Appendices A and B.

Notation. For $x \in \mathbb{R}^n$, $|x|$ denotes the Euclidean norm, and for $\mathcal{C}$ a closed subset of $\mathbb{R}^n$, $|x|_{\mathcal{C}} = \min_{y\in\mathcal{C}} |x-y|$ denotes the distance of x from $\mathcal{C}$. For $\mathcal{S}$ a subset of $\mathbb{R}^n$, $\operatorname{cl}\mathcal{S}$ and $\operatorname{int}\mathcal{S}$ are the closure of $\mathcal{S}$ and the interior of $\mathcal{S}$, respectively, and $\partial\mathcal{S}$ its boundary. For the smooth[1] dynamical system $\dot{x} = f(x)$, the value at time t of the solution passing through x_0 at time $t = 0$ will be written as $x(t, x_0)$. The more compact notation $x(t)$ will be used instead of $x(t, x_0)$, when the initial condition is clear from the context. A set $\mathcal{S}$ is said to be forward (backward) invariant for $\dot{x} = f(x)$ if each $x_0 \in \mathcal{S}$, $x(t, x_0) \in \mathcal{S}$ for all $t \geq 0$ ($t \leq 0$). The set is invariant if it is backward and forward invariant. For a locally Lipschitz function $V(t)$ we define Dini's derivative of V at t as

$$D^+V(t) = \lim_{h\to 0^+} \sup \frac{1}{h}[V(t+h) - V(t)].$$

By extension, when $V(t)$ is obtained by evaluating V along a solution $x(t, x_0)$, we denote also

$$D^+V(x_0) = \lim_{h\to 0^+} \sup \frac{1}{h}[V(x(h, x_0)) - V(x_0)]. \tag{1}$$

Note that if $\limsup = \lim$, this is simply $L_fV(x_0)$, the Lie derivative at x_0 of V along f.

2. The framework.

2.1. The problem of output stabilization and the main result.

We consider in what follows a nonlinear *single input–single output* smooth system described by[2]

$$\begin{aligned} \dot{z} &= f(z, y), \\ \dot{y} &= q(z, y) + u \end{aligned} \tag{2}$$

[1] In this paper, "smooth" means "differentiable a sufficiently large number of times" so that all of what we write makes sense.

[2] System (2) is described in the well-known *normal form* with relative degree 1 (see [22]). As discussed in section 2.2, the case of systems in normal form with a higher relative degree can be dealt with in the proposed framework.

with state $(z, y) \in \mathbb{R}^n \times \mathbb{R}$, *measured output* y, and *control input* $u \in \mathbb{R}$, and with initial conditions $(z(0), y(0))$ ranging in a known arbitrary compact set $Z \times \Xi \subset \mathbb{R}^n \times \mathbb{R}$. Associated with (2) there is a *controlled* output $e \in \mathbb{R}^p$ expressed as

$$e = h(z, y) \tag{3}$$

in which $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^p$ is a smooth function.

For system (2)–(3) the problem of semiglobal (with respect to $Z \times \Xi$) output stabilization is defined as follows. Find, if possible, a dynamic output feedback controller of the form

$$\begin{aligned} \dot{\eta} &= \varphi(\eta, y), \\ u &= \varrho(\eta, y) \end{aligned} \tag{4}$$

with state $\eta \in \mathbb{R}^\nu$ and a compact set $M \subset \mathbb{R}^\nu$ such that, in the associated closed-loop system

$$\begin{aligned} \dot{z} &= f(z, y), \\ \dot{y} &= q(z, y) + \varrho(\eta, y), \\ \dot{\eta} &= \varphi(\eta, y), \\ e &= h(z, y), \end{aligned} \tag{5}$$

the positive orbit of $Z \times \Xi \times M$ is bounded and, for each $(z(0), y(0), \eta(0)) \in Z \times \Xi \times M$,

$$\lim_{t \to \infty} e(t) = 0.$$

The problem at issue will be solved only under the following *weak minimum-phase* assumption which requires that system

$$\dot{z} = f(z, 0), \tag{6}$$

representing the zero dynamics of (2) associated with the input u and output y, has a compact attractor which is asymptotically stable. In more precise terms the assumption in question is formulated as follows.

Assumption. There exists a compact set $\mathcal{A} \subset \mathbb{R}^n$ such that

(a_1) the set $\mathcal{A}$ is locally asymptotically stable[3] for system (6) with a domain of attraction $\mathcal{D} \supset Z$;

(a_2) $h(z, 0) = 0$ for all $z \in \mathcal{A}$.

Comments on this assumption and on the proposed framework are postponed until after the next theorem, which presents the main result of the paper.

THEOREM 1. *There exists an $m > 0$, a controllable pair $(F, G) \in \mathbb{R}^{m \times m} \times \mathbb{R}^{m \times 1}$, a continuous function $\gamma : \mathbb{R}^m \to \mathbb{R}$, and for any compact set $M \subset \mathbb{R}^m$, a continuous function $\kappa : \mathbb{R}^p \to \mathbb{R}$, such that the controller*

$$\begin{aligned} \dot{\eta} &= F\eta + Gu, \qquad \eta(0) \in M, \\ u &= \gamma(\eta) + v, \\ v &= \kappa(y) \end{aligned} \tag{7}$$

solves the problem of semiglobal (with respect to $Z \times \Xi$) output stabilization.

[3]Refer to Appendix A for precise definitions regarding the notion of asymptotic stability.

2.2. Remarks on the framework and the result. The framework presented above deals with problems which can be viewed as problems of "external stabilization" of nonlinear systems, namely, problems in which the goal is to steer to zero only a few selected *external* variables, represented in the current framework by the controlled outputs e, while all other variables are simply kept bounded. In this respect the method can be used to handle systems in which there are uncontrollable internal motions that are not necessarily converging to an equilibrium but remain otherwise bounded.

The main idea pursued in the present paper to solve the problem at hand is to extend "high-gain" control paradigms, conventionally used to stabilize a minimum-phase system with respect to an equilibrium point (see [8, 33]), to the case of compact attractors. In this respect assumption (a_1) can be interpreted as a "weak" version of classical minimum-phase hypotheses for the case in which the asymptotic behavior of the zero dynamics (6) is not constrained to being an equilibrium but rather is allowed to be governed by complex bounded dynamics. Clearly, to take advantage of the fact that the trajectories of (6) with initial conditions in Z, as required in part (a_1) of the assumption, are attracted by a compact set $\mathcal{A}$, it would be desirable to have, asymptotically, y converging to zero. Of course since the controlled variable e is required to asymptotically decay to 0, it is also appropriate to assume, as done in (a_2), that $h(z,0)$ vanishes on $\mathcal{A}$. If this were to occur, in fact, then also the controlled variable e would converge to zero, and the problem would be solved. To make y converge to zero, one might wish to appeal to (somewhat standard) "high-gain" arguments and design a control law of the form $u = -ky$. However, it is well known (see, e.g., [33]) that to have y asymptotically converging to zero in a "high-gain" scheme, it is somewhat necessary that the "coupling" term $q(z,y)$ between the upper and the lower subsystem of (2) asymptotically vanishes. More specifically, it is necessary that $q(z,0)$ vanish on the set $\mathcal{A}$ to which the state z of the upper subsystem converges if y decays to zero. Now, in general, there is no guarantee that $q(z,0)$ would vanish on $\mathcal{A}$, and this is why a more elaborate controller has to be synthesized. As a matter of fact, the main result of the paper is that a suitable *dynamic* controller ensures that a property of this kind is achieved.

For the sake of simplicity, we have chosen to illustrate our theory in the case of a system of the form (2), which is rather special for a number of reasons. As a matter of fact, this system has relative degree 1 between the control input u and the measured output y, its "high-frequency gain" is equal to 1, and the dynamics (6) is assumed to possess a compact attractor which is asymptotically stable. However, the main result of the paper lends itself to the synthesis of regulators for more general classes of systems. A natural way in which assumption (a_1) can be weakened consists of viewing y as a "virtual control" of

$$\dot{z} = f(z,y)$$

and assuming the existence of a map $\alpha : \mathbb{R}^n \to \mathbb{R}$ and a compact set $\mathcal{A}$ such that

(a_1') the set $\mathcal{A}$ is locally asymptotically stable for

$$\dot{z} = f(z,\alpha(z))$$

with a domain of attraction $\mathcal{D} \supset Z$,

(a_2') $h(z,\alpha(z)) = 0$ for all $z \in \mathcal{A}$.

If this is the case, in fact, the change of variable $\tilde{y} = y - \alpha(z)$ transforms system (2) into a system in which assumptions (a_1) and (a_2) are fulfilled. The result of Theorem 1 can therefore be used, provided that the variable $\tilde{y}$ is available for feedback.

An extension of this kind is useful in handling cases of systems having relative degree $r > 1$ between the control input u and the measured output y. Consider, to this end, the case of a plant modeled by equations of the form

$$\begin{aligned} \dot{z}_1 &= f(z_1, Cz_2), \\ \dot{z}_2 &= Az_2 + B\zeta, \\ \dot{\zeta} &= q(z_1, z_2, \zeta) + u, \end{aligned} \tag{8}$$

in which $z_1 \in \mathbb{R}^{n-r+1}$, $z_2 \in \mathbb{R}^{r-1}$, $y \in \mathbb{R}$, and A, B, C is a triplet in "prime" form, with measured output

$$y = Cz_2 \tag{9}$$

and controlled output

$$e = h(z_1, z_2, \zeta). \tag{10}$$

Let $Z = Z_1 \times Z_2$ be the prescribed compact set of initial conditions for (z_1, z_2) and suppose there exists a compact set $\mathcal{A}_1$ which is locally asymptotically stable for

$$\dot{z}_1 = f(z_1, 0),$$

with a domain of attraction $\mathcal{D}_1 \supset Z_1$, and such that

$$h(z_1, 0, 0) = 0 \qquad \forall z_1 \in \mathcal{A}_1. \tag{11}$$

If this is the case, the problem of steering e to zero while keeping all internal states bounded can be easily handled in the following way. Standard backstepping arguments (see, e.g., [4] and also [11]) show the existence of matrix K such that, in the system

$$\begin{aligned} \dot{z}_1 &= f(z_1, Cz_2), \\ \dot{z}_2 &= Az_2 + BKz_2, \end{aligned} \tag{12}$$

the compact set $\mathcal{A} = \{(z_1, z_2) : z_1 \in \mathcal{A}_1,\ z_2 = 0\}$ is locally asymptotically stable for (12) with a domain of attraction $\mathcal{D} \supset Z_1 \times Z_2$. Thus, by letting $z = \mathrm{col}(z_1, z_2)$ and $\alpha(z) = Kz_2$, by virtue of (11), it turns out that assumptions (a'_1) and (a'_2) above are fulfilled by system (12) with controlled output (10). This, in view of the previous discussion, guarantees the existence of a controller solving the problem at hand provided that the variable $\tilde{y} := \zeta - Kz_2$ is available for feedback. In particular, since $\mathrm{col}(z_2, \zeta) = (y, \dot{y}, \ldots, y^{(r-1)})$, Theorem 1 guarantees the existence of a *partial state feedback* controller, namely, a controller relying not only on the knowledge of the measured output (9) but also on all its first $r-1$ derivatives with respect to time. This is not a restriction, though, because—as shown, for instance, in [14] and [33]—as long as convergence from a compact set of initial conditions is sought, all components of $\mathrm{col}(z_2, \zeta)$ can always be estimated by means of an "approximate" observer driven only by its first component Cz_2, namely, the actual measured output y.

Finally, it is worth noting that all the results presented in the paper could be generalized to treat the case in which the "high-frequency gain," which is assumed to be unitary in (2), is a generic sign-definite function of the state, namely, the case in which the dynamics of y in (2) is described by

$$\dot{y} = q(z, y) + b(z, y)u \tag{13}$$

with $b(z, y) \geq \underline{b}$ for some known $\underline{b} > 0$. Details on how this can be accomplished are rather straightforward and are not deliberately presented here, as they would only add notational complications without any extra conceptual value.

2.3. Output regulation. A special case covered by the previous setup is the one in which system (2)–(3) is described by equations of the form

$$\begin{array}{rcl} \dot{z}_1 & = & f_1(z_1), \\ \dot{z}_2 & = & f_2(z_1, z_2, y), \\ \dot{y} & = & q(z_1, z_2, y) + u, \\ e & = & h(z_1, z_2, y), \end{array} \tag{14}$$

i.e., a system with "triangular" zero dynamics. In this case, it is clear that the dynamics of z_1 is a totally *autonomous* dynamics, which can be viewed as an "exogenous" signal generator. This is the way in which the classical problem of *output regulation* is usually cast (see [23]). Depending on the control scenario, the variable z_1 may assume different meanings. It may represent exogenous disturbances to be rejected and/or references to be tracked. It may also contain a set of (constant or time-varying) uncertain parameters affecting the controlled plant.

In this context, it is important to note that the proposed framework encompasses a number of problems which have been recently addressed (see, among others, [25, 31, 11, 5, 4, 12]), all relying upon various versions of the so-called "minimum-phase" and "immersion" assumptions.

More specifically, as far as the assumption of "minimum-phase" is concerned, all the aforementioned works require that the dynamics

$$\begin{array}{rcl} \dot{z}_1 & = & f_1(z_1), \\ \dot{z}_2 & = & f_2(z_1, z_2, 0), \end{array} \tag{15}$$

with $z_1 \in \mathbb{R}^s$, $z_2 \in \mathbb{R}^{n-s}$, possess some stability property. For instance, in [31] the assumption in question is the existence of a differentiable map $\pi : \mathbb{R}^s \to \mathbb{R}^{n-r}$ whose graph

$$\mathcal{A} = \{(z_1, z_2) \in \mathbb{R}^s \times \mathbb{R}^{n-r} \ : \ z_2 = \pi(z_1)\}$$

is invariant and locally exponentially stable for (15), uniformly with respect to the exogenous variable z_1, with a domain of attraction containing the assigned compact set of initial conditions and such that

$$h(z_1, z_2, 0) = 0 \qquad \forall\, (z_1,\, z_2) \in \mathcal{A}.$$

This assumption has been substantially weakened in [5] (see also [6] and [4]) by simply asking that the positive orbit of the set of initial conditions under the flow of (15) be bounded (which is, in turn, equivalent to the existence of a compact set $\mathcal{A}$, having the properties indicated above, which is the graph of a *set-valued* map π). Of course, it is apparent that all these cases fit into the framework presented in section 2.1.

Furthermore, the framework presented in the actual literature, where the problems of output regulation are usually tackled, requires an additional assumption commonly referred to as the "immersion" assumption. The latter refers to system (15) with output $q(z_1, z_2, 0)$ which is required to be immersed into a system with prescribed properties. To mention a few of these properties, typical assumptions required immersion into a linear observable system (see [20, 31, 30]), or into a nonlinear system in canonical observability form (see [6]), or into a nonlinear system in adaptive observability form (see [11]).

Remarkably, this additional assumption is not present in the framework proposed in this work. In this respect the important conceptual result proved in Theorem 1 is that no "immersion" assumption is, in principle, necessary for the problem of output regulation to be solvable.

3. Main results.

3.1. The basic approach. In this section we overview the main steps which will be followed to prove Theorem 1. Technical proofs of the results given here are presented in Appendix B.

We consider the closed-loop system (2), (7) which, after the change of coordinates

$$\eta \to x = \eta - Gy,$$

can be rewritten as

$$\begin{aligned} \dot{z} &= f_0(z) + f_1(z,y), \\ \dot{x} &= Fx - Gq_0(z) - Gq_1(z,y) + FGy, \\ \dot{y} &= q_0(z) + q_1(z,y) + \gamma(x + Gy) + v, \end{aligned} \tag{16}$$

in which

$$f_0(z) := f(z,0), \qquad q_0(z) := q(z,0) \tag{17}$$

and

$$\begin{aligned} f_1(z,y) &:= f(z,y) - f(z,0), \\ q_1(z,y) &:= q(z,y) - q(z,0). \end{aligned}$$

Observe that we have $f_1(z,0) \equiv 0$ and $q_1(z,0) \equiv 0$ for all $z \in \mathbb{R}^n$.

Remark. In the case when the y dynamic is described by (13) (namely, the system does not have unitary "high-frequency gain"), it turns out that the change of coordinates to be considered is of the form

$$\eta \to x = \eta - G \int_0^y \frac{1}{b(z,s)}\, ds,$$

which is well defined as $b(z,y) \geq \underline{b} > 0$. In this case the resulting system exhibits all the crucial properties of (16) on which the forthcoming stability analysis is based, with, in particular, the function $q_0(z)$ defined as $q_0(z) = q(z,0)/b(z,0)$.

In what follows, system (16) is seen as a system with input v, output y, and initial conditions contained in a set of the form $Z \times X \times C$ in which $X \subset \mathbb{R}^m$ and $C \subset \mathbb{R}$ are compact sets dependent on Ξ and M. A controller of the form (7) solves the problem at issue if, for some map $\kappa : \mathbb{R}^n \to \mathbb{R}^n$, the control law $v = \kappa(y)$ is such that all trajectories of (16) originating from $Z \times X \times C$ are bounded and

$$\lim_{t\to\infty} y(t) = 0, \qquad \lim_{t\to\infty} |z(t)|_{\mathcal{A}} = 0. \tag{18}$$

As a matter of fact, since systems (2), (7), and (16) are diffeomorphic, boundedness of the trajectories of (16) with initial conditions in $Z \times X \times C$ implies boundedness of the trajectories of (2), (7) originating from $Z \times \Xi \times M$. Furthermore, by virtue of assumption (a_2), since $h(\cdot)$ is a continuous function, condition (18) implies also that $\lim_{t\to\infty} e(t) = 0$, namely, that the problem of semiglobal output stabilization is solved.

By virtue of this fact, in the following we focus our attention on system (16), and we prove that (16) controlled by $v = \kappa(y)$ has bounded trajectories and that (18) holds. To this end, for notational convenience, denote

$$p = \mathrm{col}(z,x)$$

and rewrite system (16) in the more compact form

$$\begin{aligned} \dot{p} &= M(p) + N(p, y), \\ \dot{y} &= H(p) + K(p, y) + v, \end{aligned} \tag{19}$$

in which $M(\cdot)$ and $H(\cdot)$ are defined as

$$M(p) = \begin{pmatrix} f_0(z) \\ Fx - Gq_0(z) \end{pmatrix} \tag{20}$$

and

$$H(p) = q_0(z) + \gamma(x), \tag{21}$$

and $N(\cdot)$ and $K(\cdot)$ are therefore suitable smooth remainder functions which satisfy necessarily $N(p, 0) = 0$ and $K(p, 0) = 0$ for all p. Consistently set $P = Z \times X$ so that the initial conditions of (19) range in $P \times C$. System (19) is recognized as a system in normal form with relative degree one (with respect to input v and output y) and zero dynamics given by

$$\dot{p} = M(p). \tag{22}$$

Thus, following consolidated knowledge about stabilization of *minimum-phase* nonlinear systems (see [8, 2, 33]), the capability of stabilizing (19) by output feedback is expected to strongly rely upon asymptotic properties of the zero dynamics (22). This is confirmed by the next two results showing that the existence of an asymptotically stable attractor for system (22) is sufficient to achieve boundedness of trajectories and *practical* stabilization (Theorem 2), which becomes *asymptotic* if the function $H(\cdot)$ vanishes on the attractor (Theorem 3). These results, which in the context of this paper represent building blocks for proving Theorem 1, are interesting on their own, as they represent an extension of well-known stabilization paradigms for systems with equilibria (see [33]) to the case of systems of the form (19), with zero dynamics (22) possessing compact attractors. For precise definitions of asymptotic and exponential stability used in the statement of the theorems, the reader is referred to Appendix A.

THEOREM 2. *Consider system* (19) *with* $M(\cdot)$ *and* $N(\cdot)$ *at least locally Lipschitz functions and* $H(\cdot)$ *and* $K(\cdot)$ *at least continuous functions. Let the initial conditions be in* $P \times C$. *Assume that system* (22) *has a compact attractor* $\mathcal{B}$ *which is asymptotically stable with a domain of attraction* $\mathcal{D} \supset P$. *Then for all* $\epsilon > 0$ *there exists a* $\kappa^\star > 0$ *such that for all* $\kappa \geq \kappa^\star$ *the trajectories of* (19) *with* $v = -\kappa y$ *are bounded and* $\limsup_{t\to\infty} |y(t)| \leq \epsilon$ *and* $\limsup_{t\to\infty} |p(t)|_{\mathcal{B}} \leq \epsilon$.

THEOREM 3. *In addition to the hypotheses of the previous theorem, assume that* $H(p)|_{\mathcal{B}} = 0$. *Then there exists a continuous function* $\kappa : \mathbb{R} \to \mathbb{R}$ *such that the trajectories of* (19) *with* $v = \kappa(y)$ *are bounded and* $\lim_{t\to\infty} y(t) = 0$ *and* $\lim_{t\to\infty} |p(t)|_{\mathcal{B}} = 0$. *If, additionally,* $H(\cdot)$ *and* $K(\cdot)$ *are locally Lipschitz and the set* $\mathcal{B}$ *is also locally exponentially stable for* (22), *then there exists* $\kappa^\star > 0$ *such that for all* $\kappa \geq \kappa^\star$ *the same properties hold with* $v = -\kappa y$.

For the proofs of these theorems the reader is referred to sections B.1 and B.2, respectively.

Motivated by these results (and in particular by Theorem 3), we turn our attention to the study of the zero dynamics (22) (with $M(\cdot)$ as in (20)) and to the function $H(\cdot)$ in (21) by looking for the existence of a pair (F, G) and a function $\gamma(\cdot)$ which guarantee the basic requirements behind Theorem 3 with, in particular, $H(p)|_{\mathcal{B}} = 0$. Details in this direction are presented in the next subsection.

3.2. The properties of the "core subsystem" (22). The crucial result which will be proved in this part is that, under the assumption presented in section 2.1, there is a choice of the pair (F, G) and of the map $\gamma(\cdot)$ which guarantee the existence of an asymptotically stable compact attractor $\mathcal{B}$ for (22), on which the function $H(\cdot)$ in (21) vanishes. Moreover the projection of $\mathcal{B}$ on the z coordinates coincides with $\mathcal{A}$. In view of the arguments discussed in the previous subsection, this, along with an appropriate choice of $\kappa(\cdot)$ whose existence is claimed in Theorem 3, substantially proves Theorem 1.

The result in question is proved in the next three propositions. To this end, note that the core subsystem (22) in the original coordinates (z, x) is expressed as

$$\begin{aligned} \dot{z} &= f_0(z), \\ \dot{x} &= Fx - Gq_0(z) \end{aligned} \tag{23}$$

with an initial condition in $Z \times X$. The first proposition is related to the first basic requirement behind Theorem 2, namely, the existence of a locally asymptotically stable attractor for (23).

More precisely, under the only requirement, which is that F be a Hurwitz matrix, we show the existence of a set which is *forward invariant* and *locally asymptotically stable* for (23). The set in question is described by the graph of a map.

PROPOSITION 1. *Consider system* (23) *with the z-subsystem satisfying assumption* (a_1)*, and let* (F, G) *be any pair with* F *Hurwitz. Then*

(i) *there exists at least one continuous map* $\tau : \mathbb{R}^n \to \mathbb{R}^m$ *such that the set*

$$\text{graph}(\tau|_{\mathcal{A}}) := \{(z, x) \in \mathcal{A} \times \mathbb{R}^m \ : \ x = \tau(z)\} \tag{24}$$

is forward invariant for (23)*.*

(ii) *the set* $\text{graph}(\tau|_{\mathcal{A}})$ *is locally asymptotically stable for* (23) *with a domain of attraction containing* $Z \times X$*. Furthermore, the set in question is also locally exponentially stable for* (23) *if* $\mathcal{A}$ *is such for* (6)*.*

The proof of this proposition can be found in section B.3.

Remark. Indeed, there might be many different continuous maps τ having property (i) of Proposition 1. However, it turns out that if $\mathcal{A}_0$ is any compact *subset* of $\mathcal{A}$ which is also *backward invariant* for (6), then for each $z \in \mathcal{A}_0$ there is one and only one $x_z \in \mathbb{R}^m$ such that the set $\bigcup_{z \in \mathcal{A}_0} \{(z, x_z)\}$ is invariant for (23). In particular,

$$x_z = -\int_{-\infty}^{0} e^{-Fs} G q_0(z(s, z)) ds,$$

where $z(s, z)$ denotes the value at time $t = s$ of the solution of $\dot{z} = f_0(z)$ passing through $z \in \mathcal{A}_0$ at time $t = 0$ (see [11]).

The second crucial requirement imposed by Theorem 3 is that the function $H(\cdot)$ in (21) vanishes on the asymptotically stable attractor $\text{graph}(\tau|_{\mathcal{A}})$. Here is where the precise choices of the pair (F, G) and of the map $\gamma(\cdot)$ play a role. In particular note that, by definition of $H(\cdot)$ in (21) and of $\text{graph}(\tau|_{\mathcal{A}})$ in (24), it turns out that

$$H(p)|_{\text{graph}(\tau|_{\mathcal{A}})} = (q_0(z) + \gamma \circ \tau(z))|_{\mathcal{A}}, \tag{25}$$

from which it is apparent that $\gamma(\cdot)$ should be chosen to satisfy $\gamma \circ \tau(z) = -q_0(z)$ for all z in $\mathcal{A}$. It is easy to realize that the possibility of choosing $\gamma(\cdot)$ in this way

is intimately related to the fact that the map τ satisfies the *partial* (*with respect to* $q_0(\cdot)$) *injectivity condition*

$$\tau(z_1) = \tau(z_2) \qquad \Rightarrow \qquad q_0(z_1) = q_0(z_2) \qquad \forall\, z_1, z_2 \in \mathcal{A}. \tag{26}$$

In this respect it is interesting to note that what this condition says is that we need only to reconstruct the value of $q_0(z)$ as a function of z from the knowledge of $\tau(z)$ and not the whole state z (see also the remark at the end of the subsection).

As τ is dependent on the pair (F, G), the next natural point to be addressed is if there exists a choice of (F, G) yielding the desired property for $\tau(\cdot)$. To this end is devoted the next proposition which claims that, indeed, there exists a suitable choice of (F, G), with F Hurwitz, such that the associated map $\tau(\cdot)$ satisfies the required partial injectivity condition. Besides other technical constraints on the choice of F, which will be better detailed in the proof of the Proposition 2, the main requirement on F is given by its dimension, which is required to be sufficiently large with respect to the dimension of z.

PROPOSITION 2. *Set*

$$m = 2 + 2n.$$

Then there exist a controllable pair $(F, G) \in \mathbb{R}^{m\times m} \times \mathbb{R}^{m\times 1}$*, with* F *a Hurwitz matrix, and a class-*$\mathcal{K}$ *function* $\varrho : \mathbb{R}^+ \to \mathbb{R}^+$ *such that*

$$|q_0(z_1) - q_0(z_2)| \leq \varrho(|\tau(z_1) - \tau(z_2)|) \qquad \forall\, z_1, z_2 \in \mathcal{A} \tag{27}$$

in which $\tau(\cdot)$ *is a map (associated with* F*) having the properties indicated in Proposition* 1*.*

For the proof of this proposition the reader is referred to section B.4.

Remark. Going through the proof of the previous proposition, it turns out that the pair (F, G) can be chosen as any $(2n + 2)$-dimensional real representation of the $(n + 1)$-dimensional complex pair (F_c, G_c), with $F_c = \operatorname{diag}(\lambda_1, \ldots, \lambda_{r+1})$, $G_c = (g_1, \ldots, g_{r+1})^{\mathrm{T}}$ in which g_i are arbitrary nonzero real numbers and λ_i are $n+1$ complex numbers taken arbitrarily outside a set of zero Lebesgue measure and with real part smaller than ℓ, a real number related to the Lipschitz constant of $f_0(\cdot)$ (see Proposition 4).

It turns out that the injectivity property (27) is a sufficient condition for the map $\gamma(\cdot)$ to exist. This is formalized in the following final proposition, proved in section B.5, which states that if (27) holds, then there exists a map $\gamma(\cdot)$ which makes $H(\cdot)$ vanish on the attractor $\operatorname{graph}(\tau|_{\mathcal{A}})$. The map $\gamma(\cdot)$ can be claimed, in general, to be only continuous. It is also Lipschitz in the special case in which the class-$\mathcal{K}$ function $\varrho(\cdot)$ in (27) is Lipschitz.

PROPOSITION 3. *Let* $\tau(\cdot)$ *be a continuous map satisfying* (27) *with* $\mathcal{A}$ *a closed set. Then there exists a continuous map* $\gamma : \mathbb{R}^m \to \mathbb{R}$ *such that*

$$q_0(z) + \gamma \circ \tau(z) = 0 \qquad \forall\, z \in \mathcal{A}. \tag{28}$$

If, in addition, the function $\varrho(\cdot)$ *in* (27) *is linearly bounded at the origin, then the map* γ *is Lipschitz.*

Combining the results of all the previous propositions, it appears that it is sufficient to choose the pair (F, G) of suitable dimension (with F Hurwitz) according to Proposition 2 and to choose $\gamma(\cdot)$ in order to satisfy relation (28). In fact, by doing

so we are guaranteed that the compact set $\mathcal{B} = \text{graph}(\tau|_{\mathcal{A}})$ is locally asymptotically stable for (23) with the map (21) which is vanishing on $\mathcal{B}$. This, indeed, makes it possible to apply Theorem 3 and to prove the existence of a continuous function $\kappa(\cdot)$, completing in this way the synthesis of the controller.

Remark. The reader who is familiar with recent developments in the theory of nonlinear state observers will find it interesting to compare the previous results with the design method proposed by Kazantzis and Kravaris in [27] and pursued in [28], [26], and [1]. In the framework of [27], system (23) can be identified with the cascade of an "observed" system $\dot{z} = f_0(z)$ with output $y_z = q_0(z)$ driving an "observer" $\dot{x} = Fx - Gy_z$. If the map $\tau(\cdot)$ has a left inverse $\tau_\ell^{-1}(\cdot)$, the observer in question provides a state estimate $\hat{z} = \tau_\ell^{-1}(x)$. Such a left-inverse, as shown in [1], always exists provided that the dimension of x is sufficiently large, if the pair (f_0, q_0) has appropriate *observability* properties. In the present context of output stabilization, though, left invertibility of $\tau(\cdot)$ is not needed. In fact, what the controller is expected to do is reproduce only the output $q_0(z(t))$ and not the full state $z(t)$ of the "observed system." This motivates the absence of observability hypotheses on the pair (f_0, q_0).

4. Conclusions. This paper is focused on the existence of an output feedback law that asymptotically steers to zero a given controlled variable, while keeping all state variables bounded, for any initial conditions in a fixed compact set. The proposed framework encompasses and extends a number of existing results in the fields of output feedback stabilization and output regulation of nonlinear systems. The main assumption under which the theory is developed is the existence of a state feedback control law able to achieve boundedness of the trajectories of the zero dynamics of the controlled plant. In this sense the result presented here is applicable for a wide class of nonminimum-phase nonlinear systems not tractable in existing frameworks. In the paper only results regarding the existence of the controller solving the problem at hand have been presented, while practical aspects involving its design and implementation are left to a forthcoming work.

Appendix A. Converse Lyapunov result. Consider a system of the form

$$\dot{p} = f(p), \qquad p \in \mathbb{R}^n, \tag{29}$$

in which $f(p)$ is a C^k (with k sufficiently large) function, with an initial condition ranging over a fixed *compact* set P. For system (29) assume the existence of a compact set $\mathcal{B} \subset \mathbb{R}^n$ which is asymptotically stable for (29), with a domain of attraction $\mathcal{D} \supset P$. More precisely, by setting

$$|p|_{\mathcal{B}/\mathcal{D}} = \left(1 + \frac{1}{|p|_{\partial\,\text{cl}\,\mathcal{D}}}\right) |p|_{\mathcal{B}},$$

we assume that the set $\mathcal{B}$ satisfies the following two properties:[4]

Uniform stability: There exists a class-$\mathcal{K}$ function φ such that for any $\alpha > 0$,

$$|p_0|_{\mathcal{B}/\mathcal{D}} \leq \alpha \qquad \Rightarrow \qquad |p(t, p_0)|_{\mathcal{B}/\mathcal{D}} \leq \varphi(\alpha) \quad \forall\, t \geq 0.$$

Uniform attractivity: There exists a continuous function $T : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ such that for any $\alpha > 0$ and $\epsilon > 0$,

$$|p_0|_{\mathcal{B}/\mathcal{D}} \leq \alpha \qquad \Rightarrow \qquad |p(t, p_0)|_{\mathcal{B}/\mathcal{D}} \leq \epsilon \quad \forall\, t \geq T(\alpha, \epsilon).$$

[4]In these properties the adjective "uniform" refers to independence with respect to a particular point p in a compact set $|p|_{\mathcal{B}\setminus\mathcal{D}} \leq c$.

We say that $\mathcal{B}$ is also *locally exponentially stable* for (29) if there exist $M \geq 1$, $\lambda > 0$, and $c_0 > 0$ such that

$$|p_0|_{\mathcal{B}/\mathcal{D}} \leq c_0 \qquad \Rightarrow \qquad |p(t,p_0)|_{\mathcal{B}/\mathcal{D}} \leq Me^{-\lambda t}|p_0|_{\mathcal{B}/\mathcal{D}}.$$

In this framework it is possible to formulate the following converse Lyapunov result which claims the existence of a *locally Lipschitz* Lyapunov function vanishing on the attractor. The result is not formally proved, as it can be easily deduced by the arguments presented in [34] (see in particular Theorem 22.5 and the related Theorems 22.1 and 19.2 in the quoted reference).

THEOREM 4. *Under the above uniform stability and uniform attractivity conditions, there exists a continuous function $V : \mathcal{D} \to \mathbb{R}$ with the following properties:*

(a) *There exist class-$\mathcal{K}_\infty$ functions $\underline{a}(\cdot), \overline{a}(\cdot)$ such that*

$$\underline{a}(|p|_{\mathcal{B}/\mathcal{D}}) \leq V(p) \leq \overline{a}(|p|_{\mathcal{B}/\mathcal{D}}) \qquad \forall\, p \in \mathcal{D};$$

(b) *there exists $c > 0$ such that*

$$D^+V(p) \leq -cV(p) \qquad \forall\, p \in \mathcal{D};$$

(c) *for all $\alpha > 0$ there exists $L_\alpha > 0$ such that for all $p_1, p_2 \in \mathcal{D}$ such that $|p_1|_{\mathcal{B}/\mathcal{D}} \leq \alpha$, $|p_2|_{\mathcal{B}/\mathcal{D}} \leq \alpha$, the following holds:*

$$|V(p_1) - V(p_2)| \leq L_\alpha |p_1 - p_2|.$$

If $\mathcal{B}$ is also locally exponentially stable for (29)*, then property* (a) *holds with $\underline{a}(\cdot), \overline{a}(\cdot)$ linear near the origin.*

With this result at hand, it is also possible to formulate a local input-to-state stability result for system (29) forced by an external signal. This is formalized in the next lemma.

LEMMA 1. *Let $x : \mathbb{R}_+ \to \mathbb{R}^m$ be a C^0 function. Consider the system*

$$\dot{p} = f(p) + \ell(p, x(t)) \tag{30}$$

in which $p \in \mathbb{R}^n$ and $\ell(p,0) = 0$ for all $p \in \mathbb{R}^n$. The functions $f(\cdot), \ell(\cdot)$ are C^1. Suppose that system (29) *satisfies the above uniform stability and uniform attractivity conditions. Then there exist functions $\beta(\cdot,\cdot)$ and $\gamma(\cdot)$, respectively, of class $\mathcal{KL}$ and $\mathcal{K}$, and a $d^* > 0$ such that if*

$$|p_0|_{\mathcal{B}} \leq d^* \qquad \text{and} \qquad |x(t)| \leq d^* \quad \forall\, t \geq 0, \tag{31}$$

then the right maximal interval of definition of $p(t,p_0)$ is $[0, +\infty)$, and we have

$$|p(t,p_0)|_{\mathcal{B}} \leq \max\left\{\beta(|p_0|_{\mathcal{B}}, t)\,,\, \gamma\left(\max_{\tau\in[0,t]}|x(\tau)|\right)\right\} \qquad \forall\, t \geq 0. \tag{32}$$

If the set $\mathcal{B}$ is also locally exponentially stable for (29)*, then there exist $N > 1$, $k > 0$, and $\bar{\gamma} > 0$ such that* (32) *can be modified to read*

$$|p(t,p_0)|_{\mathcal{B}} \leq Ne^{-kt}|p_0|_{\mathcal{B}} + \bar{\gamma} \max_{\tau\in[0,t]}|x(\tau)| \qquad \forall\, t \geq 0. \tag{33}$$

Proof. Pick $\beta > 0$ such that if $|p|_{\mathcal{B}} \leq \beta$, then $p \in \mathcal{D}$, and note that there exists $d_\beta > 1$ such that for all p satisfying $|p|_{\mathcal{B}} \leq \beta$,

$$|p|_{\mathcal{B}/\mathcal{D}} \leq d_\beta |p|_{\mathcal{B}}. \tag{34}$$

Pick $d^* > 0$ arbitrarily for the time being but more specifically later on. As $\ell(\cdot)$ is differentiable and $\ell(p,0) = 0$, there is an $\hat{\ell} > 0$ such that for all $|p|_{\mathcal{B}} \leq \beta$ and $|x| \leq d^*$,

$$|\ell(p,x)| \leq \hat{\ell}|x|.$$

So consider the Lyapunov function V given by Theorem 4. By using properties (b) and (c) of this theorem, setting $\alpha = d_\beta \beta$, we obtain for system (30), as long as $|p_1|_{\mathcal{B}} < \beta$ and $|x| \leq d^*$,

$$\begin{aligned} D^+V(p_1,x) &= \limsup_{h\to 0^+} \frac{1}{h}[V(p(h,p_1)) - V(p_1)] \\ &= \limsup_{h\to 0^+} \frac{1}{h}[V(p_1 + hf(p_1) + h\ell(p,x)) - V(p_1)] \\ &\leq \limsup_{h\to 0^+} \frac{1}{h}[V(p_1 + hf(p_1) + h\ell(p_1,x)) - V(p_1 + hf(p_1))] \\ &\qquad + \limsup_{h\to 0^+} \frac{1}{h}[V(p_1 + hf(p_1)) - V(p_1)] \\ &\leq \limsup_{h\to 0^+} \frac{1}{h} L_\alpha h\ell(p_1,x) - cV(p_1) \leq L_\alpha \hat{\ell}|x| - cV(p_1). \end{aligned} \tag{35}$$

Now assume (31) holds. Let $p(t,p_0)$, abbreviated $p(t)$, be the corresponding solution of (30). Let $[0,T_0)$ be its right maximal interval of definition when restricted to take values in the open set $\{p : |p|_{\mathcal{B}} < \beta\}$. Equation (35) holds for $p_1 = p(t)$ and all t in $[0,T_0)$. This implies

$$V(p(t)) \leq e^{-c(t-t_0)}V(p_0) + \frac{L_\alpha \hat{\ell}}{c} \max_{\tau\in[0,t]} |x(\tau)| \qquad \forall t \in [0,T_0). \tag{36}$$

This, in view of property (a) in Theorem 4, yields

$$|p(t)|_{\mathcal{B}} \leq |p(t)|_{\mathcal{B}/\mathcal{D}} \leq \underline{a}^{-1}(2e^{-ct}\,\overline{a}(|p_0|_{\mathcal{B}/\mathcal{D}})) + \underline{a}^{-1}\left(2\frac{L_\alpha \hat{\ell}}{c} \max_{\tau\in[0,t]} |x(\tau)|\right) \quad \forall t \in [0,T_0). \tag{37}$$

By using (34), it follows that if d^* is chosen so that

$$d^\star \leq \min\left\{ \frac{c}{2L_\alpha \hat{\ell}}\, \underline{a}\left(\frac{\beta}{3}\right), \frac{1}{d_\beta}\, \overline{a}^{-1}\left(\frac{1}{2}\, \underline{a}\left(\frac{\beta}{3}\right)\right)\right\}$$

we have

$$|p(t)|_{\mathcal{B}} < \beta \qquad \forall t \in [0,T_0).$$

From the definition of T_0, it must be infinite. So we have established that (37) holds for all $t \geq 0$ if (31) is satisfied. This proves the first part of the result. The second part of the result, namely, that under exponential stability the bound (33) holds, follows immediately by (37) by using the fact that the functions $\underline{a}(\cdot)$ and $\overline{a}(\cdot)$ can be linear near the origin. □

Appendix B. Proofs.

B.1. Proof of Theorem 2. The feedback interconnection (19) can be studied by means of arguments which are quite similar to those used in [22] to prove some of the main stabilization results of [33]. In doing this, we take advantage of the converse Theorem 4 presented in Appendix A.

Let $V : \mathcal{D} \to \mathbb{R}$ be the function given by Theorem 4. Pick a number $a > 0$ such that $C \subset B_a := \{y \in \mathbb{R} : |y| \leq a\}$ and $P \subset V^{-1}([0, a])$ (which is possible because of property (a) in Theorem 4). Define

$$\hat{c} = \max_{(p,y) \in V^{-1}([0,a+1]) \times B_{a+1}} |H(p) + K(p, y)|.$$

Also, since $N(p, y)$ is locally Lipschitz and vanishes at $y = 0$, there is a number $\hat{n}$ such that

$$|N(p, y)| \leq \hat{n}|y| \qquad \forall (p, y) \in V^{-1}([0, a + 1]) \times B_{a+1}.$$

Finally, by property (c) in Theorem 4, there is a number L_V such that

$$|V(p_1) - V(p_2)| \leq L_V |p_1 - p_2| \qquad \forall (p_1, p_2) \in V^{-1}([0, a + 1])^2.$$

Then, by choosing $v = -ky$ in the y-dynamics in (19), we get (see notation (1))

$$D^+|y| \leq -\kappa|y| + \hat{c} \qquad \forall (p, y) \in V^{-1}([0, a + 1]) \times B_{a+1}. \tag{38}$$

Also, by following along the same lines as in (35), we get

$$D^+V(p) \leq L_V \hat{n}|y| - cV(p) \qquad \forall (p, y) \in V^{-1}([0, a + 1]) \times B_{a+1}. \tag{39}$$

So now consider a solution $(p(t), y(t))$ issued from a point in $P \times C \subset V^{-1}([0, a]) \times B_a$. Let $[0, T_1)$ be its right maximal interval of definition when restricted to take values in the open set $\operatorname{int}\big(V^{-1}([0, a + 1]) \times B_{a+1}\big)$. It follows that both (38) and (39) hold for $(p(t), y(t))$ when t is in $[0, T_1)$. They give successively, for all t in $[0, T_1)$,

$$\begin{aligned}
|y(t)| &\leq e^{-\kappa t}\, a + \frac{\hat{c}}{\kappa}\left(1 - e^{-\kappa t}\right), \\
V(p(t)) &\leq e^{-ct}\, a + L_V \hat{n} \left(\frac{\hat{c}}{\kappa} \frac{1 - e^{-ct}}{c} + \frac{e^{-ct} - e^{-\kappa t}}{\kappa - c} \left[a - \frac{\hat{c}}{\kappa} \right] \right) \\
&\leq a + L_V \hat{n} \left(\frac{\hat{c}}{c\kappa} + \frac{a}{\kappa - c} \right).
\end{aligned}$$

Hence, by selecting κ to satisfy

$$\kappa > \max \left\{ 2\hat{c}, (c + 3aL_V \hat{n}), \frac{3L_V \hat{n} \hat{c}}{c} \right\},$$

we get, for all t in $[0, T_1)$,

$$|y(t)| \leq a + \frac{1}{2} \quad \text{and} \quad V(p(t)) \leq a + \frac{2}{3}.$$

This says that the solution remains in $V^{-1}([0, a + \frac{2}{3}]) \times B_{a+\frac{1}{2}}$. So from its definition, T_1 is infinite. Then, from (38), we get

$$\limsup_{t \to +\infty} |y(t)| \leq \frac{\hat{c}}{\kappa}.$$

With (39), this in turn implies

$$\limsup_{t \to +\infty} V(p(t)) \leq \frac{L_V \hat{n} \hat{c}}{\kappa}.$$

In view of property (a) in Theorem 4, the latter yields

$$\limsup_{t\to+\infty} |p(t)|_{\mathcal{B}} \le \underline{a}^{-1}\left(\frac{L_V \hat{n}\hat{c}}{\kappa}\right).$$

So the conclusion of Theorem 2 holds if we further impose that κ satisfies

$$\kappa > \max\left\{\frac{\hat{c}}{\epsilon}, \frac{L_V \hat{n}\hat{c}}{\underline{a}(\epsilon)}\right\}. \qquad \square$$

B.2. Proof of Theorem 3. The proof of this result follows by standard small-gain arguments. Let $\kappa(y) = -\alpha(y)$, where $\alpha(\cdot)$ is a continuous function such that $\alpha(0) = 0$ and $y\alpha(y) > 0$ for all $y \neq 0$. By mimicking the proof of Theorem 2 it is possible to show that for any $\epsilon > 0$ there exist a $\kappa^\star > 0$ and a $T > 0$ such that, if $|\alpha(|y|)| \ge \kappa^\star |y|$, then each trajectory of the closed-loop system issuing from the compact set $P \times C$ satisfies

$$|p(t)|_{\mathcal{B}} \le 2\epsilon \quad \text{and} \quad |y(t)| \le 2\epsilon \quad \forall\, t \ge T.$$

Observe that Lemma 1 applies to the p-component of the closed-loop solution. So let $d^\star$ be given by this lemma. By picking ϵ above satisfying $2\epsilon \le d^\star$, and by applying Lemma 1 to the p-component by picking the initial condition at time $t = T$, we obtain

$$|p(t)|_{\mathcal{B}} \le \max\left\{\beta(|p(T)|_{\mathcal{B}}, t-T)\,,\, \gamma\left(\max_{\tau\in[T,t]} |y(\tau)|\right)\right\} \qquad \forall\, t \ge T. \tag{40}$$

With the properties of the functions H and K, there exist class-$\mathcal{K}$ functions $\varrho_h(\cdot)$ and $\varrho_q(\cdot)$ such that

$$|H(p)| \le \varrho_h(|p|_{\mathcal{B}}), \qquad |K(p,y)| \le \varrho_k(|y|).$$

Clearly $\varrho_h(\cdot)$ and $\varrho_k(\cdot)$ can be taken linearly bounded at the origin if $H(\cdot)$ and $K(\cdot)$ are locally Lipschitz. We obtain, for all (p, y),

$$D^+|y| \;\le\; \varrho_h(|p|_{\mathcal{B}}) + \varrho_k(|y|) - |\alpha(y)|.$$

So let us choose $\alpha(\cdot)$ so that

$$|\alpha(y)| \ge 3\max\{\varrho_h(\bar{\gamma}^{-1}(|y|)),\, \varrho_k(|y|),\, \kappa^\star|y|\} + |y|, \tag{41}$$

where $\bar{\gamma}(\cdot)$ is a class-$\mathcal{K}$ function such that $\bar{\gamma}\circ\gamma(s) < s$ for all $s \in \mathbb{R}^+$ with γ given by Lemma 1 (see (32)). This gives

$$D^+|y| \;\le\; -|y| + \left[\varrho_h(|p|_{\mathcal{B}}) - \varrho_h(\bar{\gamma}^{-1}(|y|))\right].$$

So for the closed-loop solution, we get

$$|y(t)| \;\le\; \max\left\{\exp(-(t-T))|y(T)|\,,\, \sup_{s\in[T,t)} \bar{\gamma}(|p(s)|_{\mathcal{B}})\right\}$$

for all $t \ge T$. From this and (40) the first claim of the theorem follows by small-gain arguments. To prove the second claim of the theorem note that, if $H(\cdot)$ and $K(\cdot)$ are locally Lipschitz, the functions $\varrho_h(\cdot)$ and $\varrho_k(\cdot)$ in (41) can be taken as linear. Furthermore, by (33) in Lemma 1, the function $\bar{\gamma}(\cdot)$ also can be taken as linear. From this the claim directly follows by the previous arguments. $\square$

B.3. Proof of Proposition 1. Let $\mathcal{O}(Z)$ denote the positive orbit of Z under the flow of

$$\dot{z} = f_0(z), \tag{42}$$

namely,

$$\mathcal{O}(Z) = \mathrm{cl}\left\{ \bigcup_{t\geq 0} z(t, z_0) \ : \ z_0 \in Z \right\}.$$

The set $\mathcal{O}(Z)$ is a bounded and forward invariant set for (42), such that $\mathcal{A} \subset \mathcal{O}(Z)$. Moreover let $\hat{\mathcal{O}}(Z)$ be a compact strict superset of $\mathcal{O}(Z)$ such that $\hat{\mathcal{O}}(Z) \subset \mathcal{D}$, and define the system

$$\dot{\hat{z}} = a_0(\hat{z}) f_0(\hat{z}) \tag{43}$$

in which $a(\hat{z}) : \mathbb{R}^n \to \mathbb{R}$ is any bounded smooth function such that

$$a_0(\hat{z}) = \begin{cases} 1, & \hat{z} \in \mathcal{O}(Z), \\ 0, & \hat{z} \in \mathbb{R}^n \setminus \hat{\mathcal{O}}(Z). \end{cases}$$

Let $\hat{z}(t, z_0)$ and $z(t, z_0)$ denote the flows of (43) and (42), respectively, and note that, as a consequence of the fact that $\mathcal{O}(Z)$ is forward invariant and that systems (42) and (43) agree on $\mathcal{O}(Z)$, it turns out that

$$\hat{z}(t, z_0) = z(t, z_0) \qquad \forall\ z_0 \in \mathcal{O}(Z) \text{ and } t \geq 0. \tag{44}$$

Moreover note that for any $\hat{z}_0 \in \mathbb{R}^n$, (43) has a unique solution $\hat{z}(t, \hat{z}_0)$ which is defined and bounded on $t \in (-\infty, \infty)$.

Define now

$$\begin{array}{rcccl} \tau & : & \mathbb{R}^n & \to & \mathbb{R}^m, \\ & & z & \mapsto & \displaystyle\int_{-\infty}^{0} e^{-Fs} G q_0(\hat{z}(s, z)) ds, \end{array} \tag{45}$$

which, as a consequence of the fact that F is Hurwitz and $q_0(\hat{z}(s, z))$ is bounded and continuous in z for any $s \in \mathbb{R}$, is a well-defined *continuous* map. We show now that $\mathrm{graph}(\tau|_{\mathcal{A}}) = \{(z, \xi) \in \mathcal{A} \times \mathbb{R}^m : x = \tau(z)\}$ is a forward invariant set for (23). Pick $z_0 \in \mathcal{A}$ and $x_0 \in \mathbb{R}^m$, let $(z(t, z_0), x(t, z_0, x_0))$ denote the value at time t of the solution of (23) passing through (z_0, x_0) at time $t = 0$, and note that for all $t \geq 0$ (using (44)),

$$\begin{array}{rcl} x(t, z_0, \tau(z_0)) & = & e^{Ft}\tau(z_0) + \displaystyle\int_0^t e^{F(t-s)} G q_0(z(s, z_0)) ds \\ & = & e^{Ft} \displaystyle\int_{-\infty}^{0} e^{-Fs} G q_0(\hat{z}(s, z_0)) ds + \int_0^t e^{F(t-s)} G q_0(z(s, z_0)) ds \\ & = & \displaystyle\int_{-\infty}^{t} e^{F(t-s)} G q_0(\hat{z}(s, z_0)) ds = \int_{-\infty}^{0} e^{-Fs} G q_0(\hat{z}(s+t, z_0)) ds \\ & = & \tau(\hat{z}(t, z_0)) = \tau(z(t, z_0)). \end{array} \tag{46}$$

This, along with the fact that $\mathcal{A}$ is forward invariant for (42) and is a subset of $\mathcal{O}(Z)$, proves that $\mathrm{graph}(\tau|_{\mathcal{A}})$ is *forward invariant* for (23).

We now prove item (ii) of the proposition. To this end note that, by (46), it follows that

$$L_{a_0 f_0}\tau(z) = F\tau(z) - Gq_0(z)$$

for all $z \in \mathcal{A}$. Defining $\tilde{x} := x - \tau(z)$, the previous relation yields that $\dot{\tilde{x}}(t) = F\tilde{x}(t)$ for all $t \geq 0$ and for all initial states $x_0 \in \mathbb{R}^m$ and $z_0 \in \mathcal{A}$. This, the fact that F is Hurwitz, and that $\mathcal{A}$ is locally asymptotically (exponentially) stable for (42) immediately yield the desired result. □

B.4. Proof of Proposition 2. The result will be proved by taking the "complex" pair

$$F = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n+1}), \qquad G = (g, \ldots, g)^{\mathrm{T}} \tag{47}$$

in which $\lambda_i \in \mathbb{C}_\ell = \{\lambda \in \mathbb{C} : \mathrm{Re}\,\lambda < -\ell\}$, $i = 1, \ldots, n+1$, $\ell > 0$, and $g \neq 0$. Once we prove the result for the $(n+1)$-dimensional pair in (47), the claim of the proposition follows by taking any $(2n+2)$-dimensional "real" representation of (47).

By bearing in mind the definition of the map τ in (45) note that, as a consequence of the choice of F and G in (47), it turns out that

$$\tau(z) = \begin{pmatrix} \tau_{\lambda_1}(z) & \tau_{\lambda_2}(z) & \cdots & \tau_{\lambda_{n+1}}(z) \end{pmatrix}^{\mathrm{T}}, \qquad \tau_{\lambda_i}(z) = \int_{-\infty}^{0} e^{-\lambda_i s}\, g\, q_0(\hat{z}(s,z))ds. \tag{48}$$

We will prove next that there exists an $\ell > 0$ such that by arbitrarily choosing λ_i, $i = 1, \ldots, n+1$, in $\mathbb{C}_\ell \setminus S$, where S is a set of zero Lebesgue measure, the map τ is such that

$$\tau(z_1) = \tau(z_2) \qquad \Rightarrow \qquad q_0(z_1) = q_0(z_2) \qquad \forall\, z_1, z_2 \in \mathbb{R}^n. \tag{49}$$

More precisely we will prove that, having defined

$$\Upsilon = \{(z_1, z_2) \in \mathbb{R}^n \times \mathbb{R}^n \ : \ q_0(z_1) \neq q_0(z_2)\},$$

the set

$$S = \{(\lambda_1, \ldots, \lambda_{n+1}) \in \mathbb{C}_\ell^{n+1} \ : \ \exists\, (z_1, z_2) \in \Upsilon \ : \ \tau_{\lambda_i}(z_1) = \tau_{\lambda_i}(z_2) \quad \forall\, i = 1, \ldots, n+1\} \tag{50}$$

has a zero Lebesgue measure in $\mathbb{C}^{n+1}$ for a proper choice of ℓ. To this end the following theorem, proved in a more general setting in [1] (see also [9]), plays a crucial role.

THEOREM 5. *Let Ω and Υ be open subsets of $\mathbb{C}$ and $\mathbb{R}^{2n}$, respectively. Let $(\varpi, \lambda) \in \Upsilon \times \Omega \mapsto \delta_\tau(\varpi, \lambda) \in \mathbb{C}$ be a function which is holomorphic in λ for each $\varpi \in \Upsilon$ and C^1 for each $\lambda \in \Omega$. If, for each pair $\varpi \in \Upsilon$, the function $\lambda \mapsto \delta_\tau(\varpi, \lambda)$ is not identically zero on Ω, then the set*

$$S = \bigcup_{\varpi \in \Upsilon} \{(\lambda_1, \ldots, \lambda_{n+1}) \in \Omega^{n+1} \quad : \quad \delta_\tau(\varpi, \lambda_1) = \cdots = \delta_\tau(\varpi, \lambda_{n+1}) = 0\}$$

has a zero Lebesgue measure in $\mathbb{C}^{n+1}$.

To apply this theorem to our context we first observe the following.

PROPOSITION 4. *There exists an $\ell > 0$ such that for all $\lambda_i \in \mathbb{C}_\ell$, $i = 1, \ldots, n+1$, the map $\tau(\cdot)$ in (48) is C^1.*

Proof of Proposition 4. The map $\tau(\cdot)$ in (48) is C^1 if functions $e^{-\lambda_i s}\, g\, \partial q_0(\hat{z}(s,z))/\partial z$, $i = 1, \ldots, n+1$, are integrable on $s \in (-\infty, 0]$ for all $z \in \mathbb{R}^n$ (see [16]). Consider the expansion

$$\frac{\partial q_0(\hat{z}(s,z))}{\partial z} = \left[\frac{\partial q_0}{\partial z}\right]_{[z=\hat{z}(s,z)]} \frac{\partial \hat{z}(s,z)}{\partial z}.$$

By definition, there is a number M such that $|\hat{z}(s,z)| \leq M$ for all $s \leq 0$ and all $z \in \mathbb{R}^n$. This, along with the fact that $q_0(z)$ is C^1, shows that the first factor is bounded on $(-\infty, 0] \times \mathbb{R}^n$. As for the second factor, bearing in mind the notation introduced in section B.3, observe that

$$\frac{d}{ds}\frac{\partial \hat{z}(s,z)}{\partial z} = \left[\frac{\partial a_0(z) f_0(z)}{\partial z}\right]_{[z=\hat{z}(s,z)]} \frac{\partial \hat{z}(s,z)}{\partial z}.$$

Letting

$$\bar{f} = \max_{z \in \mathbb{R}^n} \frac{\partial a_0(z) f_0(z)}{\partial z},$$

we obtain

$$\left|\frac{\partial \hat{z}(s,z)}{\partial z}\right| \leq e^{\bar{f}|s|}$$

for all s and for all $z \in \mathbb{R}^n$. From this, the result immediately follows with $\ell = \bar{f}$. ∎

Now set $\varpi := (z_1, z_2)$ and

$$\delta_\tau(\varpi, \lambda) = \int_{-\infty}^{0} e^{-\lambda s}\, g\, [q_0(\hat{z}(s,z_1)) - q_0(\hat{z}(s,z_2))]\, ds = \tau_\lambda(z_1) - \tau_\lambda(z_2).$$

This function is C^1 in $\varpi \in \mathbb{R}^n \times \mathbb{R}^n$ and is holomorphic in $\lambda \in \mathbb{C}_\ell$ for every $\varpi \in \mathbb{R}^n \times \mathbb{R}^n$ (see [29, Chap. 19, p. 367]). Moreover as

$$\int_{-\infty}^{0} e^{-as}\, |g\, q_0(\hat{z}(s,z_1)) - g\, q_0(\hat{z}(s,z_2))\,|^2\, ds < +\infty$$

for all $\varpi \in \mathbb{R}^n \times \mathbb{R}^n$ and for all $a < 0$, the Plancherel theorem can be invoked to obtain

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} |\delta_\tau(\varpi, a+is)|^2 ds = \int_{-\infty}^{0} e^{-2as}\, |g\, q_0(\hat{z}(s,z_1)) - g\, q_0(\hat{z}(s,z_2))\,|^2\, ds \tag{51}$$

for all $a < 0$ and for all $\varpi \in \mathbb{R}^n \times \mathbb{R}^n$.

Now note that, for $\varpi = (z_1, z_2) \in \Upsilon$, we have $q_0(z_1) \neq q_0(z_2)$, and by continuity of flow with respect to time, there exists a time $t_1 < 0$ such that

$$|g\, q_0(\hat{z}(s,z_1)) - g\, q_0(\hat{z}(s,z_2))| > 0 \qquad \forall\, s \in (t_1, 0]$$

which, combined with (51), yields

$$\int_{-\infty}^{\infty} |\delta_\tau(\varpi, a+is)|^2 ds > 0.$$

This implies that, for each $\varpi \in \Upsilon$, the function $\lambda \mapsto \delta_\tau(\varpi, \lambda)$ is not identically zero on $\mathbb{C}_\ell$. Hence Theorem 5 can be applied to obtain the desired result, namely, that the set (50) has a zero Lebesgue measure.

By this result we are guaranteed that by arbitrarily picking $n+1$ complex eigenvalues in $\mathbb{C}_\ell \setminus S$ (with ℓ dictated by Proposition 4) of F defined in (47), condition (49) is satisfied. From this it is easy to show that there exists a class-$\mathcal{K}$ function satisfying (27). Define

$$\varphi(s) = \sup_{\substack{|\tau(z_1)-\tau(z_2)|\le s \\ z_1,z_2\in\mathcal{A}}} |q_0(z_1) - q_0(z_2)|.$$

This function is increasing and, as a consequence of (49), $\varphi(0) = 0$. Moreover it is possible to prove that $\varphi(s)$ is continuous at $s = 0$. Suppose that it is not; namely, as $\varphi(\cdot)$ is increasing and $\varphi(0) = 0$, suppose that there exists a $\varphi^\star > 0$ such that $\lim_{s\to 0^+} \varphi(s) = \varphi^\star$. This implies that there exist sequences $\{z_{1n}\}, \{z_{2n}\}$ in $\mathcal{A}$, such that $|q_0(z_{1n}) - q_0(z_{2n})| \ge \varphi^\star/2$ and $|\tau(z_{1n}) - \tau(z_{2n})| < 1/n$ for any $n \in \mathbb{N}$. But, as $\mathcal{A}$ is bounded, there are subsequences of $\{z_{1n}\}$, $\{z_{2n}\}$ which, for $n \to \infty$, converge to $z_1^\star$, $z_2^\star$, respectively. As $\tau(\cdot)$ and $q_0(\cdot)$ are continuous, $\tau(z_1^\star) - \tau(z_2^\star) = 0$ and $|q_0(z_1^\star) - q_0(z_2^\star)| \ge \varphi^\star/2$, which contradict (49). Hence, $\varphi(s)$ is continuous at $s = 0$. With this result at hand, define the candidate class-$\mathcal{K}$ function

$$\varrho(s) = \frac{1}{s}\int_s^{2s} \varphi(\sigma)d\sigma \; + \; s$$

which satisfies

$$\varphi(s) \le \varrho(s). \tag{52}$$

By construction this function is continuous for all $s > 0$ and, as $\varphi(s)$ is continuous at $s = 0$ and by (52), it is also continuous at $s = 0$. Moreover since, by definition of $\varphi(\cdot)$,

$$\varphi(|\tau(z_1) - \tau(z_2)|) \ge |q_0(z_1) - q_0(z_2)| \qquad \forall\, z_1, z_2 \in \mathcal{A},$$

it turns out that (52) yields

$$|q_0(z_1) - q_0(z_2)| \le \varphi(|\tau(z_1) - \tau(z_2)|) \le \varrho(|\tau(z_1) - \tau(z_2)|) \qquad \forall\, z_1, z_2 \in \mathcal{A},$$

namely, (27) is satisfied. This concludes the proof of Proposition 2. □

B.5. Proof of Proposition 3. By the result of the previous proposition we know that

$$\tau(z_1) = \tau(z_2) \qquad \Rightarrow \qquad q_0(z_1) = q_0(z_2) \qquad \forall\, z_1, z_2 \in \mathcal{A}.$$

For any $x \in \tau(\mathcal{A})$, let $[x] = \{z \in \mathcal{A} : \tau(z) = x\}$. The previous property shows that the map $q_0(\cdot)$ is constant on $[x]$. As a consequence, there is a well-defined function $\gamma_0 : \tau(\mathcal{A}) \to \mathbb{R}$ such that

$$\gamma_0(\tau(z)) = -q_0(z) \qquad \forall z \in \mathcal{A}.$$

In fact, the value $\gamma_0(x)$ at any $x \in \tau(\mathcal{A})$ is simply defined by taking any $z \in [x]$ and setting $\gamma_0(x) := -q_0(z)$. Moreover by (27), the map in question is also continuous.

Now note that, as $\mathcal{A}$ is compact and $\tau(\cdot)$ and $q_0(\cdot)$ are continuous maps, $\tau(\mathcal{A}) \subset \mathbb{R}^m$ and $q_0(\mathcal{A}) \subset \mathbb{R}$ are compact sets. From this, Tietze's extension theorem (see, for instance, Theorem VII.5.1 in [13]) can be invoked to claim the existence of a continuous map $\gamma : \mathbb{R}^m \to \mathbb{R}$ which agrees with γ_0 on $\tau(\mathcal{A})$. This implies that $q_0(z) + \gamma \circ \tau(z) = 0$ for all $z \in \mathcal{A}$ and proves the first claim of the proposition.

Furthermore, if $\varrho(\cdot)$ is linearly bounded at the origin, by compactness arguments it is possible to claim the existence of a positive $\bar{\varrho}$ such that $|q_0(z_1) - q_0(z_2)| \leq \bar{\varrho}|\tau(z_1) - \tau(z_2)|$. It follows that γ_0 is a Lipschitz function on $\tau(\mathcal{A})$. From this, the Kirszbraun theorem (see, for instance, Theorem 2.10.43 in [15]) yields the existence of a Lipschitz map $\gamma : \mathbb{R}^m \to \mathbb{R}$, with Lipschitz constant ϱ, which agrees with γ_0 on $\tau(\mathcal{A})$. This completes the proof of the proposition. □

REFERENCES

[1] V. Andrieu and L. Praly, *On the existence of a Kazantzis–Kravaris/Luenberger observer*, SIAM J. Control Optim., 45 (2006), pp. 432–456.

[2] A. Bacciotti, *Linear feedback: The local and potentially global stabilization of cascade systems*, in Proceedings of the 2nd IFAC Symposium on Nonlinear Control and System Design, Bordeaux, France, 1992.

[3] G. Bastin and M. R. Gevers, *Stable adaptive observers for non-linear time varying systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 650–657.

[4] C. I. Byrnes, A. Isidori, and L. Marconi, *Further results on output regulation by pure error feedback*, in Proceedings of the 16th IFAC World Congress, Prague, 2005.

[5] C. I. Byrnes and A. Isidori, *Limit sets, zero dynamics and internal models in the problem of nonlinear output regulation*, IEEE Trans. Automat. Control, 48 (2003), pp. 1712–1723.

[6] C. I. Byrnes and A. Isidori, *Nonlinear internal models for output regulation*, IEEE Trans. Automat. Control, 49 (2004), pp. 2244–2247.

[7] C. I. Byrnes, A. Isidori, and L. Praly, *On the Asymptotic Properties of a System Arising in Non-Equilibrium Theory of Output Regulation*, Preprint, Mittag-Leffler Institute, Stockholm, 2002–2003.

[8] C. I. Byrnes and A. Isidori, *Asymptotic stabilization of minimum-phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1122–1137.

[9] J. M. Coron, *On the stabilization of controllable and observable systems by output feedback law*, Math. Control Signals Systems, 7 (1994), pp. 187–216.

[10] E. J. Davison, *The robust control of a servomechanism problem for linear time-invariant multivariable systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 25–34.

[11] F. Delli Priscoli, L. Marconi, and A. Isidori, *A new approach to adaptive nonlinear regulation*, SIAM J. Control Optim., 45 (2006), pp. 829–855.

[12] Z. Chen and J. Huang, *Global robust servomechanism problem of lower triangular systems in the general case*, Systems Control Lett., 52 (2004), pp. 209–220.

[13] J. Dugundji, *Topology*, Allyn and Bacon, Boston, 1966.

[14] F. Esfandiari and H. Khalil, *Output feedback stabilization of fully linearizable systems*, Internat. J. Control, 56 (1992), pp. 1007–1037.

[15] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969 (reprinted in 1996).

[16] W. Fleming, *Functions of Several Variables*, 2nd ed., Springer-Verlag, New York, 1977.

[17] B. A. Francis, *The linear multivariable regulator problem*, SIAM J. Control Optim., 15 (1977), pp. 486–505.

[18] B. A. Francis and W. M. Wonham, *The internal model principle of control theory*, Automatica J. IFAC, 12 (1976), pp. 457–465.

[19] J. P. Gauthier and I. Kupka, *Deterministic Observation Theory and Applications*, Cambridge University Press, Cambridge, UK, 2001.

[20] J. Huang and C. F. Lin, *On a robust nonlinear multivariable servomechanism problem*, IEEE Trans. Automat. Control, 39 (1994), pp. 1510–1513.

[21] A. Isidori, *A remark on the problem of semiglobal nonlinear output regulation*, IEEE Trans. Automat. Control, 42 (1997), pp. 1734–1738.

[22] A. Isidori, *Nonlinear Control Systems* II, Springer-Verlag, New York, 1999.

[23] A. Isidori and C. I. Byrnes, *Output regulation of nonlinear systems*, IEEE Trans. Automat. Control, 25 (1990), pp. 131–140.

[24] R. Marino and P. Tomei, *Global adaptive observers for nonlinear systems via filtered transformations*, IEEE Trans. Automat. Control, 37 (1992), pp. 1239–1245.
[25] H. Khalil, *Robust servomechanism output feedback controllers for feedback linearizable systems*, Automatica J. IFAC, 30 (1994), pp. 1587–1599.
[26] A. J. Krener and M. Xiao, *Erratum: Nonlinear observer design in the Siegel domain*, SIAM J. Control Optim., 43 (2004), pp. 377–378.
[27] K. Kazantzis and C. Kravaris, *Nonlinear observer design using Lyapunov's auxiliary theorem*, Systems Control Lett., 34 (1998), pp. 241–247.
[28] G. Kreisselmeier and R. Engel, *Nonlinear observer for autonomous Lipschitz continuous systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 451–464.
[29] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[30] A. Serrani, A. Isidori, and L. Marconi, *Semiglobal output regulation for minimum-phase systems*, Internat. J. Robust Nonlinear Control, 10 (2000), pp. 379–396.
[31] A. Serrani, A. Isidori, and L. Marconi, *Semiglobal nonlinear output regulation with adaptive internal model*, IEEE Trans. Automat. Control, 46 (2001), pp. 1178–1194.
[32] J. Szarski, *Differential Inequalities*, Polska Akademia Nauk, Warszawa, 1967.
[33] A. Teel and L. Praly, *Tools for semiglobal stabilization by partial state and output feedback*, SIAM J. Control Optim., 33 (1995), pp. 1443–1488.
[34] T. Yoshizawa, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, Springer-Verlag, New York, 1975.

SIAM J. CONTROL OPTIM.
Vol. 45, No. 6, pp. 2299–2304

DYNAMIC PROGRAMMING FOR ERGODIC CONTROL OF MARKOV CHAINS UNDER PARTIAL OBSERVATIONS: A CORRECTION*

VIVEK S. BORKAR†

Abstract. A gap in the author's work on dynamic programming for ergodic control of partially observed Markov chains [V. S. Borkar, *SIAM J. Control Optim.*, 39 (2000), pp. 673–681] is pointed out and a correction is provided.

Key words. controlled Markov chains, ergodic control, partial observations, dynamic programming

DOI. 10.1137/05064610X

1. Introduction. Huizen Yu of MIT has pointed out a subtle gap in [1] having to do with the definition of wide sense admissible controls in control of partially observed Markov processes. The error is in the handling of Lemma 3.1 of [1]. We explain this in the next section in the context of [1] and present a corrected proof of Lemma 3.1 in section 3. The ideas of [1] were applied to a specific control problem in [4]. The results were extended to Markov processes with a Euclidean state space in [2], [3], taking a slightly modified approach. This will be commented upon later.

2. Wide sense admissibility. The framework of [1] is a controlled Markov chain $\{X_n\}$ on a finite state space S with an associated control process $\{Z_n\}$ taking values in a compact metric control space A and an associated observations process $\{Y_n\}$ taking values in a finite observation space W. The transition mechanism is

$$P(X_{n+1}=i, Y_{n+1}=j | X_m, Y_m, Z_m, m \le n) = p(X_n, Z_n, i, j) \tag{1}$$

for $n \ge 0, i \in S, j \in W$, where $p : S \times A \times S \times W \to [0,1]$ is a continuous "transition probability" function satisfying $\sum_{i,j} p(k,u,i,j) = 1, k \in S, u \in A$. X_0 has a known distribution π_0. We follow the common convention that (Y_0, Z_0) are independent of X_0. The aim is to minimize the ergodic cost

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{m=0}^{n-1} E[h(X_m, Z_m)]$$

for a prescribed $h \in C(S \times A)$ over $\{Z_n\}$ adapted to $\sigma(Y_m, m \le n), n \ge 0$, the so-called *strict sense admissible controls*. [1] introduces a relaxation of this class to the larger class of *wide sense admissible controls* defined as follows. On a possibly augmented probability space, $\{X_n\}$ above may be viewed as being given by the dynamics

$$X_{n+1} = F(X_n, Z_n, \xi_n) \tag{2}$$

for $\{\xi_n\}$ independently and identically distributed (i.i.d.) uniformly distributed on $[0,1]$ and F a measurable map $S \times A \times [0,1] \to S$ that depends on the function $p(\cdot)$ above. Suppose (X_n, Y_n, Z_n, ξ_n) are defined on a probability space $(\Omega, \mathcal{F}, \bar{P})$ such that there is another probability measure P_0 on $(\Omega, \mathcal{F})$ with the following properties:

*Received by the editors November 25, 2005; accepted for publication (in revised form) July 17, 2006; published electronically February 20, 2007.
http://www.siam.org/journals/sicon/45-6/64610.html
†School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India (borkar@tifr.res.in).

1. Under P_0, $\{Y_n\}$ is i.i.d. uniform on W and independent of $X_0, \{\xi_n\}$, and furthermore $\{Z_n, Y_0, \ldots, Y_n\}$ is independent of $\{\xi_m\}, X_0, \{Y_i, i > n\}$ $\forall n$.
2. If $\bar{P}_n, P_{0n}$ are restrictions of $\bar{P}, P_0$, resp., to $\sigma(X_0; Y_m, m \leq n; \xi_m, Z_m, m < n)$ for $n \geq 1$, then $\bar{P}_n \ll P_{0n}$ with

$$\frac{d\bar{P}_n}{dP_{0n}} = \Pi_{m=0}^{n-1} \frac{p(X_m, Z_m, X_{m+1}, Y_{m+1})}{\bar{p}(X_m, Z_m, X_{m+1})(1/d)},$$

where $d = |W|$ and $\bar{p}(i, z, j) = \sum_y p(i, z, j, y)$ $\forall i, z, j$.

Then we say that $\{Z_n\}$ is *wide sense admissible*. This includes strict sense admissible controls. The wide sense admissible control is in effect prescribed by the joint law ϕ of $(\{Y_n\}, \{Z_n\})$ under P_0. Hence it can be identified with this ϕ. Let Φ denote the set of wide sense admissible controls, i.e., of ϕ as above. The "vanishing discount" argument used later in [1] requires that we consider a pair of state-observation processes $\{\hat{X}_n, \hat{Y}_n\}$ and $\{\tilde{X}_n, \tilde{Y}_n\}$ on the same probability space with a *common* control process $\{Z_n\}$ that is wide sense admissible for both, and corresponds to a prescribed ϕ for one of them, say the latter. The way [1] tried to achieve this is as follows:

1. Given initial distributions $\hat{\pi}_0, \tilde{\pi}_0$ on S, consider

$$\hat{\Omega} = S \times S \times [0,1]^\infty \times [0,1]^\infty \times W^\infty \times A^\infty \times W^\infty$$

with the product σ-field $\hat{\mathcal{F}}$ and the probability measure

$$\begin{aligned} &\hat{P}_0(d\hat{x} \times d\tilde{x} \times d\hat{u}^\infty \times d\tilde{u}^\infty \times d\tilde{y}^\infty \times dz^\infty \times d\hat{y}^\infty) \\ &\qquad = \Gamma(d\hat{x}d\tilde{x})\ell_1^\infty(d\hat{u}^\infty)\ell_1^\infty(d\tilde{u}^\infty)\phi(d\tilde{y}^\infty, dz^\infty)\ell_2^\infty(d\hat{y}^\infty), \end{aligned} \tag{3}$$

where,

- ℓ_1, ℓ_2 are uniform distributions on $[0,1], W$, resp.,
- $\Gamma(d\hat{x} \times S) = \hat{\pi}_0(d\hat{x}), \Gamma(S \times d\tilde{x}) = \tilde{\pi}_0(d\tilde{x})$, and
- if $\phi \approx$ the law of $(Y_0', Y_1', Y_2', \ldots; Z_0', Z_1', Z_2', \ldots)$, then
  (†) $\{Y_n'\}$ are i.i.d. uniform on W and for each n, $Y_m', m > n$, is independent of $Z_m', m \leq n; Y_k', k < n$.

Condition (†) is effectively the wide sense admissibility condition of $\{Z_n\}$ for $(\{\tilde{X}_n\}, \{\tilde{Y}_n\})$. It should be noted that the definition of (3) in [1] uses $\Gamma(d\hat{x}, d\tilde{x}) = \hat{\pi}_0(d\hat{x})\tilde{\pi}_0(\tilde{x})$, i.e., the two initial conditions are independent under $\hat{P}_0$, but a dependence is introduced later on in [1].

2. Define random variables $(\hat{X}_0, \tilde{X}_0, \{\hat{\xi}_n\}, \{\tilde{\xi}_n\}, \{\tilde{Y}_n\}, \{Z_n\}, \{\hat{Y}_n\})$ canonically on this probability space so that their joint law is $\hat{P}_0$. (Thus, for example, $\hat{X}_0([\hat{x}, \tilde{x}, \hat{u}^\infty, \tilde{u}^\infty, \tilde{y}^\infty, z^\infty, \hat{y}^\infty]) = \hat{x}$.)
3. Generate $\{\hat{X}_n\}, \{\tilde{X}_n\}$ from $(\hat{X}_0, \{Z_n\}, \{\hat{\xi}_n\}), (\tilde{X}_0, \{Z_n\}, \{\tilde{\xi}_n\})$, resp., by using (2).
4. Change measure to $\hat{P}$ as follows: If $\hat{P}_n, \hat{P}_{0n}$ are the restrictions of $\hat{P}, \hat{P}_0$, resp., to $\mathcal{G}_n \triangleq \sigma(\hat{X}_0, \tilde{X}_0; \hat{\xi}_m, \tilde{\xi}_m, Z_m, m < n; \hat{Y}_m, \tilde{Y}_m, m \leq n)$ for $n \geq 0$, then $\hat{P}_n \ll \hat{P}_{0n}$ with

$$\begin{aligned} \frac{d\hat{P}_n}{d\hat{P}_{0n}} &= \Lambda_n \\ &\triangleq \Pi_{m=0}^{n-1} \left(\frac{p(\hat{X}_m, Z_m, \hat{X}_{m+1}, \hat{Y}_{m+1})}{\bar{p}(\hat{X}_m, Z_m, \hat{X}_{m+1})(1/d)} \right) \left(\frac{p(\tilde{X}_m, Z_m, \tilde{X}_{m+1}, \tilde{Y}_{m+1})}{\bar{p}(\tilde{X}_m, Z_m, \tilde{X}_{m+1})(1/d)} \right). \end{aligned} \tag{4}$$

Let $E_0[\ \cdot\ ]$ denote the expectation under $\hat{P}_0$. Lemma 3.1 of [1] claims that under $\hat{P}$, $\{\hat{X}_n, \hat{Y}_n\}, \{\tilde{X}_n, \tilde{Y}_n\}$ are two state-observation pairs with a common control $\{Z_n\}$ wide sense admissible for both. This is far from obvious. While $\{Z_n\}$ is clearly wide sense admissible for $\{\tilde{X}_n, \tilde{Y}_n\}$, it is not obviously so for $\{\hat{X}_n, \hat{Y}_n\}$. This is because, while

$$E_0\left[\frac{d\hat{P}_n}{d\hat{P}_{0n}}\middle| \tilde{X}_m, \tilde{Y}_m, m \le n; Z_m, m < n\right] \tag{5}$$

does equal

$$\Lambda_{2n} \triangleq \Pi_{m=0}^{n-1}\left(\frac{p(\tilde{X}_m, Z_m, \tilde{X}_{m+1}, \tilde{Y}_{m+1})}{\bar{p}(\tilde{X}_m, Z_m, \tilde{X}_{m+1})(1/d)}\right),$$

its counterpart

$$E_0\left[\frac{d\hat{P}_n}{d\hat{P}_{0n}}\middle| \hat{X}_m, \hat{Y}_m, m \le n; Z_m, m < n\right] \tag{6}$$

need not equal

$$\Lambda_{1n} \triangleq \Pi_{m=0}^{n-1}\left(\frac{p(\hat{X}_m, Z_m, \hat{X}_{m+1}, \hat{Y}_{m+1})}{\bar{p}(\hat{X}_m, Z_m, \hat{X}_{m+1})(1/d)}\right).$$

It is, however, of the form

$$\hat{\Lambda}_{1n} \triangleq \Pi_{m=0}^{n-1}\left(\frac{p(\hat{X}_m, Z_m, \hat{X}_{m+1}, \hat{Y}_{m+1})}{\bar{p}(\hat{X}_m, Z_m, \hat{X}_{m+1})(1/d)}\right)\zeta_n.$$

Here ζ_n is an a.s. positive random variable of the form

$$\zeta_n = \psi_n(\hat{X}_0, Z_0, \ldots, Z_{n-1})$$

for a suitable measurable function ψ_n, and is independent of $\{\hat{\xi}_k\}$ and $\{\hat{Y}_m, m \ge n\}$ under $\hat{P}_0$.

3. The correction. In order to set this right, we need the following additional assumption:

$$p(i, u, j, k) > 0\ \forall k \text{ whenever } \bar{p}(i, u, j) > 0. \tag{7}$$

In particular, this implies that the ratio $p(i, u, j, k)/(\bar{p}(i, u, j)(1/d))$ is strictly positive when defined. The following example (due to Huizen Yu) shows that some condition like this will be necessary to obtain a state-independent optimal cost for the partially observed control problem.

Example. Let $S = \{1, 2, 3, 4\}$, $A = \{a, b\}$, $W = \{q, r\}$, and let the following transition probabilities hold:

$$p(1, u, 1, q) = \frac{1}{2} \ = \ p(1, u, 2, r)\ \forall u,$$

$$p(1, u, k, \ell) = 0 \ \text{ otherwise},$$

$$p(3, u, 3, q) = \frac{1}{2} \ = \ p(3, u, 4, r)\ \forall u,$$

$$p(3,u,k,\ell) = 0 \ \text{ otherwise},$$
$$p(2,u,3,q) = 1 \ \forall u,$$
$$p(2,u,k,\ell) = 0 \ \text{ otherwise},$$
$$p(4,u,1,q) = 1 \ \forall u,$$
$$p(4,u,k,\ell) = 0 \ \text{ otherwise}.$$

The costs are $h(1,a) = 1 = h(3,b)$ and $h(i,u) = 0$ for all other choices of i,u. Let $[z]$ denote a string of the type "$zzz\cdots z$" of a random finite length. It is easy to see that if the initial law is the Dirac measure at 1, then the state trajectory is $[1],2,[3],4,[1],2,[3],4,\ldots$, with the output trajectory $[q],r,[q],r,\ldots$, with matching lengths, so that the former is exactly recovered from the latter. In other words, the process $\{\pi_n\}$ of conditional laws of X_n given $\sigma(Y_m, m \le n)$ for $n \ge 0$ remains Dirac. Similarly if the initial law is a Dirac measure at 3, the state process is exactly known from the observations. A similar reasoning applies if the initial laws are Dirac at 2 or 4. In particular, the set of Dirac measures is invariant for the nonlinear filter, which recursively generates $\{\pi_n\}$. The optimal cost is zero in all these cases, as we can choose control b in state 1 and a in state 3. On the other hand, if the initial law is $\pi'(\{1\}) = \pi'(\{3\}) = \frac{1}{2}$, then $\{\pi_n\}$ is of the form $[\pi'],\pi'',[\pi'],\pi'',\ldots$, where $\pi''(\{2\}) = \pi''(\{4\}) = \frac{1}{2}$. A similar statement holds if the initial law is π'' instead. Thus $\{\pi',\pi''\}$ is also invariant for the nonlinear filter. In this case the optimal cost will be strictly positive: when $\pi_n = \pi'$, choosing a with probability q and b with probability $1-q$ will yield a cost of $\frac{1}{2}$ regardless of the choice of q. Note that the transition probabilities of $\{X_n\}$ are independent of the control and the resulting state process is an irreducible aperiodic, and therefore ergodic, Markov chain. But the resulting process of conditional laws is a nonergodic Markov process.

The correction is as follows: We replace (3) by

$$\begin{aligned}&\hat{P}_0(d\hat{x}\times d\tilde{x}\times d\hat{u}^\infty\times d\tilde{u}^\infty\times d\tilde{y}^\infty\times dz^\infty\times d\hat{y}^\infty)\\ &\qquad = \hat{\pi}(d\hat{x})\tilde{\pi}(d\tilde{x})\ell_1^\infty(d\hat{u}^\infty)\ell_1^\infty(d\tilde{u}^\infty)\phi(d\tilde{y}^\infty,dz^\infty)\ell_2^\infty(d\hat{y}^\infty),\end{aligned} \tag{8}$$

and (4) by

$$\begin{aligned}\frac{d\hat{P}_n}{d\hat{P}_{0n}} &= \frac{\Gamma(\{(\hat{X}_0,\tilde{X}_0)\})}{\hat{\pi}(\{\hat{X}_0\})\tilde{\pi}(\{\tilde{X}_0\})}\\ &\quad\times\Pi_{m=0}^{n-1}\left(\frac{p(\hat{X}_m,Z_m,\hat{X}_{m+1},\hat{Y}_{m+1})}{\bar{p}(\hat{X}_m,Z_m,\hat{X}_{m+1})(1/d)}\right)\left(\frac{p(\tilde{X}_m,Z_m,\tilde{X}_{m+1},\tilde{Y}_{m+1})}{\bar{p}(\tilde{X}_m,Z_m,\tilde{X}_{m+1})(1/d)}\right).\end{aligned} \tag{9}$$

Define yet another probability measure $\bar{P}_0$ as follows: If $\bar{P}_{0n} \triangleq$ the restriction of $\bar{P}_0$ to $\mathcal{G}_n$, then

$$\frac{d\bar{P}_{0n}}{d\hat{P}_{0n}} = \Pi_{m=0}^{n-1}\left(\frac{p(\tilde{X}_m,Z_m,\tilde{X}_{m+1},\tilde{Y}_{m+1})}{\bar{p}(\tilde{X}_m,Z_m,\tilde{X}_{m+1})(1/d)}\right).$$

Then, using (7), we have

$$\frac{d\hat{P}_n}{d\bar{P}_{0n}} = \frac{\Gamma(\{(\hat{X}_0,\tilde{X}_0)\})}{\hat{\pi}(\{\hat{X}_0\})\tilde{\pi}(\{\tilde{X}_0\})}\Pi_{m=0}^{n-1}\left(\frac{p(\hat{X}_m,Z_m,\hat{X}_{m+1},\hat{Y}_{m+1})}{\bar{p}(\hat{X}_m,Z_m,\hat{X}_{m+1})(1/d)}\right),$$

and under $\bar{P}_0$, $\{\hat{Y}_n\}$ is i.i.d. uniform on W and independent of $\hat{X}_0, \{\hat{\xi}_n\}$, and furthermore $\{Z_n, \hat{Y}_0, \ldots, \hat{Y}_n\}$ is independent of $\{\hat{\xi}_m\}, \hat{X}_0, \{\hat{Y}_i, i > n\}$ $\forall n$. This establishes the wide sense admissibility of $\{Z_n\}$ for $\{(\hat{X}_n, \hat{Y}_n)\}$ and completes the proof of Lemma 3.1 of [1].

Let $\tau = \min\{n \geq 0 : \hat{X}_n = \tilde{X}_n, \hat{Y}_n = \tilde{Y}_n\}$. As in [1], we assume that

$$M \stackrel{\Delta}{=} \sup E[\tau] < \infty, \tag{10}$$

where the supremum is over all Γ and all wide sense admissible $\{Z_n\}$. Let $\{Z_n\}$ above be chosen to be optimal for the α-discounted cost when the initial law is $\hat{\pi}$. Also define

$$\begin{aligned}
X_n^* &\stackrel{\Delta}{=} \tilde{X}_n I\{n \leq \tau\} + \hat{X}_n I\{n > \tau\},\\
\bar{X}_n &\stackrel{\Delta}{=} \hat{X}_n I\{n \leq \tau\} + \tilde{X}_n I\{n > \tau\},\\
Y_n^* &\stackrel{\Delta}{=} \tilde{Y}_n I\{n \leq \tau\} + \hat{Y}_n I\{n > \tau\},\\
\bar{Y}_n &\stackrel{\Delta}{=} \hat{Y}_n I\{n \leq \tau\} + \tilde{Y}_n I\{n > \tau\},\\
\xi_n^* &\stackrel{\Delta}{=} \tilde{\xi}_n I\{n \leq \tau\} + \hat{\xi}_n I\{n > \tau\},\\
\bar{\xi}_n &\stackrel{\Delta}{=} \hat{\xi}_n I\{n \leq \tau\} + \tilde{\xi}_n I\{n > \tau\}.
\end{aligned}$$

Note that

- $\{\bar{\xi}_n, \xi_n^*\}$ are i.i.d. uniform and independent of $\hat{X}_0, \tilde{X}_0$ under both $\hat{P}, \hat{P}_0$;
- under $\hat{P}_0$, $\{\bar{Y}_n\}, \{Y_n^*\}$ are i.i.d. uniform and independent of $\{\bar{\xi}_n\}$, $\{\xi_n^*\}$, $\hat{X}_0$, $\tilde{X}_0$;
- under $\hat{P}_0$, $\{Z_n, \bar{Y}_0, \ldots, \bar{Y}_n\}$ (resp., $\{Z_n, Y_0^*, \ldots, Y_n^*\}$) for each n is independent of $\bar{Y}_m, m > n, \{\bar{\xi}_k\}$ (resp., $Y_m^*, m > n, \{\xi_k^*\}$);
- $\{\bar{X}_n\}$ (resp., $\{X_n^*\}$) are generated recursively by (2) from $(\hat{X}_0, \{\bar{\xi}_n\}, \{Z_n\})$ (resp., $(\tilde{X}_0, \{\xi_n^*\}, \{Z_n\})$).

Furthermore, one verifies that

$$\begin{aligned}
\frac{d\hat{P}_n}{d\hat{P}_{0n}} \stackrel{\Delta}{=} & \frac{\Gamma(\{(\bar{X}_0, X_0^*)\})}{\hat{\pi}(\{\bar{X}_0\})\tilde{\pi}(\{X_0^*\})} \Pi_{m=0}^{n-1} \frac{p(\bar{X}_m, Z_m, \bar{X}_{m+1}, \bar{Y}_{m+1})}{\bar{p}(\bar{X}_m, Z_m, \bar{X}_{m+1})(1/d)}\\
& \times \Pi_{m=0}^{n-1} \frac{p(X_m^*, Z_m, X_{m+1}^*, Y_{m+1}^*)}{\bar{p}(X_m^*, Z_m, X_{m+1}^*)(1/d)}
\end{aligned}$$

for $n \geq 1$. In particular, this shows by the same arguments as above that $(\{\bar{X}_n\}, \{\bar{Y}_n\}, \{Z_n\})$ as well as $(\{X_n^*\}, \{Y_n^*\}, \{Z_n\})$ are state-observation-control triples for which $\{Z_n\}$ is simultaneously wide sense admissible. The rest follows exactly as in [1], leading to Theorems 4.1 and 4.2 of [1], in particular the dynamic programming equation for ergodic control given by (4.1) there. Theorem 4.1 (ii) of [1] yields the existence of an optimal wide sense admissible control $\{Z_n\}$ of the form $Z_n = v(\pi_n), n \geq 0$, for some measurable $v(\cdot)$, where $\pi_n \stackrel{\Delta}{=}$ the regular conditional law of X_n given $Y_m, Z_m, m \leq n$. Plugging this back into (2.4) of [1] (which is the recursive nonlinear filter describing the evolution of $\{\pi_n\}$), one inductively sees that this control is in fact strict sense admissible. This implies that the value function remains unchanged if we move from strict sense admissible over to wide sense admissible controls, and thus justifies the above relaxation.

In the corresponding developments for Euclidean state space worked out in [2], [3], the counterpart of (7) is already assumed. (In fact, it is available free for the sampled nondegenerate diffusions considered later in [2] and [3].) The coupling time τ gets replaced by a "coupling time at the pseudo-atom." In this case, however, $\Gamma(d\hat{x}d\tilde{x})$ need not be absolutely continuous with respect to $\hat{\pi}(d\hat{x})\tilde{\pi}(d\tilde{x})$ in general. But the approach taken in [3] works with independent initial conditions and therefore does not run into this difficulty.

Acknowledgments. I thank Aristotle Arapostathis, Amarjit Budhiraja, and Huizen Yu for extensive discussions.

REFERENCES

[1] V. S. Borkar, *Average cost dynamic programming equations for controlled Markov chains with partial observations*, SIAM J. Control Optim., 39 (2000), pp. 673–681.

[2] V. S. Borkar, *Dynamic programming for ergodic control with partial observations*, Stochastic Process. Appl., 103 (2003), pp. 293–310.

[3] V. S. Borkar and A. Budhiraja, *A further remark on dynamic programming for partially observed Markov processes*, Stochastic Process. Appl., 112 (2004), pp. 79–93.

[4] V. S. Borkar, S. K. Mitter, and S. Tatikonda, *Optimal sequential vector quantization of Markov sources*, SIAM J. Control Optim., 40 (2001), pp. 135–148.